联机分析处理数据立方体代数*

裴 健 柴 玮 赵 畅 唐世渭 杨冬青

(视觉与听觉信息处理国家重点实验室 北京 100871)
(北京大学计算机科学与技术系 北京 100871)

摘要 数据立方体是多维数据库和以多维分析为基础的联机分析处理技术的核心机制.文章提出了一个支持多维数据库和多维分析的关于数据立方体的代数,从而为数据仓库及联机分析处理的语义描述提供了理论基础.同时,文章还论述了数据立方体的一些应用,以证明该工具所具有的强大功能.

关键词 数据立方体,代数,多维数据库,联机分析处理.

中图法分类号 TP311

数据仓库及联机分析处理技术是近年来数据库技术领域的一个研究重点和热点.数据仓库是面向主题的、集成的、包含历史数据的、数据相对稳定的支持决策制定过程的数据集合[1].从逻辑上讲,数据仓库是一个多维数据库.联机分析处理(online analytical processing,简称OLAP)以多维分析为基础,刻画了在管理和决策过程中对数据进行多层面、多角度的分析处理的要求[2].数据立方体(data cube)[3,4]是数据仓库和联机分析处理的核心概念之一.

然而,到目前为止,数据仓库和联机分析处理技术还缺乏充分的理论基础,多维模型缺乏相应的形式化理论.对数据仓库和联机分析处理的功能和操作通常都是以非形式化的方法来描述的,不同的厂商和人员对相同的用语往往赋予不尽相同的含义,妨碍了数据仓库及联机分析处理技术的深入应用.

本文提出了一种关于数据立方体的代数,支持数据仓库中的多维数据概念模式和在此之上的联机分析处理操作,为数据仓库及联机分析处理的语义描述提供了理论基础.我们提出的模型具有以下优点:

- 数据立方体是唯一的基础概念,简洁实用;
- 能够描述联机分析处理的操作,功能完备;
- 该代数的表达能力与关系代数相同.

本文给出了数据立方体的概念、有关重要性质和运算,并讨论了有关的应用.在第1节中,非形式化地讨论了多维数据模型、联机分析处理和数据立方体的一些性质,综述了相关的工作和文献.第2节给出了数据立方体代数的基本概念,证明了数据立方体可以作为多维模式的一个计算模型,并讨论了数据立方体的一些重要性质.第3节给出了数据立方体代数的操作.第4节讨论了数据立方体代数的应用.

1 问题背景及有关工作

在数据仓库的多维数据模式和联机分析处理中,要求在逻辑上采用多维的方式来组织和处理数据.根据数据分析的需求,要确定多维模式中的一些属性作为对数据对象性质的观察角度,称为维(dimension),维往往决

---

* 本文研究得到视觉与听觉信息处理国家重点实验室开放课题基金资助.作者**裴健**,1969年生,博士生,主要研究领域为数据仓库,联机分析处理,数据挖掘技术,信息系统集成技术.**柴玮**,女,1972年生,硕士生,主要研究领域为数据库,数据仓库,数据挖掘技术.**赵畅**,女,1973年生,硕士生,主要研究领域为数据库技术.**唐世渭**,1939年生,教授,博士生导师,主要研究领域为数据库理论与技术,空间信息处理,数字图书馆.**杨冬青**,女,1945年生,教授,博士生导师,主要研究领域为数据库理论与技术,空间信息处理技术,数字图书馆.

本文通讯联系人:裴健,北京 100871,北京大学计算机科学与技术系数据库研究室

本文1998-12-09收到原稿,1999-02-08收到修改稿

定着数据对象的属性.同时,反映数据对象特性的属性称为指标(measure).这样的结构称作数据立方体(data cube).实际上,并没有什么一般的判据来区分维和指标,所有划分都是根据分析的当前需要而进行的,是相对的和暂时的.维还可以有层次结构,例如,日期可以按日 月份-季度-年度组织,连续的维也可以使用基于属性的归纳(AOI)[5]等方法形成层次结构.

运用多维数据模型,可以进行多种联机分析处理操作,如向上综合(roll-up)、向下考察(drill-down)、旋转(pivoting)、局部分析(slicing-dicing)等.因此,联机分析处理的过程就是根据数据分析的要求,从原始数据中构造各种数据立方体,并对立方体执行有关的操作,再把结果返回给用户的过程.

在实现中,出于对效率的考虑,常常需要预先计算一些维上的聚集,以使在回答有关查询时,能够直接使用聚集数据,而不需要从原始数据开始计算.对于基于大规模数据集的数据立方体而言,预先计算关键聚集,对性能改善具有重要作用.显然,在具有 n 个维的原始数据上可能的聚集共有 2^n 种.上一层的聚集可以从下一层的聚集中由再对其中的一个维进行聚集而得到.一个聚集也称为数据立方体的一个节点(cuboid).

