地质调查大数据研究的主要问题分析

严光生[1,2], 薛群威[2], 肖克炎[3], 陈建平[4], 缪谨励[2], 余海龙[2]
YAN Guangsheng[1,2], XUE Qunwei[2], XIAO Keyan[3], CHEN Jianping[4], MIAO Jinli[2], YU Hailong[2]

1.中国地质调查局,北京 100037;
2.中国地质调查局发展研究中心,北京 100037;
3.中国地质科学院矿产资源研究所,北京 100037;
4.中国地质大学(北京),北京 100083
1. *China Geological Survey, Beijing* 100037, *China;*
2. *Development and Research Center, China Geological Survey, Beijing* 100037, *China;*
3. *Institute of Mineral Resources, Chinese Academy of Geological Sciences, Beijing* 100037, *China;*
4. *China University of Geosciences (Beijing), Beijing* 100083, *China*

摘要:地质调查大数据包含地质调查工作中产生的多来源、多模态地质数据,以及公共服务与支撑管理产生的数据。一些与数据和计算有关的地质问题,限于当时的信息技术条件,没有得到很好的解决,解决这类地质问题及信息数据共享问题是地质调查大数据处理技术的基本目标。在地质调查大数据处理技术中,应当积极开展多类型地质数据采集器、新型非易失性存储技术、分布式计算、内存计算技术产品开发与应用,然后集中开展、深度分析与挖掘、可视分析技术产品开发与应用,最终形成地质调查大数据处理技术体系与产品线,以产品应用推动资源共享,提升地质调查信息化服务品质。

关键词:地质调查;大数据;地质数据采集器;分布式计算;内存计算;深度挖掘
中图分类号:P628 **文献标志码:**A **文章编号:**1671-2552(2015)07-1273-07

Yan G S, Xue Q W, Xiao K Y, Chen J P, Miao J L, Yu H L. An analysis of major problems in geological survey big data. *Geological Bulletin of China*, 2015, 34(7):1273-1279

Abstract: Geological survey big data include geological data produced by multi-source and multi-mode from geological survey as well as data from public service and management support. Under the technical condition of limited information, some geological problems based on data and computation have not yet been solved perfectly. The resolution of these problems and the sharing of data constitute basic objectives of the processing technique of geological survey big data. In the processing technique of geological survey big data, geological data acquisition unit, non-volatile memory, distributed computation and in-memory computing should be applied first, followed by deep discovering and visual analytics. The architecture and product line for the processing technique of geological survey big data will eventually ensure informationization service quality of geological survey.

Key words: geological survey; big data; geological data acquisition unit; distributed computation; in-memory computing; deep discovering

地质调查的过程同时是地质调查信息的处理过程。信息科技60年迅猛发展提升着地质信息处理能力,影响着地质调查面貌,渗透到地质调查思维。

1 数据的产生

地质调查是人们对地球表层有目的的探测与探索。从数据角度看,地质调查以多来源、多模态

收稿日期:2015-02-03;**修订日期:**2015-06-11
资助项目:国家自然科学基金项目(批准号:41372066)和中国地质调查局项目(编号:1212011121243)
作者简介:严光生(1963-),男,博士,研究员,从事矿产资源评价和空间地球化学研究。E-mail: yguangsheng@mail.cgs.gov.cn

数据展现地球表层现状与发展过程。从系统角度看,地质调查是参与人、数据处理机、地球构成的“人-机-地”系统。地质调查数据的产生情况如表1所示。其中网络信息与管理信息来自人机交互系统,地球信息来自机地交互系统和人地交互系统。

地质调查数据产生的位置与时间表现出整体的规律性和局部的随机性。地质调查不但产生地质观测与探测数据,还产生服务、管理及其参与人数据,表现出确定性与不确定性交织的复杂状态。

2 基本认识

地质调查大数据是地质调查工作和信息科学技术发展、融合到一定程度的结果。源动力来自于2个方面:①地质调查业务不断调整拓展,大量新型技术方法的应用,数据共识基本形成;②地质调查信息化服务需求日渐增强,亟需从独占走向共享、从粗放走向精细。

地质调查大数据试图解决以下3类问题:

