大数据时代的地学知识图谱研究

周成虎^{1,2*}, 王华^{2†}, 王成善³, 侯增谦⁴, 郑志明⁵, 沈树忠⁶, 成秋明³, 冯志强⁷, 王新兵⁸, 闫海荣⁹, 樊隽轩⁶, 胡修棉⁶, 侯明才¹⁰, 诸云强¹

1. 中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室, 北京 100101;
2. 中国科学院大学, 北京 100049;
3. 中国地质大学(北京)地球科学与资源学院, 北京 100083;
4. 中国地质科学院地质研究所, 深地动力学自然资源部重点实验室, 北京 100037;
5. 北京航空航天大学人工智能研究院, 北京 100191;
6. 南京大学地球科学与工程学院, 内生金属矿床成矿机制研究国家重点实验室, 南京 210023;
7. 中国石化石油勘探开发研究院, 北京 100083;
8. 上海交通大学电子信息与电气工程学院, 上海 200240;
9. 清华大学自动化系, 北京 100084;
10. 成都理工大学沉积地质研究院, 成都 610059

* 通讯作者, E-mail: zhouch@lreis.ac.cn

† 通讯作者, E-mail: wangh@lreis.ac.cn

收稿日期: 2020-11-30; 收修改稿日期: 2021-01-04; 接受日期: 2021-03-01; 网络版发表日期: 2021-05-19

国家自然科学基金项目(批准号: 41421001、42050101、42050105)资助

摘要 进入21世纪以来, 地球科学研究正进入一个以建立新知识体系为核心和大数据驱动为手段的重大转折时期, 从传统的百科全书式的学科知识体系到计算机可理解与可操作的知识图谱是地学知识研究的一次革命性跃迁。地学知识图谱在采纳一般知识表达的图模式基础上, 拓展地学知识所特有的时空特征, 融合图、文、数等地学要素, 从而建立全域地学知识表达模型; 发展联邦式群智协同地学知识图谱构建方法, 协同全球地球科学家, 实现高质量的专业知识图谱构建; 发展基于深度解析的多模态地学数据动态知识图谱构建方法, 从海量的地学文献资料中提取地学知识, 实现最新、最全的动态地学知识图谱构建。全面而系统的地学知识图谱不仅可以深化现有的地学大数据分析, 而且可以推进大数据驱动的高精度地质时间轴构建、规则与数据驱动的智能地图编制、地学知识演化与推理分析等研究, 将进一步拓展数据与知识双重驱动的地学研究新方向, 开辟地球科学、信息科学和数据科学交叉的新领域, 实现地学研究的源头创新和时空大数据研究的重大理论突破。

关键词 地学知识图谱, 全域地学知识表达模型, 联邦群智协同, 高精度地质时间轴

中文引用格式: 周成虎, 王华, 王成善, 侯增谦, 郑志明, 沈树忠, 成秋明, 冯志强, 王新兵, 闫海荣, 樊隽轩, 胡修棉, 侯明才, 诸云强. 2021. 大数据时代的地学知识图谱研究. 中国科学: 地球科学, 51(7): 1070–1079, doi: [10.1360/SSTe-2020-0337](https://doi.org/10.1360/SSTe-2020-0337)

英文引用格式: Zhou C, Wang H, Wang C, Hou Z, Zheng Z, Shen S, Cheng Q, Feng Z, Wang X, Lv H, Fan J, Hu X, Hou M, Zhu Y. 2021. Geoscience knowledge graph in the big data era. Science China Earth Sciences, 64(7): 1105–1114, <https://doi.org/10.1007/s11430-020-9750-4>

1 引言

新一轮科技革命与产业变革正在全球兴起, 人类正迈进大数据、云计算、人工智能、区块链与物联网融合发展的新科技时代, 以大数据和深度学习为基础的第二代人工智能在图像识别、语音翻译等领域得到广泛应用(Guo等, 2014; Guo, 2017a, 2017b). 大数据不仅正在改变人类生活、生产和思维的方式, 而且正在引导科学研究从实验、理论、模拟的科学范式, 进入到大数据驱动的第四科学范式(Tansley和Tolle, 2009).