目前,对于多维数据模型和联机分析处理操作的描述大都是非形式化的,甚至是通过示例来进行的,因此,很容易造成误解和歧义.在系统设计和实现过程中,迫切需要有一种简便的、形式化的语义描述方法,这就是本文的研究动机.

我们认为,支持多维数据模型和联机分析处理的语义描述的形式化工具应该具有以下特征:

- 概念简单,易于理解和使用;
- 能够简洁地描述联机分析处理的各种操作;
- 支持维的层次等概念;
- 既支持单步操作,也支持过程性操作;
- 与实现无关.

文献[3]非形式化地提出了数据立方体的概念,认为是 SQL 中 group by 子句的一般化,并引入了关键字 *ALL* 来表示聚集的属性.文献[6]给出了多维数据库的一个非形式化模型,但对维的结构和数据立方体的复杂操作却没有进行深入的讨论.值得指出的是,在文献[6]中,因为没有把数据立方体的节点和数据立方体的概念进行严格区分,所以带来了处理上的复杂性,而且在有关数据立方体间的连接等操作的讨论上也存在着不足.文献[4]使用了和文献[3]一致的数据立方体的概念,研究了在数据立方体中选择哪些节点进行实体化(materialization)的问题,并给出了一个贪心算法.

然而,目前的文献均没有形式化地研究数据立方体及有关运算的理论.因此,对许多关键问题,尚无有效的形式化工具.

2 基本概念和重要性质

2.1 预备知识

为了方便地讨论数据立方体,需要对所涉及的域进行扩充,增加两个元素:⊥和 *ALL*.其中⊥对应于空值,*ALL* 的直观意义对应于文献[3]中引入的关键字 *ALL*.

定义 2.1(域的扩充). 域 D 的扩充域 $D^+=D\cup\{\perp,ALL\}$,并且,若$<$是 D 上偏序关系,则⊥是$(D^+,<)$上最小元,*ALL* 是$(D^+,<)$上最大元.即$\forall\ d\in D^+$,$\perp<d$ iff $d\neq\perp$,$d<ALL$ iff $d\neq ALL$.

定义 2.2(聚集函数). $aggr:2^{D_1\times\cdots\times D_n}\to D'_1\times\ldots\times D'_m$ 称为 $D_1\times\ldots\times D_n$ 上(到 $D'_1\times\ldots\times D'_m$)的聚集函数,当且仅当存在一个函数 $aggr_{assist}:(D'_1\times\ldots\times D'_m)\times(D'_1\times\ldots\times D'_m)\to D'_1\times\ldots\times D'_m$(称为聚集辅函数),使得$\forall\ T\in 2^{D_1\times\cdots\times D_n}$,$T_1\subseteq T$,有 $aggr(T)=aggr_{assist}(aggr(T_1),aggr(T-T_1))$.

2.2 数据立方体

定义 2.3(数据立方体). 数据立方体是一个六元组,$cube=(Dom,D,Mdom,M,f,aggr)$,其中,

(1) $Dom=dom_1\times\ldots\times dom_n$,$n>0$,称为维的域,$dom_i(1\leqslant i\leqslant n)$都是域;

(2) $D=\{d_1,\ldots,d_n\}$,称为维标识集,$d_i(1\leqslant i\leqslant n)$是 $dom_i(1\leqslant i\leqslant n)$的标识;

(3) $Mdom=mdom_1\times\ldots\times mdom_m$,$m>0$,称为指标的域,$mdom_i(1\leqslant i\leqslant m)$都是域;

(4) $M=\{m_1,\ldots,m_m\}$,称为指标标识集,$m_i(1\leqslant i\leqslant n)$是 $mdom_i(1\leqslant i\leqslant m)$ 的标识;

(5) $f:Dom\rightarrow Mdom$ 是 Dom 到 $Mdom$ 上的部分映射,称为立方体的基(cube base);

(6) $aggr$ 是 $Mdom$ 上的聚集函数.

记$(Dom,Mdom,aggr)$为数据立方体 $cube$ 的特征(signature).

当维和指标的值域都明确时,可以使用维的标识来代表维的名称和维的域,使用指标的标识代表指标的名称和指标的域,这时,数据立方体可以简记为四元组 $cube=(D,M,f,aggr)$.

在数据立方体的定义中,涉及了标识与域的对应.因此,存在着换名和域的次序交换的问题.从本质上看,换名和交换域的次序不影响数据立方体的代数性质.容易证明,换名和交换域的次序是全体数据立方体集合上的一个等价关系.在下文中,我们将在全体数据立方体关于换名和交换域的次序这一等价关系所形成的等价类上讨论数据立方体的性质和运算.

