

地学信息共享与知识发现

甘甫平¹, 刘兴起^{1,2}, 王永江^{1,3}

(1. 中国地质大学, 北京 100083; 2. 中国科学院青海盐湖研究所, 青海西宁 810008;
3. 航空物探遥感中心, 北京 100083)

摘 要: 地学信息共享是地学研究的重要内容,也是地学(数字地球)发展的趋势之一。促使或完成地学信息共享的核心技术主要为技术标准化规范、空间数据库、互操作(元数据)以及传输信道技术等。知识发现是建立在共享的地学信息基础之上,能有效地实现知识的增值以及知识的再生。知识发现与信息共享二者共同促进,相互发展。

关键词: 地学信息共享; 核心技术; 知识发现; 知识的增值

中国分类号: P-37 **文献标识码:** A **文章编号:** 1008-858X(2000)03-0062-07

0 引言

地球是一个复杂、开放、多变的巨系统,各层圈间(包括大气圈、小圈、生物圈等)相互作用、相互影响。传统地学研究着眼于这一巨系统的各个分支,如地球化学、地球物理等,相对地趋向精深、细的微观深入分析。但随着资源危机、环境恶化、灾害频发等一系列日趋严峻问题的出现,人类不得不面对地学的可持续发展研究以及人地协调关系的研究。这些研究是建立在地学各分支学科的基础之上,进行学科交叉渗透、整体综合,以理解和模拟不同时空尺度下的地球各圈层的演变过程及它们之间的相互作用机制,而揭示全球变化规律、全球变化的区域响应模式以及预测环境承载力、资源承载力。

显然,传统地学难以适应这一研究模式与方向的转变,难以对传统地学各分支领域内的海量数据进行综合分析,也不会以全局的观点整合各分支学科的知识进行演绎推理。也就是说,传统地学研究难以实现各个尺度下各个应用方向的地学信息共享。因此,必须探寻和开拓新的领域。“数字地球”以及“中国数字地球”^[1]这一概念的提出,能有效地解决这一难题。但所有这一切必须建立在地学信息共享的基础之上。通过信息共享,可以从不同的角度、不同的应用目的、不同的研究层次来获取相应的知识,以供决策和实现信息增值目的,并反作用于地学的研究,最终加速“数字地球”以及“中国数字地球”实现。

收稿日期: 2000-02-24

作者简介: 甘甫平(1974-),男,博士生,主要从事遥感地质地理信息系统研究

1 地学信息共享

地学信息共享是有条件的, 不仅体现在利益的分配上,更重要的是地学信息具有容量大、多尺度、异地分布、类型繁多、属性不一的特点,这些均成了信息共享潜在的障碍。与此同时,地学数据库、地学空间数据处理系统、管理信息系统、办公自动化系统、信息传输系统等也具有异种硬件、异种软件、异种网络环境、异种开发平台、异种组织和部门的分布格局。因此必须采用相应的技术,对地学数据进行标准化,对数据库进行集成化改造并改善和提高信道的传输容量、传输速率等。

1.1 地学信息特征

信息是表现事物特征的一种普遍形式,是事物性质特征和状态的表征,是一切有用的知识^[2]。地学信息是指地物实体的性质、特征和状态的表征,是对地物物质、能量不均衡的描述。

地学信息来源于地学各分支。比如地球化学、地球物理、矿床、构造地质、水文地质等。从逻辑结构看,表征地学信息的数据可抽象为空间、属性、时间特征。但地学数据除其共性外,还普遍具有多源性、尺度变换性(特征的时空尺度)、不确定性等(图 1)。从图 1^[3]可以了解到: (1)多源即地学数据可来自不同的部门、来自不同的研究领域、来自不同的研究区域以及不同的研究层次; (2)特征的时空尺度是地学信息的重要特征,是地学信息观察、表示、分析和传输的详细程度的描述。在探寻模式或过程随比例变化的规律时,必须注意到许多地学过程并不是按比例线形变化或均匀变化^[4]。 (3)不确定性存在于地学数据生命周期的每一个阶段,从数据采集到数据再现,再到数据分析和最后结果;而且数据间具有渐变的特征。在地学信息分析中,不确定性主要表现为类型属性的不确定性、位置属性的不确定性、空间相关的不确定性^[4]。如果对这些问题重视不够,将会极大的影响地学信息共享的有效性,影响决策的公正性,影响科研结果的可靠性,同时对地学信息共享技术的设计也可能造成许多障碍。

