

基于知识图谱的深地数据智能检索算法实现

杨强根¹, 刘洪光^{2*}, 施霞珍^{3*}, 周永贵⁴, 张培星⁴, 施昌辉⁵

1.中国地质大学(武汉)信息化工作办公室, 湖北武汉 430074

2.航天宏图信息技术股份有限公司武汉分公司, 湖北武汉 430073

3.武汉蓝天绿野咨询设计有限公司, 湖北武汉 430024

4.航天宏图信息技术股份有限公司, 北京 100195

5.湖北溪岸节能环保科技有限公司, 湖北武汉 430056

摘要: 深部地质数据是地球科学研究的基础性战略资源, 在保障国家能源安全、防灾减灾、促进地质研究等方面起着重要作用。为了科学、有效地发挥深部地质数据的效能, 对深部地质数据开展有效检索和智能检索, 本研究引入知识图谱技术, 结合深部地质数据特点, 理清深部地质知识的关联关系, 创新性地研究、实现深部地质数据智能检索算法。该算法在进行验证分析的同时, 开展其在深地科学大数据管理与服务系统平台的应用。研究结果表明: 本算法能够为深部地质智能检索业务提供有效的支持, 显著提高深部地质检索业务的准确性和可靠性, 为深部地质其他方面的业务应用提供了新的思路和方法。

关键词: 知识图谱; 本体; 深部地质数据; 智能检索; 算法。

中图分类号: PP542+.5

收稿日期: 2025-9-30

Implementation of Knowledge Graph-Based Intelligent Retrieval Algorithm for Deep Geological Data

Yang Qianggen¹, Liu Hongguang^{2*}, Shi Xiazhen^{3*}, Zhou Yonggui⁴, Zhang

¹ **作者简介:** 杨强根(1973-), 男, 高级工程师, 博士, 从事 GIS 以及网络信息技术应用研究. ORCID: 0000-0001-9215-8128. E-mail: yangqg@cug.edu.cn

^{2*} **通信作者:** 刘洪光(1985-), 男, 高级工程师, 硕士, 从事 GIS、自然灾害监测预警风险评估以及信息技术应用研究. ORCID: 0009-0004-8127-9054. E-mail: 308287460@qq.com

^{3*} **通信作者:** 施霞珍(1984-), 女, 高级工程师, 硕士, 从事生态保护及环境咨询行业. ORCID: 0009-0003-5362-5522. E-mail: 395649178@qq.com

Peixing⁴, Shi Changhui⁵

1. Network and Information Center, China University of Geosciences, Wuhan 430074, China

2. Piesat Information Technology Co., Ltd. Wuhan Branch, Wuhan 430073, China

3. Wu Han Lan Tian Lv Ye Consulting And Design Co., Ltd., Wuhan 430024, China

4. Piesat Information Technology Co., Ltd., Beijing 100195, China

5. Hubei Xian Energy Saving And Environmental Protection Technology Co., Ltd., Wuhan 430056, China

Abstract: Deep geological data serves as a fundamental strategic resource for Earth science research, playing a crucial role in ensuring national energy security, disaster prevention and mitigation, and advancing geological studies. To scientifically and effectively utilize the potential of deep geological data, this study introduces knowledge graph technology to enable efficient and intelligent retrieval of such data. By analyzing the characteristics of deep geological data, we clarify the interrelationships of deep geological knowledge and innovatively research and implement an intelligent retrieval algorithm for deep geological data. While validating the algorithm, we also apply it to a deep Earth science big data management and service platform. The results demonstrate that this algorithm provides effective support for intelligent retrieval of deep geological data, significantly improving the accuracy and reliability of retrieval operations. Moreover, it offers new ideas and methods for other applications in deep geological research.

Key words: knowledge graph ; ontology ; deep geological data ; intelligent retrieval ; algorithm.

1 引言

中国社会经济的快速发展，对能源资源、地质安全提出了深刻的需求。为满足以上需求，同时深刻认识和了解我国大陆深部地质结构和能源资源情况，自 2016 年开始，中国地质调查局开始了深部地质调查工程。该工程通过综合地球物理探测、钻探等技术手段，系统地揭示地壳及上地幔结构与动力学过程。经过数年的发展，深部地质调查工程积累了存储海量、内容丰富的深部地质数据。如何发挥深部地质数据的效能，成为深部地质调查工程研究的课题之一。这其中一项重要的工作为逐步建设、完善深地科学大数据管理与服务系统平台。其中，深部地质数据检索是深地科学大数据

管理与服务系统平台重要的组成部分，也是地质业务工作人员常用的功能模块之一。

在深地科学大数据管理与服务系统平台建设过程中研究发现：受限于数据的体量和技术的发展，常用的传统检索功能存在以下劣势和不足。（1）检索对象单一有限。检索对象为指定数据库中单一的表格或与表格通过数据表外键关联的其他表格；同时，表格中的数据均为规范化的结构数据。使得检索过程无法对相关的多源、异构、海量的全部数据进行全面检索。（2）检索对象关联性弱。检索对象关联点仅为数据表的外键，该关联点为工作人员对结构化数据初步了解后，在数据库构建之初建立的。检索过程无法扩展、应用多源、异构、海量的全部数据之间的逻辑关联关系。（3）检索对象语义理解深度不足。现有的检索过程主要采用匹配策略，是对检索对象的理解局限于直接表达语义的文字，无法深层次理解不同文字表达的类似、相近、相同的语义内容。（4）检索方式智能性差。有限的检索对象、弱化的关联关系、语义理解浅薄，仅能反馈有限的检索结果，无法获得对全部数据的全面、综合、智能的检索结果。

近些年来，为提升业务数据效能，提高数据检索水平，多位专家、学者在地质以及其他多个业务中，开展人工智能技术和知识图谱技术在数据应用的研究工作。包括：分析地质领域知识（包括地质知识、矿床知识、地质构造知识、地震灾害知识、地质灾害知识、地球科学知识等），确定其本体和实体，研究合适、科学的地质知识图谱构建方法，为地质知识图谱的后续应用提供支持（齐浩，2020；廖振鑫，2021；张航，2021；邱芹军等，2022；贺海霞，2023；段雨希等，2024；许鑫冉，2024；刘洪光等，2025）。基于知识图谱技术，面向业务知识的具体业务应用（如数据预测评价、数据推荐、数据查询等），针对性地研究、完善相应的方法，以提升具体业务应用的效能（张彬钰，2023；李诚等，2024；肖克炎等，2025）。岩土数据，或者融合机器学习技术，形成数据驱动场地表征；或者结合大数据技术，分析数据间的遗传特征及其基本属性；或者采用均值法或加权均值法，分析岩土参数数据分布和场地相似性。最终充分利用岩土数据，推动提升岩土数据的应用效能（Kok-Kwang Phoon & Wengang Zhang, 2023 ; Dongsheng Liu et al., 2022 ; Liang Han et al., 2022）。构建地学领域知识图谱的本体，考虑本体层级信息量、本体拓扑结构等因素，研究相关度算法，提高地学领域知识的相关度计算效能（陈小红等，2017）。以上的研究结论和成果，为开发建设优良

