第 31 卷 第 6 期 海 洋 学 报 Vol 31, No 6
2009 年 11 月 ACTA OCEANOLOGICA SINICA November 2009

海洋多维数据仓库构建研究

季民[1], 靳奉祥[1], 李婷[1], 赵相伟[1]

(1. 山东科技大学 测绘科学与工程学院, 山东 青岛 266510)

摘要: 空间数据仓库的数据集成能力以及对复杂数据分析、高层决策的支持可为异质异构海洋数据的集成和综合应用提供方法论。以海洋多维数据仓库构建为目标, 对维、维层次、海洋多维数据模型等概念进行了形式化描述和定义, 并且以海洋渔业主题分析为例, 对渔业生产的事实、维度、维层次及维层次关系进行识别。针对时空维的复杂度, 以折衷的混合多维数据模型结构对海洋多维数据仓库进行构建。

关键词: 数据仓库; 多维数据库; 海洋渔业; 混合模型

中图分类号: TP311; P208 文献标志码: A 文章编号: 0253-4193(2009)06-0048-06

1 引言

人们在长期的海洋观测和渔业生产过程中积累了丰富的海洋数据和渔业生产统计数据, 这些数据是人类认识海洋、揭示鱼群洄游规律的重要数据来源, 但是现有的数据均以不同的格式、不同的尺度、不同的空间基准存储在不同的专题数据集中, 从而在不同程度上限制和阻碍了海洋数据的综合应用及海洋现象发现的过程。为了更好地适应当今社会对各级尺度下海洋问题的研究, 需要有 1 个统一的信息视图将大量历史的、现实的海洋数据按照相应的研究主题转换成统一的格式, 从而实现异质异构数据的集成、存储和管理。空间数据仓库的数据集成能力以及对复杂数据分析、高层决策的支持为该问题的解决提供了方法论[1]。

要构建海洋空间数据仓库, 需要根据面向海洋主题问题, 通过元数据驱动机制, 将异质异构的海洋数据经过清洗、抽取和变换, 以多维方式组织在仓库的综合数据集中; 在具体的分析过程中可按照主题问题, 将任一维与其他维进行组合, 以多维方式显示数据, 供人们从不同角度、多方位地认识复杂的海洋世界。对海洋多维数据仓库的构建并没有一种现成的或万能的建模方法, 本文在总结多维数据库相关概念的基础上, 对海洋多维数据模型进行了形式化定义。以海洋渔业主题分析为例, 进行了事实、维度及维层次的识别, 并且构建了以混合模型为主的海洋多维数据仓库。

2 多维数据库基本概念

2.1 数据立方体

数据立方体(data cube)是多维数据仓库中组织和存储数据的重要手段[2], 它采用多维立体数据存储方式来取代传统的平面数据存储方式, 为进行多维数据分析提供根本保证。在理论上立方体可以具有 n 维。图 1 中的每个三维立方体直观地反映了渔获量在 3 个维度(渔业公司、渔区、渔种)上的变化和对比。若增加第四维——时间维, 则可将其看作三维结构的立方体随时间维的变化序列。按照这样的组织方式, 可以把任何 n 维数据的显示看成是一个 $n-1$ 维"立方体"的序列。

收稿日期: 2009-04-10; 修订日期: 2009-07-22。

基金项目: 现代工程测量国家测绘局重点实验室开放课题资助项目(TJES0805); 海洋溢油鉴别与损害评估技术国家海洋局重点实验室开放基金资助项目(200903); 国家"八六三"计划项目(2009AA12Z147); 海岛(礁)测绘技术国家测绘局重点实验室资助项目(2009B14)。

作者简介: 季民(1970—), 男, 山东省齐河县人, 副教授, 博士, 从事空间数据组织和 GIS 系统集成研究。E-mail: jimin@sdust.edu.cn; jamesjimin@hotmail.com