(1)过去计划经济体制下,地质信息资料分割保存,形成信息孤岛,数据信息顺畅流动困难,信息与数据共享问题一直是制约地质调查发展的瓶颈。

(2)在以往地质调查工作中,存在一些与数据和计算相关的地质问题,由于当时信息技术条件的限制,没有得到解决,或者解决效率、精度不能令人满意。这一类问题普遍存在于地质调查具体工作中。

(3)地质调查信息化服务产品类型不足,生产周期偏长,需求响应欠准确、欠及时。这一类问题已经逐渐成为地质调查工作的焦点问题。

地质调查大数据是一个三元组$<\Omega, f_\Omega, R_f>$,Ω是一个巨数据集,f_Ω是定义在Ω上的处理技术方法集,R_f是f_Ω上的关联关系。通常,巨数据集Ω的计数测度只增不减,包含地质调查产生的数据。处理技术方法集f_Ω的操作对象是地质调查产生的数据,操作基础是信息技术,尤其是新一代信息技术,是地质调查大数据处理技术的核心元素。关联关系R_f定义解决地质调查问题的思路逻辑,体现地质调查大数据的功用与质量。

地质调查大数据是“用”出来的,应从地质调查大数据处理技术研究与开发起步,解决技术应用中的具体问题。当应用达到相当的广度、深度后,一些有关地质调查大数据的共性科学问题会浮现或抽象出来,这时就是在更高层面上解决问题的时刻了。

3 地质调查大数据处理的技术问题

结合地质调查数据处理一般流程和大数据处理技术的特殊性,把地质调查大数据处理主要技术问题分为采集与传输、存储与管理、计算模式与系统、分析与挖掘、可视分析、隐私与安全6个方面。

(1)采集与传输

对应地质调查数据的产生,地质调查大数据外延显然更大,采集对象更加全面地覆盖地质调查工作。同时,通过实时或准实时数据通信获取地质调

表1 地质调查数据的产生情况

Table 1 The generation of geological survey data

类 别	工 作	产生场所	产生源	数据含义	数据形态
地球信息	地质填图编图	野外、室内	填图编图人	地质体、地质现象、地质过程等	数值、影像等
	地球物理方法	野外	物探设备	电性、磁场、重力、地震波、声波、放射性等	数值等
	地球化学方法	室内、部分野外	化探设备	元素指标等	数值等
	地质科学实验	室内	实验仪器	特定指标等	数值、影像等
	地质钻探	野外、室内	钻探设备、断面扫描设备等	连续断面等	影像、数值等
	地质监测	野外	监测仪器、监测人员	监测指标等	数值、影像等
	地质资料综合	室内	地质调查人员	文档、图件等	特定等
网络信息	公共服务	计算机网络	客户端	访问热点等	数据等
管理信息	内部管理	信息系统、内部网络	参与人等	管理指标等	数据等

表 2　地质调查大数据采集与传输技术及主要问题

Table 2　The acquisition and transmission techniques and problems in geological survey big data

工 作	采集技术	传输技术	主要问题
地质填图编图	野外地质数据采集技术等	移动通信技术、卫星通信技术等	扩展采集技术、集成加密压缩通信技术
地球物理方法	物探设备数据采集模块等	移动通信技术、卫星通信技术等	改造物探设备、集成加密压缩通信技术
地球化学方法	化探测试设备数据采集模块等	光纤通信技术等	改造化探测试设备、集成加密压缩通信技术
地质科学实验	实验仪器数据采集模块等	光纤通信技术、卫星通信技术等	改造实验仪器、集成加密压缩通信技术
地质钻探	连续断面扫描技术等	光纤通信技术等	研发采集与传输产品
地质监测	监测仪器数据采集模块等	移动通信技术、卫星通信技术等	融合数据加密压缩技术、集成加密压缩通信技术
地质资料综合	自动识别技术等	光纤通信技术等	研发采集与传输产品
公共服务	日志技术计算环境监测技术等	光纤通信技术、移动通信技术等	研发集成采集与传输模块
内部管理	数据触发技术等	光纤通信技术、移动通信技术等	研发集成采集与传输模块