的深部地质数据检索，规避传统检索的弊端提供了多方面的参考。同时，由于以上研究成果未面向深部地质数据应用，需对其进行客观的分析和合理的采用。

综合分析、研究、参考以上专家学者的研究结论和成果，面向深部地质数据管理的业务工作需要，本研究创新性地知识图谱技术、深部地质数据、深部地质业务工作内容进行综合、全面、科学、有效的融合。充分分析、发掘多源、异构、海量的深部地质数据之间的时空关联性、语义关联性和内在关联性，构建深部地质知识图谱。基于深部地质知识图谱，通过扩展多源、异构、海量的深部地质数据检索对象，强化深部地质数据之间的关联关系，深化深部地质数据间的语义理解程度，构建并实现深部地质数据智能检索算法。通过该算法，不仅能够获得深部地质数据全面综合、精准有效、科学智能的检索反馈，而且便于地质业务工作人员深入了解深部地质数据与地质业务检索之间有机的关联关系，显著提高地质业务工作人员的工作水平。基于知识图谱的深地数据智能检索算法的实现，一方面，可以大幅发挥深部地质数据的效能，补充完善包含优良深部地质数据检索的深地科学大数据管理与服务系统平台；另一方面，为今后深部地质业务工作提供算法参考、技术参考和理论参考。

2 深部地质知识图谱构建

充分且广泛搜集深部地质科学术语等与深部地质领域相关的多样化数据，形成构建深部地质知识图谱的基础。基于该数据，分析、研究深部地质科学专业数据之间的空间关系、属性关系、时间关系、语义关系和分级关系，构建深部地质探测领域本体。基于深部地质探测领域本体中的专业数据类别领域本体和地理位置领域本体中各个概念的信息量与拓扑结构，分析、评估概念之间的关联关系和关联程度，结合与深部地质领域相关的多样化数据，构建深部地质知识图谱。

2.1 知识图谱构建方法

知识图谱是一种高效的知识挖掘、组织和表示方式，能够有效、科学、完善地解决海量、庞大、复杂多学科知识数据的自动获取、形式表达、计算推理等问题。其以有向图（节点-边）的形式组织知识数据，采用形式化且机器能够理解的语言进行表达，并能支持知识数据的计算推理。由于知识图谱具有强大的知识表达、查询和推理能力，其已在多个领域的智能问答、智能搜索、个性化推荐

等方面得到广泛应用。地学领域知识图谱（Geoscience Knowledge Graph, GKG）是知识图谱技术在地学领域的延伸。

构建知识图谱的方法主要有三种：自顶向下、自底向上、二者混合。这三种方法中，自顶向下法较好体现概念间层次；自底向上法更新快、支持大数据量的知识图谱构建；混合方法灵活性强（黄恒琪等，2019）。综合考虑深部地质科学的业务需求和专业数据间的关联关系及数据体量，采用自顶向下的方法构建深部地质知识图谱。首先，广泛、大量收集深部地质领域相关的海量、多源、异构的业务数据，内容覆盖文献资料、领域专著、地学数据库等多个维度；然后，通过自顶向下的方式建立本体（模式层），实现地学概念、复杂关系及规则等的形式化表达，为实例层的构建及知识的计算推理奠定统一的模式；最终，在本体（模式层）约束下进行自底向上的大规模实体（数据层）自动抽取填充，通过结构化和形式化的方式组织、表示、存储地学知识，构建完成深部地质知识图谱。深部地质知识图谱构建方法如错误!不能识别的开关参数。所示。

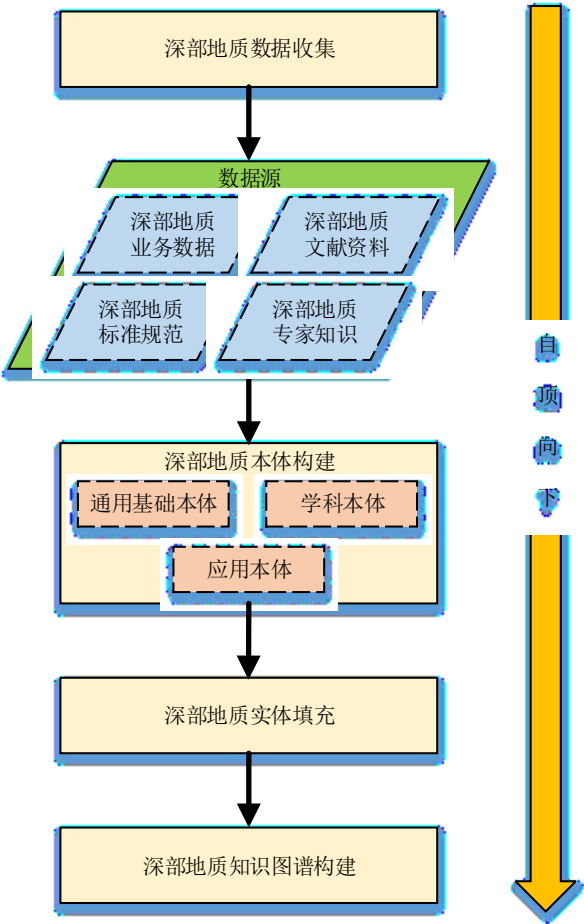


图 1 深部地质知识图谱构建方法

Fig. 1 Deep Geological Knowledge Graph Construction Methodology

2.2 知识图谱构建流程

深部地质知识图谱构建流程包括数据收集、本体构建、实体填充、知识图谱构建等流程。

(1) 数据收集

采用资料收集、数据电子化、数据拷贝、数据库直接读取、数据服务接入、网络爬虫等技术手段，参考《地质学汉语叙词表 2010》、《深地探测科学数据中心标准数据分类与编码方案》、《地球物理学名词》、全国科学技术名词审定委员会术语在线平台、地质资料馆中各类地质报告及文献和专家判识，充分、广泛、大量地收集、汇聚深部地质领域相关数据。收集的深部地质数据主要包括：深反射地震数据、宽角反射/折射地震数据、宽频带地震数据、短周期台阵数据、大地电磁测深数据、重力数据、磁力数据、地应力监测数据、科学钻探监测数据、同位素填图与示踪数据、高温高压试验数据、地球化学试验数据、地球动力学模拟数据、深部地质工程专题数据、深部地质科学研究专题数据、深部地质其他专题数据、元数据等多种数据。

