文章编号:1671-9352(2007)09-0051-05

# 企业空间数据仓库的构建方法

梁银[1,2],张虹[1]

(1. 中国矿业大学 环境与测绘学院,江苏 徐州 221008; 2. 徐州师范大学 计算机科学与技术学院,江苏 徐州 221116)

**摘要**:在现有企业数据仓库多维模型的基础上,结合空间数据的特性和不同决策处理需求,提出了3种构建企业空间数据仓库的方法.通过原型系统的验证,说明这些方法是有效可行的.

**关键词**:企业空间数据仓库;空间OLAP;空间维层次;空间度量

**中图分类号**:TP311.13 **文献标志码**:A

# Constructing methods for enterprise spatial data warehouse

LIANG Yin[1,2], ZHANG Hong[1]

(1. School of Environment Science and Spatial Informatics, China University of Mining and Technology, Xuzhou 221008, Jiangsu, China; 2. Department of Computer Science and Technology, Xuzhou Normal University, Xuzhou 221116, Jiangsu, China)

**Abstract**: Three methods for constructing an enterprise spatial data warehouse based on an existing multidimensional data model of enterprise data warehouse were proposed by considering the unique features of spatial data and requirements of different decision-making processes. The validation of this prototype system shows that these methods are effective and feasible.

**Key words**: enterprise spatial data warehouse; spatial OLAP; spatial dimension hierarchy; spatial measure

## 0 引言

目前不少企业都建立了自己的数据仓库,并从中获得了很好的收益.很多研究表明存储在企业数据仓库中的多粒度数据有80%的与空间地理有关[1],如供应商与经销商地址、客户地址、营业网点分布等.但在企业数据仓库的设计和实施过程中通常把这些空间数据泛化为非空间数据进行处理,不是以地图的方式,丢失了很多空间特性,如用字符串"淮海西路123号"来表示某营业网点的地址.近年来,随着卫星勘测系统、遥感系统、全球定位系统、医学影像以及其他计算机化的数据搜集工具的广泛使用,已获得了大量的地理空间数据,并存储在相关的空间数据库、地理信息系统GIS,以及其他空间信息仓库中.如何把这些空间信息集成到企业数据仓库中,为企业决策提供更丰富完善的分析环境,提高信息的可视化及空间分析能力,已成为一个亟待解决的问题.

目前已建立的企业数据仓库一般不能存储、操作空间数据,而目前空间数据的管理通常由地理信息系统GIS来实现.因此,将企业数据仓库和GIS 2种技术结合起来来构建企业空间数据仓库.构建企业空间数据仓库是利用空间信息提高决策支持的一种有效的方法.一方面已经比较成熟的数据仓库技术可以提供有效的数据访问方法和管理大量数据,而且大多数联机分析操作如切片、切块、旋转、上卷、下探等,以及管理聚集数据获得的经验都可以扩展到空间数据仓库中管理空间数据.另一方面,用于管理空间数据的GIS技术得到了很大发展,特别是在空间索引结构、空间存储管理、空间查询与空间分析、以及空间信息可视化等方面已进行了深入研究,并得到了广泛应用,把GIS技术引入到企业数据仓库中,可以有效支持空间决策.

收稿日期:2007-04-30

基金项目:江苏省自然科学基金资助项目(BK2005021)

作者简介:梁银(1970- ),女,讲师,博士研究生,研究方向:空间数据仓库.Email:liangyinq86@163.com

然而,构建企业空间数据仓库并不是企业数据仓库与GIS的简单耦合,在概念模型、物理存储、查询优化等方面都需要新的技术支持,目前尚有很多技术问题需要解决. 本文重点阐述企业空间数据仓库的3种构建方法及相关技术,并应用到我们研制的企业空间数据仓库原型系统ESDW中.

## 1 企业空间数据仓库的构建

企业空间数据仓库的模型与现有数据仓库模型一样,可以采用多维模型,物理存储也同样可以采用星型或雪花型模式,但为了管理空间数据类型,在企业数据仓库的维和度量中需要增加空间信息. Han在文献[2]中最早提出了空间数据仓库的框架,扩展了数据仓库中数据立方体的维和度量的概念,引入了空间维和空间度量,将维分为非空间维、空间到非空间维和空间到空间维3种,将度量分为数值度量和空间度量2种. 因此,在本文的研究工作中,在将空间信息融合进企业现有的数据仓库中,采用下列3种方式构建企业空间数据仓库:(1)把空间信息作为多维模型中的维引入;(2)把空间信息作为分析主题引入;(3)在维和度量中都包含空间信息.

