·争鸣与探索·

环境信息数据仓库建设及其相关的技术应用

厉　青, 王　桥, 申文明, 吴传庆
(中国环境保护总局信息中心, 北京　100029)

摘　要: 阐述了环境信息数据仓库建设的需求和体系框架, 介绍了环境信息数据仓库建设的关键技术和基于数据仓库的决策支持系统, 列举了环境信息数据仓库在 2000 年—2001 年间对我国西部 12 个省、自治区和直辖市的大范围生态环境调查课题中的应用价值。

关键词: 环境信息; 数据仓库; 管理; 决策
中图分类号: X830.3　　文献标识码: C　　文章编号: 1006-2009(2004)02-0036-03

Environmental Information Data Warehouse Development and Other Technology Application

LI Qing, WANG Qiao, SHEN Wen-ming, WU Chuan-qing
(*Information Center of Chinese State Environmental Protection Bureau, Beijing* 100029, *China*)

Abstract: The need and framework of environmental information data warehouse development was discussed. Its key technology and decision-support system based upon environmental information data warehouse were introduced. The application of environmental information data warehouse in large-scale ecological environmental investigation of west region in 2000 to 2001 was introduced.

Key words: Environmental information; Data warehouse; Management; Decision

环境信息数据仓库(Data Warehouse, DW, 简称数据仓库)是 20 世纪 90 年代信息技术迅速发展的热点, 它是一种高效的数据存储系统, 可根据主题通过专业模型对不同来源数据库中的原始业务数据进行抽取与聚集[1]。数据仓库技术的实现为解决环境领域中存在的"数据量大, 有用信息量少"的问题, 以及环境信息的共享与互操作、分析与综合提供了有效途径和方法。

1　数据仓库的建设

1.1　需求分析

在数据仓库构建之前首先要进行用户需求分析, 涉及到的用户主要包括国家及地方各级环境保护部门的领导和技术人员。其中, 各级技术负责人和技术人员是数据仓库的主要服务对象。需求分析的建立可采用直接征询或亲自查阅相关环境信息建设的有关技术规范、报告和要求等方式。调查内容包括各级用户对数据仓库的使用要求, 一般会涉及到下列需求: 存储和管理工作中获得的大量原始环境数据; 编制相关的环境信息报告和图件; 进行环境综合分析与评价; 存储和管理综合研究成果信息; 建立共享数据库机制, 为其他相关部门提供信息服务。

1.2　数据仓库的体系结构

数据仓库体系结构是建立在传统的数据库管理系统上, 它们对传统的数据库进行了二次加工, 形成了不同级别的数据种类。数据仓库的体系结构模型包括: 环境源数据及其数据库、环境数据的组织、加工以及在此基础上的用户分析工具。

1.2.1　环境源数据及其数据库

环境保护工作是一项复杂的系统工程, 与环境相关的信息涉及到环境学、气象学、地理学、地质学等多学科, 数据来源十分丰富。主要有环境背景数

收稿日期: 2003-09-15; 修订日期: 2003-12-13

作者简介: 厉　青(1970—), 女, 江苏淮安人, 工程师, 博士, 从事环境遥感、环境信息的应用研究。

据(地形、地貌、地表质地、气候状况等) 、环境监测数据(大气污染监测、水质污染监测、土壤和固弃物监测、生物和生态监测、噪声污染监测、放射性污染监测、电磁辐射监测等) 、环境统计数据、地面调查数据(社会经济、生物和生态、生态环境建设项目、环境污染等) 、环境管理数据(决策管理、环境监理等) 。这些数据源主要存贮在环境数据库中, 是环境数据仓库体系结构中的基础。

1.2.2　环境数据组织

数据仓库的一个重要特征是面向主题[2], 故数据仓库要面向环境保护各个主题来组织数据。由于环境信息具有区域性, 数据仓库又是一个空间数据仓库, 所以在数据的组织中包括各个主题信息地、物的空间属性。在数据仓库中, 数据的组织是以业务工作的主题为主线, 针对不同的主题组织相应的数据体进入数据仓库中。