现代对地观测系统实现小时级的全球监测, 各类自动观测台站网络实现了对全球降水、生物量等要素的持续观测, 数字化出版文献资料实现了随时随地可接入, 地球科学研究已经进入了一个全球覆盖、全天候监测、全要素观测的大数据时代. 同时, 作为典型的数据密集型科学, 地球科学在数据集成与共享、数据挖掘与知识发现等方面面临诸如数据混杂、机理缺乏的空间统计分析的挑战, 大数据的众多潜在优势在地球科学相关研究中尚未得到充分发挥, 知识驱动的地球科学大数据分析的理论与方法亟待发展, 构建全域地学知识图谱、探讨地学知识演化等, 是当代地学知识研究的前沿领域和战略重点(翟明国等, 2018; 汪品先和翦知潜, 2019). 人工智能是大数据价值挖掘与提升的关键, 而知识图谱则是人工智能的重要基石之一, 是实现统计表征与物理表征融合的核心基础. 本文从地学知识表达的图模型、地学知识图谱构建方法和地学知识图谱应用等方面, 阐述和讨论面向地学大数据分析的地学知识图谱研究的关键科学问题与前沿方向, 展望地球时空大数据分析和知识发现的研究.

2 地学知识表达模式

地学是研究地球各圈层形成、演化及其相互作用的科学, 包括大气科学、海洋科学、地理学、地质学和地球物理学等分支学科, 其学科知识体系复杂多样(孙鸿烈, 2017). 如何兼顾地学现象与过程的多尺度时空特征, 融合图、文、数等多种信息载体表征, 构建跨越地球科学各分支学科的地学知识表达模式是地学知识图谱研究的基础和出发点.

2.1 地学知识体系与地学知识图谱

地球系统是一个复杂的巨系统, 空间上绵延数万里, 时间上跨越数十亿年. 认知地球的地学知识宏观庞大, 构建完整的地学知识体系是一个系统性知识工程, 也是地球科学家不懈的追求. 作为一门学科知识, 地学知识具备一般知识所应具有的特性, 遵循常见的知识分类体系, 如经济合作与发展组织将知识分为关于事物和现实的知识、关于自然规律和原理的知识、关于技能和诀窍的知识以及关于人力资源的知识等四类(经济合作与发展组织, 1997); 美国心理学家、教育家本杰明·布鲁姆将知识分为事实性知识、概念性知识、程序性知识和元知识等四类(经济合作与发展组织, 1997). 这些知识工程的研究成果, 均可成为地学知识体系建设的基础.

但与其他学科知识相比, 多尺度时空特性则是地学知识构成的基本要素. 一方面, 地球科学家在思考和研究问题时, 一般需要锁定研究对象的空间范围和时间跨度, 如地质学家常采用百万年作为重大地质事件度量的基本时间单位, 采用洲际的空间尺度研究板块运动; 另一方面, 限于技术条件, 地球科学所使用的观测和分析方法与途径也是多尺度的, 如古生物学家的化石标本、地质年代学家的“金钉子”剖面, 均具有有限的、可观测的空间尺度. 因此, 采纳一般知识分类体系、融合时空特征、综合运用图文数各要素载体, 以树状知识架构为主体, 融合网状、知识链等架构, 构建一个开放式、可拓展的地学知识框架, 则是地学知识体系建设的核心和创新之处.

地学知识图谱是以结构化的图模式, 将相关知识有效组织起来形成计算机可理解和可计算的一种知识体系. 知识图谱概念和雏形可以追溯到20世纪60年代, 并在图书情报领域得到广泛应用. Boyack等(2005)用节点和线条构建了一幅知识图谱, 展示出80万篇科技文献之间存在的内在联系和学科关系; Auer等(2007)采用语义网络等方法, 推动开放型链接数据库DBpedia建设. 语义网络描述客观世界中的概念、实体及其相互关系的方法也是地学知识图谱表达的有效方法(唐杰, 2020; 张雪英等, 2020); 2012年谷歌正式发布知识图谱引擎(Singhal, 2012), 建成了包含6亿多实体、180多亿条属性或关系的知识节点, 创建了用于从非结构化网络文本中获取事实信息的新一代知识图