定义 2.4(同基数据立方体集). $f:Dom\rightarrow Mdom$ 是域集 Dom 到域集 $Mdom$ 上的部分映射,$CUBE_f=\{(Dom,D,Mdom,M,f,aggr)\mid(Dom,D,Mdom,M,f,aggr)$是数据立方体$\}$称为基为 f 的同基数据立方体集.

显然,以下定理成立.

定理 2.1. 同基数据立方体在换名和交换域的次序等价意义上唯一.

由定理 2.1,$CUBE_f$ 可简记为$\langle Dom,Mdom,f,*\rangle$.

2.3　数据立方体与多维模式

多维模式是多维数据库和数据仓库的数据逻辑模型,我们将证明数据立方体可以作为多维模式的实现抽象,它描述了在多维模式上的数据分析过程.正是由于数据立方体形式化地刻画了多维模式以及其上的多维分析,数据立方体代数才具有重要的理论意义.

关于多维模式的形式定义,有多种基本等价的方法,文献[7~9]分别作了讨论.这些定义虽然在描述上有所差异,但其本质是类似的.因此,我们采用文献[8,9]的定义.

定义 2.5(**多维模式**). 一个 n 维模式是一个四元组 $MDS(A,K,D,F)$,其中 $A=\{a_1,\ldots,a_m\}$ $(m\geqslant n)$,称为属性集;每个属性是一个域(domain),$K\subseteq A$,称为码集;$D=\{d_1,\ldots,d_n\}$,称为维标识集;$F:D\rightarrow 2^A$,称为分维函数,且满足:

(1) 对于任意的 $1\leqslant i,j\leqslant n,i\neq j,F(d_i)\cap F(d_j)=\varnothing$,且

(2) $\bigcup\limits_{d\in D}F(d)\subseteq A$,且

(3) 对于任意的 $1\leqslant i\leqslant n,F(d_i)\cap K\neq\varnothing$,且

(4) $(\bigcup\limits_{d\in D}F(d))\cap K=K$.

把 $F(d)(d\in D)$称为一个维表,记 $fact=\{K\cup(A-\bigcup\limits_{d\in D}F(d))\}$,称为事实表,$A$ 称为 MDS 所对应的关系模式.

与数据立方体换名和交换域次序的情况相类似,在多维模式上也存在着维标识的换名问题.同样地,维标识换名不影响多维模式的代数特性.因此,我们在维标识换名这一等价关系所形成的等价类上讨论有关性质.

以下定理表明,一个多维模式唯一对应着一个数据立方体的集合,该集合中的所有数据立方体具有相同的基.换言之,多维模式描述了一个数据结构,而对应的数据立方体集合可以视为该数据结构及在该数据结构上可能进行的多维分析操作的代数实现.

定理 2.2. 令 $CUBE_f/\cong$和 $\sum MDS=\{MDS(A,K,D,F)\}/\cong$分别为全体同基数据立方体集和全体多维模式关于换名和交换次序的等价关系所形成的商集,$\forall\ MDS(A,K,D,F)\in\sum MDS/\cong$,定义 $\Psi(MDS(A,K,D,F))$如下:$\langle Dom,Mdom,f,*\rangle\in CUBE_f/\cong$,

(1) $Dom=domain(F(d_1))\times\ldots\times domain(F(d_n))$,其中 $domain(F(d_i))$是 $F(d_i)$的域,是 $F(d_i)$所包含属性的卡氏积;

(2) 在 MDS 中,$fact=\{K\cup(A-\bigcup\limits_{d\in D}F(d))\}$,记 $fact-K=\{t_1,\ldots,t_l\}$,则 $Mdom=domain(t_1)\times\ldots\times$

$domain(t_i)$，其中 $domain(t_i)$是属性 t_i 的域；

(3) 定义 $f:Dom\rightarrow Mdom$，$\forall\ r=\langle key_{d_1},\ldots,key_{d_n},t_1,\ldots,t_l\rangle$是 MDS 中事实表中的记录，其中 key_{d_i}是属于 $F(d_i)$的码，则$\forall\ dom_1,\ldots,dom_n,dom_i\in domain(F(d_i)),1\leqslant i\leqslant n,dom_i|_{K\cap F(d_i)}=key_{d_i},f(dom_1,\ldots,dom_n)=\langle t_1,\ldots,t_l\rangle$.

Ψ 是 $\sum MDS$ 上的一个映射.

2.4 数据立方体的重要性质——数据立方体实体化

所谓实体化是指，预先执行某些计算，存储计算结果，以使在数据分析时，直接可以使用计算好的结果，而不需要从原始数据当中计算. 在本文中，我们所讨论的实体化都是指针对维的聚集.