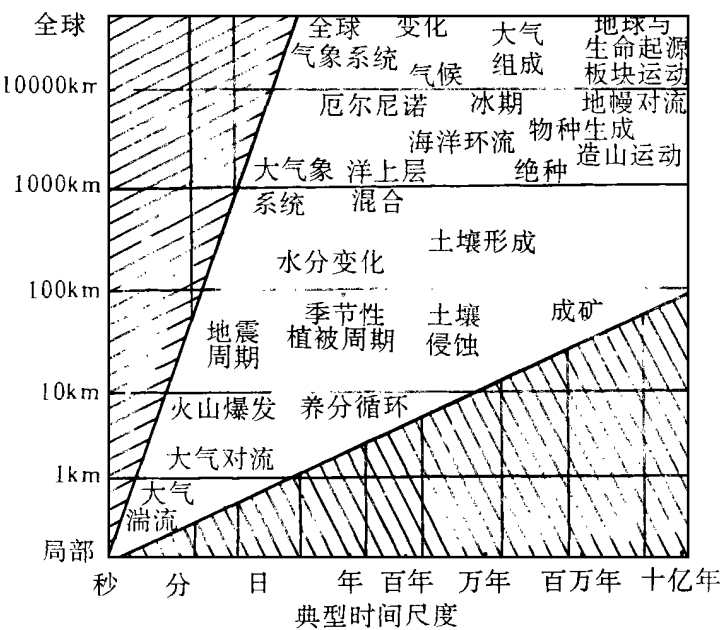


图 1 地球系统过程: 特征的时空尺度

Fig. 1 The earth system procedure Time and space rules of characteristics.

1.2 地学信息共享技术

借助 GIS 技术,大多数国家都建立了分门别类、类型较为齐全的地学数据库和图形库。比如美国联邦地调所 (USGS)建立了美国资源库、全国煤炭数据库、全国水文数据存储检索系统、海洋地质数据库、地球化学与岩石数据库以及全美数字制图数据库 (NDCDB)。这些数据库存入了全美数万个矿床、矿点以及数十万处钻孔和野外露头数据^[5,6]。要对这繁多的数据库实行动态管理,以便从中获取有用的信息和知识是相当不容易的。为此目前研究的热点与重心是标准化技术、空间数据仓库技术、互操作技术 (元数据) 以及传输信道技术等。

1.2.1 标准化技术

地学空间数据的标准化、规范化以及 GIS 的建设,是地学信息共享的一个关键问题。目前从事地学数据共享研究的机构有国际标准化组织 ISO 的 TC211 小组、欧洲标准化组织 CEN / TC287、CERCO、美国联邦地理数据委员会 FGDC、Open GIS 协会 OGC 等。其研究内容涉及 GIS 技术标准化的各个方面,如网络环境下共享地理数据和互操作的开放式协作标准、数字地理信息以及地理信息的对象结构标准等。

在具体的地理信息标准化实践中,其一是标准文本的起草和研究,即采用基于大量理论与实践基础之上的规范化文本来指导实践;其二是建立在实践活动之上的标准或标准体系的制定^[7]。我国已出台了城市地理信息系统标准化指南^[8]。有了技术标准化的支撑,地理信息系统能兼容来自多方面的信息资源,按地理空间 (Geo-spatial) 坐标进行数据管理、查询和检索;通过地学分析、模拟和预报等软件系统进行加工与决策,提供多层次的信息服务^[9]。

1.2.2 空间数据仓库

即使建立在技术标准化之上的数据库,往往可能是异质的操作平台或不同的应用软件,或因为地学“比例”变换、地学信息不确定性等造成数据标准不统一。为从多学科、多角度、综合地去感知世界,很有必要将来自不同的多个应用系统、不同学科的相关数据转换成统一的格式,集成、存储在一起,构成一个空间数据仓库,基本构想见图 2^[10]。