1.2.3　环境数据加工及其组成

数据仓库的数据来源于不同的面向具体领域应用的数据管理系统, 由于这些数据之间存在着不可避免的冗余和数据格式以及数据标准的差异, 为了优化数据仓库的分析功能, 源数据必须经过适当加工处理以最适宜的方法进入数据仓库。环境数据加工主要包括提炼、转换和空间变换。其中数据提炼包括数据的提取, 需要按照环境数据的特点, 对不需要的信息进行数据项的重构和删除, 要对环境数据库中各专题表的字段值进行解码和翻译, 补充缺少的信息、检索数据的完整性及相容性; 数据转换需统一数据的编码及数据结构、给数据加上时间标志、根据不同需要对数据进行多种运算以及语义转换; 空间变换主要指空间坐标和比例尺的统一、赋予环境数据以空间属性。

数据仓库的一般数据组成包括: 基本数据、综合数据、历史数据、元数据。基本数据就是上述经过变换后的源数据。数据仓库由元数据组织结构, 元数据是一种说明数据的数据。在数据仓库的建设中, 其历史数据由基本数据通过元数据的时间机制生成, 综合数据也由元数据的综合机制生成。数据仓库中的元数据在进行数据信息决策支持系统(DSS) 分析时, 起着定位数据仓库目录的作用, 当数据从业务环境向数据仓库环境传送时, 指导从基本数据到综合数据、到历史数据转变的算法选择。

1.2.4　客户端分析工具

数据仓库是一种支持共享和互操作的系统, 也是一种开放的支持环境保护领域中各种类型客户使用的网络化信息系统。网络的设计模式可采用 C/S 结构, 通过客户端的应用程序界面, 客户不仅可利用 GIS 中一般的查询和分析服务, 也可利用与专题信息有关的强大的分析工具。按功能划分, 数据仓库的客户端分析工具包含查寻、验证、挖掘等功能。

2　数据仓库建设的关键技术

数据仓库建设中的关键技术主要涉及到数据库和数据挖掘技术, 并在很大程度上取决于对环境数据库中知识的发现与挖掘[3]。比较适用的数据挖掘技术主要有: 规则发现技术与决策树分类技术、神经网络技术、模糊发现方法、统计方法、基于网页的三维空间可视化技术。

目前大部分数据挖掘工具主要采用规则发现技术与决策树分类技术来发现数据模式和规则, 其核心是某种归纳算法, 其优点是规则和决策树可读。神经网络技术主要是对环境数据中的非线性数据和含噪声数据的有效信息提取。模糊发现方法主要是应用模糊逻辑进行数据查询和排序。统计方法主要适合于分析现有信息, 而不善于从原始数据中发现数据模式和规则。基于网页的三维空间信息的可视化技术是研究地球系统的一种高新技术手段, 这种技术用 Web 作为传输机制, 把多维地理信息转化为直观的三维可视化模型, 模拟现实世界, 使人们更快更准地做出决策。基于 Web 的可视化包括 3 个部分: 数据集、应用模型和用户(据 Rohrer and Swing, 1997) 。其中数据集是经过变换后的源数据, 称为多维数据集, 它包括空间三维属性数据、时间属性数据以及一些专题数据。

3　基于数据仓库的决策支持系统

基于数据仓库的决策支持系统与传统的决策支持系统有很大的差异。在传统的决策支持系统中, 数据库、模型库、知识库的设计和实现是独立的, 缺乏内在的统一性, 而三者在数据仓库和数据挖掘组成的新的 DSS 构架中却得到了有效的统一。数据决策支持系统的解决方案见图 1。