定义 2.6(维聚集). 数据立方体 $cube=(Dom,Mdom,f,aggr),Dom=dom_1\times\ldots\times dom_n,aggr:2^{Mdom}\rightarrow AGG$，其中 AGG 是聚集函数 $aggr$ 的值域，f 关于 $dom_i(1\leqslant i\leqslant n)$的聚集是以下映射：

$f':dom_1\times\ldots\times dom_{i-1}\times\{ALL\}\times dom_{i+1}\times\ldots\times dom_n\rightarrow AGG$，

$\forall\ (d_1,\ldots,d_{i-1},d_{i+1},\ldots,d_n)\in dom_1\times\ldots\times dom_{i-1}\times dom_{i+1}\times\ldots\times dom_n$，

$f'((d_1,\ldots,d_{i-1},ALL,d_{i+1},\ldots,d_n))=aggr(\{m\mid \exists\ d_i\in dom_i,f((d_1,\ldots,d_{i-1},d_i,d_{i+1},\ldots,d_n))=m\})$.

定义 2.7(实体化单元集). 数据立方体 $cube=(Dom,Mdom,f,aggr)$，如下定义 $cube$ 的实体化单元集 $M(cube)$：

(1) $f\in M(cube)$；

(2) $\forall\ g\in M(cube)$，d 是 g 的任意一个维，g 关于 d 的聚集 $g'\in M(cube)$.

$M(cube)$是满足以上条件的最小集合，其中的元素称为数据立方体的实体化单元(cuboid).

定理 2.3. 对于任意给定数据立方体 $cube$，其实体化单元集确定且唯一.

定理 2.4. 数据立方体 $cube=(Dom,Mdom,f,aggr)$，$|Dom|=n$，则 $|M(cube)|=2^n$.

定义 2.8. 在 $M(cube)$上定义如下关系：$\prec_1$：$f\prec_1 g$ 当且仅当 g 是 f 在一个维上的聚集. 记$\prec$为$\prec_1$的传递闭包，称为 $M(cube)$上的聚集关系.

定理 2.5. $\prec$是 $M(cube)$上的聚集关系，$\langle M(cube),\prec\rangle$是一个布尔代数.

2.5 数据立方体的重要性质——数据立方体的同态

为了研究数据立方体代数上的运算，必须引入数据立方体同态的概念. 数据立方体的同态在刻画数据立方体的运算和应用中具有重要的意义.

定义 2.9(数据立方体同态). 设有数据立方体 $cube_1=(Dom_1,Mdom_1,f_1,aggr_1)$和 $cube_2=(Dom_2,Mdom_2,f_2,aggr_2),aggr_1:2^{Mdom_1}\rightarrow AGG_1,aggr_2:2^{Mdom_2}\rightarrow AGG_2$，若$\exists$ 映射 $\xi:(Dom_1\rightarrow Dom_2)\cup(Mdom_1\rightarrow Mdom_2)\cup(AGG_1\rightarrow AGG_2)$，使得

(1) $\forall\ d\in Dom_1,\xi(f_1(d))=f_2(\xi(d))$，且

(2) $\forall\ T\subseteq Dom_1,\xi(aggr_1(\{f_1(t)\mid t\in T\}))=aggr_2(\{f_2(\xi(t))\mid t\in T\})$，

则称 $cube_1=(Dom_1,Mdom_1,f_1,aggr_1)$关于 ξ 同态于 $cube_2=(Dom_2,Mdom_2,f_2,aggr_2)$，或 ξ 是两个数据立方体的一个同态(映射)，记作 $cube_1\overset{\xi}{\sim}cube_2$，简记为 $cube_1\sim cube_2$. 若 ξ 是单射，则称为单同态；若 ξ 是满射，则称为满同态；若 ξ 是双射，则称为同构，记作 $cube_1\overset{\xi}{\cong}cube_2$，或简记为 $cube_1\cong cube_2$.

3 数据立方体的操作

本节给出能够支持联机分析处理的数据立方体操作的代数描述.

3.1 指标和维转换

定义 3.1. $aggr$ 是 $D_1\times\ldots\times D_n$ 上的聚集函数，若对于某个 $D_i,1\leqslant i\leqslant n$，有$\forall\ (d_1,\ldots,d_{i-1},d_{i+1},\ldots,d_n)\in D_1\times\ldots\times D_{i-1}\times D_{i+1}\times\ldots\times D_n,\forall\ d',d''\in D_i,aggr((d_1,\ldots,d_{i-1},d',d_{i+1},\ldots,d_n))=aggr((d_1,\ldots,d_{i-1},d'',d_{i+1},\ldots,d_n))$，则称 $aggr$ 与 D_i 无关.