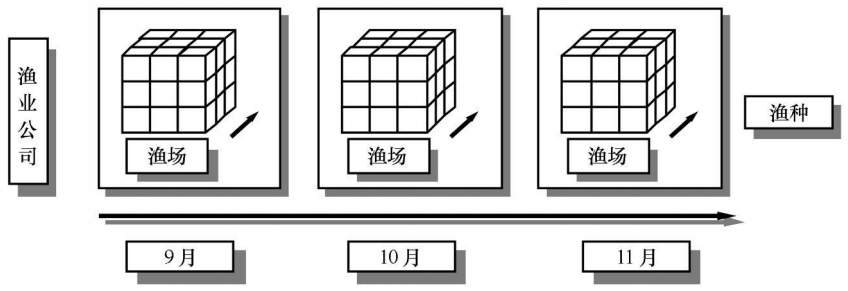


图 1　渔业生产渔获量四维立方体表达

2.2　维

在多维数据库中维是一个必要而又与众不同的概念[2]。多维数据模型的一个主要目标就是利用维为事实的访问提供尽可能多的途径, 每维代表一个统一的访问数据仓库的信息路径。

2.3　维层次

在实际中数据立方体的维常常具有多个属性, 这些属性可按细节程度组织成层次形式[3], 例如时间维层次为日、周、月、季、年。维层次能清楚地体现下钻和上钻操作。对象间的层次关系可抽象为 3 类[4]: 特化/ 泛化关系、聚集关系以及成员关系。无论维层次结构中的哪种关系, 低维层所对应的现实世界中的对象在逻辑上都包含于(或"小于") 高维层所对应的对象[5]。

2.4　维成员

维的 1 个取值称为该维的 1 个成员, 若维是多层次的, 则维成员是在不同维层次中取值的组合[6-7], 例如时间维有日、月、年 3 个层次, 分别在日、月、年上取 1 个值组合起来就得到了时间维的 1 个维成员, 即"某年某月某日"。

2.5　事实

在海洋渔业领域, 事实代表着要进行分析的且我们感兴趣的模式或事件, 在大多数多维数据模型中, 事实由其组合在一起的维值的隐式定义, 只有当特定维值的组合没有造成空穴时, 1 件事实才会存在, 然而有些模型将事实看成具有独特个性的一级对象。大多数多维模型也要求将每件事实与每维最低级别的 1 个维值建立联系, 但是有些模型放宽了这种要求。每件事实都有一定的粒度, 该粒度是由不同维值组合的层次所确定的, 如 4 个维度(渔业公司、渔区、渔种、时间) 在各自不同的层次级别上的取值就确定了在渔业生产这件事实的不同渔获数据的综合程度。

2.6　度量

度量包括两部分内容: 一是事实的数量属性, 如产品销售中的价格或利润; 另一个是公式, 通常是 1 个简单的综合函数, 比如 sum, 它能够将几个度量值组合为 1 个。在 1 个多维数据库中度量一般代表用户想去优化的事实的属性。

2.7　多维数据操作

(1) 切片(slice) 和切块(dice)。在多维数据的某一维上选定一维值的动作称为切片。切片的结果可使原多维数据降维。切块与切片类似, 在多维数据的某一维上选定 1 指定维值区间的动作称为切块。选定多维数据的 1 个三维子集的动作也可以称为切块。

(2) 下钻(drill-down) 和上钻(drill-up)。下钻和上钻是运用维的层次结构对数据进行综合程度分析表达的相反过程。从较高的综合层次跳转到较低的综合层次去的分析方法叫下钻, 反之叫上钻。

(3) 旋转(rotate)。旋转是改变 1 张表格或页面显示的维方向。例如, 旋转可能包含交换行和列或是把行维变换到列维中去。

3　海洋多维数据模型形式化定义

数据立方体是数据仓库和联机分析处理的核心概念之一。众多研究者对以数据立方体为核心的多维数据建模进行了卓有成效的研究。文献[7] 根据海洋渔业数据的特点构建了 1 个栅格化的、面向对象的多维数据立方体模型, 为了表达属性维的层次结构, 模型中引入了类和对象的概念, 并且以四元组进行表达。文献[8] 提出了通过正交指令来实现多维数据概念模型, 这个模型定义代数及微积分并使其等价, 它能够清楚地分离出结构和内容, 允许采用相当简单和透明的方法来定义数据处理语言, 特别在表示数据立方体操作符时十分简单。文献[9] 提

