地质科学大数据及其利用的若干问题探讨

吴冲龙^{①②*}, 刘刚^{①②}, 张夏林^①, 何珍文^①, 张志庭^①

① 中国地质大学计算机学院, 地质过程与矿产资源国家重点实验室, 武汉 430074;

② 智能地学信息处理湖北省重点实验室, 武汉 430074

* 联系人, E-mail: wucl@cug.edu.cn

2015-12-24 收稿, 2016-03-07 修回, 2016-03-08 接受, 2016-05-16 网络版发表

国家自然科学基金(41172300, 41101368)和国家高技术研究发展计划(2008AA121602, 2012AA121401)资助

摘要 地质科学大数据是一种时空大数据, 其特征与一般大数据有相似之处, 也有显著差别. 为了应对挑战, 需要引进大数据理论、方法和技术, 开展对地质科学大数据的统合和利用. 采用大数据技术直接在海量时空数据和文本数据中挖掘知识, 能突破采样随机性和样本空间狭小、仅凭少量观测数据和固有模式进行判断, 以及传统数据分析方法的限制, 有可能取得地质科学的新发现. 目前常用的一系列数学地质方法和一般空间数据挖掘方法, 经改造可用于地质科学时空大数据挖掘. 大数据时代地质科学发展所涉及的关键科学技术问题, 包括结构化数据与半结构化非结构化数据、大数据与小数据、混杂性数据与精确性数据、模型与数据、静态勘查模型与动态监测模型等的一体化存储管理和处理, 数据挖掘与数据分析的结合、相关关系与因果关系的统一, 地质科学大数据的深度挖掘和可视化. “玻璃地球”是地质科学大数据的有效载体, 开展“玻璃地球”建设是解决上述各种科学技术问题的最佳途径之一.

关键词 大数据, 地质时空数据, 玻璃地球, 云技术, 数据挖掘

面对大数据时代, 地质科学将如何发展? 各国学术界和产业部门都在考虑这个问题, 并采取了应对的策略, 其中最具代表性的是美国地质调查局. 美国地质调查局根据当前的新形势、新挑战, 在连续几次制定的科学战略及10年规划中^[1~3], 明确提出了基于大数据的核心科学体系, 并制定了核心科学体系的科学战略目标. 其中, 目标一是为临界带描述、认识及制图提供调查研究技术, 包括填制四维地质图等来提高时间和空间分辨率, 以及展示复杂地质结构的动态变化; 目标二是通过科学服务扩展调查成果的应用, 包括实现信息管理和发布现代化, 改进调查工作、数据分析、可视化和信息处理; 目标三侧重于学科交叉和综合的技术方法, 如数据的整合、综合性数据的观测、合成产品的开发以及建立学科交叉科

学的方法途径等, 提高信息的覆盖面、科学质量、实用性和及时性.

地球科学是数据密集型科学^[4,5], 美国地质调查局的科学战略作为一项科学大数据战略. 其核心科学体系以临界带(包括岩石圈、水圈、生物圈和大气圈)为重点研究对象, 以传统地质学、地理学和生物学为基础, 采用数据密集型的工作方法, 实现数据及相关信息的管理和存档, 进而实现数据高效便捷的集成与融合, 促进交叉学科或综合学科的发展, 以解决复杂性的科学问题和社会问题. 其工作方法包括三个层面: 第一个层面是数据获取和保存, 包括数据采集和数据库; 第二个层面是数据操作, 包括数据挖掘和数据分析; 第三个层面是知识发现和归纳, 包括对地球系统的认识、理解和解释. 美国地质调查局的

引用格式: 吴冲龙, 刘刚, 张夏林, 等. 地质科学大数据及其利用的若干问题探讨. 科学通报, 2016, 61: 1797–1807

Wu C L, Liu G, Zhang X L, et al. Discussion on geological science big data and its applications (in Chinese). Chin Sci Bull, 2016, 61: 1797–1807, doi: 10.1360/N972015-01035

核心科学体系战略,是要建立一个把气候变化、土地利用变化、生态系统、能源和矿产、环境健康、自然灾害和水资源等7个科学战略领域无缝整合起来,更有效地解决各种问题的组合式科学框架,以便将数据、科学和技术方法及模型,组织到恰当的时空尺度中去,驱动科学的合成,促进对整个地球系统运转的理解,增强对临界带的科学认识和决策支持。

资源、环境是人类生存和社会发展面临的两大主题。为了应对严峻的挑战,需要在已有成就的基础上,引进大数据理论、方法和技术,开展对地质科学大数据的统合应用,争取地质科学的新发现。为此,本文拟就所涉及的若干问题做些初步探讨。

1 地质科学大数据的特征

地质科学大数据是一种时空大数据,主要产生于基础地质、矿产地质、水文地质、工程地质、环境地质、灾害地质的调查、勘查和相应的地质科学研究过程中,能源、矿产的开发利用和环境、地灾的监测、防治过程中,以及各类地基、空基对地遥感观测活动中。数据采集的具体手段,包括露头地质观测、勘查工程、地球物理探测、地球化学探测、遥感和物理测试、化学分析。因此,地质科学大数据具有科学大数据的显著属性。所有地质科学研究和资源勘查生产的对象,就其形成过程和分布状况而言,均具有庞大的时间与空间范围和复杂的层次结构。所采集的地质数据,在时态特征上有静态和动态之分,在聚集方式上有间歇性集中积累和连续性分散积累之分。其中,各类地质调查、勘查和勘察数据是静态的间歇性集中积累,而各类资源开发、监测和对地遥感观测数据是动态的连续性分散积累。

科学大数据具有高维度(high dimension)、高计算复杂性(high complexity)和高不确定性(high uncertainty)^[5],地质科学大数据是其中典型的一类。由于地质对象演化的时间漫长、空间庞大,各种地质作用影响因素众多,过程曲折反复,而且地质体深埋于地下,存在着“参数信息不完全、结构信息不完全、关系信息不完全和演化信息不完全”的状况,这种高维、高计算复杂性和高不确定性更为显著。地质科学大数据在形式上具有显著的多类、多维、多量、多尺度、多时态和多主题特征,与社会生活和商业活动所产生的大数据有一定差别,但贴合4V特征^[6],即体量大而完整(volume)、类型多且关联(variety)、聚集快却