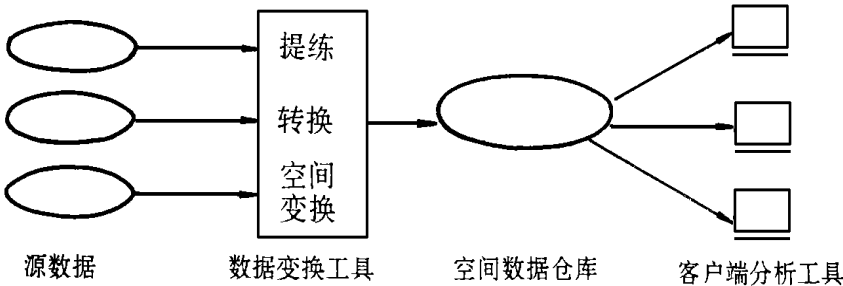


图 2 空间数据仓库功能构想

Fig. 2 Imagination of spatial data warehouse's function

空间数据库是根据主题,以集成的方式实现数据的变化与增值,其数据具有时空属性。空间数据仓库可采用多维数据库概念模型。如 Langran 的基于栅格加变长列表数据^[11],空间时态对象关系数据库^[12]。后者在模型中保存一个事件列表数据,利用时间拓扑指针以及空间变化指针对空间

实体位置或特征变化进行分析、管理

空间数据仓库的建立使 GIS 应用的信息环境分为操作环境和分析环境。操作数据库负责数据的日常操作应用,分析数据库则根据主题通过专业模型抽取和凝集而成,专用于各种复杂分析,为高层决策者服务^[10]。但专门开辟一个分析环境具有重复性,浪费了大量的资源并占据大量的内存。为此,可根据主题以及各自应用的需要,利用程序激活并组织成相关的动态虚拟的空间数据仓库,以满足分析、提取、研究的需要,其生存周期仅为被激活的这段时间,即取即用,用完即卸载。这对接口软件提出了较高的要求,可借用 Java 对网络上资源管理的相似性进行实现。

1.2.3 互操作性

空间数据仓库是一种面向主题进行数据集成、综合管理的数据库,分析环境和操作环境相分离。而互操作是自底而上将已有的系统和应用集成,是不同应用(包括软硬件)之间利用一个稳定的接口动态地、直接地相互调用^[13],是不同系统之间信息的共享或交换^[4]。

互操作性研究的长期目标是开发能够自动提取和更新元数据的方法^[4]。元数据是“关于数据的数据”。通过对元数据的管理,能够了解信息的名称、位置、作者或信息来源、数据格式、分辨率等,能够找到不同系统和媒体之间的共同性。

从技术的角度,数据共享强调数据的互操作性^[7]。对于数据的转换,主要是基于相关表、转换器以及转换标准;相应地建立标准数据字典及标准的特征登记(Feature registry)^[7]。这仅仅是数据的集成,基于 GIS 互操作还必须包括对处理过程的集成,其交流是双向的。

实现互操作其关键是达成统一的协议,通过共同的接口,从而实现信息的交换与共享。

1.2.4 传输信道技术

地学信息具有空间属性,对传输带宽、传输速率、信道容量要求较高。WWW 服务器化在 Internet 上的应用使数据量巨大,具有复杂结构的地学图件、影像图等地学信息汇聚传输成为可能,也使地学之间信息共享成为可能。

WWW (World Wide Web) 是 Internet 上的多媒体信息服务器,Internet 上多媒体信息传输、检索、获取的一个套件,包括 URL 技术、HTTP 技术、HTML 技术等。URL 技术确定信息源地址;HTML 技术组织和处理信息;HTTP 技术传输信息。借助 WWW 上的浏览器,比如 Netscape Explore、Internet Assistant 和 Websurfer 等软件,可全面实现信息文本由传统的平面、静止、孤立的、无时间、无空间变化的信息载体变为立体、变动、关联的超时间、超空间的多媒体信息载体^[9]。这是地学信息多媒体网络化的发展趋势,比如 Web-GIS,将促进多媒体电子地图、多媒体数字模拟以及虚拟现实技术的发展。

2 知识发现

知识发现,是指从数据库中发现知识(KDD),亦称数据发掘(Data Mining,即 DM)。其目的是把大量的原始数据转换成有价值的知识,用以描述过去的趋势和预测未来,是决策支持的过程。也就是从数据库中提取隐含的、感兴趣的、高水平的模式^[14]。空间数据知识发现是建立在地学信息共享的基础之上,单独或综合地运用统计、归纳、聚类、空间分析、探测性数据分析、Rough 集方法、云理论、模式识别、决策树以及神经网络、遗传算法等^[14]去挖掘数据库中隐含的丰富的知识用

