渤海激光单分子海洋油气化探综合评价与预测系统数据仓库设计方案

初晓璐[1], 杨作升[1], 张　勇[1], 刘　展[2], 李谷祺[1]

(1. 中国海洋大学 河口海岸带研究所, 山东 青岛 266003; 2. 中国石油大学(华东), 山东 东营 257061)

摘要: 以渤海区域地学数据为例, 阐述了渤海激光单分子海洋油气化探综合评价与预测系统数据仓库的设计方案。该数据仓库采用了以渤海激光单分子海洋油气化探数据为中心的星架数据管理模式, 并实现了空间数据与属性数据的一体化存储、数据的可视化抽取与可视化挖掘等关键性技术。

关键词: 激光单分子海洋油气化探; 数据仓库; 数据抽取; 数据挖掘
中图分类号: P208　　文献标识码: A　　文章编号: 1000-3096(2007)06-0015-05

数据仓库(Data Warehouse, 简称 DW)概念起源于 20 世纪 80 年代中期, 而后又经过被誉为"数据仓库之父"的 Prism Soultion 公司副总裁 W. H. Inmon 加以定义与发展。1900 年对数据仓库所作的定义为:"数据仓库是面向主题的、集成的、稳定的、不同时间的数据集合, 用于支持经营管理中的决策过程。"[1,2] 相对于传统的数据库(Data Base, 简称 DB)侧重于数据的操作而言, DW 则侧重于数据的分析处理, 是一种针对海量数据进行管理与分析的技术。近年来, 数据仓库技术得到了迅速的发展和应用, 而将 DW 技术引入到油气化探数据的分析管理中, 并与 GIS 融合, 则可为油气化探的研究提供一种新的有效方法。中国油气化探历经几十年, 积累了大量的油气化探数据资源, 但数据库的建设工作却相对比较薄弱。国土资源部开发了省级化探数据库, 提供了基本的数据处理功能, 主要是侧重于固体矿产资源调查的化探数据管理, 但它们都是基于关系数据模式的, 缺乏空间数据的管理功能。2000 年中国新星石油公司化探中心首次建立了中国主要含油气盆地油气化探数据库, 收集并录入了 30 年来中国主要含油气盆地或地区油气化探数据及相关信息, 为系统研究中国区域油气地球化学场特征提供了数据资源和基础。

面向主题是数据仓库最重要的特征之一[3]。地学数据仓库是激光单分子海洋油气化探综合评价与预测系统的基础。作者根据国家和相关行业已经公布的数据标准, 并考虑到渤海地学数据的实际情况, 结合国内外数据仓库的建设经验, 紧紧围绕油气化探数据这一主题, 建立了渤海激光单分子海洋油气化探综合评价与预测系统数据仓库, 为数据仓库应用于海洋油气化探预测提供了新的思路。

1　渤海激光单分子海洋油气化探综合评价与预测系统数据仓库结构

"激光单分子海洋油气化探综合评价与预测系统"是海洋多学科、多技术的综合体。其数据大都属于空间数据, 并且具有多源性和多样性。多源性表现为数据来源于多个部门、多种采集系统, 数据具有多种比例尺, 数据结构也存在着较大的差异; 多样性则表现为数据种类多, 其内容包括了物探、化探、地质、油气等。这些数据主要有 3 个方面的特点: (1) 存在大量的描述性数据, 难以简单的量化; (2) 名称多且杂, 特别是地质数据; (3) 存在大量的图形, 若采用直接方式存储地质图形的话, 一方面会占用大量的存储空间, 另一方面, 图形内部各元素和图形之间的关系也无法表示。对于空间数据的分析与管理, 地理信息系统(GIS)是一种可以借助的技术, 库中只存放矢量数据及数据关系, 在取出数据时, 再经 GIS 合成处理成相应的图形。因此, 可以将 GIS 与数据仓库技术有机地结合起来, 形成 GIS 数据仓库, 从真正意义上实现海底数据的智能化管理。

根据"激光单分子海洋油气化探综合评价与预测系统"所涉及的数据特点及应用需求, 设计"激光单分

收稿日期: 2004-04-05; 修回日期: 2005-03-20
基金项目: 国家 863 计划资助项目(2002AA615160)
作者简介: 初晓璐(1982-), 女, 山东掖县人, 在读硕士研究生, 研究方向为地理信息系统在地学中的应用, E-mail: cxlu222@hotmail.com

子海洋油气化探综合评价与预测系统”数据仓库时应考虑到数据的特殊性,在把握面向主题的前提下,以 GIS 作为工具平台,借鉴一般数据仓库的设计方法,将数据仓库中的数据组织以空间为核心进行组织。依据这样的指导思想,作者设计了如图 1 所示的适于“海洋油气化探分析”的 GIS 数据仓库模型。