查数据,为达到地质调查工作服务目标提供基础。地质调查大数据采集与传输的基本目标是应收、尽收、速收。

地质调查大数据采集与传输技术多样,也相对成熟,主要问题在于技术集成与产品化设计生产方面。地质调查大数据采集与传输技术及主要问题如表2所示。

(2)存储与管理

地质调查工作中,不同类型的应用对存储系统的性能、可靠性等指标有不同的要求,这在存储与管理中并不是新问题,但地质调查大数据的大体量、高复杂度放大了达到这些技术指标的难度,导致“存储墙”问题越来越严重。

地质调查数据处理应用中存在2个突出的问题:①数据体量增加到一定程度后,系统停止运转;②读写外部存储碎片数据时,系统效率极低。这2个问题首先与地质调查数据处理使用的计算机存储硬件有直接关系。当前,地质调查数据处理中使用的内部存储硬件主要是SRAM/ DRAM工艺的,外部存储硬件主要是磁盘。SRAM工艺存储密度限制片上存储容量增长,SRAM/ DRAM工艺高静态功耗阻碍存储层次发展,SRAM/ DRAM 工艺对粒子和射线撞击产生的软错误问题没有抵抗能力,纠错电路限制存储容量增加并引起功耗[1]。磁盘是计算机体系结构中唯一还在使用的机械单元,与其他电子存储单元在访问效率、延迟等指标上存在量级上的差距,如表3所示[2]。上述2个问题与地质调查数据处理使用的软件体系也有关系。地质调查应用主要的存储与管理软件是文件系统与数据库,当前,地质调查数据存储与管理整体设计不明确,绝大部分应用没有进行针对性较强的存储与管理设计优化,造成整体效率与具体应用效率都有待提高。地质调查大数据存储与管理的基本目标是软硬件优化升级,大幅提升效率。

地质调查大数据的存储与管理技术与当前使用的技术有较大的区别,内容更丰富,结构也更复杂,但效率普遍提高,有些技术能够达到量级上的提高。地质调查大数据的存储与管理技术如表4所示。

(3)计算模式与系统

计算模式在以往地质调查数据处理中很少提及,但计算模式是地质调查大数据处理的核心问题之一。所谓地质调查大数据计算模式,就是根据地质调查大数据的数据特征和计算特征,从多样性的地质调查大数据计算问题和需求中提炼并建立抽象或模型。在地质调查工作中,与数据计算有关的业务非常多,业务目标不同,对数据计算响应的时

表 3　各类存储操作响应时间

Table 3　The response time of some storage devices

操 作	时间
机械磁盘一次寻址定位	4ms
从机械磁盘顺序读取1MB数据	2ms
从SSD磁盘顺序读取1MB数据	0.3ms
从内存中读取1MB数据	十几微秒

表4 地质调查大数据存储与管理的主要问题

Table 4 The storage and management problems and analysis of geological survey big data

应用问题	对策分析	存储与管理技术
文件系统分层小而多导致文件读写效率低下	高可用、读写分离、读写效率均衡	非易失性存储技术、大内存大缓存技术、分布式文件系统、分布式缓存技术、设备冗余技术等
数据库查询规模过大导致响应时间过长	分布式、高效索引、高可用	分布式数据库、分布式缓存技术、专用类型数据库等
空间分析计算规模过大导致系统无响应或崩溃	新型存储结构	大内存技术、计算与存储融合技术、并行技术等

限要求也不同,操作的数据不同,数据的计算方式就不同,因而需要甄别不同计算模式,分类分析地质调查大数据处理中的计算模式。地质调查大数据计算模式主要类型与特点如表5所示。

地质调查大数据处理的大部分对象是空间数据,关联关系复杂,当前主流的批处理计算难以从根本上解决可行性与效率问题;内存计算[3]在计算机体系结构层面解决地质调查大数据处理的问题,具有广谱性,可以很容易地与其他计算模式结合,形成具有优异计算性能的应用系统;随着内存价格的不断下降和新型非易失性存储器的发明,服务器可配置的内存容量不断提高,采用内存计算完成高速的地质调查大数据处理有了现实的可能性。内存计算是地质调查大数据处理技术发展的重要趋势。