杂乱(velocity)和价值大但稀疏(value)。地质数据既有结构化的,也有半结构化的和非结构化的,通常呈碎片化状态以文本、图形和图像方式堆积着。例如,大量的野外露头描述数据、钻孔岩芯描述数据和各种地质调查、勘查报告,以及大量地质图件、素描和照片,长期以来是以纸质形式存储和管理的,即便已经建立的众多关系数据库、空间数据库和对象关系数据库,也主要是存储和管理那些表格化和矢量化了的结构化数据,而文字描述、记录和总结等,都是以大字段的文本方式和栅格图形方式入库的,很少进行规范化处理和结构化转换。缺乏一种能够有效地一体化存储和管理结构化、半结构化和非结构化数据的方式。此外,地质科学大数据与社会生活和商业活动大数据的差异,还表现在数据采集和处理方法,以及信息感知、数据挖掘和知识发现的方法上。

地质科学时空大数据的上述特征,决定了其采集、存储、管理、处理和应用方式与方法的特殊性和挑战性。为了充分发挥地质科学大数据的作用,一方面可以借鉴并采用一般大数据技术,如实时数据的采集管理、大数据的存储管理体系架构和分析处理体系架构;另一方面又要根据实际情况研发适用的专业大数据技术,如针对地质时空数据的分布式并行时空索引、顾及时空关系和地质语义的预调度方法等。评判一个传统算法是否适合改造成大数据环境的算法,其评判指标主要包括算法的任务可分解性、数据可分解性和数据流分段关联性。因此,对于传统地质数据处理方法的改造也应该从这三方面着手,尽可能发掘任务的并行性,数据的几何分解性,降低数据流分段关联性,使其适应分布式计算和高性能计算为主的大数据环境,如图1所示。

2 地质科学大数据的管理

地质科学时空大数据在存储管理方面的特殊性,突出表现在以下几个方面。

首先,由于地质体、地质结构、地质资源、地质环境和地质灾害的形成、发展和演变,具有庞大的时空范围和众多的影响因素,为了研究其中的地质作用和成矿、成藏和成灾机理和过程,需要实现对研究区地下-地上、地质-地理、空间-属性数据的一体化采集、存储、管理和处理,以便从系统的观念出发,进行整体分析、关联分析、控制分析和动态分析。其次,由于地质体、地质结构和地质过程的极端复杂性、不

图1 地质大数据技术借鉴与改造技术路线

Figure 1 Technology roadmap of referring and remoulding of big data technologies for geological big data

可见性和数据采集的抽样方式, 导致出现“结构信息不全、关系信息不全、参数信息不全、演化信息不全”的状况, 需要对地质数据进行三维、动态、多尺度、多细节层次的可视化建模, 以便形象、直观地感知地质对象并提高认知能力和水平. 因此, 基于“主题信息管理法、信息交互本体法、信息分析综合法、行为功能模拟法和系统整体优化法”的地质信息科学方法论^[7], 建立能够有效地实现结构化、半结构化和非结构化数据一体化、静态数据与动态数据一体化、地质数据与地质模型一体化存储管理的地质信息系统, 就显得十分的必要和重要.

值得指出, 上述三维可视化地质模型不但是地质体、地质结构及其动态变化的形象表达, 而且是一种地质时空大数据的集成化载体, 即地质数据-地质模型的一体化存储, 而其构建过程, 又在相当大的程度上实现了结构化、半结构化和非结构化数据的一体化, 以及静态数据与动态数据一体化. 因此, 目前世界各国正在大力开展的“城市、矿山、油田和水电工程的三维地质信息系统建设, 以及相应的三维地质建模和“玻璃地球”^[8-10]建设, 正是实现地质科学大数据一体化和集成化存储、管理的重要途径和基本方式.

所谓“玻璃地球”, 是一种由地质信息和地理信息相结合、存储于计算机网络上、可供多用户访问和开展地质、资源和环境决策分析的三维可视化虚拟浅层地壳^[9]. “玻璃地球”是“数字地球”向地下的延伸, 主要差别在于其管理和处理的对象是地下地质体及其属性. “玻璃地球”采用数据与模型一体化存储的构

建方式, 通过整合、集成、管理各类地质数据, 进而全面、系统地描述和表达地壳浅层和深部的地质结构、成分及其空间关系, 同时提供各种地质数据分析与挖掘、地球科学研究、行政决策支持与公众服务. 因此, “玻璃地球”是地质科学大数据的有效载体, 其首要问题是海量多源多尺度地质数据的一体化管理.

从技术角度看, 由于地质体和地质现象时空变化的语义复杂性、过程的非线性和不确定性、变化信息的多维、多尺度和实时性, 而目前所采用的准动态实时GIS时空数据模型, 难以应对高动态监测数据流的存储、管理的需求, 无法支持多传感器接入、多粒度时空变化和时空过程多层次复合的语义表达, 也难以支持相关动态建模和模拟的需要. 为此, 有必要在借鉴实时GIS时空数据模型^[11]的同时, 引入地质时空对象及地质时空对象版本的概念和方法, 提出一种包含时空过程、几何特征、尺度和语义的“多层次、多粒度、多版本的地质时空数据模型”. 然后, 基于可扩展的插件式开发平台系统, 把WebGIS和TGIS技术结合起来, 进而利用云平台、NoSQL和Hadoop技术, 实现对地质时空大数据模型的动态管理. 同时, 借助OPC(过程控制的对象链接和嵌入, Object Linking and Embedding for Process Control)技术(OPC Foundation. OPC Complex Data Specification Version 1.00; <http://www.opcfoundation.org>.2003)来实时接入观测数据, 实现对多维地质时空事件的实时响应. 必要时, 还可以考虑从系统论的角度出发, 采用“多版本时空对象进化数据模型”.