“Knowledge Vault”(Dong等, 2014; 陆锋等, 2017), 极大地推动了知识图谱技术方法的发展和应用。

自2017年以来, 美国将开放知识网络(Open Knowledge Network)作为一项国家科技战略予以重点推动(NSTC, 2018)。2019年美国国家自然科学基金会(NSF)资助的43项学科融合加速先导项目(总预算3900万美元)中, 包含了21项以开放知识网络为主题的项目, 并计划继续投入更大的资金支持该方向(<https://www.nsf.gov/pubs/2019/nsf19050/nsf19050.jsp>; <https://www.nsf.gov/od/oia/convergence-accelerator/index.jsp>)。其中, 美国爱达荷大学主持的深时知识图谱(Deep Time Knowledge Graph)项目, 通过引入最新的语义网模型建立机器可读的深时“语言”, 以此联通国际地质年代标准和区域地质年代标准, 统一管理不同版本的地质年表的年代概念, 并使用其开展网络环境下深时数据的探索和分析(Ma等, 2020)。目前, SocialWiki联合豆瓣时间等多家青年媒体共同发起“人类知识谱系建设计划”, 试图通过全球的协作, 绘制更加广泛的学科图谱, 帮助人们探索和学习知识。这些研究和实践都将有助于促进地质知识图谱的研究。

2.2 全域地质知识图谱自适应表达模型

知识表达是构建计算机可理解、可计算的知识图谱的基础, 也是知识交流过程中的重要步骤。常见的知识表达方式包括自然语言、结构化表格和图形图像等, 其中, 自然语言表达一般是定性的、含糊的描述, 结构化表格难以刻画地质实体间的时空关系, 图形图像则不足以精确描述复杂地质过程。因此, 面对地质大数据的分析和复杂多样的地质知识体系的整合, 迫切需要发展一种综合性的形式化表达语言(许珺等, 2010)。以有向图为代表的图模式有效地建立了知识对象间的语义关联, 可以采纳作为地质知识表达的基本模式, 并通过时空拓展, 成为地质知识图谱的一种有效表达模式。例如, 在地理学科领域, 学者们提出了YAGO2(Hoffart等, 2013)、GeoKG(Wang等, 2019)等地理知识表达模型, 以有向图结构记录地理知识(Balatore等, 2015)。其中, YAGO2是通用知识图谱YAGO的升级版, 增添了时间和空间表达的谓词描述, 进而记录了每个实体的时间信息和空间信息; GeoKG则从地理学学科角度出发, 提出了以地理实体为核心围绕时间、空间、属性、状态、改变和关系的地理知识表

达模型, 实现了地理实体演化过程的形式化描述。

现有的地质知识表达模型一般将时间、空间和部分属性信息进行简单的补充或记录, 难以表达复杂的全域地质知识(张雪英等, 2020; Oramas等, 2017)。如何从图结构本质出发, 融合地质知识中复杂的时空特征、计算属性及地质知识关系与规则, 构建跨时空维度的地质知识图谱表达模型, 则是一项具有重要挑战的基础研究。基于此, 我们提出一种由复杂时空信息表达模型(实体对象表达模型)和地质实体对象关系(边)构成的全域地质知识图谱自适应表达的基本模型(图1)。该模型仍然采用头实体/实体组合、尾实体/实体组合和离散化关系/映射函数三元组资源的描述框架, 但在节点上拓展了时间和空间特征, 在边关系上融入复杂模型; 在地质场景上增加了智能识别和对象知识对齐, 对底层地质知识图谱的自适应表达形式、存储结构等进行重组, 以适应特定场景和特定领域高效推理的需求, 从而达到地质知识图谱在全域的自适应表达。

3 地质知识图谱构建方法

归纳和总结人类已有科学知识一直是人类科学研究活动的组成之一, 如编撰学科经典与辞书、大百科全书等。科学计量学的奠基人Price在20世纪60年代运用统计方法, 以科学论文数据库为基础, 发现科学知识指数增长规律, 这成为知识图谱应用的开拓性工作之一(陈悦等, 2008)。

3.1 知识图谱构建策略与方法

知识图谱构建策略一般分为“自下而上”和“由上而下”两种。开放知识图谱的构建多采用“自下而上”的策略, 自动地从各文本数据中抽取概念或实体以及它们之间的关系, 如谷歌的Knowledge Vault。专业领域知识图谱多采用“由上而下”的策略, 事先确定好需要构建的本体和实体, 采用专家方式构建, 所构建的知识图谱的专业水准高。中国科学家发起和正在推进实施的“深时数字地球”大科学计划, 组建了18个学科工作组, 采用“由上而下”的策略, 推动建设固体地球科学的学科知识体系、知识规则, 目的是为下一步全域地质知识图谱建设提供基础。