定义 3.2(指标-维转换). 数据立方体 $cube=(D_1\times\ldots\times D_n,M_1\times\ldots\times M_m,f,aggr)$,若 $aggr$ 与指标 $M_i(1\leqslant i\leqslant m)$无关,则定义关于 M_i 的指标-维转换操作如下:$MD(cube,M_i)=(D_1\times\ldots\times D_n\times M_i,M_1\times\ldots\times M_{i-1}\times M_{i+1}\times\ldots\times M_m,f',aggr')$,其中,$\forall\ (d,m')\in D_1\times\ldots\times D_n\times M_i$,若 $f(d)=(m_1,\ldots,m_{i-1},m',m_{i+1},\ldots,m_m)$,则

(1) $f'((d,m'))=(m_1,\ldots,m_{i-1},m_{i+1},\ldots,m_m)$,且

(2) $aggr'(\{(m_1,\ldots,m_{i-1},m_{i+1},\ldots,m_m)\})=aggr(\{(m_1,\ldots,m_{i-1},m',m_{i+1},\ldots,m_m)\})$.

定义 3.3(维-指标转换). 数据立方体 $cube=(D_1\times\ldots\times D_n,M_1\times\ldots\times M_m,f,aggr)$,定义关于维 $D_i(1\leqslant i\leqslant n)$的维-指标转换操作如下:

$$DM(cube,D_i)=(D_1\times\ldots\times D_n,M_1\times\ldots\times M_m\times D_i,f',aggr').$$

其中

$$\forall\ d=(d_1,\ldots,d_{i-1},d_i,d_{i+1},\ldots,d_n)\in D_1\times\ldots\times D_n,f'(d)=(f(d),d_i),$$

且

$$\forall\ (m,d)\in M_1\times\ldots\times M_m\times D_i,d\in D_i,aggr'(\{(m,d)\})=aggr(\{m\}).$$

维-指标转换简称为 DM 转换.

维-指标转换把一个维加入指标中去,但并不同时删除该维.这时,有一个维与一个指标相同,实质上是一种复制,这与指标-维操作是不一样的.维-指标操作的复制方式是由指标对维的函数依赖所决定的.

容易证明,指标-维转换和维-指标转换操作满足以下性质.

定理 3.1. $MD(cube,M_i)$是一个数据立方体.

定理 3.2. $DM(cube,D_i)$是一个数据立方体.

定理 3.3. 数据立方体 $cube=(D_1\times\ldots\times D_n,M_1\times\ldots\times M_m,f)$,$op_1,op_2\in\{MD,DM\}$,$d_1,d_2\in\{D_1,\ldots,D_n,M_1,\ldots,M_m\}$,则当 $d_1\neq d_2$ 且 $op_1(cube,d_1)$和 $op_2(cube,d_2)$均有定义时,$op_2(op_1(cube,d_1),d_2)=op_1(op_2(cube,d_2),d_1)$.

定理 3.3 表明,当涉及不同的维时,指标和维之间的转换操作可以按任意次序执行,结果不变.这个性质为联机分析的维变换(pivot)操作的并行和并发提供了理论保证.

3.2　指标和维的退化

如果聚集函数与指标中的某个属性无关,即在分析过程中不再需要使用该属性时,显然可以直接去掉该属性,这可通过指标退化操作来实现.

定义 3.4(指标退化). 数据立方体 $cube=(D_1\times\ldots\times D_n,M_1\times\ldots\times M_m,f,aggr)$,若 $aggr$ 与指标 $M_i(1\leqslant i\leqslant m)$无关,则如下定义关于 M_i 的指标退化操作 $Deg(cube,M_i)=(D_1\times\ldots\times D_n,M_1\times\ldots\times M_{i-1}\times M_{i+1}\times\ldots\times M_m,f',aggr')$,其中$\forall\ d\in D_1\times\ldots\times D_n$,若 $f(d)=(m_1,\ldots,m_{i-1},m',m_{i+1},\ldots,m_m)$,则

(1) $f'(d)=(m_1,\ldots,m_{i-1},m_{i+1},\ldots,m_m)$,且

(2) $aggr'(\{(m_1,\ldots,m_{i-1},m_{i+1},\ldots,m_m)\})=aggr(\{(m_1,\ldots,m_{i-1},m',m_{i+1},\ldots,m_m)\})$.

定理 3.4. $Deg(cube,M_i)$是一个数据立方体.

维的退化要比指标的退化复杂一些,因为必须保持指标对退化后维的函数依赖.

定义 3.5(基于聚集的维退化). 对于 $cube=(D_1\times\ldots\times D_n,M,f,aggr)$,$aggr:M\to AGG$,定义维 $D_i(1\leqslant i\leqslant n)$基于聚集的退化操作如下:

$$Deg(cube,D_i)=(D_1\times\ldots\times D_{i-1}\times D_{i+1}\times\ldots\times D_n,AGG,f',aggr_{assist}),$$

其中

$$\forall\ d=(d_1,\ldots,d_{i-1},d_{i+1},\ldots,d_n)\in D_1\times\ldots\times D_{i-1}\times D_{i+1}\times\ldots\times D_n,$$
$$f'(d)=aggr(\{f((d_1,\ldots,d_{i-1},d_i,d_{i+1},\ldots,d_n))\mid d_i\in D_i\}).$$

定理 3.5. $Deg(cube,D_i)$是一个数据立方体.