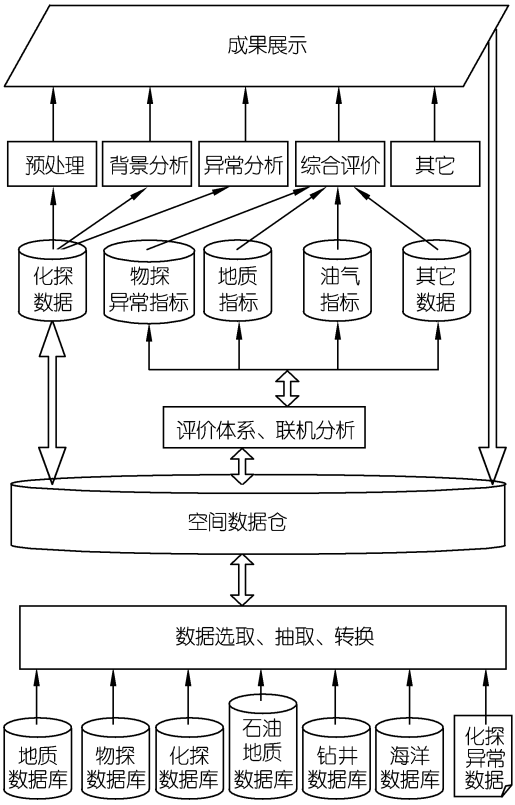

图 1 渤海激光单分子海洋油气化探综合评价与预测系统数据仓库逻辑模型

Fig. 1 The logic model of data warehouse of laser single molecule benthal geochemical prospect of petroleum and gas appraisement and forecast system in the Bohai Sea

适于“渤海激光单分子海洋油气化探综合评价与预测系统”的 GIS 数据仓库模型基本上是由六大层组成的,由下向上依次为:基础数据库层、数据核心层、数据共享层、数据抽取层、应用层和数据成果层。该模型的下部三层是数据仓储层,它以常规的关系数据库为基础,以数据仓库技术为核心实现数据的高效管理。最底层的专业数据层主要由常规的现存的专业数据库组成,它是“海洋油气化探分析”数据仓库的基础数据库系统,所有的原始数据均存储在这一层,这样做可以充分利用已有的数据库资源;第二层是数据核心层——数据仓核心仓库,是数据仓储的数据来源,主要是从各专业数据库获取。它并不是对底层数据库的简单堆积,而是要按照统一的数据标准对基础数据库中的数据进行提取、转换、汇总等,以实现高效、自动地进行数据的管理和使用。数据仓储在系统中起承上启下的作用,它将底层基础数据提炼转换形成统一的数据体,供上层模块使用及外部共享;第三层是数据共享池,它提供数据共享的工具,为上层的数据应用及外部数据调用准备数据。而上部三层则是数据应用层,也就是通常所说的数据仓库的数据挖掘和分析。它包括海底油气资源评价、海洋工程评价、海洋区域经济评价等解释及决策方法模块,可以称之为数据仓库系统的方法库或函数库或工具盒。其中数据抽取层是数据仓库系统功能的一个重要体现,它提供可视化的数据抽取工具,根据应用层的需求从数据共享池中利用 GIS 技术为相应的评价与解释模块可视化地提取所需的数据,如海底油气资源评价方面,如根据储层描述、储量计算、油藏模拟、经济评价等方法对数据的需求,数据抽取工具应能自动地或人机交互可视化地从数据共享池中提取所需要的数据(地质、构造、物探、化探、钻井、海底地形等),通过评价模块形成对目标地质单元的认识,对可能获取的油气资源储量的评价,以便为最终进行勘探决策提供可靠的依据。最顶层是数据产生层,它利用 GIS 工具为应用层提供数据解释与评价结果的图件与表格的制作工具,其数据将被存储到数据仓共享池中,作为数据仓库的成果数据。

2 主要技术路线

本系统采用 Oracle9i 作为基础数据库平台与数据仓库平台[4,5],以 Delphi7.0 为开发工具,结合 MAPX 控件进行开发。主要技术路线与技术要点如下。

2.1 基础数据库数据标准

参照国家和相关行业的数据标准[6-9],作者采用了这样一套解决方案[10,11]。此方案既避免了数据的冗余,又实现了空间数据库与属性数据库的相互对应。

2.1.1 空间图层划分与命名

在 GIS 中空间数据是分图层管理的,作者采用了下列分组码命名规则来对图层文件进行命名,它可以保证多幅拼接后每个图形信息及相应属性信息的独立性,防止图层名重复出现。图层名编码结构如图 2。