### 1.1 包含空间维的多维模型

#### 1.1.1 概念模型

多维模型中包含空间维,不包含空间度量. 空间维可以只有一个,也可以有多个,但若有多个空间维时,需要考虑它们之间的拓扑关系. 每个空间维中包含与几何对象有关的描述属性和几何属性. 如果空间信息只是作为分析数据对象性质的观察角度时,则把它作为空间维处理.

例如,查找指定位置的商店在2006年A类产品的销售额,需要把商店位置设置为空间维,如图1(a)所示. 在空间维商店位置中,商店名称、所在城市和商店地址是商店的描述属性,商店位置是商店地理位置的几何属性. 这是只有一个空间维的示例. 如果模型中涉及到多个空间维,需要考虑空间维之间的空间连接,这时空间连接谓词要在事实表中指定. 例如,为了分析各个城市中客户的购买行为,这就关系到2个空间维:客户的地理位置是一个空间维,用点对象表示;另一个空间维是城市,用面对象表示. 在商品销售事实表中指定这两个空间维的空间连接谓词是包含Contains,如图1(b)所示.

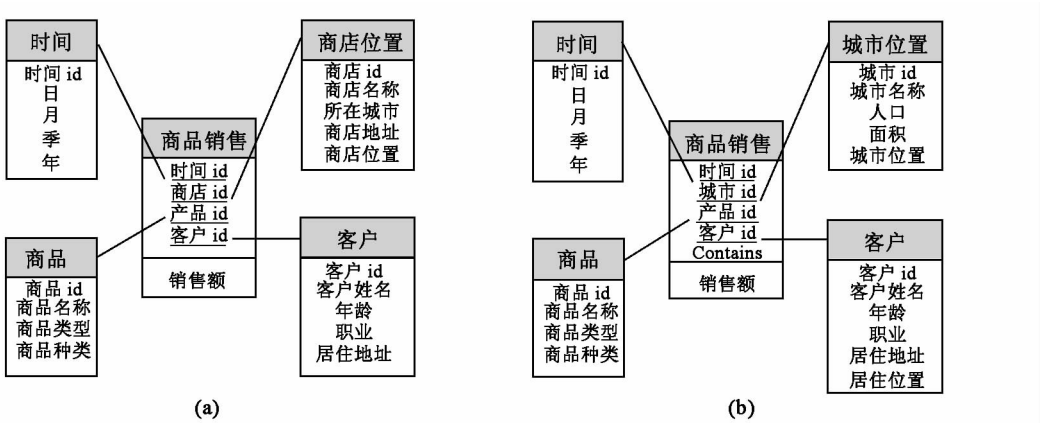


图1 包含空间维的星型模式

(a) 包含一个空间维的星型模式;(b) 包含2个空间维的星型模式

Fig.1 Star scheme with spatial dimension

(a) A star schema with a spatial dimension; (b) A star schema with two spatial dimensions

在这种模型中,多维分析的目标是销售额,商店、客户或城市的地理位置只是作为分析销售额的不同角度.

#### 1.1.2 空间维的层次表示

在企业数据仓库中,每个维可由一个或多个属性组成,维内的多个属性之间可能形成层次关系,表示不同的综合度,如日-月-年,商品-小类-大类等,各个维中不同维层次的组合可以预先进行物化,以提高OLAP的响应速度,且由较低维层次的聚集结果可以导出较高维层次的聚集结果. 而在空间数据仓库中,空间维不同于非空间维的情况,主要是由于大部分空间维上的层次划分比较复杂,其概念层次是相对的. 在设计阶段维层次可能很多,也可能是未知的,尤其是用户的某些预定义区域、或任意创建的ad-hoc查询区域,都需要基于地图进行在线分组计算. 所以不能直接应用OLAP操作中常用于提高系统性能的视图物化技术. 为了解决这个问题,我们采用了以下2种技术:

(1)对于某些应用存在一些默认分组,比如,南京市的销售点是一个分组,上海的销售点是另一个

分组等. 在这种情况下空间维可以根据默认分组分为多个层次,如果对每个默认分组进行物化,则与这些分组相关的查询可以直接得到结果.