总体上看,地质调查大数据处理需要面向实际工作,提供多种计算模式的服务。

(4)分析与挖掘

地质数据定量分析一直是地质调查重要的工作内容,但赵鹏大[4]认为,目前地球科学的分析和可视化方法已经远远落后于创造数据的能力。地质调查大数据分析与挖掘技术首先要解决的问题是地质调查工作区多来源、多模态、多时态数据的相关性和模式分析,这样的分析可以克服个体的波动性,发现更多可靠的、隐藏的模式和知识。地质调查大数据分析与挖掘的技术问题还表现在以下4个方面:

第一,以往地质调查数据分析的一个重要方法是采样,当数据体量比较大时,可以通过采样技术把数据规模变小[5]。很显然,在很多地质问题中,采样意味着信息的丢失。如果不运用采样技术,考虑对地质调查大数据全集进行分析,意味着需要分析的数据量急剧膨胀与增长,其面临的技术问题就是体量巨大的数据如何分析。

第二,以往地质调查数据分析方法集中于线性空间中的统计方法,以及一些初级的非线性方法,在小样本上运用这些方法获取局域的地质特征。当在地质调查大数据上运用这些方法时,令人不安的结果往往是方法收敛早于数据规模波动,有必要针对地质调查大数据的一些方法进行改造,或者提出新的方法,这是地质调查大数据分析与挖掘面临的另一项技术问题,即深度分析。

表5 地质调查大数据主要计算模式

Table 5 The computing mode of geological survey big data

计算模式	主要实例	特点
查询计算	秒级响应查询	数据复杂、体量巨大、大承载
批处理计算	物化探数据分析、网站日志分析、物探遥感自动解译等	计算规模较大、可离线
流式计算	地质调查管理信息分析、网站日志分析等	数据运动、计算固定、实时性
迭代计算	地质问题模式学习、地质过程模拟等	循环调度
图形计算	空间数据分析、图形编辑等	关联关系复杂、数据规模大、计算规模大
图像计算	遥感影像分析、断面影像分析、资料影像分析、图像预处理等	数据规模大、大规模矩阵计算、易并行
内存计算	普适	高实时性、高普适性

第三,集合了地质描述、地质数据、地质图表、地质认识的地质调查资料价值巨大,目前,地质调查资料处理技术主要集中在前处理和检索查询方面,而地质调查综合资料的深度分析与挖掘是地质调查大数据分析与挖掘面临的重要技术问题。

第四,地质矿产资源评价、地质环境监测预警等重要业务在不断发展,对数据与计算技术的要求更精致、更敏捷,实现对这些重要业务的有效技术支撑是地质调查大数据分析与挖掘面临的基本技术问题。

地质调查大数据分析与挖掘的另一类技术问题源于地质调查信息化服务。网络形式的公众服务必然产生公众访问数据,这些数据以某些特定形式存储,对这些数据的分析与挖掘有助于优化服务系统配置,提高服务质量与效率。在这方面,一些通用的分析与挖掘技术可以派得上用场,而具有地质调查特色的分析与挖掘技术是需要重点关注的,如地质调查空间数据热点技术、地质图块的快速检索技术、地质图块与地质资料快速匹配技术等。

(5)可视分析

有别于一般的数据处理工作,地质调查工作中很大一部分地质问题是地质专业技术人员在空间数据或图件基础上进行综合分析后解决的,这种解决问题的模式为地质调查大数据可视分析技术提供了可能。可视分析就是通过交互可视界面来进行分析、推理和决策的过程[6],本质也是知识发现。可视分析与一般分析与挖掘的不同在于,其不依赖于数学模型,而是一种探索式分析,这与很多地质问题的解决模式一致。

地质调查大数据可视分析的技术基础是多年积累的地质体建模、地质过程建模、地质调查数据可视化及其交互的技术,可视分析就是在这些技术的基础上,克服高维性、不确定性和异构性,研究开发从复杂地质调查数据中抽取有效特征的方法,通过探索式分析完成地质调查大数据中知识发现,其基本技术流程如图 1 所示。