以上数据呈现海量、多源、异构的特点，同时以结构化、半结构化、非结构化的形式存在。存在形式包括纸质文档、网格数据、Excel 数据表格、矢量数据、栅格数据、专业文件格式（如 SEG-D 格式、seg-y 格式、SAC 格式、WXL 格式等）等多种形式。

参考《中国地质调查局地质调查技术标准（DD2006-05）地质信息元数据标准》、《地球深部探测数据汇交规范（内部试行）》、《深地探测科学数据中心标准数据分类与编码方案》、《地质调查项目管理暂行办法》（中地调发〔2016〕158 号）、《成果地质资料管理技术要求》（DD 2010-06）等标准，采用 ArcGIS、QGIS、PostgreSQL9.2 数据库、达梦数据库、金仓数据库等软件对收集到的深部地质数据进行标准化、规范化处理，从而为深部地质知识图谱的构建提供科学、稳定、可靠、合理的数据基础。

(2) 本体构建

深部地质领域本体是深部地质知识图谱的概念模型、逻辑基础，其对深部地质领域实体进行约束。按照本研究中深部地质领域本体的领域范围，基于地学知识图谱本体的内容和功能定位，将深

部地质领域本体划分为 3 个层次：通用基础本体层、学科本体层和应用本体层。通用基础本体层包含与具体的地学子学科和应用无关的本体，用于支撑学科本体和应用本体的构建，主要包括时间本体、空间本体等；学科本体层包含地学不同子学科的本体，用于引用、复用通用基础本体，同时支撑相应学科的应用本体的构建；应用本体层面向具体应用，构建时引用通用基础本体以及涉及的学科本体。

基于标准、规范的深部地质数据，提炼深部地质知识图谱的概念，形成深部地质知识图谱抽象的本体结构，同时以实体、关系的形式进行表示。深部地质知识图谱本体的构建主要包括地学领域本体框架构建、领域边界确定、关系建模。

1) 地学领域本体框架构建。地学领域本体框架是深部地质知识图谱本体构建的基础和总体架构。构建该框架，需掌握地质学、地理名称和土地科学等地学内容具有检索意义的基本名词和基本术语。参考《地质学汉语叙词表 2010》中的地学概念、全国地质资料馆的各类地质报告及文献中规范的主题词，系统性地构建地学领域本体框架。该框架为四级树结构，第二层级本体涵盖地学领域的 31 个具体类别，包括：普通地质学、构造地质学、地层学、地球物理、地球化学、岩石学、矿物学、矿床学、地球物理勘探、环境地质学等。

2) 领域边界确定。依据《深地探测科学数据中心标准数据分类与编码方案》，综合深部地质数据特点，深部地质数据划分为 5 个大类，16 个中类。深部地质数据类型划分详见**错误!不能识别的开关参数。**。本研究对深部地质数据中主要专题类别（3 个大类，13 个中类）涉及的关键概念进行细化，明确、完善大类、中类、细化类别的从属关系。大类包括深地探测、深地监测、深地实验；中类包括深反射地震、宽角反射/折射地震、宽频带地震、短周期台阵、大地电磁测深、重力、磁力、地应力监测、科学钻探监测、同位素填图与示踪、高温高压试验、地球化学实验、地球动力学模拟。

表 1 深部地质数据类型划分

Table 1 Classification of Deep Geological Data Types

大类名称	中类名称
深地探测	深反射地震

	宽角反射/折射地震
	宽频带地震
	短周期台阵
	大地电磁测深
	重力
	磁力
	地应力监测
深地监测	科学钻探研究
	同位素填图与示踪
	高温高压试验
深地实验	地球化学试验
	地球动力学模拟
	深地工程专题
深地专题	深地科学研究专题
	深地其他专题
深地其它	

3)关系建模。基于地学领域本体框架,分析标准化、规范化的深部地质各类实体间的层次关系,以上下级为拓扑方向将所有实体词关联,将地学领域的通用知识与深部地质领域的特有概念相结合,构建完整的层次化树状结构。该层次结构划分为五个层级。

(3) 实体填充

基于本体构建内容,详细细化深部地质领域实体,抽取标准化、规范化的深部地质数据,以填充每个专题类别下一级细化类别的实体数据。部分专题细化后不同专题类别抽取、填充的实体详见

错误!不能识别的开关参数。。

表 2 深部地质知识图谱实体填充(部分)

Table 2 Deep Geological Knowledge Graph Entity Population (Partial)

本体	实体
深反射地震	地震反射，地球深部，二维地震测线，SEG Y 数据，SPS 文件，人工源激发，叠加偏移剖面
宽角反射/折射地震	地震宽角反射，地震测深，地震折射，地球深部，折射反演速度，人工源激发
宽频带地震	流动地震台站，固定地震台站，宽频带地震，30s~50hz 或更宽
短周期台阵	短周期地震，小间距、高密度观测台阵，原始连续地震记录，1s-20hz 及以上
大地电磁测深	宽频大地电磁，时间序列，EDI 数据，远参考，长周期大地电磁，时频转换

(4) 知识图谱构建

基于数据收集、本体构建、实体填充等各个阶段的成果，深部地质知识库结合图数据库，构建完成深部地质知识图谱。深部地质知识本体是深部地质知识图谱的关键部分，其地学领域本体包含 5 个层级：第一层级本体 1 类（深部地质探测领域本体）；第二层级本体 31 类；第三层级本体 10603 类；第四层级本体 4399 类；第五层级本体 150 类。例如，“深部地质探测领域本体”属于一级层级本体，“地球物理学”属于二级层级本体，“地震学”属于三级层级本体，“深反射地震”、“宽角反射/折射地震”等属于四级层级本体，四级本体的描述词构成五级层级本体。深部地质探测领域知识本体如错误!不能识别的开关参数。所示。