(2)使用空间索引树(如 $R$-tree)在最细空间粒度上构建分组层次[3],作为空间维的分层,每个空间维需要建立一棵空间索引树. 例如,图 2 描述了商店的空间数据及其相应的 $R$-tree,根据索引树中的路径,可以创建数据立方体,也可以自动生成空间维上的概念层次,如图 3 所示. 在建立空间索引树的基础上,便可以有效利用企业数据仓库中现有的物化视图技术,实现空间数据仓库中视图的选择、预计算和物化. 这种方法不仅可以保持企业数据仓库的星型模式,而且提供了处理空间数据的能力.

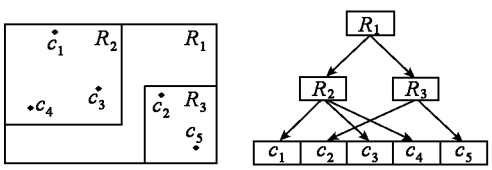


图 2　空间数据及 $R$-tree

Fig.2　The spatial data and the corresponding $R$-tree

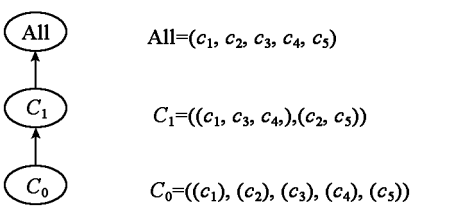


图 3　$R$-tree 表示的空间维的层次

Fig.3　Spatial dimensional hierarchy for the $R$-tree in Figure 2

另外,通过使用空间索引树表示空间维上的概念层次,可以有效完成空间 OLAP 的 Ad-hoc 查询处理. 为了从空间索引树的中间结点直接获得聚集结果,减少结点的访问次数,以提高查询效率,还可采用 AR-tree, aRB-tree 和 aCR-tree 等空间索引树作为空间维的分组层次.

### 1.1.3　物化视图的选择

物化视图的代价由存储代价、维护代价和计算代价 3 部分组成,为了使物化视图的代价最小化,在这种模型中,可以考虑把用户查询分为空间和非空间 2 部分,非空间部分是查询中除去空间操作和操作对象后得到的查询,可能包含传统的选择 – 投影 – 连接操作、比较操作和聚集操作,而空间部分仅包含已定义好的空间连接操作,并可通过维护 1 组指向空间对象的指针来完成. 一般在线计算空间操作的代价比存储空间视图本身小. 因此,仅物化查询的非空间部分,而在线计算空间部分,这样可以减少物化视图的总代价.

## 1.2　包含空间度量的多维模型

### 1.2.1　概念模型

在多维模型中的事实表中包含空间度量,没有空间维. 空间度量可以表示成指向一个或多个空间对象的指针集合,是多维分析的目标,可以通过非空间维来进行分析. 为了在某些非空间维上进行上卷操作,还需要定义空间聚集函数.

例如,用户想分析 2006 年购买 A 类产品的客户在哪些城市,或消费超过 5 万元的客户所在的城市时,需要在多维模型中使用空间度量来表示客户的地理位置,并且还须定义一个合并聚集函数,把相邻的城市合并为一个大的空间对象,如图 4 所示. 其中城市位置是空间度量,union 是空间聚集函数,当进行上卷操作时,一组连通的空间对象会合并为一个新的空间对象. 在这种模型中城市位置是多维分析的主题,用户可以得到城市的地理位置对商品销售的影响等信息,而且城市位置的其他描述属性,如人口、商店个数等都可用于决策处理.

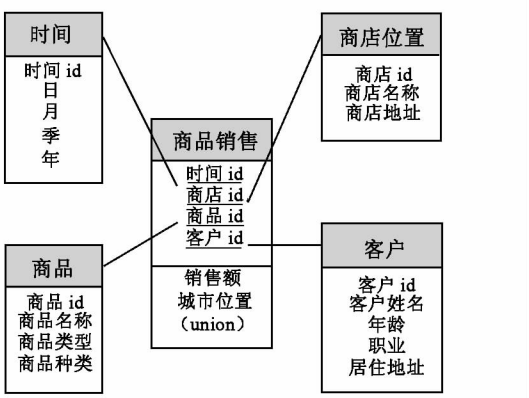


图 4　包含空间度量的星型模式

Fig.4　Star scheme with a spatial measure

### 1.2.2　空间度量的计算

空间度量类似于数值度量,空间数据的聚集函数根据计算性质也可以分为 3 类[4]:(1)空间分配型(spatial distributive)聚集函数,包括 convex hull, union, intersection, length 等函数;(2)空间代数型(spatial algebraic)聚集函数,例如求 $n$ 个点的中心(center)和几何体重心(gravity)都是空间代数型聚集函数;(3)空间整体型(spatial holistic)聚集函数,例如等分割(equi-partition)和最短距离(min-distance)等函数.