图 1　可视分析技术流程

Fig. 1　The flow chart of visual analytics

当前的机器智能在复杂地质数据的视觉识别和理解方面远不及人脑智能,而超过 50%的智能与视觉识别有关。因此,对地质特征模型可视化结果的修正和判断,体现出人脑智能和机器智能的差异,其中蕴含的就是新知识。探索式可视分析以人脑智能向机器智能的转换、机器智能向人脑智能的展示为主线,实现地质认识的螺旋式进化,在这一方面,需要解决的技术难题包括:①对地质体和地质现象的数据或参数的输入常常存在谬误和不精确,因而人脑智能转换为机器智能是多人协同、反复修正的过程;②地质调查大数据环境下,各种可视分析方法需要具备可扩充性和容纳不同对象的能力,这样才能支持多来源、多时态的地质大数据处理。

(6)隐私与安全

隐私是不愿意被他人知道或他人不便知道的敏感信息[7]。地质调查大数据以服务为重要目标,因此存在服务参与人的隐私问题。

安全指不受威胁,没有危险、危害、损失[8]。信息安全指采取技术和管理的安全保护手段,保护软硬件与数据不因偶然的或恶意的原因而遭到破坏、更改、显露[9]。数据安全一直是地质调查信息化工作的重要内容。

地质调查大数据隐私与安全面临的新情况和带来的威胁与挑战如表 6 所示。地质调查大数据隐私与安全需要法律、政策、管理与技术共同维护,从技术层面,当前主要隐私和安全保护方法包括:文件访问控制技术、基础设备加密、匿名化保

表 6　地质调查大数据隐私与安全问题

Table 6　Privacy and security question of geological survey big data

新情况	威胁或挑战
数据的细节越来越丰富	数据泄露风险增大,可能涉及人身安全
黑客也采用大数据分析技术发动攻击	数据处理过程的安全问题严重
位置信息暴露严重	人的移动模式与身份识别之间存在联系
分布式存储与分布式计算模式增多	数据传输与交互产生更多泄漏丢失隐患
数据多元化及关联性提高	单一数据的安全隐私保护方法脆弱
高价值数据增多	APT攻击可能性增大

护技术、加密保护技术、数据失真技术、可逆置换算法等。

4 科学问题

地质调查大数据的研究处于探索阶段，地质调查大数据处理技术开发也处于起步阶段，因而对地质调查大数据研究中的科学问题准确描述还十分困难，这里把科学问题讨论限制在地质大数据范畴，也就是地球信息的范畴，探索性地提出2个带有普遍性与根本性的问题。

（1）地质大数据仅仅是地质世界的数据映象，还是具有反映地质世界规律能力的数据世界？

地质科学一直以来以科学实验、知识归纳为主要研究手段，近几十年来数值模拟也有一些发展。赵鹏大[10]指出，当前已经进入大数据时代，大数据成为新的科学范式（数据密集型科学研究的“第四范式”[10]），是一场技术革命、颠覆性创新[11]。那么，地质大数据的关联关系与地质世界的因果关系之间存在怎样的联系？或者表述为：地质大数据的相关性研究与地质科学研究功效一致，还是仅仅是地质科学研究的补充？

在简单的、封闭的系统中，基于小数据的因果关系是很容易做到的，但对复杂的、开放的巨系统，传统的因果关系是不是奏效很难说。地质大数据进行的关联分析是“知其然而不知其所以然”的，相关性表达2个或多个地质变量取值之间的某种规律性，严格地说，统计相关性是无法检验逻辑上的因果关系的[12]。不过一些支持者，如Mayer-Schönberger[13]在《大数据时代》一书中指出了大数据时代处理数据理念的三大转变，即要全体不要抽样，要效率不要绝对精确，要相关不要因果。也就是“数据-信息-知识-智慧”要让位于“数据-价值”的研究思路。

当前，面对地质大数据，地质科学研究人员有可能从中查找、分析或挖掘所需要的信息、知识和智慧，甚至无需直接接触所研究的对象。

（2）地质大数据关系网络的本质特征是什么？

地质数据之间复杂的网络关系是地质大数据的存在形式，深入分析地质大数据关系网络，才能把握地质大数据的本质。针对大型复杂的研究区域，地层、物探、化探、影像等数据之间的关系如何定量表达，这一系列表达关系的变量中是不是存在一些整体上有规律的部分，这是地质大数据背后的关系网络研究的重要内容。