作为一个新兴的研究领域, 知识图谱的构建方法

图 1 全域地学知识图谱自适应表达模型

快速发展, 自动化和智能化的构建方法不断出现, 极大地提高了知识图谱构建的能力。常规的以人工输入与编辑为主的传统方法, 依然得到广泛的采用, 并不断改进, 特别是专家知识对齐比对的自动化与智能化水平在提升, 如Cyc系统(Lenat, 1995), 基于众包思想、网络化协同编辑的Freebase和Wikidata等(Chah, 2017; Mitraka等, 2015), 基于大规模、半结构化或非结构化的文本资料提取知识的DBpedia、YAGO、NELL、PROSPERA等(Ansari等, 2019; Carlson等, 2010; Nakashole等, 2011)。这些方法和系统为地学知识图谱的构建

奠定了良好的基础。

地学知识图谱以地球科学共同认可的知识体系为基础, 是对地球科学领域内的所有知识节点(包括已知的各种现象与事实、基本的概念与定义、自然原理与规律、观测与分析技术方法等)以及这些知识点之间的相互关系进行清晰、明确的阐释, 是一种可以为机器所理解的地球科学知识库和“推理机”。根据地学知识的来源和可靠性等, 对于相对稳定和成熟的知识体系, 特别是领域的事实性知识和概念性知识多采用人机协同编辑导入策略, 如群智协同构建方法; 对于散

布在大量文献资料中的知识,特别是出版的学术论文、图书、研究报告,多采用文本文献资料挖掘与知识发现的方法,如网络文本解析与知识发现方法。随着知识图谱的进一步发展,领域专家知识和文本动态知识将会融合发展,相互补充和支撑,形成混合型的构建体系。

3.2 群智协同的地质知识图谱构建

地质知识体系覆盖面广,大部分知识和经验都来自于地质专家,采用传统的方法将这些专家的知识经验和人工地输入到计算机中,需要高度的协同和昂贵的成本,特别是不确定性或模糊性较高的地质知识,其形式化表达极其困难,从而导致这方面研究进展缓慢。随着智能移动终端的快速发展和普及,基于移动互联网的群智协同提供了一种通过聚集群体智慧解决问题的新模式。依靠群体力量,采集结构化模式,可以实现地质知识图谱的实时更新,并且还可借助群智对机器挖掘出的知识进行校验。为此,我们提出以专家知识为核心的联邦式群智地质知识图谱构建框架(图2),其中需求重点关注和解决的关键技术方法问题如下。

(1) 群智协同构建的规模庞大的地质知识图谱中,不可避免地产生不同专家的知识之间的矛盾和冲突,需要有效地甄别新旧知识的迭代更新、不同理论体系之间的观点碰撞等。例如,针对异构语义环境下的知识创建与整合以及协同开发工作中冲突的发现和解决, [Gangemi等\(2007\)](#)提出了协同本体设计模型(Collaborative Ontology-Design Ontology, CDO); [Chen等\(2008\)](#)将知识协同构建冲突分为协同冲突、一致性冲突和异常冲突等三类,并提出Command Package Pool方法以实现冲突的自动检测。本框架采用知识贡献度与可信度评价的冲突知识融合修正方法,实现群智输入的地质知识冲突的自动检测。

(2) 研究和建立可持续的共建共享模式,是构建全域地质知识图谱的重要机制保障。例如,对于参与知识图谱群智构建的知识拥有者,宣称自身知识并申明知识产权,但又不愿意公开知识细节的场景,需要研究和发展具备隐私保护能力的知识存储证明方法,推动“我的知识我做主”的知识共享模式。为了激励学科专家积极主动地参与到群智协同构建过程中,需要建立群智协同下的知识价值贡献评估方法以及学科专家信誉评价模型,实现对专家知识贡献的评估,建立有效的

图2 联邦式群智协同地质知识图谱构建框架

群智协同的激励机制。

(3) 全球地球科学家大规模的协同, 需要有效地解决地学领域知识多语言表示问题, 研究和建立基于规则、统计、深度学习的多源知识图谱的一致性融合的技术和方法。

3.3 基于深度解析的多模态地学数据动态知识图谱构建

海量结构化与非结构化、出版发行和内部使用的地学文本文献中, 隐含了大量的地学知识, 特别是最新的动态知识。运用机器学习、深度学习等现代技术方法, 从数据感知、实体识别和关系抽取等方面实现从海量文本文献资料中提取并更新地学知识, 是构建动态的地学知识图谱的重要途径, 也是现代地学知识发现和图谱构建的重要途径。针对此, 我们提出基于深度解析的多模态地学数据动态知识图谱构建方法(图3)。