但是空间度量又不同于数值度量,主要体现在以下 4 个方面:(1)数值度量是简单类型,它的语义只局限于定量描述,而空间度量类型复杂,约束性强;(2)数值度量的聚集结果是新的数值,而空间度量的聚集是一组相关的空间对象指针,只有当这些空间指针所指向的空间对象是相邻的时,才合并为

一个新的空间对象;(3)数值度量的计算时间较小,而空间度量的计算开销较大,更需要预先计算一些空间 Cuboid,并以视图的形式加以存储,才能满足用户对响应时间的需求;(4)空间度量占用的存储空间远大于数值度量,一个空间度量可能占用几兆字节.

因此,计算并存储所有空间度量是不现实的,根据不同的应用需求,可以采用以下 3 种计算空间度量的方法:

(1)空间指针汇集.用一组指针表示参与聚集的空间对象. 这种方法占用空间少,但对于 OLAP 需要在线计算空间度量,适用于需要空间聚集操作的对象个数较少的情况.

(2)空间度量的近似计算.预计算并存储空间度量粗略的近似值,虽然会降低精确度,但可以减少存储空间和计算时间,是一种比较常用的方法. 由于用户在进行决策分析时,更注重的是趋势变化,在这种情况下,粗略的数据即能满足需求,很少需要使用精确数据. 因此,目前这种方法已得到了广泛的研究,主要提出了最小外接矩形的方法[5]、基于旋转最小外接矩形的方法、基于多级抽点的方法、基于数据精度转换的方法等近似计算方法[6].

(3)空间度量的选择物化.选择一部分参与聚集的空间对象预先进行计算并存储,这样不但可以得到精确的结果,而且可以减少在线计算的时间,一般可以采用空间贪心算法,指针相交算法和对象连结算法等.

### 1.3 包含空间维和空间度量的多维模型

这种模型综合了以上 2 种模型,既包含空间维,也包含空间度量,而且都可以不只一个. 空间维可以使用空间索引树来作为分组层次;空间度量可以表示成空间对象的指针集合,也可以通过空间拓扑操作来获得,同时空间度量可以通过空间和非空间维来进行分析. 当存在多种空间信息,其中有的是作为分析目标,有的是作为分析数据的观察角度时,需要使用这种多维模型. 图 5 表示的星型模式是用于分析最靠近高速公路和居民区的商店位置,其中,高速公路、商店位置和居民区位置是空间维,多个空间维涉及多元拓扑关系,在商品销售事实表中指定了空间维间的操作是 Distance;空间度量通过空间拓扑操作 Min-distance 来获得.

## 2 ESDW 原型系统

基于第 1 节中提出的 3 种构建空间数据仓库的方法,我们研制了 ESDW 企业空间数据仓库原型系统,并应用于煤矿瓦斯管理空间数据仓库系统. 该原型系统采用 3 层体系结构,如图 6 所示.