出了 1 种支持多维分析语义描述的形式化工具——数据立方体代数, 它的概念清晰、简单而易于理解。文献[10] 通过对多维数据库和多维视图的区分, 以形式化方法给出了多维数据库的定义及导出方法。文献[11] 以编序和映射为基础, 提出了一种能够充分表达数据仓库复杂数据结构和语义的多维数据模型, 并且提供 1 个以 OLAP 操作为核心的操作代数, 支持层次结构间的复杂聚集操作序列。利用层次结构间聚集函数的约束, 提出定义层次游标的方法来解决维层次结构的选择问题[12]。基于前面众多研究者的成果, 结合海洋数据的特点, 本文给出如下的海洋多维数据库模型的形式化定义。

定义 1: 维是进行海洋数据观察的角度, 它是由维关键字、维属性和维层次组成的三元组, 即 $d_k = < D\_key, D\_attr, D\_h >$, 其中 D_key 是维的关键属性, h 代表维层次(hierarchies); D_attr 是维的描述性属性, 非关键属性; D_h 是维层次结构。维的值域为 $dom(d_k) = \{dm_1, \cdots, dm_i, \cdots, dm_n\}$, 其中 dm_i 是维成员。

定义 2: 维层次是进行数据观察的细节程度, 根据维属性的取值, 可对维成员进行分组形成多个层次 $h(d_k) = \{h_1, h_2, \cdots, h_n\}, n > 0$。各层次的值域是维的值域的子集, 即 $dom(h_i) \subset dom(d_k)$ (当 n 等于 1 时, 表示该维中只有 1 个层次, 否则为多个层次)。

定义 3: 维层次关系是维层次之间的一种二元关系, 即对于任意 2 个维层次 h_i 和 h_j, 如果有 1 个函数 $F: dom(h_i) \rightarrow dom(h_j)$, 则称 h_i 和 h_j 满足层次关系 $\leqslant$, 记为 $h_i \leqslant h_j$, F 称为层次关系的分类函数。维层次关系主要分为 3 类: 特殊/泛化关系、聚集关系、成员关系。维层次关系结构记为 $\theta = \{(h_i, h_j, \varphi_{ij}) / h_i, h_j \subset h(d_k), h_i \leqslant h_j, \varphi_{ij}$ 称为层次关系分类函数$\}$。

定义 4: 维层次链描述的是维层次上的有序关系, 对于 $\{h_1, h_2, \cdots, h_m\} \subset h(d_k)$, 若 $h_1 \leqslant h_2 \leqslant \cdots h_{m-1} \leqslant h_m$, 则称$\{h_1, h_2, \cdots, h_m\}$为维 d_k 的一个层次链。

定义 5: 度量(measures) 是海洋多维数据模型中的事实属性, 它依赖于维, 并且和维共同反映多维事实对象, 它是由属性和操作组成的二元组, $M = < M\_attr, M\_oprt >$, 其中 M_attr 是度量属性, 可分为基本度量和导出度量; M_oprt 是度量可以执行的操作, 如 sum, average 等操作, 其值域为 $dom(M)$。

定义 6: 海洋多维数据模型由三元组描述, 即 $R = < M, D, E >$, 其中,

(1) $M = \{m_1, m_2, \cdots, m_k\}$, 称为度量属性集合;

(2) $D = \{d_1, d_2, \cdots, d_n\}$, 为所有维的集合, d_i 称为维;

(3) 度量属性集合 M 函数依赖于维集合 D, 即 D 和 M 之间有函数 $F: dom(d_1) \cdots dom(d_n) \rightarrow dom(m_1) \cdots dom(m_k)$, 其中 $dom(d_i)$ 是维 d_i 的值域, $dom(m_j)$ 是度量属性 m_j 的值域, 而 E 是依赖函数的集合。