若图名超过 3 个汉字时则取前两个汉字和最后一个汉字的拼音首字母。若出现重名时,则前两位不变,第三位改为数字顺序号。

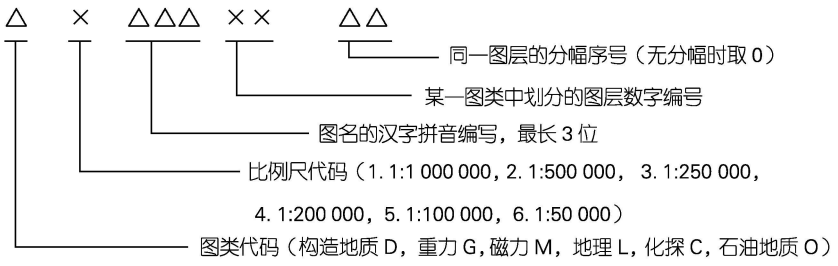

图 2　图层命名规则

Fig. 2　The rule of layer coding

对于每个图层中的点、线、面等图元，除给出了唯一的图元编码外，还按有关规定定义了其空间属性（点符号、线形、线色等）。

2.1.2　属性数据库命名、编码规则

属性数据通常用来反映与空间实体对应的属性，一般是通过分类、量算、命名、统计等方法得到的。简单地说，它就是对应空间实体属性的数据库。

每个图层的图元（点、弧段、多边形）的性质、意义等通过属性数据描述，这些数据的集合构成了一个图层的属性文件，为保证其唯一性与相应空间图层的相关性，采用了图 3 所示的图层属性文件命名规则。编码方式与图层名编码相同，识别码采用字符，取属性表主要含义的一个汉语拼音的首字母。对于图层属性表中每一个属性字段均按 GB9649-88 中定义进行了编码。

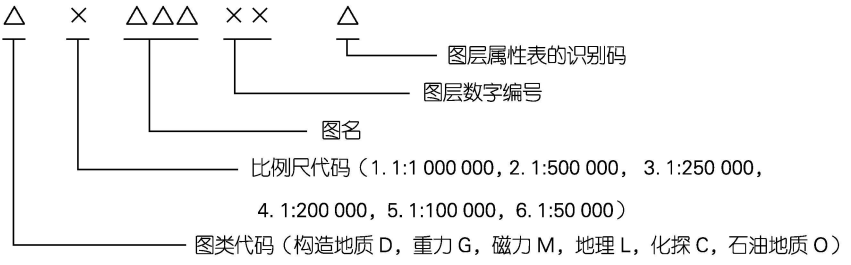

图 3　图层属性表命名规则

Fig. 3　The rule of attribute table coding

2.2　空间数据与属性数据的一体化存储

目前的大多数 GIS 产品，空间数据存储和管理是基于文件系统的，而空间数据的描述信息即属性数据则由关系数据库进行管理，两者之间用关键字或指针联系。空间数据和属性数据的分离管理造成数据完整性和一致性不能保证、缺乏数据动态增长能力和数据优化管理、系统查询能力和分析效率低下，以及数据共享和并行处理无保证。地学数据仓库的物理表现为一个多维数据库，不仅要对超大数据库进行动态管理，还要实现多进程、多线程、内存缓冲、快速索引、数据完整性和一致性保证、并发控制、安全和恢复机制及分布式处理机制。这些都要求空间数据与属性数据统一在一个数据库管理系统下。另外管理地学数据的数据仓库系统除了具有普通数据库所具有的功能外，还应具有如下特征：（1）支持空间数据操作；（2）查询语言具有可扩充性；（3）空间数据的有效存储及组织。所以数据仓库系统必须引入新的技术来满足对空间数据管理的要求。目前，主要的做法有 2 种：（1）数据库管理系统与空间操作实现模块分开。空间操作由 GIS 操作模块实现，数据的查询和存取采用双库结构，由通用的 DBMS 和空间数据管理软件包实现。（2）以当前的关系数据库技术为基础，按要求进行扩充和完善。

在本系统开发中，作者采用 Oracle9i 存放空间数据与属性数据。在基础数据库中，将空间数据与属性数据分开存放在 Oracle9i 中，二者通过关键字段——图元编码来连接[12]。在数据仓库中，充分利用 Oracle9i spatial 的功能，将抽取后的空间数据与属性数