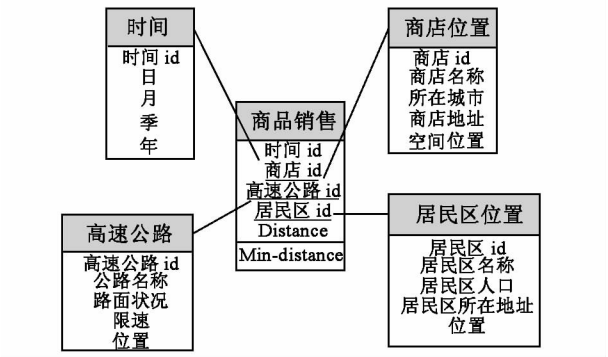


图 5 包含空间维和空间度量的星型模式

Fig.5 A star scheme with spatial dimension and spatial measure

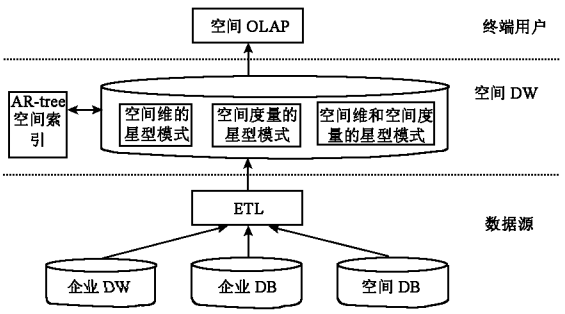


图 6 ESDW 体系结构

Fig.6 Architecture of ESDW

第一层通过 ETL 工具,将多种类型的空间和非空间数据根据需要进行抽取,然后转换成统一的数据格式,再根据用户的需求把空间和属性数据集成在一起,存储到空间数据仓库中.

第二层采用 SQL Server2000 数据库存储空间和属性数据,并通过 ArcSDE8.3 管理空间数据. 在 ESDW 中提供默认分组和 AR-tree 空间索引树作为空间维的分组层次,这 2 种技术在建立模型时,可以选择其一,也可以同时选择. 同时 ESDW 也提供了空间指针汇集和最小外接矩形 2 种计算空间度量的方法,用户可以根据查询结果精确度的要求选择其中一种方法.

在空间数据仓库中,支持只有空间维,只有空间度量,既有空间维也有空间度量的 3 种星型模式,具体使用哪种模式更能表达用户的需要,由设计者根据以下情况而定:

(1)如果空间信息只需要可视化,3 种模型都可以实现. 例如,查询 2006 年 A 类产品销售额在 50 万元以上的商店位置 store location,这时 store location 可以作为空间维,也可以作为空间度量来实现.

(2)需要比较不同地理区域上的数据,或在某些指定的地理区域上进行数据分析时,空间信息只能以维的形式引入.

(3)空间对象需要进行聚集处理,如,用户查询某日销售额大于 50 万元的地区,需要把满足条件的相邻区域合并为一个大的区域,这时空间信息只能以空间度量的形式集成.

星型模式中维的建立可以通过维设置向导完成. 维分为时间维、空间维和一般维,由于空间维可分为非空间维、空间到非空间维和空间到空间维 3 种,因此,对于空间维可以选择维层次属性是空间层次或非空间层次.

第三层是用户层,提供查询输入和结果显示. 用户可以通过类似于现有系统的图形用户界面提交查询,也可以通过选择地图上的任一区域提交联机分析请求. 联机分析结果可以采用报表、图表的方式显示,也可以通过 GIS 图形显示组件,叠加到背景地图中,提供更加直观的显示.

## 3　结束语

为了在现有的企业数据仓库中集成空间信息,以提高空间决策分析能力,本文提出了 3 种构建企业空间数据仓库的方法,具体讨论了相应的多维模型和关键技术,并应用到本工作研制的企业空间数据仓库原型系统 ESDW 中. 通过 ESDW 原型系统的验证,说明所提方法和技术是有效可行的. 目前正在进一步完善 ESDW 系统的功能.

致谢 最后要感谢东南大学计算机科学与工程学院的孙志挥教授对论文的悉心指导,从而使论文能得以顺利完成.

**参考文献:**

[1] BIMONTE S, TCHOUNIKINE A, MIQUEL M. Towards a spatial multidimensional model[C]// Proceedings of the 8th ACM international workshop on Data warehouse and OLAP. New York: ACM Press, 2005:39-46.

[2] HAN J, STEFANOVIC N, KOPERSKI K. Selective materialization: An efficient method for spatial data cube construction [C]// Proceedings of the 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin Heidelberg: Springer-Verlag, 1998: 144-158.

[3] RAO F Y, ZHANG L, YU X L, et al. Spatial hierarchy and OLAP-favored search in spatial data warehouse[C]// Proceedings of the Sixth ACM International Workshop on Data Warehousing and OLAP. New York: ACM Press, 2003: 48-55.

[4] SHEKHAR S, CHAWLA S. Spatial databases: A tour[M]. Prentice Hall: New Jersey, 2003.

[5] STEFANOVIC N, HAN J, KOPERSKI K. Object-based selective materialization for efficient implementation of spatial data cubes[J]. IEEE Transactions on Knowledge and Data Engineering, 2000, 12(6):938-958.

[6] 童云海, 谢昆青, 唐世渭. 空间数据仓库模型和空间 Cube 计算方法[J]. 计算机科学, 2002, 29(10): 1-5.

(编辑:孙培芹)