定义 7: 海洋数据仓库是 1 组多维数据集合, 记为$(< M_1, D_1, E_1 >, < M_2, D_2, E_2 >, \cdots, < M_m, D_m, E_m >)$。

定义 8: 维聚集是对海洋多维数据集合 $R = < M, D, E >$ 的维集合 D 中的维度 d_i 进行聚集操作, 可表示为 $Dagg(R, d_i, \Psi)$, 其中 Ψ 是针对各度量属性的聚集函数集合, 维聚集的结果是去掉维度 d_i 的新的多维数据集合。

定义 9: 层次聚集是对海洋多维数据集合 $R = < M, D, E >$ 的维度 d_i 的层次链 $h_1 \leqslant h_2 \leqslant \cdots h_{m-1} \leqslant h_m$ 进行聚集操作, 可表示为 $Hagg(R, d_i, h_k, \Phi)$, 其中 Φ 是将各度量属性的聚集函数集合沿层次链经过 k 次聚集, 层次聚集的结果是去掉维度 d_i 的层次链的前 k 个层次而得到的新的多维数据集合。

4 海洋多维数据仓库构建

海洋多维数据模型的形式化定义为多维数据仓库的构建提供了理论基础。根据数据仓库构建的过程, 海洋多维数据仓库的构建需要经过概念模型设计、逻辑模型设计和物理模型设计等几个阶段。海洋数据仓库概念模型设计的目标是根据主题分析需求对事实、度量、维度和层次进行确定, 并且建立面向主题的信息集合的概念性定义。逻辑模型设计是采用相适应的多维数据模型, 如星型模型等, 对多维数据进行逻辑表达。在完整的逻辑设计基础上就可以进行海洋多维数据仓库的物理模型设计。下面就以海洋渔业主题为例对渔业多维数据库构建的各层次模型设计进行实践应用。

4.1 事实及维度的识别

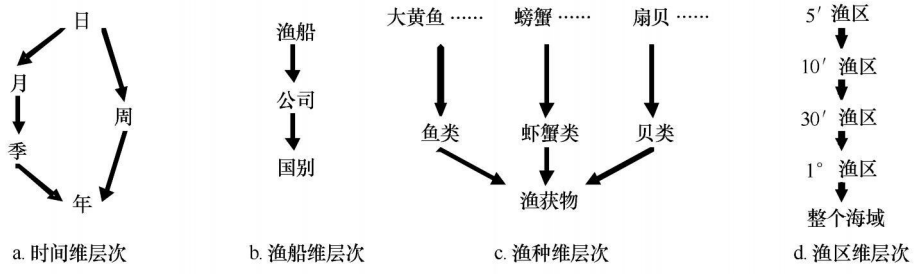
根据面向的海洋渔业主题的特点, 海洋分析决策人员关心的问题包括渔业资源的评估、渔场的时空分布预测、海洋环境因素与渔场关联规则的提取等, 这些问题的解决无不涉及历史渔获统计数据, 为此我们可以将渔获产量的统计作为事实, 其中产量和投网次数作为度量, 具体事实表结构如图 2 所示,

其中作业时间、渔获种类、作业方式、渔船所属公司等可作为事实的维度。对其他渔业问题进行分析时, 除了考虑渔获事实之外, 还要考虑同期的海洋环境要素的观测事实, 因此在事实表中还应增加温度、盐度、叶绿素浓度的度量数据。由于各类度量数据的量纲、大小等的差异比较大, 为了便于以后数据的分析和关联规则的提取, 我们可对各度量数据分别进行归一化处理, 规一化后的数据的取值区间为[0, 1], 具体归一化的公式为 $f(x) = (x - x_{min})/(x_{max} - x_{min})$, 其中 x_{max}, x_{min} 分别为相应度量数据的最大值和最小值。