定理 3.6. 数据立方体 $cube=(D_1\times\ldots\times D_n,M_1\times\ldots\times M_m,f)$,则

$$Deg(DM(cube,D_i),D_i)=cube.$$

3.3 选 择

选择操作使得使用者可以根据需要从数据立方体的基当中选出一个子集,以这个子集为基,形成一个新的数据立方体.

定义 3.6(选择). 数据立方体 $cube=(D,M,f,aggr)$,p 是 $D\times M$ 上的谓词(布尔函数),定义选择函数 $\sigma_p(cube)=(D,M,f',aggr)$,其中$\forall\ d\in D, f'(d)=f(d)$ iff $p(d,f(d))$.

定理 3.7. $\sigma_p(cube)$是一个数据立方体.

3.4 具有相同特征的数据立方体的集合运算

从直觉上看,特征是数据立方体的"模式",因此,具有相同"模式"的数据集(基)可以进行并、交、差等传统集合操作.

定义 3.7(数据立方体的集合运算). 数据立方体 $cube_1=(D,M,f_1,aggr)$和 $cube_2=(D,M,f_2,aggr)$,$\forall\ d\in D$,若 $f_1(d)$和 $f_2(d)$均有定义,则有 $f_1(d)=f_2(d)$,那么

(1) $cube_1\bigcup cube_2=(D,M,u,aggr)$,其中$\forall\ d\in D, u(d)=\begin{cases}f_1(d), \text{ if } f_1(d)\text{有定义}\\ f_2(d), \text{ if } f_1(d)\text{无定义且 } f_2(d)\text{有定义}\end{cases}$;

(2) $cube_1\bigcap cube_2=(D,M,g,aggr)$,其中$\forall\ d\in D, g(d)=f_1(d)$, if $f_1(d)$有定义且 $f_2(d)$有定义;

(3) $cube_1-cube_2=(D,M,h,aggr)$,其中$\forall\ d\in D, g(d)=f_1(d)$, if $f_1(d)$有定义且 $f_2(d)$无定义.

定理 3.8. $cube_1\bigcup cube_2$,$cube_1\bigcap cube_2$ 和 $cube_1-cube_2$ 均是数据立方体.

3.5 数据立方体的积与连接

数据立方体的积把两个数据立方体拼接起来,形成一个更大规模的数据立方体.

定义 3.8(数据立方体的积). $cube_1=(D_1,M_1,f_1,aggr_1)$和 $cube_2=(D_2,M_2,f_2,aggr_2)$是数据立方体,数据立方体的积 $cube_1\times cube_2=(D_1\times D_2,M_1\times M_2,f,aggr)$,其中$\forall\ (d_1,d_2)\in D_1\times D_2, f((d_1,d_2))=(f_1(d_1),f_2(d_2))$当且仅当 $f_1(d)$和 $f_2(d)$均有定义;$\forall\ \{(m_1,m_2)\}\in 2^{M_1\times M_2}, aggr(\{(m_1,m_2)\})=(aggr_1(\{m_1\}),aggr_2(\{m_2\}))$.

定理 3.9. $cube_1\times cube_2$ 是数据立方体.

连接是数据立方体的一个重要操作,它使得两个数据立方体根据一定的要求连接成一个数据立方体.

定义 3.9(数据立方体关于谓词的连接). 数据立方体 $cube_1=(D_1\times D,M_1\times M,f_1,aggr_1)$和 $cube_2=(D\times D_2,M\times M_2,f_2,aggr_2)$,$\theta$ 是 $D_1\times D\times D_2\times M_1\times M\times M_2$ 上的谓词,两个数据立方体关于谓词 θ 的连接是 $cube_1\ \triangleright\!\!\triangleleft_{\theta}\ cube_2=(D_1\times D\times D_2,M_1\times M\times M_2,f,aggr)$,其中

(1) $\forall\ (d_1,d,d_2)\in D_1\times D\times D_2, f((d_1,d,d_2))=(f_1((d_1,d)),f_2((d,d_2))|_{M_2})$当且仅当 $f_1((d_1,d))$和 $f_2((d,d_2))$均有定义,且 $f_1((d_1,d))|_M=f_2((d,d_2))|_M$,且 $\theta(d_1,d,d_2,f_1((d_1,d)),f_2((d,d_2))|_{M_2})$;

(2) $\forall\ \{(m_1,m,m_2)\}\in 2^{M_1\times M\times M_2}, aggr(\{(m_1,m,m_2)\})=(aggr_1(\{(m_1,m)\}),aggr_2(\{(m,m_2)\}))$.