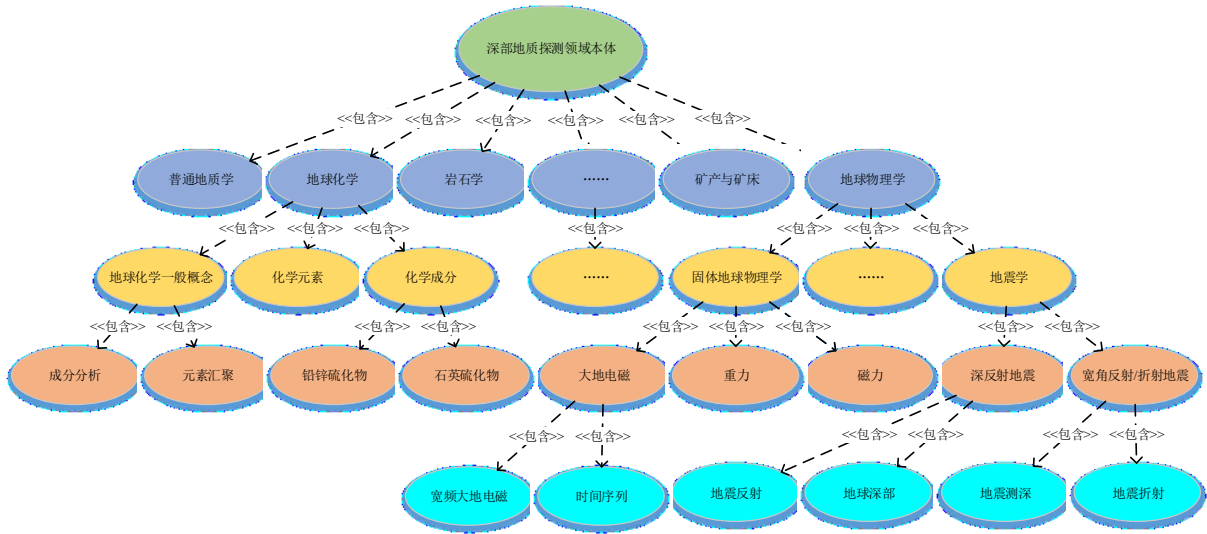


图 2 深部地质探测领域知识本体

3 深部地质数据智能检索算法实现

构建完成的深部地质知识图谱，理清深部地质领域本体中各个概念的关联关系和相关的信息量；同时，通过深部地质领域实体链接，实现深部地质知识与深部地质数据的关联。深部地质知识图谱为各项深部地质业务应用（特别是深部地质数据检索）提供科学、稳定、可靠的支持。

为改变传统检索固有的弊端和不足，适应深部地质业务工作内容，满足深部地质数据检索的检索对象关联性高、检索方式智能化多样化等需求，本研究基于深部地质知识图谱，研究并实现深部地质数据的智能检索算法。以期深刻理解检索关键词的语义信息，将检索关键词精准地映射到深部地质数据中，建立检索与数据之间的有机关联关系。不仅能够显著发挥深部地质数据的效能，而且为今后深部地质业务工作提供算法参考、技术参考和理论参考。

3.1 算法流程

针对深部地质知识图谱中本体和实体之间的树状结构关联关系，通过编制合适的编码规则，以数据表的形式将深部地质知识图谱存储于数据库中，从而推理深部地质数据条目间的关联性，进而利用深部地质数据间关联度的排序，实现深部地质数据智能检索算法，完成深部地质数据的智能检索。

深部地质数据智能检索算法的框架涉及到数据层、知识层、应用层。数据层主要为多源、异构、海量的深部地质数据实体，是深部地质数据智能检索算法的数据基础。知识层主要包括：深部地质数据概念、关系、规则等内容，是深部地质数据本体；不同类型词向量元素，进行词向量相关度计算。应用层是深部地质数据智能检索算法在深地科学大数据管理与服务系统平台的具体体现，即为深部地质数据检索模块。深部地质数据智能检索算法框架如**错误!不能识别的开关参数**所示。

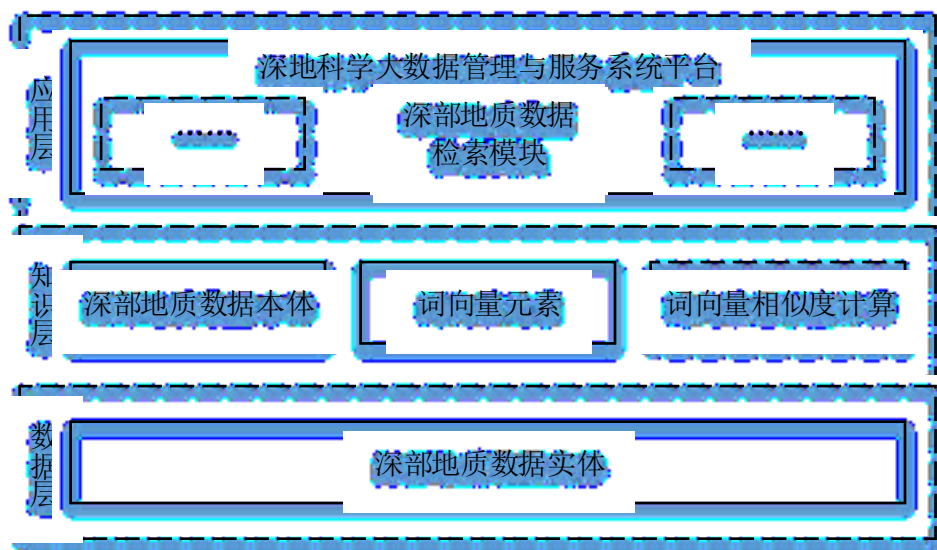


图 3 深部地质数据智能检索算法框架

Fig. 3 Deep Geological Data Intelligent Retrieval Algorithm Framework

深部地质数据智能检索算法的流程主要包括词语映射、节点相似度计算、词向量相关度计算。

(1) 词语映射。按照深部地质科学业务逻辑内容，对深部地质业务的数据条目名称、专题类别和关键词进行词语切分，从分词中提 1 取关键词干，形成本体可以识别的词向量，并将获得的词干的词语映射到本体中对应的节点。通过图谱结构（本体树状结构），基于第一层级的本体，计算下属各个层级每个节点的概率值。(2) 节点相似度计算。通过图谱结构（本体树状结构），对词向量元素进行分类统计，获取不同类型词向量元素的分类统计结果；按照词向量元素分类节点，计算获取节点相似度；不同类型的词向量，分别获取节点相似度计算结果的最大值。(3) 词向量相关度计算。综合词向量元素分类统计结果和节点相似度，计算出词向量相关度，量化词向量之间的关联程度，从而服务于深部地质数据的智能检索业务。深部地质数据智能检索算法流程如**错误!不能识别的开关参数**。所示。