4　应用

我国西部生态数据仓库建设是国家环保总局于 2000 年—2001 年间对西部 12 个省、自治区和

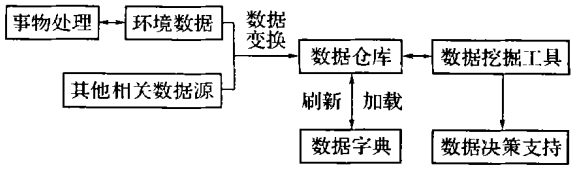

图 1　数据决策支持系统的解决方案

直辖市的大范围生态环境调查的一个应用研究课题, 研究区域涉及的数据量大, 类型众多。有 3 个时期(20 世纪 80 年代后期、90 年代中期、2000 年) 的遥感影像和解译数据、有大量的地面调查数据(社会经济、生物和生态、生态环境建设项目、环境污染等) 和生态环境背景数据(地形、地貌、地表质地、气候状况等) , 还有生态环境评价等。大批数据既具空间特性又具时间特性。数据源的存贮形式也不一, 与遥感相关的数据以文件形式存贮, 其他数据主要以数据库形式存贮。

为了更好地提取有用信息, 研究时应用了数据仓库技术。在具体实施中遵循由主题组织数据的原则, 按照生态环境背景、遥感影像数据库、土地利用及土地覆被空间数据库、典型案例生态环境数据库及相关元数据库等方式组织数据, 并对数据进行规范整理, 将它们转换到统一的平台(SuperMap 的工作空间) 上集中管理。各专题数据组成相关的数据集, 并采用 SuperMap 公司的 SDX(空间数据引擎) 平台和 SQL 数据库, 使空间数据和属性数据都存贮在数据库中, 以有利于数据的提取和提炼。在数据提炼中, 根据不同的应用目的, 采用了相关的数据挖掘方法, 形成基础应用模型。

5　结论

数据仓库对环境信息建设的意义十分重大, 可以大大提高环境保护工作的效率。环境数据大多为空间数据, 它们的量纲不一、形式多样, 既有定量测量数据, 又有定性的文字描述。解决环境问题的方法多种多样, 这在很大程度上与数据的不确定性、经验性、间接性和完整性等因素有关, 而且每一种方法均能产生与其他方法不同的数据源。

[参考文献]

[1]　王建平, 刘　琪. 数据仓库在信息处理中的作用[DB/OL]. http: //electron. cetin. net.

[2]　周丽娟, 柳　池, 刘大昕. 关于数据仓库若干关键技术的研究[J]. 微机发展, 2002, 12(1): 29－31.

[3]　李德仁, 王树良, 史文中, 等. 论空间数据挖掘和知识发现[J]. 武汉大学学报(信息科学版) , 2001, 26(6): 491－499.

(上接第 33 页)

引起监测数据的突变。以下是常遇到的几种情况:

(1) 监测浓度值为负数;

(2) 监测浓度值连续出现 3 个以上恒定的不变值;

(3) 单个监测子站的某项污染物浓度值出现 1 个、数个极高或极低的值, 而相同时段内其他子站无此现象出现;

(4) 某一子站的某项污染物浓度变化趋势与其他子站明显不一致。

3　异常突变值的判断与处理

异常突变值的判断及处理程序见图 1。

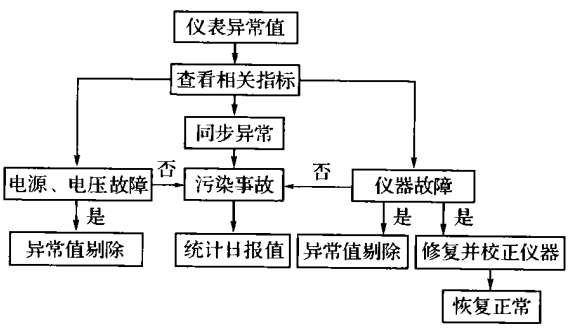

图 1　异常突变值的判断及处理程序

[参考文献]

[1]　陈　伟, 吴　楠. 环境空气自动监测系统的常见故障及排除[J]. 环境监测管理与技术, 2002, 14(1): 36.

[2]　李　劲, 周英涛, 张　勇, 等. 环境空气自动监测故障与维修特例[J]. 中国环境监测, 2002, 18(3): 55－57.