图 2　海洋渔业事实表划分

4 2　维层次及层次关系的识别

维层次是多维数据建模中一个非常重要的概念, 而且若没有维的层次关系, 就无法进行上钻、下钻的分析操作。根据渔获事实和海洋环境要素事实数据分析的需要, 对时间维、渔船维、渔种维、渔区维的维层次结构定义如图 3 所示。



a. 时间维层次　b. 渔船维层次　c. 渔种维层次　d. 渔区维层次

图 3　各维层次结构

4 3　多维模型结构的实现

在多维数据模型中由于星形模型中维表不能有效地表达维的层次结构, 并且在雪花模型中当维层次过多时又会增加数据查询的复杂度, 因此人们提出了折衷方案的混合模型。在混合模型中只有对最复杂的维才进行标准化。在海洋渔业主题分析中, 由于渔区维和时间维层次的复杂度, 因此可选用混合模型对多维数据进行组织。

4 3 1　度量数据的可计算性和事实表划分

为了保证事实表中数据分析的有效性, 相关表中的数据粒度必须一致, 这又会引伸出另外的 1 个问题, 即度量数据的可计算性。可计算性通常是指度量数据的可加性, 包括充分可加、不可加、半可加。例如, 渔获产量数据是充分可加的, 即对相同粒度下的渔获数据可以进行充分的汇集, 而平均网产是不可加的; 半可加性是指在满足一定条件下度量数据是可加的。数据的可加性是 1 个抽象概念, 有时我们对度量数据进行统计意义上的处理, 如求均值、期望、方差、标准差、中位数、极差等; 有时要按照一定的运算模型, 如局域运算、邻域运算、全局运算等。因此, 我们采用可计算性来替代传统的可加性概念, 可计算性概念应该具有更广泛的内涵。

在考虑海洋渔业度量数据的运算时一般都希望数据是完全可加的, 这就需要汇总数据在时空粒度上保持一致, 为此我们将渔业生产事实和海洋环境观测事实划分为不同的事实表。针对渔业生产事实, 可进一步划分为历史生产事实表和近期生产事实表。近期生产事实表主要是指最近 5 a 的渔业生产汇总数据, 当数据超出时限后, 将其进一步汇总, 并且把它存放入历史生产事实表中。

4 3 2　事实表和维表的连接

在构建渔业多维数据模型的过程中, 为了防止

维层次过多产生的查询复杂化问题,我们只对时空维进行规范化,而对其他维则采用非标准化的形式,将时空等维度与事实表相连,由此形成如图 4 所示的多维混合数据结构模型。数据仓库模型设计的最终目标是形成事实表和维表连接所构成的数据结构图(星形图或雪花图)[6],而此处我们考虑到海洋渔业时空数据分析的复杂度,采用星形图和雪花图折衷的方案——混合数据模型结构来设计海洋渔业数据仓库模型。经过数据标准(如数据类型、约束条件、索引)、主键、外键、属性等的定义,将图 4 所示的模型结构转译成数据仓库的物理结构。在物理实现过程中,当不同事实表的某些维度的粒度与维层次一致时,可使维表指针指向共享维。