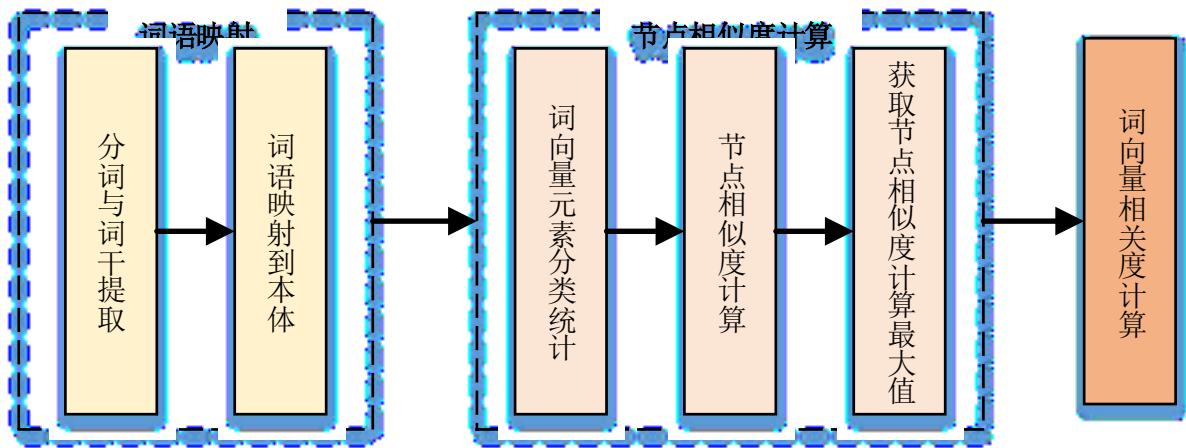


图 4 深部地质数据智能检索算法流程

Fig. 4 Intelligent Retrieval Algorithm for Deep Geological Data

3.2 算法实现

以构建完成的深部地质知识图谱和算法流程为基础，逐步实现深部地质数据智能检索算法。该算法实现的关键包括分词与词干提取、词向量元素分类统计、节点概率值和信息量值计算、节点相似度计算、词向量相关度计算。

(1) 分词与词干提取

深部地质数据中，数据条目的名称和专题类别中的内容主要为短语，而且存在和生成词向量无关的信息，同时关键词中也有无法在本体节点中进行匹配的词语。针对该情况，对数据条目的内容进行筛选和处理，通过分词提取关键词干，并将词干映射到本体中相应的节点中，并进行相应的编码。采用开源的 `jieba` 工具包进行分词，利用本体库和地质词典对分词过程进行约束，不仅能解决分词过程中存在的分词歧义、分词结果无法与本体建立映射等问题，而且能更好地控制分词的粒度。分词与词干提取的效果详见错误!不能识别的开关参数。。

表 3 分词与词干提取的效果对比

Table 3 Comparison of Tokenization and Stemming Effects

分词内容	无词典分词的词向量	有词典分词的词向量
中低温岩浆热液脉状金矿床	中低温，岩浆，热液，脉状，金矿床	中低温，岩浆热液，脉状金矿床
生物地球化学沉积矿床	生物，地球，化学，沉积，矿床	生物，地球化学，沉积矿床

(2) 词向量元素分类统计

词向量由两组本体节点组成。两组待对比词向量中，将存在父子节点关系的节点中等级最高的节点作为分类节点。统计每个分类节点下包含的词向量中元素数量，计算出每种分类的占比结果 ω_i ($i=1, \dots, n$)。

词向量元素分类统计具体流程为：合并两个词向量，并将所有本体节点向上一级到其父节点。若父节点为第一层级节点，则此本体节点保持不变。找到某一类节点中等级最高的，选中某一个节点，依次计算其与其他节点的最小公共父节点。若最小公共父节点为第一层级节点，则将被对比元素添加到新的列表中；找到非第一层级节点的结果中等级最高的父节点，作为分类节点。对新列表中的剩余元素按以上过程循环计算，直到没有剩余元素为止。

(3) 节点概率值和信息量值计算

将位于第一层级本体节点为整个本体的最高级的原始父节点。规定其概率值为 1，向下逐步采用均等概率分布的方法为每个本体节点分配概率值。再利用节点概率值计算节点的信息量值。

某个本体节点 C_i 的概率值 $P(C_i)$ 和信息量值 $I(C_i)$ 的计算公式如下：

$$P(C_i) = \frac{1}{m_1 \times m_2 \times m_3 \times \dots \times m_n} \quad (1)$$

$$I(C_i) = -\lg P(C_i) = -\lg \frac{1}{m_1 \times m_2 \times m_3 \times \dots \times m_n} \quad (2)$$

式中： m_1 为第一层级的节点数量，取值为 1； m_2 为第二层级的节点数量； m_3 为第三层级的节点数量； m_n 为节点 C_i 的所在层级的节点数量；节点数量的取值为整数。

通过以上公式中可以看出：本体节点的等级越低，概率值越小，信息量值越高。

(4) 节点相似度计算

对于两个本体节点 C_i 和 C_j ，同时考虑本体的层次为树状结构，则二者的相似度 $\text{sim}(C_i, C_j)$ 定义为：该两个本体节点共享的最低层次父节点的最大信息量值除以两个本体节点的信息量值之和。节点相似度的值介于[0,1]的区间范围之内，且选取节点相似度计算的最大值。

节点相似度 $\text{sim}(C_i, C_j)$ 的计算公式如下：

$$\text{sim}(C_i, C_j) = \frac{I(\text{sup}(C_i, C_j)) - 2\lg(P(\text{sup}(C_i, C_j)))}{I(C_i) + I(C_j) - \lg P(C_i) - \lg P(C_j)} \quad (3)$$

式中： $\text{sup}(C_i, C_j)$ 为节点 C_i 和 C_j 共享的最低层次父节点，该父节点提供节点 C_i 和 C_j 共享的最大信息量； $I(\text{sup}(C_i, C_j))$ 为节点 C_i 和 C_j 共享的最低层次父节点的信息量值； $I(C_i)$ 为节点 C_i 的信息量值； $I(C_j)$ 为节点 C_j 的信息量值； $P(\text{sup}(C_i, C_j))$ 为节点 C_i 和 C_j 共享的最低层次父节点的概率值； $P(C_i)$ 为节点 C_i 的概率值； $P(C_j)$ 为节点 C_j 的概率值。

(6) 词向量相关度计算

对于两组词向量，综合全部词向量元素分类统计结果和其中分类节点相似度计算结果，计算出两组词向量的相关度。其中，词向量元素分类统计结果作为计算的权重。

词向量相关度 $V(\text{correl})$ 的计算公式如下：

$$V(\text{correl}) = \text{sim}_A \omega_A + \text{sim}_B \omega_B + \dots + \text{sim}_J \omega_J \quad (4)$$