为了实现海量地质时空大数据的高效管理、调度和应用, 还需要发展完善的高效时空索引技术. 在目前的时空数据库中, 通常缺失并行时空索引的一体化与时空索引结构并行化, 严重阻碍了大数据时代时空数据库中分布式并行缓存机制、并行预调度与调度机制、四维时空数据快速检索调度、大规模时空分析等一系列瓶颈问题的有效解决. 因此, 加强具有可并行化结构的时空索引方法的探索研究, 并开展对时空索引的分布式和并行化的一体化研究, 是十分必要的. 近期提出的分布式并行时空索引(DPSI)多层次理论架构和基于间隔关系算子的并行时空索引(IPSI)方法^[12,13], 突破了高维度下树形索引层次结构的局限性, 实现了主从模式下的分布式并行时空索引(MSDPSI), 以及对等模式下的分布式并行时空索引(PPDPSI). 实践结果表明, 该项研究成果显著提升了分布式并行计算环境下的数据并行时空索引性能, 能够推动地质时空大数据时空索引技术的发展.

同时, 根据我国地质工作的部门分工和行政分级体制, 为了有效地存储管理地质科学大数据, 还需要建立分布式的数据云存储、云管理和云服务体系.

3 地质科学大数据的挖掘

数据挖掘是从大数据中发现知识的主要方法^[14,15]. 地质科学的时空大数据挖掘, 是从地质数据库或数据仓库中寻找并抽取隐含的知识和时空关系, 发现有用的特征和规律的一种理论、方法和技术, 在地质规律研究、成矿预测、资源评价、环境保护和地质灾害防治领域都有重要的应用. 大数据技术以数据为中心, 在全体数据中挖掘知识, 可突破采样随机性和样本空间狭小, 以及仅凭少量观测数据和模式进行判断的限制, 而且能基于大型数据库和数据仓库来实现. 这无疑是一个巨大的进步. 数据挖掘与数据分析的差别在于, 所使用的数据量从MB级或GB级转化为TB级甚至PB级, 数据类型从结构化数据扩大到文本、视频、图片等半结构化或非结构化数据, 评判前提从有假设条件过渡到没有明确假设条件, 所发现的知识具有先前未知性、有效性和实用性.

3.1 时空数据挖掘的基本内容

(i) 时空数据挖掘的任务. 时空数据挖掘的基本任务, 是在不同的时间与空间概念层次挖掘出各种类型的知识, 并用相应的知识模型表示出来, 然后

根据所采用的知识表示方法设计出相应的推理模型, 为不同领域、不同层次、具有不同应用需求的用户提供行之有效的辅助决策支持. 李德仁等人^[16]曾把知识表示方法概括为: 基于规则的表示法、基于逻辑的知识表示、基于关系的知识表示、基于模型的知识表示、基于本体的知识表示、面向过程的知识表示、面向对象的知识表示, 以及语义网络表示、脚本表示和模拟表示, 等等.

具体地说, 就是在地质时空数据库和数据仓库的基础上, 利用统计学、模式识别、人工智能、集合论、模糊数学、云理论、机器学习、可视化等相关技术和方法, 以及各种相关信息技术手段, 从海量多类多层次的时空数据、属性数据中提取未知的、有用的和可理解的可靠知识, 从而揭示出蕴含在地质科学大数据背后的相关关系和演化趋势, 实现新知识的自动或半自动获取, 为资源预测、发现和环境评价、减灾提供决策依据. 简言之, 就是直接从地质时空数据库和数据仓库中发现知识, 并提供相关的决策支持.

(ii) 时空数据挖掘的内容. 地质时空大数据挖掘的内容, 就是从地质数据库和数据仓库中可能发现的知识类型. 根据一般空间数据挖掘理论^[16], 地质时空大数据挖掘的内容有: (1) 空间几何知识, 例如几何度量及其统计规律; (2) 时空特征规则, 例如特征出现的位置与拓扑关系; (3) 时空分布规律, 例如时间节律、分带性及其组合规律; (4) 时空趋势规则, 例如变化趋势及其控制因素; (5) 时空异常规则, 例如地质异常及其控制因素; (6) 时空分形规则, 例如奇异值、分形与多重分形; (7) 面向对象的知识, 例如对象子类中的普遍性特征; (8) 时空协同定位规则, 例如特定对象的布尔空间子集; (9) 时空分类规则, 例如空间实体特征、变化和属性差异; (10) 时空聚类规则, 例如时间、空间和状态聚类; (11) 时空关联规则, 例如对象实体时空共生条件及拓扑关系; (12) 时空依赖规则, 例如实体间或属性间函数依赖关系及变化; (13) 时空区分规则, 例如多维时空中实体分布与区分; (14) 时空演化规则, 例如空间目标随时间的变化; (15) 时空预测规则, 例如属性特征的时空差异性预测; (16) 时空决策规则, 例如目标时空特征和行为的应对规则.

3.2 时空数据挖掘的基本方法

李德仁等人^[16]曾把时空数据挖掘的理论方法归纳为: 确定集合论、扩展集合论、仿生学、可视化和

决策树等五种类型。时空数据挖掘方法的大量涌现,使得地质科技人员有可能从传统的数据分析向数据挖掘发展,即直接从地质数据库或数据仓库中进行知识发现。但是,目前数学地质领域常用的一系列统计学方法^[17-19]、分形与多重分形方法^[20]和一般空间数据挖掘方法^[16],经过筛选和适当改造可用于地质科学时空大数据的挖掘。这些方法的改造所面对的难题和挑战,主要在于对地质对象的多尺度、多时态、多参数、多模态和不确定性特征的适应性、预处理和升维问题。其中的适应性是指数据挖掘从基于一般数据库和数据仓库,转为基于GIS,以及由基于GIS转为基于三维地质模型;预处理是指对大数据集中的错误值或进行数据清洗、数据转换、连续数据的离散化、空值的替代、数据子集的随机抽取和预挖掘模型的建立;升维是指海量数据的计算模型,随实体模型由一维转为二维,二维转为三维,甚至由三维转为四维。这些问题需要通过深入的研究加以解决,下面拟结合数学地质领域的研究成果,对可用于地质科学大数据挖掘的常用方法做一些补充说明。