第一, 基于多源地学数据深度解析的非结构化数据感知。针对海量的文本、图片、数据表格、地图等

非结构化的文本资料进行分类, 并对同一来源的数据标记其关联属性, 如地图的名称、区域、制图指标、经纬度范围、时间等, 实现文本关联与多源数据感知; 在标记的基础上, 对所处理的文本进行分割、纯文本提取、文本分词和句法分析等, 特别是采用一定已有规则知识消除文本中的非实质性的半结构化文本; 采用文本匹配和统计学习等方法, 对不同来源、具有一定相似性的图-文-数进行标记关联, 特别是通过基于规则过滤以及神经网络模型, 对文本描述中关键词信息的抽取。与一般的文本数据分析相比, 地学文献资料中常常包含大量的地图、表格等多种成分, 因此在知识获取时, 需要深度解析地图, 需要理解地图符号的地学意义、识别各类空间关系等, 并结合地学知识体系特定文本中的地学对象进行规整。

第二, 基于关键词的实体对象与知识抽取。深度学习的键之一是有大量优质的训练语料样本, 而依靠人工选取、标注训练语料的方式难以实现, 发展高效、可信的无监督学习算法至为关键, 如基于关键词图模型的对象抽取。在该类算法中, 采用基于海量文

图3 基于深度解析的多模态地学数据动态知识图谱构建框架

本分词结果和词频-逆向文档频率(TF-IDF)算法量化地质领域实体的统计特征, 依据排序结果搭建常用词表, 构建海量文本的语言网络图; 基于图特征大小和向量匹配排序, 在语言网络图上寻找具有重要作用的词或者短语, 筛选出文本中的关键词, 抽取相应的对象实体, 构建图文数-关键词的索引关系, 完成地质知识中图表-文本描述-数字间的匹配, 从而实现基于非结构化文本的地质知识提取。

第三, 知识歧义消除与动态地质知识图谱构建。针对由多数据源引起的知识歧义和冲突问题, 以地质知识图谱中实体概念特定的语义时空关联为特征通过深度强化学习进行分类和聚类, 以分别解决一词多义和多词同义问题; 并以信源属性为特征, 通过特征学习训练信源可信度完成属性对齐以消除知识冲突。

4 地质知识图谱的典型应用方向

地质知识图谱的广泛应用可以推动地球科学与信息科学、数据科学的交叉融合, 促进学科发展。基于知识驱动的时空地质大数据分析有助于实现更加精准的地质分析, 推动基于统计表征和物理表征的地质大数据综合分析; 基于已有地质知识库和知识引擎, 可以推动地质知识体系研究, 理解地质知识演化的特征, 发现新的地质知识, 形成新的研究突破点和创新点; 将地质知识和地图编制知识融合, 可以推动地图制图的智能化与自动化发展; 将地质知识与地球系统模型结合, 可以推动矿产资源的探测与预测研究。以下重点阐述地质知识图谱三个典型的应用方向。

4.1 地质知识推理

地质知识推理是指从地质知识图谱中的实体概念间关系出发, 经过计算机推理, 建立地质实体间的新关联, 理解地质知识体系演化特征, 发现地质新知识。目前常用的知识推理方法包括符号推理和统计推理。符号推理的核心是利用相关规则, 从已有的实体关系推理出新的实体关系, 并检测其中的可能的逻辑冲突, 如Wang和Li(2016)提出对文本语料的上下文信息分析方法TEKE; 统计推理则是根据以往的经验和分析, 利用机器学习等方法发展知识图谱中新的实体之间关系, 并利用最大化后验概率等统计方法对推理假设进行验证或推测, 如Tay等(2017)建立的用于学习知识图