式中：两组词向量的全部词向量经过元素分类统计后，确定有 $A, B, \dots, J$ 类； sim_A 为 A 类节点相似度； ω_A 为 A 类节点计算权重，即为该类节点在两组词向量的全部词向量中的占比； sim_B 为 B 类节点相似度； ω_B 为 B 类节点计算权重，即为该类节点在两组词向量的全部词向量中的占比； $\dots$ ； sim_J 为 J 类节点相似度； ω_J 为 J 类节点计算权重，即为该类节点在两组词向量的全部词向量中的占比。

3.3 算法性能

一般传统的检索主要采用的是匹配策略，即通过设置的检索关键字与指定数据中的字符进行直接的匹配比对；比对完全一致，则输出检索反馈结果。检索反馈结果的数量，取决于关键字的长度、使用频率和数据中关键字出现的次数、位置等因素。检索反馈结果数据为单一的结构化数据。

相较于传统的检索，深部地质数据智能检索算法，不仅能够输出同等数量的匹配检索结果，而且还能根据深部地质数据之间的时空关联性、语义关联性和内在关联性，全面、综合检索多源、异构、海量的全部数据，进一步输出更多数量的检索反馈结果。检索反馈结果的形式也进一步扩展到包括结构化数据在内的多种形式的深部地质数据。

4 深部地质数据智能检索算法验证

为验证深部地质数据智能检索算法的效能，设置检索关键词，分别进行匹配检索和智能检索。基于检索结果，做进一步的分析研究，为后续开展的深部地质数据检索的开发建设提供支持。

4.1 算法结果

设置两组检索关键词“深反射地震”和“锆石测年”，分别进行匹配检索和智能检索。匹配检索和智能检索的结果对比详见**错误!不能识别的开关参数。**。从匹配检索和智能检索的结果对比表格中可以看出，采用深部地质数据智能检索算法的智能检索相较于传统的匹配检索，检索结果数量显著提高。其中，“深反射地震”检索结果的命中率由匹配检索的 9 条提升为智能检索的 31 条，检索数量命中率提高了 2.4 倍；“锆石测年”检索结果的命中率由匹配检索的 0 条提升为智能检索的 3 条，检索数量命中率突破了 0。

表 4 匹配检索和智能检索的结果对比

Table 4 Comparison of Matching Retrieval and Intelligent Retrieval

检索关键词	匹配检索		智能检索	
	结果数量	结果内容	结果数量	结果内容
“深反射地震”	9 条	华南深反射地震单炮记录（2016-2018），	31 条	华南深反射地震单炮记录（2016-2018），东北额尔古纳-虎林剖面，华南秦岭剖面深地震反射数据，东北剖面深地震反射数据，华北剖面深地震反射数据，华南剖面深地震反射数据，华南深地震反射剖面图集（2016-2018），青藏高原剖
				南深反射地震单炮记录（2016-
				2018），东北额尔古纳-虎林剖面，华南
				万载-永春剖面短周期地震数据
				（2018），秦岭剖面深地震反射数据，华
				南岩石圈深部纵波速度各向异性图
				（2021），东北剖面深地震反射数据，华
				北剖面深地震反射数据，华南岩石圈三维纵波速度异常图（2021），华南剖面深地震反射数据，青藏高原可可西里剖面及青
				藏高原腹地 EW 剖面，华南深地震反射剖

面深地震反射数据， 松辽盆地及周缘深地 震反射剖面数据集， 华南深反射地震剖面 数据集（2016-2018）				面图集（2016-2018），武夷廊带宽频地 震波形数据（2017-2019），南岭宽频地 震波形数据（2017-2020），华南剖面宽 频带流动地震台阵探测实验，青藏高原剖 面深地震反射数据，青藏高原 NS 剖面 （羌塘盆地及藏北），东南沿海南岭州- 萍乡剖面，东南沿海 NE 剖面（东山-霞 浦，安溪-屏南）及青藏高原东北缘红原- 景泰剖面，松辽盆地及周缘深地震反射剖 面数据集，华南万载-永春剖面地壳结构 （2020 年），华南深反射地震剖面数据 集（2016-2018），东南沿海 NW 剖面 （厦门-宜丰）及（台州-阜阳），青藏高 原拉萨地块 EW 向剖面，华南岩石圈密度 模型（2020），华南岩石圈磁化率模型数 据（2020），中国东北重力推断的岩浆岩 体属性表，华南岩石圈磁化率模型 （2020），华南岩石圈密度模型数据 （2020），惠州多要素城市地质调查项目 2019 年东南沿海密集台阵原始数据，航 磁数据向下延拓方法（2021），惠州多要 素城市地质调查项目 2019 年东南沿海宽 频带原始数据
“锆石测	0 条	/	3 条	紫金山锆石 U-Pb 同位素年龄数据表

年”	(2017 年)，川口铅石 U-Pb 定年数据
	(2020 年)，钦杭成矿带 LA-ICP-MS 铅
	石 U-Pb 同位素 (2020)

4.2 结果分析

(1) 检索关键词为“深反射地震”时，匹配检索，仅仅是检索出深部地质数据条目名称中与“深反射地震”相关明显的深地数据，并对其他相关的深部地质数据进行检索。对于同样的检索词，采用深部地质数据智能检索算法的智能检索，除检索出深部地质数据条目名称中与“深反射地震”相关明显的深地数据以外，还根据算法计算出的相关度数值，检索出其他相关的深部地质数据。当相关度数值 ≥ 0.40 ，检索出的深部地质数据包括：标题以地名命名的东北额尔古纳-虎林剖面、青藏高原可可西里剖面及青藏高原腹地 EW 剖面等地震剖面数据，一些涉及地震相关的纵波数据。如若进一步考虑较为宽泛的相关度值的范围（相关度数值 ≥ 0.25 ），还会检索到一些磁法和重力相关的深地领域内容。通过智能检索，可以实现对地质业务工作人员感兴趣条目的有机补充。“深反射地震”智能检索结果及相关度数值详见**错误!不能识别的开关参数。**。

表 5 “深反射地震”智能检索结果及相关度数值

Table 5 "Deep Seismic Reflection" Intelligent Search Results and Relevance Scores

检索关键词	智能检索	
	结果内容	相关度
“深反射地震”	华南深反射地震单炮记录（2016-2018）	1.0
	东北额尔古纳-虎林剖面	0.75
	华南万载-永春剖面短周期地震数据（2018）	0.67
	秦岭剖面深地震反射数据，华南岩石圈深部纵波速度各向异性图（2021），东北剖面深地震反射数据，华北剖面深地震反射数据，华南岩石圈三维纵波速度异常图（2021），华南剖面深地震反射数据，青藏高原可可西里剖面及青藏高原腹地 EW 剖	0.60