(i) 基于概率论的数据挖掘方法。这是一种针对属性空间分布的随机性和不确定性的数据挖掘方法。在地质对象空间中,各种地质作用、地质过程和地质现象,及其所呈现的、可被感知的特征,在很大程度上受概率法则的支配和影响。在数学地质领域中常用的各种数理统计方法,都可利用来进行空间数据挖掘。其中包括:回归分析法、因子分析法、判别分析法、聚类分析法、证据权重法、趋势分析法、时间序列法,以及克立格分析法,等等。

(ii) 基于扩展集合论的数据挖掘方法。地质空间对象的特征在许多方面具有显著的不确定性,仅有基于传统集合论的各种数据挖掘方法,还不能满足要求。因此,需要发展一系列基于扩展集合论的空间数据挖掘方法。近年来,这方面的研究成果比较多,诸如模糊数学、粗糙集理论和云模型等。这些新的理论与方法,在地质空间数据挖掘方面的应用有很好的前景。

(iii) 基于仿生学的数据挖掘方法。这是模仿生物的智能行为和遗传规律的一种空间数据挖掘方法。已有的实践表明,这类方法对于从复杂的地质数据中发现潜在的知识,实现对复杂地质结构和地质过程的认知,具有重要的价值。主要方法有:人工神经网络法、蚁群算法和演化算法。

(iv) 其他定量数据挖掘方法。在利用数据挖掘进行地质异常定量识别和圈定时,还常用如下计算方法:地质复杂系数(C 值)法、组合熵(H 值)法、地质相似系数(S 值)法、分维(F 值)法、地质关联度(R 值)法、混沌特征提取法、块褶积滤波、计算几何、多重分形、决策树和机器学习方法,等等。

(v) 文本数据的挖掘方法。文本数据挖掘简称为文本挖掘(Text Mining, TM),是数据挖掘领域的一个重要分支。它以非结构化或半结构化的文本数据作为挖掘对象,利用定量计算和定性分析法,从中寻找信息的结构、模型、模式等具有潜在价值的知识的过程^[21]。海量的定性描述和文字记录的存在,是地质数据的一大特点。在以往的地质规律研究和资源预测评价中,通常只从这些描述性数据中提取少量的“特征参数”,并由此建立各种认知模式,然后凭借模式进行判断和预测。文本挖掘可以在全部描述性记录中发现知识,可突破数据量和认知模式的限制,有可能全面揭示文本中隐含的地质时空结构和地质因素间的相关关系,发现新的知识。目前,文本地质数据挖掘已有成功应用实例^[22],但其理论、方法和技术还不成熟。由于普遍存在着多时态、多尺度和多主题特性,在一个文本数据集中往往隐藏着多种特征信息和多种主题知识,需要综合应用多种挖掘技术,才能从中获取完整的知识。

(vi) 可视化法。采用数据可视化方式来进行数据挖掘,主要是用于直观而形象地将空间数据所隐含的特征显示出来,帮助人们通过视觉分析来寻找其中的结构、特征、模式、趋势、异常现象或相关关系等空间知识。为了确保这种方法行之有效,必须设置功能强大的三维可视化工具和辅助分析工具,其中包括矢量剪切工具和模型透视工具。

近期,时空数据挖掘已经开始从单机、单库挖掘向在线挖掘发展^[19]。在线数据挖掘以多维视图为基础,强调通过网络执行效率和对用户命令的及时响应;以时空数据仓库为直接数据源,通过在线时空数据查询及与OLAP、决策分析、数据挖掘等分析工具的配合,可完成时空信息提取和知识发现。其中,海量地质时空数据网络传输、互操作和深度挖掘技术,亟待进一步研发和应用。

4 地质科学大数据的可视化

在进行地质现象和地质过程分析、地质矿产资源

评价和开发利用决策时,对于大量的不确定因素,要依靠研究人员本身的知识 and 经验进行定性理解、定量估算和关系描述,并结合时空数据模型和时空分析模型来进行分析、预测、评估和辅助决策.从数学逻辑角度看,这是一种半结构化或不良结构化甚至非结构化问题,而数据可视化正是描述、表达和理解各种半结构化甚至非结构化问题的关系和模型的最佳方法和手段.正因为如此,地质科学领域比地理科学领域,更加强调实现“体三维”可视化,即实现地质体内部的结构和成分的三维可视化.由于地质工作对象的特殊性,地质数据可视化所涉及的内容也更为丰富,而表现形式也更为复杂.

地质时空大数据的可视化,集中体现在多维地质模型构建和“玻璃地球”^[23]建设上.“玻璃地球”的关键技术是地质体和地质结构的三维动态可视化建模.地质时空数据可视化从应用角度可分为表达三维可视化、分析三维可视化、过程三维可视化、设计三维可视化和决策三维可视化等五类^[24](图2).作为地质大数据的载体,“玻璃地球”只有实现了五类可视化,才能够具有支持大数据挖掘的价值.

其中,表达可视化泛指原始数据和计算成果以图形或图像形式在屏幕或其他介质上的显示,是空间决策支持认知过程可视化的基础,贯穿于其他各类可视化之中;分析可视化泛指在可视化环境中进行的各种地质空间决策分析,是空间决策支持认知过程可视化的核心;过程可视化是指在体三维环境中开展各种地质过程的可视化动态模拟,以及地质作用的可视化虚拟仿真,因而是使三维静态地质模型转变为四维动态地质模型的关键步骤;设计可视化是指在体三维可视化环境中进行各种地质工程设计,是使地质工程设计从二维方式转变为三维方式的基础;决策可视化是指在体三维乃至四维可视化环境中,进行矿产资源潜力或工程地质条件评价、矿产资源开发和地质工程设计的多方案比较选优,以及地质灾害和污染事件的预警、防治决策和应急预案的制定,等等.