5 地质调查大数据处理技术开发方法

地质调查大数据处理技术的开发既要继承以往地质数据处理技术开发的一些方法与技巧，更要考虑在开发思路上的不同之处。

（1）为了降低成本，提高能效，地质调查大数据相关系统需要摆脱传统的通用体系，趋向专用化架构技术[14]，适度抽象有利于整体把握地质调查大数据处理技术的一致性与协调。

（2）以往地质调查数据处理技术和系统更多地是面向3S技术，一些系统是3S技术在地质调查工作中的应用，对地质调查服务与应用环境、性能等方面缺乏深入工作。地质调查大数据处理技术开发只关注3S技术已经不够，还要关注信息科学与技术的发展，尤其是计算机科学与技术的发展。

（3）重视具体应用软件开发、忽视软硬件集成开发是以往地质调查数据处理系统开发时的缺陷。地质调查大数据处理技术开发不仅要关注软件，还要关注系统，尤其是信息系统与物理系统结合、信息技术与自动化技术结合。

（4）以往地质调查数据处理系统开发偏重系统功能的实现，地质调查大数据处理技术开发只关注功能实现已经不可能了，还要关注性能与复杂度，开发难度提升。

（5）地质调查大数据处理所涉及的数据与计算规模是空前的，必须有精准的需求分析，以及完整、可靠的技术设计，在严格的技术监督下逐步展开，否则可能造成经费和时间的浪费。

6 结 语

地质调查大数据处理技术是地质调查信息化服务的技术核心，以信息化服务产品体系推动资源共享是当前的重要目标。地质调查大数据研究，应从分析以往解决得不理想的地质问题入手，充分利用新一代信息技术，更新当前数据处理环境，在新环境下提出合理、有效的解决方案。另外，考虑在数据体量增大、类型复杂、响应时间有要求的情况下，针对以往解决得不理想的地质问题，着重进行地质数据的智能分析与深度挖掘，考虑合理、有效的解决方案。

致谢:成文过程中国土资源部姜作勤研究员和中国地质调查局发展研究中心杨东来研究员给予了有益的指点和建议,野外调研期间得到了中国地质调查局南京地质调查中心徐震宇、陈基炜和中国地质调查局成都地质调查中心张建龙、李富的协助,在此一并表示感谢。

参考文献

[1]孙广宇,王鹏,张超.基于新型非易失存储的存储结构[J].中国计算机学会通讯,2014,10(4):18-25.

[2]Scott C. Latency Numbers Every Programmer [EB/OL](2015-01-28)[2015-02-03]http://www. eecs. berkeley.edu/~rcs/research/interactive_latency.html. 2015.

[3]哈索.亚历山大·蔡尔.内存数据管理[M]. 北京: 清华大学出版社, 2012.

[4]赵鹏大. 大数据时代需重视数字地质研究.中国国土资源报, 2013.

[5]覃雄派,王会举, 杜小勇, 等. 大数据分析——RDBMS 与 MapReduce 的竞争与共生[J]. 软件学报, 2012, 23(1): 32-45.

[6]Thomas J J, Cook K A. Illuminating the Path: the Research and Development Agenda for Visual Analytics[C]//IEEE CS, 2005.

[7]Blum A, et al. A learning theory approach to noninteractive database privacy[J]. J. ACM, 2013, 60(2): 1-25.

[8]Bainbridge W S. Privacy and property on the net: Research questions [J]. Science, 2003, 302(5651): 1686-1687.

[9]Weiser M. Critical Issues and Information Security and Managing Risk[C]//The 9th International Conference on Computing and Information Technology (IC2IT 2013), Springer, 2013.

[10]Hey A J G. The Fourth Paradigm: Data-intensive Scientific Discovery[M]. Microsoft Research, 2009.

[11]赵鹏大. 大数据时代的地学研究[C]//湖北地质科技论坛, 2014.

[12]李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊, 2012, 27(6): 647-657.

[13]Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think[M]. Houghton Mifflin Harcourt, 2013.

[14]程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(90): 1889-1908.