面	
华南深地震反射剖面图集（2016-2018），武夷廊带宽频地震波形数据（2017-2019），南岭宽频地震波形数据（2017-2020）	0.57
华南剖面宽频带流动地震台阵探测实验	0.56
青藏高原剖面深地震反射数据，青藏高原 NS 剖面（羌塘盆地及藏北），东南沿海南岭州-萍乡剖面，东南沿海 NE 剖面（东山-霞浦，安溪-屏南）及青藏高原东北缘红原-景泰剖面，松辽盆地及周缘深地震反射剖面数据集	0.50
华南万载-永春剖面地壳结构（2020 年），华南深反射地震剖面数据集（2016-2018）	0.40
东南沿海 NW 剖面（厦门-宜丰）及（台州-阜阳），青藏高原拉萨地块 EW 向剖面	0.375
华南岩石圈密度模型（2020）	0.35
华南岩石圈磁化率模型数据（2020），中国东北重力推断的岩浆岩体属性表，华南岩石圈磁化率模型（2020），华南岩石圈密度模型数据（2020），惠州多要素城市地质调查项目 2019 年东南沿海密集台阵原始数据	0.29
航磁数据向下延拓方法（2021）	0.26
惠州多要素城市地质调查项目 2019 年东南沿海宽频带原始数据	0.25

（2）检索关键词为“锆石测年”时，由于该内容属于同位素测年的范畴，且数据库中条目的名称、关键词和摘要没有出现“锆石测年”的内容，所以采用对深部地质数据条目名称中与“锆石测年”相关明显的深地数据进行直接检索的匹配检索是检索不出任何结果的。对于同样的检索词，采用深部地质数据智能检索算法的智能检索，计算出相关度数值 ≥ 0.20 ，则可根据本体推理出三条锆石同位素定年的相关数据。通过智能检索，不仅提高了检索结果的适用范围，而且能够得到更为全

面的检索结果。“锆石测年”智能检索结果及相关度数值详见错误!不能识别的开关参数。。

表 6 “锆石测年”智能检索结果及相关度数值

Table 6 "Zircon U-Pb Dating " Intelligent Search Results and Relevance Scores

检索关键词	智能检索	
	结果内容	相关度
“锆石测年”	紫金山锆石 U-Pb 同位素年龄数据表（2017 年）	0.30
	川口锆石 U-Pb 定年数据（2020 年）	0.20
	钦杭成矿带 LA-ICP-MS 锆石 U-Pb 同位素（2020）	

4.3 验证说明

综合算法结果内容和结果分析内容，可以明显地看出：创新性采用知识图谱技术构建的深部地质数据智能检索算法相较于传统的匹配检索，检索结果数量显著提高，检索结果范围大幅扩大，检索结果内容更加丰富。算法结果内容和结果分析内容，直接证明了本研究中的深部地质数据智能检索算法的优势性、有效性和可行性。

深部地质数据智能检索算法的应用，不仅能够提高地质业务工作人员的检索工作水平，而且能够扩大地质业务工作人员对深部地质数据的理解范围。通过该算法，为不同地质业务知识储备程度的业务工作人员，提供全面、综合、科学、充分的深部地质业务数据支持和智能检索技术支持。

5 结论与展望

5.1 研究结论

深部地质数据是地球科学研究的基础性战略资源，呈现多源、异构、海量的特点，其在保障国家能源安全、防灾减灾、促进地质研究等方面具有重要作用。为充分、有效、合理地利用好深部地质数据，本研究引入知识图谱技术，结合深部地质数据特点，研究深部地质数据智能检索算法，并在深部地质业务工作中展开应用。本研究取得的研究成果如下：

（1）深部地质知识图谱。充分、广泛搜集深部地质科学术语等与深部地质领域相关的多样化数据，分析、研究深部地质科学专业数据之间的空间关系、属性关系、时间关系、语义关系、分级关

系和关联程度，采用自顶向下的方法，构建深部地质知识图谱，为深部地质数据智能检索提供逻辑基础、关系基础和关联基础。

(2) 深部地质数据智能检索算法。基于深部地质知识图谱及其包含的本体和实体之间的树状结构关联关系，利用合适的编码规则，将深部地质知识图谱以数据表形式存储于数据库中。以深部地质数据条目间的关联性和关联度排序为基础，对深部地质数据进行词语分词、词干提取、词向量分类统计，逐步计算获取深部地质领域本体节点的概率值、信息量值及相似度，最终获取词向量相关度，从而构建完成深部地质数据智能检索算法，以服务于深部地质数据的智能检索业务。

(3) 深部地质数据智能检索算法验证及应用。基于深部地质科学术语等与深部地质领域相关的多样化数据，设置检索关键词“深反射地震”和“锆石测年”，分别进行匹配检索和智能检索。根据检索结果的对比情况，验证了深部地质数据智能检索算法的有效性、科学性；并将深部地质数据智能检索算法应用于深地科学大数据管理与服务系统平台的研发工作中，开发建设出平台重要组成部分——深部地质数据管理和分发系统的数据集资源查询检索功能模块。

5.2 研究展望

深部地质数据因其获取来源的多样性、数据结构的差异性、物理存储的海量性，使其呈现出一定的复杂性。本研究分析、构建的深部地质知识图谱和深部地质数据智能检索算法，虽然满足了当前深部地质业务工作对深部地质数据检索的业务应用需求，但是受限于多种因素，导致本次研究内容存在着一定的不足，需要在今后的研究中做进一步的优化和完善。

(1) 动态本体更新机制待引入。本研究中，本体是通过分析、研究广泛搜集到的深部地质领域相关的多样化数据获取的。该过程耗费时间长、适应灵活性弱，难于适应深部地质数据迅速汇聚增长、深部地质数据业务本体快速更新的趋势要求。为顺应以上趋势，持续提供优良的深部地质检索服务，今后需研究、引入动态本体更新机制，进而持续补充、完善深地科学大数据管理与服务系统平台，发挥深部地质数据效能。

(2) 语义理解待增强。现有的语义理解是通过本体库、地质词典约束的深部地质数据的语义，从而造成语义的范围和深度有限。随着深部地质数据的汇聚、扩充，有限的语义理解，后续难于提

供优良的深部地质数据检索支持。今后，一方面扩充深部地质数据相关的词典、标准、规范等说明性资料，另一方面积极融合深度学习技术，持续不断地深化对深部地质数据的理解程度。