谱中实体关系的自适应鲁棒转移矩阵模型puTransE。

无论是符号推理, 还是统计推理, 都未能有效地表达地质知识所具有时空依赖性与非平稳性(异质性), 尚未充分利用地质大数据的多模态特征间关联进行建模, 难以实现对地质知识多属性的挖掘与预测。针对此问题, 我们提出基于实体属性和时空关联的双向推理概念模型。其中, 针对地质实体演化的时间跨度大、时间节点信息较多等特点, 采用因果卷积和扩张因子实现对长时间历史标注信息的特征学习, 避免传统的递归神经网络在应对长序列输入时可能会出现梯度消失或梯度爆炸的问题; 针对地质实体的空间关系及其对应的拓扑模型, 建立基于异构图构建和图卷积神经网络的地学实体空间特征学习机制; 采用多任务学习模型的动态时空关系建模, 可实现对地质实体的时间和空间特征的深度提取, 实现属性与时空关联之间的双向关联, 利用属性反推时空关联。

4.2 基于地质知识图谱的高精度地质时间轴构建

统一时空框架是开展地球科学研究的前提条件和基石, 建立时间分辨率达到万年级精度的统一深时时间框架则是地球科学家的梦想。面对用化石“首现”定义“金钉子”的方法所面临的质疑与挑战(Aubry等, 1999; Walsh等, 2004; Lucas, 2018; Davydov, 2020), 基于地层学、古生物学、地质年代学、天文旋回和同位素年代学等相关学科知识建立地质知识图谱, 融合海量的动态数据, 运用现代大数据分析技术与方法, 构建高精度地质时间轴, 将会在传统的“金钉子/层型”和地质年表基础上开启国际地层学研究的后层型(Post-Stratotype)阶段。其中, 重点解决以下两方面的科学问题与基础突破。

(1) 发展对地层学、古生物学、地质年代学、天文旋回和同位素年代学的基本概念、术语、规范、技术方法及其相互关系进行自动提取的智能方法, 制定基于国际标准规范语言, 设计建立适合地球科学知识图谱描述的编码规范, 构建机器可理解的地层学、古生物学、地质年代学、天文旋回和同位素年代学的知识图谱本体层, 特别是时间本体。

(2) 发展基于高性能计算和人工智能技术的时间对比与标定、重大生物环境事件的时间和速率的厘定、地质时间轴构建的人工智能算法等, 构建地质时间轴的知识体系与知识图谱, 对有关的知识节点(包括

基本概念、对象、现象、过程、标准、方法等)以及这些知识点之间的相互关系进行清晰、明确的阐释,分步建立不同地质时期的地质时间轴,如古生代十万年到万年级别的新一代地质时间轴,并逐步集成,最终形成跨越整个地球演化史的地质时间轴。

4.3 智能地图编辑与制图

地图作为空间信息可视化表达和传输的方式历史悠久,并随着人类进步和科技发展而不断演进(郭仁忠和应申, 2017; 陈军和陈晋, 2018)。地图制图是一项耗时耗力的复杂任务,涉及地图设计、地图编辑、制图综合、打印输出等多个环节。研究和发展自动化、智能化的编图系统一直是地图学科领域的重点。20世纪80年代初到90年代中期,借助于符号主义人工智能(第一代人工智能)的发展,国内外出现了研究地图制图专家系统的热潮(熊伟, 2019),但其中的地图制图知识工程瓶颈问题没有解决,距离地图制图的智能化仍有很大差距。以地学知识图谱为驱动的地图编制则是实现智能编图的一个新途径,也是地学知识图谱最为直接的应用。

基于地学知识图谱的智能编图是指在地图制图基本规律和规则的约束下,以地学知识和制图知识为制图工作的基本驱动力,以数据智能为自主决策的基本保障,实现自主完成各种制图资源包括数据、模型、模板、方法等的判断、调配,自主决策学习时机,自主判定产品质量及优化策略等,推动自动化、智能化的地图编制。智能化地图编制的核心任务可分为由知识驱动的制图流程制定的外环任务和由数据智能辅助的地图编制操作的内环任务。外环任务以流程知识为驱动,确定整个制图流程;内核任务是地图制图的核心,包括数据处理、总体设计、内容组织、符号化表达、制图输出等一系列地图制图的操作,是实现地图制图的基础。内环任务在外环任务的驱动下,实现智能化地图制图。

5 结论与讨论

地学知识图谱是地球科学研究的前沿方向,构建地学知识图谱是地球科学界的一项系统性知识工程。本文从地学知识的特点和实际应用出发,提出一种由复杂时空信息表达模型(实体对象表达模型)和地学实