目前,上述各种面向对象、面向过程、具有空间认知能力的可视化技术,在国际上发展很快,在国内也已经开始应用于区域地质调查、城市地质调查、工程地质勘察设计、矿产和水文地质勘查、矿山和油田资源开发、矿权管理、储量核算、水利水电工程和地质灾害勘察治理等专业领域的决策支持中^[25-31],显著地提高了地矿资源的分析、评价、管理和辅助决策

水平.今后的发展方向,是研发三维动态显示技术和三维透视技术,提供具有沉浸感、动态、交互的环境和工具,进一步开展盆地和造山带地质过程分析、工程地质条件和地质灾害危险性评价、能源和矿床形成演化分析,以及盆地和造山带地质过程、油气成藏和金属成矿过程和地质灾害孕育发生过程的三维动态模拟和仿真.进而,把基于各类地质调查、勘查和勘察数据所构建的三维静态地质模型,与那些来自各类资源开发、环境监测和对地遥感观测的动态地质模型耦合起来,再加上传感器、互联网、云技术和智能计算,使“玻璃地球”和“玻璃国土”,向“智慧地球”和“智慧国土”转化.

5 地质科学大数据的利用

大数据的出现并存在着,是一种客观事实.大数据概念的提出,代表一个信息时代、一个思维方式和一个技术潮流^[5].其特点^[32]在于,所处理的数据集合“不是随机样本,而是全体数据”,所容许的数据品质“不是精确性,而是混杂性”,所揭示的数据内涵“不是因果关系,而是相关关系”.这三个特点也是长期困扰地质资源、地质环境和地质灾害预测、评价、管理、监控、预警和决策领域的难题和难点.大数据理论、方法和技术的引进,对于突破采样随机性和样品空间狭小、大量良莠难分的非结构化和半结构化数据无法利用,以及可靠的作用机理、因果关系和动力学模型缺乏,仅凭少量观测数据和固有模式进行判断、预测等限制,无疑有极大的好处.

以金属矿产成矿预测为例.这是找矿勘探的一个必要步骤,对于降低投资风险,提高命中率有重要意义.矿床的类型繁多,其形成分别与一定的地质背景和地质作用相联系.地质背景包括地壳区域构造单元和地层单元、深部的上地幔和岩石圈状况和地表的古地理、古气候环境;地质作用包括构造作用、岩浆作用、沉积作用、变质作用、风化作用.这些地质背景和地质作用在广阔的空间范围内,在漫长的地质历史中,是动态变化和反复叠加的,只有在多种条件都有利的情况下才能形成大型矿床.所形成的矿床同样会经历多期次的改造,或者进一步叠加成矿,或者遭受破坏.长期以来的科学研究和经验积累,形成了一个矿床学和成矿预测学学科,专业人员都是在一定的理论和方法指导下,凭借已有的知识和经验并采用定性或定量的方法进行预测找矿.

图2 多维多尺度“玻璃地球”建设与五个可视化

Figure 2 Five visualization technologies for the construction of multi-dimensional and multi-scale Glass Earth

最初,无论是定性或定量方法,都是根据典型矿床的勘探资料,从成矿规律研究揭示的因果关系理论出发,抽提出若干个特征性的“找矿标志”,形成一种“成因模型”或“成矿模式”,然后用这种“模型”进行

矿床预测. 这种方法是行之有效的,人们利用这种方法找到了许多重要的矿床. 但是,随着浅表的、易于发现的矿床陆续被找到后,这些“成因模型”的局限性也逐步显露出来了,其预测的准确率不断降低. 往

往出现这样的情况,地质专业队伍在一些地方根据露头观测、物探、化探和遥感,甚至配合少量钻探,识别出了多种“找矿标志”,却始终没有找到预期的矿床.而在另一些地方,地质专业队伍在十几年乃至几十年内几进几出,花费了巨大的人力物力和财力却一无所获,最后又因为一个偶然的机会有了重大发现,取得了重大的突破.这种情况说明了,已有的理论、知识和“模型”存在着不足或缺陷.

为了弥补这些不足或缺陷,人们开始从各自掌握的勘探成果中进行总结,于是出现了大量新的矿床“成因模型”,甚至每个矿床都有自己的“成因模型”.结果可想而知,“成因模型”的价值受到了质疑,许多人又回到了由经验构成的“客观模型”或“存在模型”——用系列剖面图和平面图客观地展示矿床赋存的位置、形态和范围,围岩的岩性、岩相和分布,以及蚀变的类型和分布.这种“存在模型”对于为了扩大已知矿床范围而进行的就矿找矿,无疑是有效的.但对于广阔的未知区而言,这种方法显得无从下手.在这种情况下,追求“相关关系”而不追求“因果关系”的矿床统计预测方法^[17]和多重分析预测方法^[19]等被提上了议事日程.人们还发现了一种特殊规律,即矿床平均品位越高者规模越小,而平均品位越低者规模越大,并通过统计分析得出了一系列矿种的“品位-吨位模型”,并用于矿床预测中.

随后的研究表明,矿床是地质异常的产物^[33]、矿床在混沌边缘分形生长^[34],求异和相似类比相结合,传统的成矿动力学和混沌动力学相结合,是进一步解决成矿预测问题的可能途径.类似于郭华东等人^[5]对于“胡焕庸线”形成机理的空间认知研究所指出的,这类问题的解决“势必需要建立一个巨型复杂非线性模型系统,其中包括各种过程的机理模型,各种模型的相互耦合,各种尺度上的模型相互作用,以及极其复杂的模型触发、传递、反馈机制”.反观以往的预测方法和预测模型,都是根据已有知识和经验,采用少量的特征参数构成的,通过露头地质观测、物探、化探、遥感和钻探所获得的海量记录数据,多数被过滤掉了.由于既有知识的不完善,就造成了预测成功率不高的结果.特别是面对找矿的新领域、新类型和新深度,可用于建模的相关知识更加有限.为了发现新的知识,认识新的成矿规律,我们需要面对的是全部的原始记录数据,而不是人为抽取的少量特征数据.这样我们所要处理的将是具备超高维度、超高计

算复杂性和超高不确定性的多尺度、多变量、多时态、多相关的空间和属性大数据.因此,成矿规律和成矿预测问题的进一步解决,必然要依托于“玻璃地球”承载的地质科学大数据系统,并基于第四科学范式进行研究,即在“玻璃地球”平台上对海量的地质时空数据和矿产资源开采数据进行虚拟汇聚和预处理,分别对各种矿床类型进行相关数据时空的表征和建模,然后采用适宜的大数据技术进行挖掘和分析.