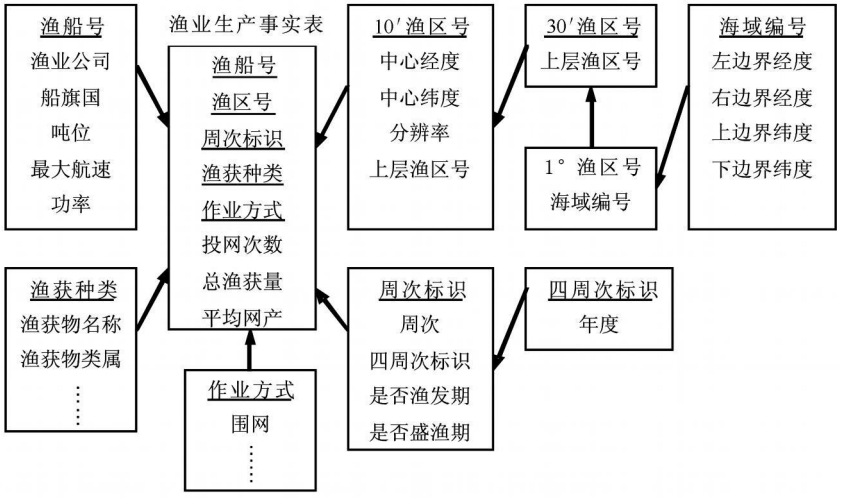


图 4 渔业生产数据多维模型结构

5 结语

本文从海洋多维数据仓库构建的需求出发,对多维数据库的相关概念进行了分析论述,并且对维、维层次、维层次关系、维层次链、度量、维聚集、层次聚集、海洋多维数据模型及海洋数据仓库等概念进行了形式化定义,在此基础上,以海洋渔业主题分析为例,对渔业多维数据库的概念模型和逻辑模型进行设计,通过对渔业生产的事实、维度、维层次及维层次关系的识别,针对时空维的复杂度,以折衷的混合多维数据模型结构构建了海洋多维数据仓库。

参考文献:

[1] 季民. 海洋渔业 GIS 时空数据组织与分析[D]. 青岛:山东科技大学, 2004
[2] PEDERSEN T B, JENSEN C S. Multidimensional database technology[J]. Computer, 2001, 12: 40—46
[3] 林友芳, 黄厚宽, 田盛丰. 铁路货运数据仓库多维视图的组织及其物化策略[J]. 铁道学报, 2001, 23(2): 8—12
[4] TRYFONA N, BUSBORG F, CHRISTIANSEN J G. StarER: a conceptual model for data warehouse design[G]// Proc ACM 2nd Int Workshop Data Warehousing and OLAP (DOLAP 99). New York: ACM Press, 1999: 3—8
[5] 杨云, 虞健飞, 张恒喜. 数据仓库多维建模的扩展 ER 模型[J]. 计算机工程与应用, 2003, 18:45—48
[6] 马刚. 采用数据仓库技术实现贷款管理 DSS[D]. 大连: 大连理工大学, 2000
[7] 苏奋振. 海洋渔业资源时空动态研究[D]. 北京: 中国科学院地理科学与资源研究所, 2001
[8] GYSSENS M, LAKSHMANAN L V S. A foundation for multi-dimensional databases[G]// Proceedings of the 23rd International Conference on VLDB. San Francisco, CA: Morgan Kaufmann Publishers Inc, 1997: 106—115
[9] 裴健, 柴玮, 赵畅, 等. 联机分析处理数据立方体代数[J]. 软件学报, 1999, 10(6): 561—569
[10] 陈微, 仲萃豪. 一种多维数据库和多维视图模型[J]. 计算机研究与发展, 1999, 36(2): 214—218
[11] 李建中, 高宏. 一种数据仓库的多维数据模型[J]. 软件学报, 2000, 11(7): 908—917
[12] 迟忠先, 李艳红, 张春涛, 等. OLAP 核心技术——数据立方体的研究现状与展望[J]. 计算机工程, 2002, 28(10): 316—318
[13] 苏奋振, 周成虎, 杨晓梅, 等. 海洋地理信息系统理论基础及关键技术研究[J]. 海洋学报, 2004, 26(6): 22—28

The research on marine multidimensional data warehouse construction

JI Min[1], JIN Feng-xiang[1], LI Ting[1], ZHAO Xiang-wei[1]

(1. *Geomatics Collage, Shandong University of Science and Technology, Qingdao* 266510, *China*)

Abstract: Spatial data warehouse has the ability to integrate the heterogeneous marine data, and gives the support of complex data analysis and high-level decision-making. In order to construct such a marine multi-dimensional data warehouse, the definitions of these concepts were offered, such as a dimension, a dimension hierarchy, a marine multidimensional data model. Taking a marine fishery theme as an example, fishery facts, the dimension, the dimension hierarchy, and the relationship between the dimension hierarchies were identified. Considering the complexity of spatial-temporal dimension, the marine multidimensional data warehouse was constructed using compromise hybrid data model structure.

Key words: data warehouse; multidimensional database; marine fishery; hybrid data model