定理 3.10. $cube_1\ \triangleright\!\!\triangleleft_{\theta}\ cube_2$ 是数据立方体.

定理 3.11. $cube_1\times cube_2=cube_1\ \underset{\text{ture}}{\triangleright\!\!\triangleleft}\ cube_2$.

定理 3.9 表明,数据立方体的积是数据立方体连接的特例.

4 数据立方体代数应用

4.1 数据立方体代数的表达能力

表 1 对所讨论的数据立方体运算进行了小结.显然,还可以在数据立方体代数上定义许多运算,但上节定义的数据立方体的操作已经足以描述联机分析处理的操作,包括维变换(pivot),roll-up,drill-down 和 slice and dice 等.

表 2 非形式化地把数据立方体代数的运算与关系代数的运算进行了比较.显然,一个关系可以视为所有属性均是维且指标为{1}的数据立方体.于是,关系代数的运算就都可以使用数据立方体代数的运算来表示.

表 1　数据立方体代数的运算

单个数据立方体内的运算	相同特征的数据立方体运算	任意特征的两个数据立方体间的运算
指标-维转换	集合运算	积
维-指标转换		连接
指标退化		
基于聚集的维退化		
选择		

表 2　数据立方体代数与关系代数的运算比较

	关系代数	数据立方体代数
传统集合运算	并、交、差、卡氏积等	并、交、差等
特殊关系运算	选择、连接、投影等	指标和维的转换、退化、选择、积、连接等

另一方面,数据立方体可以使用关系表来作为数据结构.可以证明,数据立方体代数可以用关系代数来实现.我们有以下重要定理.

定理 4.1. 数据立方体代数与关系代数的表达能力相同.

证明:首先,以一个 n-元关系为基,以计数函数为聚集函数,可以构造一个数据立方体.逐个考虑关系代数上最小操作集的每个操作.可以验证,这些操作都可以用数据立方体代数运算的复合来表示.因此,数据立方体代数的表达能力不弱于关系代数.

其次,把一个数据立方体 $cube=(D,M,f,aggr)$ 实现为一个关系表 $D\times M$,其上有函数依赖 f.可以验证,数据立方体代数的每个运算都可以实现为关系代数操作的复合.因此,数据立方体代数的表达能力不强于关系代数.

由以上两点,定理得证.

4.2　数据立方体代数的应用

数据立方体代数是支持多维分析语义描述的一个强有力的数学工具,可以应用于联机分析处理语言和系统的分析设计、数据仓库的规范说明等方面.此外,借助数据立方体代数,还可以从理论上严格地分析研究多维分析和数据挖掘中的一些问题.

本节借助数据立方体代数来分析基于属性的归纳方法,以说明数据立方体代数的简单应用.

基于属性的归纳(AOI)[5,10]是数据挖掘中的一种有效的概念描述算法.它通过对维的数据进行分析,利用概念层次或数据分布的事实进行概念归纳描述.基于属性的归纳算法可以表述为数据立方体代数中的一个递归函数 AOI,只要阈值条件不满足,将不断递归执行.

文献[11]给出了如下基于属性的归纳方法的算法描述.

```
begin
  While 数据集中行的数目大于阈值 do begin
        选定某个属性 A;
        {对属性 A 执行一般化(generalization)步骤}
        if 在 A 的层次概念结构中存在一个更高(一般)的级别
        then 用高一级的值代替 A 的值;
        归并相同的行;
  end; {while}
end.
```

显然,在选定属性 A 时有许多不同的方法,因此,有可能导致归纳的结果不一致.同时,归纳的过程是否一定终止,这些都是需要用形式化方法予以解决的问题.

定义4.1. 满射集$\{C_i \rightarrow C_{i+1} | 1 \leqslant i \leqslant l, l \in N\}$称为域集$\{C_1, \ldots, C_l\}$上的一个有限概念层次结构，当且仅当$\{C_i \rightarrow C_{i+1} | 1 \leqslant i < l\}$中没有双射且$|C_l| = 1$. 其中域$C_1$称为起始域，域$C_l$称为终结域. 这时，显然总有$|C_{i+1}| < |C_i|$.

引理4.2. 对于任意域C，若$|C|$有限，则必存在以C为起始域的概念层次结构.

显然，在实际应用中，任意数据集的值域总是有限的. 因此，一定存在或可以构造赖以进行归纳的有限概念层次. 可以证明以下定理.

定理4.3. AOI算法终止.