体对象关系(边)构成的地学知识图谱自适应表达模型,为构建全域地学知识图谱确定了基本架构;建立了联邦式群智协同地学知识图谱构建和基于深度解析的多模态地学数据动态知识图谱构建的框架,为建立地学知识图谱确定了两种策略和关键方法;阐述了地学知识推理、高精度地质时间轴构建和智能地图编辑与制图等三个典型的地学知识图谱应用方向,为进一步开展地学知识图谱应用提供了思路与指导。

地学知识图谱研究处于初期的起步阶段,无论是地学知识体系的结构化梳理与语义表达,还是大规模知识图谱构建与大数据分析应用等,均需要开展深化研究,特别是以下几个核心问题亟待探索。

(1) 融合地图要素的地学知识表达。目前的知识表达多采用“节点-边”的基本架构模式,尚未能够有效地融合地学研究的各类专题地图,较难解析复杂的地学现象或事件的时空关系。本文提出的拓展模型,在实现地图要素的智能化提取和分析方面,尚存在一定的难度。

(2) 面向应用主题的知识图谱快速重构。目前,大部分的知识图谱都是在已有的场景下进行构建,如何建立以“时间-空间-主题”为核心的知识图谱快速重构和自适应性分析,则是知识图谱应用的关键。

(3) 多元大数据的知识发现与地学知识图谱更新。目前,动态知识图谱的构建多是基于文献资料,尚未实现从更多来源的数据中发现知识,检测新知识的可靠性和可信度,快速更新知识图谱。

致谢 本文得到许多国内外同行的指导,在此难以一一署名致谢,深表歉意。特别感谢参加2019年第242期双清论坛“数据驱动地球科学:从传统到数据时代”以及深时数字地球系列研讨会的专家和学者。本文的主题思想来源于大家共同的思考,所以从这个意义上讲,这篇文章应该是一个集体之作。感谢国家自然科学基金委地学部姚玉鹏研究员、张朝林研究员等在基金委原创探索性项目的总体设计和研究方向上的指导。

参考文献

- 陈军, 陈晋. 2018. GlobeLand30遥感制图创新与大数据分析. 中国科学: 地球科学, 48: 1391–1392
- 陈悦, 刘则渊, 陈劲, 侯剑华. 2008. 科学知识图谱的发展历程. 科学学研究, 26: 449–460