除了金属矿产资源的预测、评价,地质时空大数据的可能利用方向还包括:① 油气资源预测、评价;② 油气、矿产和水资源开发监管与保护;③ 降水及海水入侵智能监测与动态模拟;④ 地下水污染智能监控与污染源追踪;⑤ 工程地质条件及其变化智能评价与管控;⑥ 地下工程与地质体相互作用智能感知与测控;⑦ 地下建筑安全监测与预警;⑧ 地质灾害孕育智能感知、监测与预警;⑨ 土地和土壤智能监管、决策等等.

长期连续不断的地质勘查工作,积累了海量的基础地质数据.但多年来地质资料服务对象和服务领域主要是集中在地矿行业和传统的地勘行业(冶金、有色、煤炭、石油等),进入21世纪以后,地质工作与经济建设的关系比以往任何时期更为密切,地质工作渗透到了社会经济生活的各个方面.农业地质、城市地质、灾害地质、环境地质、旅游地质等新型领域的出现,极大地拓展了服务领域,与此同时,社会公众对地质信息的需求,也越来越迫切.为了满足社会对地质资料的需求,中国地质调查局开展了“地质云”^[35,36]建设.所谓“地质云”,就是充分利用历史的和现势的点源地质、地理、矿产、环境、地灾、地球物理、地球化学、遥感等结构化和非结构化数据,以应用和服务为主线,以新知识发现为目的,在数据采集、数据整合、数据传输、数据挖掘、知识发现等相关技术研发与集成基础上,基于专业局域网环境建立的分布式地质大数据交换与服务共享平台,即地质云平台.目前的地质云平台构建,是基于云架构体系对已有的中国地质调查信息网格平台进行升级,即在网格技术、云计算技术综合研究的基础上,对中国地质调查信息网格平台的功能进行抽象、分类、细化和重用性包装^[37].在“地质云”平台的框架下,以中国地质调查局发展研究中心、中国地质调查局各专业中心、大区中心为骨干结点,建成地质调查信息服务集群体系;以全国地质资料馆、省级地质资料馆为骨

干结点, 建成地质资料服务集群体系; 开展数据资源整合集成, 形成信息资源合理布局、服务分级、上下联动的地质资料信息服务新局面。其中包括引进和应用云GIS技术, 以期获得大规模计算能力, 解决海量空间数据的分布式存储、处理任务划分、查询检索、互操作、虚拟化等关键问题。

地质时空大数据统合和利用, 涉及科学研究的第四范式^[38]和地质数据科学^[39]的一系列理论、方法和关键技术。其中包括: 建立地质时空大数据的一体化空间参考体系; 进行数据的空间基准、时态、尺度和语义的一致性处理和数据融合; 探索对各类静态地质勘查数据的进行集成化、结构化、可视化转换, 并与各类分布式动态地质观测数据进行一体化存储、管理的“玻璃地球”方式; 研究地质时空大数据存储与智能处理和数据挖掘技术。

6 认识与结论

基于上述分析, 可以得到以下几点基本认识:

(1) 地质科学大数据是一种时空大数据, 既具有一般大数据的“4V”特征, 又有独特的性质。开展地质时空大数据的统合与利用, 是积极应对社会可持

续发展所面对的资源、环境新形势、新挑战的需要。

(2) 以密集型数据为基础, 在全体数据中挖掘知识, 可以突破采样随机性和样本空间狭小, 以及仅凭少量观测数据和模式进行判断的限制, 而且能基于大型数据库和数据仓库来实现。这对于地质资源、地质环境和地质灾害的预测、评价有重大意义。

(3) “玻璃地球”是地质时空大数据的有效载体, 开展“玻璃地球”建设, 能够使非结构化、半结构化的碎片式地质数据, 向结构化、集成化转变, 并实现数据与模型的一体化存储、管理和可视化, 提升对复杂地质体、地质结构和地质过程的认知能力。

(4) 地质时空大数据的统合和利用, 涉及一系列理论、方法和技术问题。主要包括地质-地理的一体化空间参考体系; 静态结构勘查模型与动态变化监测模型的耦合; 数据挖掘、数据融合和地质大数据智能化处理技术。

(5) 地质时空大数据利用的领域十分宽广。采用大数据技术可以有效地揭示隐藏在海量地质时空数据背后的结构、特征和相关关系等未知的知识。基于大数据技术, 只要获取并掌握海量的地质时空数据, 就可能导致地质科学重大新发现。

参考文献

- 1 Kempthorne D, Myers M D. Facing Tomorrow's Challenges—U.S. Geological Survey Science in the Decade 2007–2017. In: U.S. Geological Survey Circular 1309. Reston, Virginia: U.S. Geological Survey, 2007
- 2 Linda G, Belnap J, Goldhaber M, et al. Geology for a Changing World 2010–2020—Implementing the U.S. Geological Survey Science Strategy. In: U.S. Geological Survey Circular 1369. Reston, Virginia: U.S. Geological Survey, 2011
- 3 Bristol R S, Euliss N H, Booth N L, et al. Science Strategy for Core Science Systems in the U.S. Geological Survey, 2013–2023. Reston, Virginia: U.S. Geological Survey, 2012
- 4 Zhao P D. Big data era calls for data science in various fields of science (in Chinese). Chin Sci Tech Award, 2014, 183: 29–30 [赵鹏大. 大数据时代呼唤各科学领域的科学数据. 中国科技奖励, 2014, 183: 29–30]
- 5 Guo H D, Wang L Z, Chen F, et al. Scientific big data and digital earth (in Chinese). Chin Sci Bull, 2014, 59: 1047–1054 [郭华东, 王力哲, 陈方, 等. 科学大数据与数字地球. 科学通报, 2014, 59: 1047–1054]
- 6 Mayer-Schönberger V, Cukier K. Big Data: A Revolution that Will Transform How We Live, Work and Think. London: John Murray, 2013
- 7 Wu C L, Lin G, Tian Y P, et al. Introduction to Geological Information Science and Technology (in Chinese). Beijing: Science Press, 2014. 18–21 [吴冲龙, 刘刚, 田宜平, 等. 地质信息科学与技术概论. 北京: 科学出版社, 2014. 18–21]
- 8 Carr G R, Andrew A S, Denton G J, et al. The “Glass Earth”—Geochemical frontiers in exploration through cover. Aust Inst Geosci Bull, 1999, 28: 33–40
- 9 Wu C L, Weng Z P, Liu G, et al. Discussion on “Glass Land” of China (in Chinese). Geol Sci Tech Inf, 2013, 31: 1–8 [吴冲龙, 翁正平, 刘刚, 等. 论中国“玻璃国土”建设. 地球科技情报, 2013, 31: 1–8]
- 10 Wu C L, Liu G. Current situation, existent problems, trend and strategy of the construction of “Glass Earth” (in Chinese). Geol Bull China, 2015, 34: 1281–1287 [吴冲龙, 刘刚. “玻璃地球”建设的现状、问题、趋势与对策. 地质通报, 2015, 34: 1281–1287]
- 11 Gong J Y, Li X L, Wu H Y. Spatio-temporal data model for real-time GIS (in Chinese). Acta Geod Cartogr Sin, 2014, 43: 226–232 [龚健雅, 李小龙, 吴华意. 实时GIS时空数据模型. 测绘学报, 2014, 43: 226–232]
- 12 He Z W, Kraak M-J, Huisman O, et al. Parallel indexing technique for spatio-temporal data. ISPRS-J Photogramm Remote Sens, 2013, 78: 116–128