定义4.2. 设需要对数据立方体$cube_0 = (D, M, f, aggr)$实施归纳，为每个维上表示有限概念层次的数据立方体指定一个唯一的聚集函数，若$cube'$由$cube$经过在一个维上的一步归纳而得到，则记为$cube \succ_1 cube'$，称为一步归纳关系. 集合$CUBE_{\succ}$是满足以下条件的最小集合：

(1) $cube_0 \in CUBE_{\succ}$；

(2) $\forall\, cube \in CUBE_{\succ}$，$\{cube' | cube \succ_1 cube'\} \subseteq CUBE_{\succ}$.

记$\succ$为$CUBE_{\succ}$上$\succ_1$的传递闭包，称为归纳关系.

定理4.4. $(\mathrm{CUBE}_{\succ}, \succ)$是一个格.

算法 AOI可能因采用不同的属性选取策略而在满足阈值时停机在不同的结果上. 因此，归纳结果可能有所不同. 但是，定理4.4表明，不同的归纳结果是同一个格中的不同元素，因此，具有Church-Roser性质，可以合流. 合流性保证了基于属性的归纳方法在数据语义上的一致性.

5 结 论

本文提出了一个支持多维分析语义描述的形式化工具——数据立方体代数，讨论了其基本概念、重要性质、有关运算和应用. 数据立方体代数对于联机分析处理的研究具有重要的理论价值.

目前，我们正在以数据立方体代数为理论工具，进行深入的研究工作，并取得了一些阶段性的结果. 我们的主要方向包括以下几个方面：

· 以该理论模型为基础，设计和实现支持联机分析处理的多维数据库语言，并考察不同的底层机制对语言实现的影响；

· 以数据立方体代数为理论基础，研究联机分析处理的集成理论，扩展联机分析处理操作，支持大规模的多级分析综合和数据集成挖掘；

· 研究基于数据立方体的数据仓库设计理论，探索多维模式的半自动设计方法.

参考文献

1 Inmon W H. Building the Data Warehouse. 2nd ed., New York: John Wiley and Sons, Inc., 1996

2 Codd E F, Codd S B, Salley C T. Beyond decision support. Computer World, 1993,27(30):87～89

3 Gray J, Chaudhuri S, Bosworth A *et al*. Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Data Mining and Knowledge Discovery, 1997,1(1):29～53

4 Harinarayan V, Rajaraman A, Ullman J D. Implementing data cube efficiently. In: Jagadish H V, Mumick Inderpal Singh eds. Proceedings of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1996. 205～216

5 Han J, Fu Y. Exploration of the power of attribute-oriented induction in data mining. In: Fayyad U M *et al* eds. Advances in Knowledge Discover and Data Mining. Cambridge: AAAI/MIT Press, 1996. 399～421

6 Agrawal R, Gupta A, Sarawagi S. Modeling multidimensional databases. In: Gray Alex, Larson Per-ke eds. Proceedings of the 13th International Conference on Data Engineering. Birmingham: IEEE Computer Society Press, 1997

7 Gjyssens M, Lakshmanam L V S. A foundation for multi-dimensional databases. In: Proceedings of the 23rd International Conference on Very Large Databases. San Fransisco: Morgan Kaufmann Publishers, Inc., 1997. 106～115

8 Pei J, Tang S, Yang D *et al*. Encapsulate multi-dimension objects in a data warehouse environment. In: Chen J *et al* eds. Proceedings of the 27th International Conference on Technology of Object Oriented Languages and Systems. IEEE

Computer Science Press, 1998. 362～371

9 Pei J, Tang S, Yang D *et al*. An algorithm for star schema construction based on query example. In: Rabi *et al* eds. Proceedings of the International Conference on Information Technology (ICIT'98). New York: MiGrall-Hill Press, Inc., 1998

10 Han J, Cai Y, Cercone N. Data-driven discovery of quantitative rules in relational databases. IEEE Transactions on Knowledge and Data Engineering, 1993,(5):29～40

11 Heinonen O, Mannila H. Attribute-oriented induction and conceptual clustering. Series of Publications C, No. C-1996-2. Department of Computer Science, University of Helsinki, 1996

An Algebra for Online Analytical Processing Data Cube

PEI Jian　CHAI Wei　ZHAO Chang　TANG Shi-wei　YANG Dong-qing

(*National Laboratory on Machine Perception　Beijing University　Beijing* 100871)
(*Department of Computer Science and Technology　Beijing University　Beijing* 100871)

Abstract　Data cube is the central mechanism in multi-dimension database and online analytical processing (OLAP) based on multi-dimensional analysis. In this paper, an algebra for OLAP data cube is proposed, which supports multi-dimensional database and analysis. It can be the theoretical foundation of semantic specification of data warehousing and OLAP manipulations. Some applications of the new mathematical tool are presented as well to show the power of the contribution.

Key words　Data cube, algebra, multi-dimensional database, online analytical processing.