- 郭仁忠, 应申. 2017. 论ICT时代的地图学复兴. 测绘学报, 46: 1274–1283
- 经济合作与发展组织, 著. 杨洪进, 薛澜, 译. 1997. 以知识为基础的经济. 北京: 机械工业出版社
- 陆锋, 余丽, 仇培元. 2017. 论地理知识图谱. 地球信息科学学报, 19: 723–734
- 孙鸿烈. 2017. 地质大辞典. 北京: 科学出版社
- 唐杰. 2020. 浅谈人工智能的下一个十年. 智能系统学报, 15: 193–198
- 汪品先, 翦知湣. 2019. 探索南海深部的回顾与展望. 中国科学: 地球科学, 49: 1590–1606
- 熊伟. 2019. 人工智能对测绘科技若干领域发展的影响研究. 武汉大学学报: 信息科学版, 44: 101–105
- 许珺, 裴韬, 姚永慧. 2010. 地质知识图谱的定义、内涵和表达方式的探讨. 地球信息科学学报, 12: 496–502
- 翟明国, 杨树锋, 陈宁华, 等. 2018. 大数据时代: 地质学的挑战与机遇. 中国科学院院刊, (8): 825–831
- 张雪英, 张春菊, 吴明光, 闫国年. 2020. 顾及时空特征的地理知识图谱构建方法. 中国科学: 信息科学, 50: 1019–1032
- Ansari G A, Saha A, Kumar V, Bhambhani M, Sankaranarayanan K, Chakrabarti S. 2019. Neural program induction for KBQA without gold programs or query annotations. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: AAAI Press. 4890–4896
- Aubry M P, Berggren W A, van Couvering J A, Steininger F. 1999. Problems in chronostratigraphy: Stages, series, unit and boundary stratotypes, global stratotype section and point and tarnished golden spikes. *Earth-Sci Rev*, 46: 99–148
- Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. 2007. DBpedia: A nucleus for a web of open data. In: Aberer K, Choi K S, Noy N, Allemang D, Lee K I, Nixon L, Golbeck J, Mika P, Maynard D, Mizoguchi R, Schreiber G, Cudré-Mauroux P, eds. The Semantic Web. ISWC 2007, ASWC 2007. Lecture Notes in Computer Science, vol 4825. Berlin, Heidelberg: Springer. 722–735
- Ballatore A, Bertolotto M, Wilson D. 2015. A structural-lexical measure of semantic similarity for geo-knowledge graphs. *ISPRS Int J Geo-Inform*, 4: 471–492
- Boyack K W, Klavans R, Börner K. 2005. Mapping the backbone of science. *Scientometrics*, 64: 351–374
- Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka Jr. E R, Mitchell T M. 2010. Toward an architecture for never-ending language learning. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010. Atlanta, Georgia, USA, July 11–15, 2010. 1306–1313
- Chah N. 2017. Freebase-triples: A methodology for processing the freebase data dumps. arXiv:1712.08707v1 [cs.DB]
- Chen Y, Zhang S, Peng X, Zhao W. 2008. A collaborative ontology construction tool with conflicts detection. In: Fourth International Conference on Semantics, Knowledge and Grid. Los Alamitos: IEEE Computer Society. 12–19
- Davydov V I. 2020. Shift in the paradigm for GSSP boundary definition. *Gondwana Res*, 86: 266–286
- Dong X, Gabrilovich E, Heitz G, Horn W, Lao N, Murphy K, Strohmann T, Sun S, Zhang W. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 601–610
- Gangemi A, Presutti V, Catenacci C, Lehmann J, Nissim M. 2007. C-ODO: An OWL meta-model for collaborative ontology design. In: Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge (CKC 2007) at the 16th International World Wide Web Conference (WWW2007). Banff
- Guo H D. 2017a. Big data drives the development of Earth science. *Big Earth Data*, 1: 1–3
- Guo H D. 2017b. Big Earth data: A new frontier in Earth and information sciences. *Big Earth Data*, 1: 4–20
- Guo H D, Wang L, Chen F, Liang D. 2014. Scientific big data and Digital Earth. *Chin Sci Bull*, 59: 5066–5073
- Hoffart J, Suchanek F M, Berberich K, Weikum G. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artific Intell*, 194: 28–61
- Lenat D B. 1995. CYC: A large-scale investment in knowledge infrastructure. *Commun ACM*, 38: 33–38
- Lucas S G. 2018. The GSSP method of chronostratigraphy: A critical review. *Front Earth Sci*, 6: 191
- Ma X G, Ma C, Wang C B. 2020. A new structure for representing and tracking version information in a deep time knowledge graph. *Comput Geosci*, 145: 104620
- Mitraka E, Waagmeester A, Su A, Good B. 2015. Wikidata: A central hub for linked open life science data. In: the Biocuration 2015 Conference. Beijing
- Nakashole N, Theobald M, Weikum G. 2011. Scalable knowledge harvesting with high precision and high recall. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. 227–236
- NSTC (National Science and Technology Council). 2018. Open Knowledge Network. 8
- Oramas S, Ostuni V C, Noia T D, Serra X, Sciascio E D. 2017. Sound and music recommendation with knowledge graphs. *ACM Trans Intell Syst Technol*, 8: 1–21
- Singhal A. 2012. Introducing the Knowledge Graph: Things, not

- strings. Google Blog. <https://www.blog.google/products/search/introducing-knowledge-graph-things-not/>
- Tansley S, Tolle K. 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research
- Tay Y, Luu A T, Hui S C. 2017. Non-parametric estimation of multiple embeddings for link prediction on dynamic knowledge graphs. In: Proceedings of the Thirty First Conference on Artificial Intelligence (AAAI). Menlo Park: AAAI. 1243–1249
- Walsh S, Gradstein F, Ogg J. 2004. History, philosophy, and application of the Global Stratotype Section and Point (GSSP). *Lethaia*, 37: 201–218
- Wang S, Zhang X, Ye P, Du M, Lu Y, Xue H. 2019. Geographic Knowledge Graph (GeoKG): A formalized geographic knowledge representation. *ISPRS Int J Geo-Inform*, 8: 184
- Wang Z, Li J. 2016. Text-enhanced representation learning for knowledge graph. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. Menlo Park: AAAI. 1293–1299

(责任编辑: 李新)