据自动关联起来,形成了空间数据与属性数据的一体化存储。这样做的优点是真正实现了空间数据与属性数据的统一,能够实现数据的时时更新,便于数据的维护和用户使用。

2.3 以激光单分子海洋油气化探数据为事实表的数据组织构架

数据仓库中的数据管理模式与一般意义上的数据库有着本质的差异,目前商用数据仓库的数据管理模式研究得比较多,也比较成熟,但是对基于空间数据的地学数据仓库的管理模式还没有可引用的,还处于研究中。在本系统中,作者采用了以主题数据为中心的星架管理模式,即以化探数据为事实表,以其它数据(断层、油气、磁力异常、重力异常)为维度表的星架结构,如图 4 所示。

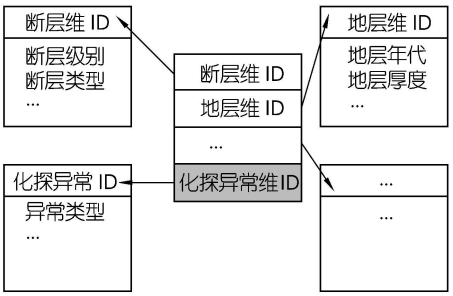

图 4 激光单分子海洋油气化探数据组织构架

Fig. 4 The framework of data in laser single molecule berrthal geochemical prospect of petroleum and gas

数据在数据仓库中是以事实表、维度表的形式存储的,这种存储形式非常利于以各种数据模型直接对数据进行分析,使数据仓库形成一个集成系统为用户提供各种分析服务。

2.4 数据可视化抽取与数据可视化挖掘

2.4.1 数据可视化抽取

数据抽取是数据仓库获取数据的重要工具,它是指用户按照应用主题,将选定区域中与主题(油气化探评价)相关的数据从基础数据库(源数据库)中抽出,并按照主题对数据的要求,经过重新组织后再存储在数据仓库中。它包括空间图层抽取和属性抽取。目前,数据仓库抽取主要在商业领域应用较多,处理的数据也仅仅局限于没有空间标识的商业数据。

数据可视化抽取就是根据主题要求,将与主题相关的数据由基础数据库抽取到数据仓库中,并且使之存储到数据仓库中。它一直是空间数据库,特别是专业性较强的应用系统开发中的技术难点,目前还没有成熟的通用方法。可行的技术方案有两个:一是利用 GIS 控件技术在 GIS 平台上进行二次开发;二是使用基于数据仓库的 GISOLAP(GIS On Line Analytical Processing,联机分析处理)技术[13],GISOLAP 模型的实现是在数据仓库的基础上,首先通过属性数据仓库专业的分析工具——OLAP,产生属性数据的分析结果,然后通过数据绑定技术,把产生的分析结果与空间数据连接,再通过 GIS 的空间分析、空间定位以及可视化等功能,把综合分析的结果通过电子地图和其他形式直观地表示出来。

在本系统中,作者根据地学行业的工作要求,设计了按工区抽取和按工程项目抽取两种方案。要采用 GIS 技术进行开发,必须对空间数据进行处理,根据系统的实际要求,使用 MAPX 控件技术,开发出了一套基于图形的抽取工具,为空间图层数据抽取设计了点抽取、线抽取和面抽取等工具包;并且为属性数据抽取设计了通过选定区域空间图层索引相关的属性数据库进行数据提取的工具,使得空间数据的属性数据会自动抽取出来。这样基本实现了基于空间数据的可视化抽取技术,形成空间数据和属性数据的统一。

2.4.2 数据可视化挖掘

一般来说,数据挖掘是从大型数据库或数据仓库中发现并提取隐藏在其中的模式和关系的过程,目的是帮助分析人员寻找数据间潜在的关联,发现被忽略的要素,是数据仓库优越性的重要体现[14]。目前国内,数据挖掘仅仅应用在商业中,例如保险、银行等行业,所挖掘的数据也仅仅是和空间数据无关的商业数据。

数据挖掘工具的设计要解决好两个问题:一是要合理地设计方法库和模型库,它们由常规的统计分析方法和专业数据解释与评价方法组成,构成数据挖掘的基础。方法库和模型库的基本单元是算法,这些算法是通用的,各种分析处理方法是由不同的算法组成,故方法库和模型库又可以理解为系统的函数库或工具盒,它们的调用和参数赋值是通过消息驱动来实现的。使用过程中,既可直接选择方法、设置参数进行分析处理,又能间接地利用数据档案自动完成;二是要实现数据的可视化提取,实现数据分析与评价方法所需数据的自动或人-机交互式可视化提取。