以进行科学研究、决策分析以及宏观管理。知识、信息和数据的关系如图 3^[15]。在知识发现的过程中,必须加强空间数据模型以及可视化的研究。空间数据模型是知识发现的基础,可视化可揭示知识之间的动态关联,实现知识的再发现。

2.1 空间数据模型

地学现实具有复杂性、动态性、模糊性、不确定性,其形式多样、时空尺度不一、结构复杂、边界无规律、数据海量等特征,因此杂乱无章。但在特定的空间条件下,地学信息相互间存在着相互映射(imaging)关系,即主要素相互映射,局部映射整体,现在映射过去,现在过去映射未来等^[2]。与此同时,地学信息还是环境变迁的载体,即具有地学记忆性,如地质记忆中的岩石记忆、地质构造记忆、地貌记忆、第四纪松散沉积物的记忆以及树木年轮和树干形状的记忆^[2]。这些映射关系与记忆性质启示我们可以利用特定的应用模型去研究地学。在知识的发现过程中,有助于进行演绎推理以及全球化战略的研究。

建立在信息共享机制之上的数据库,蕴藏着地学实体物质流和能量流不平衡动态演变的信息流。这些信息流在不同的尺度下有着不同的动态响应,在不同的研究领域也有着不同的表现形式。而空间数据模型可以根据不同的参数再现这一过程。它可以从物质能量相互映射的动态关系中,从物质的复杂记忆中、从地学信息的相互关联中、以及从地学可视化的虚拟显示中去探寻异常与不确定性,去描述数据的多样性与可靠性,从而获取知识。空间地学数据模型目前一般为基于栅格的体素 Voxel(Volume element)模型与基于矢量的等值面(Iso-surface)模型,以及地学信息模型和时空四维数据模型。前述的数据仓库也可以认为是这类模型的特类。

2.2 可视化

建立在知识发现之上的空间数据模型的可视化,不仅仅是满足一般的查询、分析、管理功能,不仅仅是以图表、文字来进行三维的显示,而必须要求注重于问题的过程和起初的洞察,利用计算机的想象技术把信息传递到视觉,由此得到知识。虚拟现实是可视化的最高境界。对地球史上生物集群灭绝的模拟,可以综观全球变化;对地球层圈动力作用的模拟,可以演示地球的形成及变化趋势以及对地质灾害如火山爆发、地震的预测和防治。但这一切都需要海量数据的共享以及高深技术的支撑,这还有很长的一段路要走。

可视化与空间数据模型是双向的动态链接,相互制约、相互支撑、共同发展。但它们都离不开源数据的支持,源数据是知识发掘的土壤。

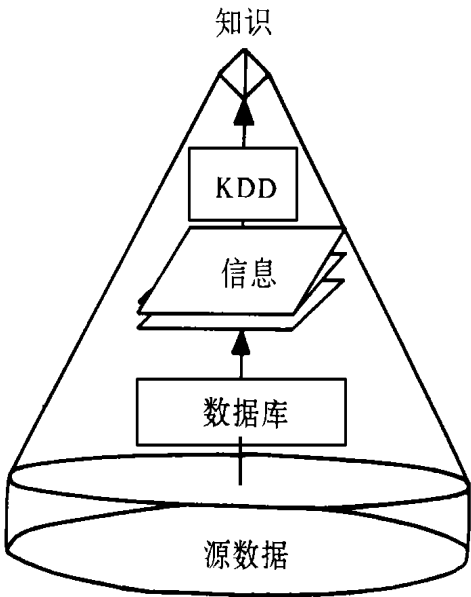


图 3 数据—信息—知识金字塔
Fig. 3 Data- information- knowledge delta

3 讨论

地学信息共享是地球科学发展的一个重要趋势。加强共享技术的研究是共享的必要条件,但信息共享的目的与对象却制约着共享技术的走向。因此必须清醒认识到地学信息共享不仅仅要面向科研,还必须面向大众,面向社会;不仅仅要面向地学本身,还必须面向其他各学科;不仅仅要面向资源的开发,还必须面向人地协调与可持续发展;不仅仅要面向国内,还必须面向国际共享。在实现地学信息的共享中,除了努力开发设计自己的共享技术外,还必须大力建设地学空间数据库,尤其是加强源数据的开发。也就是说,必须加强地球科学的研究。这是非常重要的也是必须的。否则,地学信息的共享将是无米之炊。提供准确可靠、国际承认以及居国际领先地位的源数据,才能在信息共享的资源竞争中占据一席之地。