- 13 Zheng Z F. Distributed parallel spatio-temporal index techniques (in Chinese). Doctor Dissertation. Wuhan: China University of Geosciences, 2014 [郑祖芳. 分布式并行时空索引技术研究. 博士学位论文. 武汉: 中国地质大学, 2014]
- 14 Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3rd ed. Burlington: Morgan Kaufmann, 2012
- 15 Fayyad U M. Advances in Knowledge Discovery and Data Mining. Menlo Park CA: AAAI/MIT Press, 1996
- 16 Li D R, Wang S L, Li D Y. Spatial Data Mining Theories and Applications (in Chinese). 2nd ed. Beijing: Science Press, 2013 [李德仁, 王树良, 李德毅. 空间数据挖掘理论与应用. 第二版. 北京: 科学出版社, 2013]
- 17 Zhao P D, Li Z J, Hu W L. Statistical Prediction of Mineral Deposit (in Chinese). Beijing: Geological Publishing House, 1983 [赵鹏大, 李紫金, 胡旺亮. 矿床统计预测. 北京: 地质出版社, 1983]
- 18 Agterberg F. Geomathematics: Theoretical Foundations, Applications and Future Developments. Heidelberg: Springer, 2014
- 19 Zhao P D, Wu C L, Guo T L, et al. Oil & Gas Geological Anomaly Analysis and Evaluation of the Lower Paleozoic in South China (in Chinese). Beijing: Science Press, 2010 [赵鹏大、吴冲龙、郭彤楼, 等. 中国南方下古生界油气地质异常分析与评价. 北京: 科学出版社, 2010]
- 20 Cheng Q M. A new technique for quantifying anisotropic scale invariance and for decomposition of mixing patterns. *Math Geol*, 2004, 36: 345–360
- 21 Feldman R, Dagan I. Knowledge discovery in textual data-bases (KDT). In: Usama M F, Ramasamy U, eds. Proceeding of the First International Conference on Knowledge Discovery and Data Mining (KDD-95). Montreal: AAAI, 1995. 112–117
- 22 Chen J P, Hou C B, Wang G W, et al. Research on text data mining in quantitative assessment of mineral resources (in Chinese). *Comp Tech Geoph Geoch Expl*, 2005, 27: 263–266 [陈建平, 侯昌波, 王功文, 等. 矿产资源定量评价中文本数据挖掘研究. 物探化探计算技术, 2005, 27: 263–266]
- 23 Gore A. The digital Earth: Understanding our planet in the 21st century. *Aust Surveyor*, 1998, 43: 89–91
- 24 Wu C L, He Z W, Weng Z P, et al. Property, classification and key technologies of three-dimensional geological data visualization (in Chinese). *Geol Bull China*, 2011, 30: 642–649 [吴冲龙, 何珍文, 翁正平, 等. 地质数据三维可视化属性、分类和关键技术. 地质通报, 2011, 30: 642–649]
- 25 Wu C L, Mao X P, Wang X P, et al. Model construction and software development of 3D hydrocarbon pool-forming dynamics (in Chinese). *Petrol Geol Exp*, 2001, 23: 301–311 [吴冲龙, 毛小平, 王燮培, 等. 三维油气成藏动力学建模与软件开发. 石油实验地质, 2001, 23: 301–311]
- 26 Wu C L, Niu R Q, Liu G, et al. Construction aim and solution of the urban geological information system (in Chinese). *Geol Sci Tech Inf*, 2003, 22: 67–72 [吴冲龙, 牛瑞卿, 刘刚, 等. 城市地质信息系统建设的目标与解决方案. 地质科技情报, 2003, 22: 67–72]
- 27 Wu L X, Yin X R, Zhong Y P. Restudy on digital mine: Characteristics, framework and key technologies. *J China Coal Soc*, 2003, 28: 1–7 [吴立新, 殷作如, 钟亚平. 再论数字矿山: 特征、框架与关键技术. 煤炭学报, 2003, 28: 1–7]
- 28 Wu C L, Tan Z H, Li W Z, et al. Research and development of geohazard: A geological hazard exploration information system of the Three Gorges reservoir area (in Chinese). *Hydrogeol Eng Geol*, 2006, 123–128 [吴冲龙, 谭照华, 李伟忠, 等. 三峡库区地质灾害勘察点源信息系统的研发. 水文地质工程地质, 2006, 123–128]
- 29 Zhang X L, Wu C L, Weng Z P, et al. Research and application of the digital mine software QuantyMine (in Chinese). *Earth Sci*, 2010, 35: 303–310 [张夏林, 吴冲龙, 翁正平, 等. 数字矿山软件(QuantyMine)若干关键技术的研发和应用. 地球科学, 2010, 35: 303–310]
- 30 Zhang Z T. Research on 3D geological modeling and three-dimensional process visualization technology of basin fault block structure (in Chinese). Doctor Dissertation. Wuhan: China University of Geosciences, 2011 [张志庭. 盆地断块构造三维建模与过程可视化技术研究. 博士学位论文. 武汉: 中国地质大学, 2011]
- 31 Li Z L. Research on the intelligent dynamic construction method of fine 3D ore body model (in Chinese). Doctor Dissertation. Wuhan: China University of Geosciences, 2011 [李章林. 精细三维矿体模型智能化动态构建方法研究. 博士学位论文. 武汉: 中国地质大学, 2011]
- 32 Mayer-Schönberger V, Cukier K. Big data: A revolution what will transform how we live, work, and think (in Chinese). Hangzhou: Zhejiang People's Publishing House, 2013. 261 [维克托·迈尔-舍恩伯格, 肯尼斯·库克耶, 著. 盛杨燕, 周涛, 译. 大数据时代: 生活、工作与思维的大变革. 杭州: 浙江人民出版社, 2013. 261]
- 33 Gray J. On eScience: A transformed scientific method. In: Hey T, Tansley S, Tolle K, eds. The Fourth Paradigm: Data-intensive Scientific Discovery. Washington DC: Microsoft Press, 2009
- 34 Zhao P D, Chi S D, Chen Y Q. Finding out the geological anomalies: A basis of metallogenic prognosis (in Chinese). *Geol J China Univ*, 1996, 2: 361–373 [赵鹏大, 池顺都, 陈永清. 查明地质异常: 成矿预测的基础. 高校地质学报, 1996, 2: 361–373]
- 35 Yu C W. Fractal Growth of Mineral Deposits at the Edge of Chaos (in Chinese). Hefei: Education Press, 2006 [於崇文. 矿床在混沌边缘分形生长. 合肥: 教育出版社, 2006]
- 36 Chen J P, Li J, Cui N, et al. The construction and application of geological cloud under the big data background (in Chinese). *Geol Bull China*, 2015, 34: 1260–1265 [陈建平, 李婧, 崔宁, 等. 数据背景下地质云的构建与应用. 地质通报, 2015, 34: 1260–1265]
- 37 Wang X, Li J C, Chen H, et al. Big data and geological data information services. *Geol Bull China*, 2015, 34: 1309–1315 [王翔, 李景朝, 陈辉, 等. 大数据与地质资料信息服务: 需求、产品、技术、共享. 地质通报, 2015, 34: 1309–1315]
- 38 Lü X, Li J Q, Gong A H, et al. Study and implementation of the key technologies of China Geological Survey information grid platform based on cloud architecture. *Geol Bull China*, 2015, 34: 1323–1332 [吕霞, 李健强, 龚爱华, 等. 基于云架构的中国地质调查信息网格平台关键技术研究. 地质通报, 2015, 34: 1323–1332]
- 39 Zhao P D. Digital geology and mineral resources evaluation (in Chinese). *J Geol*, 2012, 36: 225–228 [赵鹏大. 数字地质与矿产资源评价. 地质学刊, 2012, 36: 225–228]