数据可视化挖掘是数据仓库的难点,目前还没有成熟的通用方法。在本课题中,作者采用 Oracle9i Spatial 的 OLAP 技术,利用神经网络、灰色理论和模糊数学等综合评价模型,结合 MAPX 控件进行可视化挖掘,寻找和油气化探有关的因素,基本上解决了数据可视化挖掘的技术难点。

在系统开发中,通过以下三个步骤实现了数据挖掘:(1)数据准备。根据数据挖掘(数据处理与评价)的需求从数据仓库中使用数据抽取工具包,可视化地提取和油气化探分析与评价主题相关的数据,存放在

数据仓库的临时层中,以便下一步进行调用。(2) 数据格式转换。根据挖掘工具(油气化探处理与评价模型)对数据格式的要求,使用数据格式转换工具,将不同格式的数据转换为挖掘工具(油气化探处理与评价模型)所要求的格式。(3) 模型挖掘。数据挖掘是数据仓库最重要的应用之一。数据挖掘利用数据挖掘工具在数据中查找模型,这个搜寻过程可以由系统自动执行,自底向上搜寻原始事实以发现它们之间的某种联系,也可以加入用户交互过程,由分析人员主动发问,从上到下地找寻以验证假定的正确性。

本系统首次提出了基于地学数据仓库的可视化挖掘,是对 GIS 在地学领域应用的一项补充和尝试。

3 结论

作者参考国内外先进的数据仓库设计方案,根据国家和相关行业已经公布的数据标准,建立了以渤海激光单分子海洋油气化探数据为主题的数据仓库结构。DW 技术的引入实现了空间数据与属性数据的一体化存储,并且通过使用 MAPX 控件技术,结合以主题数据为中心的星架管理模式,完成了数据的可视化抽取和可视化挖掘,为数据仓库在地学中的应用提供了新的思路。

参考文献:

[1] Inmon W H. 数据仓库[M]. 王志海译. 北京: 机械工业出版社, 2003.

[2] Inmon W H, Hackathorn R D. Using the Data Warehouse[M]. 北京: 机械工业出版社, 2002.

[3] 王广清, 杨学良. 数据仓库技术及其在电信计费领域应用的探讨[J]. 计算机工程与应用 1999, 9: 98-102.

[4] 石丽, 李坚. 数据仓库与决策支持[M]. 北京: 国防工业出版社, 2003.

[5] 飞思科技产品研发中心. Oracle 数据仓库构建技术[M]. 北京: 电子工业出版社, 2003.

[6] GB9649-88, 地质矿产术语分类代码(上、中、下)[S].

[7] DZ/T0127-94, 固体矿产矿点(矿床地质数据文件格式)[S].

[8] DZ/T0126-94, 固体矿产(钻孔地质数据文件格式)[S].

[9] DDZ 970, 资源评价工作中地理信息系统工作细则[S].

[10] 张勇, 杨作升. 渤海区域地质信息管理系统数据模型[J]. 海洋学报, 2002, **24**(4): 76-81.

[11] 张勇, 杨作升. 渤海区域地质信息管理系统数据库设计方案[J]. 海岸工程, 2002, **21**(4): 1-5.

[12] 张勇, 杨作升. 利用 MAPX 实现空间数据库与属性数据库的挂接[J]. 青岛海洋大学学报, 2003, **33**(1): 87-94.

[13] 林杰斌, 刘明德. 数据挖掘与 OLAP 理论与实务[M]. 北京: 清华大学出版社, 2003.

[14] 陈京民. 数据仓库与数据挖掘技术[M]. 北京: 电子工业出版社, 2002.

Data warehouse design plan for the Bohai Sea laser single molecule benthal geochemical prospect of petroleum and gas appraisement and forecast system

CHU Xiao-lu[1], YANG Zuo-sheng[1], ZHANG Yong[1], LIU Zhan[2], LI Gui-qi[1]

(1. Institute of Estuarine and Coastal Studies, Ocean University of China, Qingdao 266003, China; 2. China University of Petroleum (East China), Dongying 257061, China)

Received: Apr., 5, 2004

Key words: laser single molecule benthal geochemical prospect of petroleum and gas; data warehouse; data extraction; data mining

Abstract: Taking the geologic data in the Bohai Sea as an example, the data warehouse design plan for the Bohai Sea laser single molecule benthal geochemical prospect of petroleum and gas appraisement and forecast system is described in the this article. This data warehouse adopted astral frame data management mode, whose center is the data of the Bohai Sea laser single molecule benthal geochemical prospect of petroleum and gas. Some key technologies are realized, such as unitary storage of spatial data and attribute data, visual data extraction and data mining and so on.

(本文编辑: 刘珊珊)