(3) 知识图谱构建待完善。基于深部地质科学术语等与深部地质领域相关的多样化且比较充足的数据，本研究采用自顶而下的方法构建了深部地质知识图谱。深部地质知识图谱的这种构建方法，对人工干预的依赖性偏强，同时导致深部地质领域本体难于更新，很难适应深部地质数据迅速扩充、深部地质知识图谱快速更新的发展趋势。为更好地满足深部地质业务工作对深部地质知识图谱的需求，今后需引入自底向上和自顶向下与自底向上相结合的知识图谱构建方法，并加强以上方法在深部地质知识图谱构建过程中的应用研究。

(4) 智能检索算法待优化。智能检索算法的构建过程中，相关的深部地质实体间的层次化树状结构，结构清晰，易于理解，查询效率高；节点概率值、节点信息量值和节点相似度的计算简洁明了。由于，未能充分考虑深部地质数据的复杂程度，从而在一定程度上忽略了同级深部地质实体间的关联关系，同时部分降低了节点相似度计算的健壮性和准确性。在今后的研究中，需基于本研究成果，进一步研究深部地质实体间网状结构、其它相似度度量方法在智能检索算法构建过程中应用的适用性和可能性。

References

- Chen , X. H. , Pan , M. , Shi , Y. L. , et al. , 2017 . Research on Semantic Relevance Calculation Based on Domain Constraint Ontology . *Science Technology and Engineering* , 17(19):187-189. doi:10.3969/j.issn.1671-1815.2017.19.034
- Duan , Y. X. , Qiu , Q. J. , Tian , M. , et al. , 2024 . Geological map-oriented knowledge graph construction and intelligent Q&A application. *Chinese Journal of Geology* , 59(2):588-602. doi:10.12017/dzgx.2024.041
- Han , L. , Wang , L. , Ding , X. M. , et al. , 2022 . Similarity quantification of soil parametric data and sites using confidence ellipses . *Geoscience Frontiers* , 13(1):1-13. doi: 10.1016/j.gsf.2021.101280
- He , H. X. , 2023 . Construction and application of earthquake knowledge Graph (A Thesis Submitted for the Degree of Master) . Lanzhou : Lanzhou Jiaotong University .
- Huang , H. Q. , Yu , J. , Liao , X. , et al. , 2019 . Review on Knowledge Graphs . *Computer Systems & Applications* , 28(6):6-

10. doi:10.15888/j.cnki.csa.006915
- Li , C. , Mao , X. , Lu , W. J. , et al. , 2024 . Research on natural language place name query method . *Science of Surveying and Mapping* , 49(4):187-195. doi:10.16251/j.cnki.1009-2307.2024.04.018
- Liao , Z. X. , 2021 . The Construction And Application Of Knowledge Graph For Geology (A Thesis Submitted for the Degree of Master) . Chengdu : University of Electronic Science and Technology of China .
- Liu , D. S. , Liu , H. L. , Wu , Y. , et al. , 2022 . Characterization of geo-material parameters: Gene concept and big data approach in geotechnical engineering , *Geosystems and Geoenvironment* , 1(1):1-9. doi: 10.1016/j.geogeo.2021.09.003
- Liu , H. G. , Yang , Q. G. , Shi , X. Z. , et al. , 2025 . Construction Methodology of the Modeling the Risk Assessment for Typhoon-Collapse Disaster Chain for Public Safety , *Earth Science* , 50(10):4060-4065. doi: 10.3799/dqkx.2025.153
- Phoon , K. K. , Zhang , W. G. , 2023 . Future of machine learning in geotechnics , *Georisk : Assessment and Management of Risk for Engineered Systems and Geohazards* , 17(1), 7-22. doi: 10.1080/17499518.2022.2087884
- Qi , H. , 2020 . The Construction of Ontology-based Earth Science Knowledge Graph (A Thesis Submitted for the Degree of Master) . Nanjing : Nanjing University .
- Qiu , Q. J. , Wu , L. , Ma , K. , et al. , 2023 . A Knowledge Graph Construction Method for Geohazard Chain for Disaster Emergency Response . *Earth Science* , 48(5):1876-1887. doi: 10.3799/dqkx.2022.313
- Xiao , K. Y. , Li , C. , Tang , R. , et al. , 2025 . Big data intelligent prediction and evaluation . *Earth Science Frontiers* . doi:10.13745/j.esf.sf.2025.4.58
- Xu , X. R. , 2024 . Research on the Construction Method of Knowledge Graph of Complex Geological Structure Based on Knowledge Reasoning (A Thesis Submitted for the Degree of Master) . Chengdu : University of Electronic Science and Technology of China .
- Zhang , B. Y. , 2023 . Research on Geographic Data Recommendation Methods Based on Knowledge Graphs (A Thesis Submitted for the Degree of Master) . Beijing : Beijing University of Civil Engineering and Architecture .
- Zhang , H. , 2021 . Construction and Application of knowledge Graph in the Field of Debris Flow Disaster (A Thesis Submitted for the Degree of Master) . Kunming : Yunnan Normal University .

中文参考文献

- 陈小红,潘懋,史艳丽,等,2017.基于领域约束本体的语义相关度计算研究.科学技术与工程,17(19):187-189.
- 段雨希,邱芹军,田苗等,2024.面向地质图的知识图谱构建及智能问答应用.地质科学,59(2):588-602.
- 贺海霞,2023.地震灾害知识图谱构建与应用研究(硕士学位论文).兰州:兰州交通大学.
- 黄恒琪,于娟,廖晓,等,2019.知识图谱研究综述.计算机系统应用,28(6):6-10.
- 李诚,毛曦,路文娟,等,2024.基于知识图谱的自然语言式地名查询方法研究.测绘科学,49(4):187-195.
- 廖振鑫,2021.面向地质领域的知识图谱构建研究及应用(硕士学位论文).成都:电子科技大学.
- 刘洪光,杨强根,施霞珍,等,2025.面向公共安全的台风-崩塌灾害链风险评估模型构建方法.地球科学,50(10):4060-4065.
- 齐浩,2020.基于本体的地球科学知识图谱的构建(硕士学位论文).南京:南京大学.
- 邱芹军,吴亮,马凯,等,2023.面向灾害应急响应的地质灾害链知识图谱构建方法.地球科学,48(5):1876-1887.
- 肖克炎,李程,唐瑞,等,2025.大数据智能预测评价.地学前缘,32(4):2-15.
- 许鑫冉,2024.基于知识推理的复杂地质构造知识图谱构建方法研究(硕士学位论文).成都:电子科技大学.
- 张彬钰,2023.基于知识图谱的地理数据推荐方法研究(硕士学位论文).北京:北京建筑大学.
- 张航,2021.泥石流灾害领域知识图谱的构建与应用(硕士学位论文).昆明:云南师范大学.