Discussion on geological science big data and its applications

WU ChongLong^{1,2}, LIU Gang^{1,2}, ZHANG XiaLin¹, HE ZhenWen¹ & ZHANG ZhiTing¹

¹ State Key Laboratory of Geological Processes and Mineral Resources, School of Computer Science, China University of Geosciences, Wuhan 430074, China;

² Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430074, China

The geological big data is a kind of spatio-temporal big data whose characteristics has both similarities and marked differences with the general big data. Due to long-term evolution of geological objects with huge space and influence of various geological process, its development undergoes a complicated process. Geological bodies are buried in deep underground so that geological data has the feature of incomplete information of parameter, structure, relationship and evolution, which has high computation complexity with high dimension and significant uncertainty. In order to address those challenges, the theories, methods and techniques of big data should be introduced to integrate and utilize geological science big data. It is necessary to set up unified spatial reference system for geological spatio-temporal big data to realize data consistency processing and fusion with regards to spatial standard benchmark, tense, scale and semantics. Various kinds of integrated storage and management methods for static geological exploration data and dynamic distributed geological observation data will be developed. On the one hand, current big data techniques can be adapted, such as realtime data storage and management, big data storage and management architecture, and analysis and processing architecture. On the other hand, specialized big data techniques based on actual situation should be developed, such as distributed parallel spatio-temporal indexing for geological data, and pre-dispatch method considering spatio-temporal relationship and geological semantics. To judge a traditional algorithm whether is suitable for being transformed into big data environment, main criteria include algorithm task decomposability, data decomposability and data flow segmented relevance. Hence, the transformation of traditional geological data processing methods will be from these three aspects to excavate the task parallelism, geometry decomposability and reduce the data flow segmented relevance as much as possible to adapt to the distributed computing and high performance computing of big data environment. The big data technique, which may be a breakthrough in geological science, can mine knowledge from massive spatio-temporal and text data directly in spite of the limits in geological sampling methods such as the randomness and sparsity of sample space, the judge relied on inadequate observational data and fixed mode, and traditional data analysis methods. A series of mathematical geology methods and spatial data mining methods which are frequently used nowadays can be used in geological spatio-temporal big data mining after being remolded. The key science and technology issues referred in geological science development in big data era can be listed as follows—the integrated storage management and handling of structured, semi-structured and unstructured data, big and small data, hybrid and precision data, model and data, static exploration model and dynamic monitoring model, the combination of data mining and data analysis, the unification of correlation and causality, deep mining and visualization of geological science big data, and so on. Taking metal mineral metallogenic prediction as an example, geological big data utilization and analysis are studied. In terms of a global view, Glass Earth is the underground extension of Digital Earth. Glass Earth is a kind of 3D visualized virtual shallow crust, which is an aggregation of geological and geographical information and stored on the computer network to be accessed by multi-user and provide analysis service and decision making to the research and applications of geology, resources and environment. The Glass Earth is the effective vehicle of geological science big data and the construction of the Glass Earth is one of the best ways to solve all the aforementioned scientific and technical problems.

big data, geological spatio-temporal data, Glass Earth, cloud technique, data mining

doi: 10.1360/N972015-01035