诚然,地学信息共享的必然结果是从中获取有用的知识。知识发现在地学信息之中占据重要的地位。21世纪是信息的社会,对信息的占有将会显得举足轻重,但从信息中获取知识却是一个信息增值的问题,不容忽视。因此地学除了切切实实地进行基础性研究用以获取重要的源数据外,还必须加强信息化的步伐。

参考文献:

- [1]徐冠华,孙枢,陈动泰.迎接“数字地球”的挑战[J].遥感学报,1999,3(2): 85-89.
- [2]承继成.地球信息科学基础理论纲要(上)[J].地球信息科学,1996,2: 15-24.
- [3] ESSC(Earth System Sciences Committee), NASA Advisory Council. Earth System Science—A Closer View[R]. Washington: NASA, 1988.
- [4] UCGIS. Research Priorities for Geographic Information Science[J]. Cartography and Geographic Information System, 1996, 23(3): 115-127.
- [5]姜作勤,俞宏全.全美地质调查所数字制图 GIS应用与地学信息管理的现状与趋势[J].地质信息技术,1992,8(2): 10-43.
- [6]吴冲龙.计算机技术与地矿工作信息化[J].地学前缘,1998,5(2): 343-355.
- [7]陈常松,何建邦.地理信息共享中标准与技术问题的一般性评论及其需求[J].遥感技术与应用,1998,13(2): 37-43.
- [8]蒋景瞳,刘若梅.城市地理信息系统标准化问题[J].地理信息世界,1998,4: 1-4.
- [9]陈述彭,何建邦.地理信息系统的基础研究[J].地球信息,1998,4: 5-7.
- [10]赵霏生,杨崇俊.空间数据仓库的技术与应用[J].地理信息世界,1998,4: 5-7.
- [11] Langran, G. Time in Geographic Information Systems[M]. London: Taglor Francis, 1992.
- [12]马劲松,黄杏元. GIS空间时态对象关系数据库模型研究[J].地理信息世界,1998,4: 8-9.
- [13]黄裕霞,陈常松. GIS的互操作[J].地理信息世界,1998,4: 10-12.
- [14]邱凯昌,李德仁.空间数据发掘和知识发现的框架[J].武汉测绘科技大学学报,1997,22(4): 328-332.
- [15] Hatesky-Shapiro G. An Overview of Knowledge Discovery in Database: Recent Progress and Challenges[A]. Ziarko W. Rough Sets, Fuzzy Sets and Knowledge Discovery[C]. New York: Springer-Verlag, 1994.

Sharing of Geosciences Information and Knowledge Discovery

GAN Fuping¹, LIU Xing-qí^{1,2}, WANG Yong-jiang^{1,3}

(1. *China University of Geoscience, Beijing 100083, China;*

2. *Institute of Salt Lakes, The Chinese Academy of Sciences, Xining 810008, China;*

3. *Aero Geophysical Remote& Sensing Center, Beijing 100083, China)*

Abstract Sharing of geosciences information is an important content of studying the geosciences and is a trend of developing Geoscience. It must rely on many key techniques such as the regularity standard of technique, mass database, metadata and information transport in order to share effectively the information which came from different fields of geosciences. Furthermore, knowledge discovery can reproduce new knowledge based on sharing geosciences, and enhance the sharing capability at the same time.

Key words Sharing of geosciences; Key technique; Knowledge discovery; Reproduction of knowledge

声 明

1. 为适应我国信息化建设需要,扩大作者学术交流渠道,本刊已加入《中国学术期刊(光盘版)》和“中国期刊网”。作者著作权使用费与本刊稿酬一次性给付。如作者不同意将文章编入该数据库,请在来稿时声明,本刊将作适当处理

2. 本刊已入万方数据资源系统(China Info)数字化期刊群