第4卷第2期 软件学报 Vol. 4, No. 2
1993年4月 JOURNAL OF SOFTWARE Apr., 1993

统计与科学数据库上的代数操作

李建中 孙文隽 丁华福

(黑龙江大学,哈尔滨 150080)

ALGEBRAIC OPERATORS ON STATISTICAL AND SCIENTIFIC DATABASES

Li Jianzhong, Sun Wenjun and Ding Huafu

(*Heilongjiang University, Harbin* 150080)

Abstract According to the requirement of the statistical and scientific database applications, the operators on the databases are defined in the paper. The operators are based on the semantic data model MICSUM[3] and form the algebra on the set of C—relations, atomic summary tables and compound summary tables which are the basic operation objects in MICSUM[3]. The algebra called MS algebra. The MS algebra extends the relational algebra with extended operation semantics, the extended set of operation objects and new algebraic operators that support the statistical analysis queries. MS algebra is a theoretical framework for statistical and scientific database query languages.

摘要 针对统计与科学数据库的应用要求,本文以语义数据模型MICSUM2[3]为基础,以C—关系、原子统计表和复合统计表[3]为操作对象,定义了统计与科学数据库上的操作.这些操作构成了C—关系、原子统计表和统计表集合上的代数,简称MS代数.MS代数从两个方面扩展了关系代数,一是MS代数操作具有更丰富的语义和更广泛的操作对象;二是MS代数包括很多支持统计分析查询的新代数操作.MS代数是构造对用户友好的统计与科学数据库查询语言的理论基础.

§1. 引 言

用于统计分析的数据库称为统计数据库[1,2].科学数据库是存储科学实验或模拟结果的数据库[1,2].科学数据库与统计数据库有类似的结构和应用目的,所以,我们把两者放在一起来研究,统称为SSDB.SSDB管理系统是一种数据库管理系统,它提供搜集、存储、维护和查询统计与科学数据的能力,并且支持这类数据上的统计分析.统计分析可以是简单的综合统计(如求

本文1990年9月3日收到,1991年1月23日定稿.本题目由中国科学院管理、决策和信息系统开放实验室资助.作者**李建中**,教授,主要研究领域为数据库理论与技术.**孙文隽**,助研,1986年毕业于哈尔滨工业大学,主要研究领域为数据库理论与技术.**丁华福**,助研,1989年硕士毕业于哈尔滨工业大学,主要研究领域为图形工程学.

和、计数、求中值等),也可以是高级的统计技术(如回归分析等).SSDB在国民经济计划、商业金融、医疗卫生、科学研究、制造业、国防等各领域都有重要应用.SSDB有很多面向事务处理数据库管理系统难以支持的特点[1,2],需要新的数据模型.这种数据模型应该支持复杂数据类型并具有丰富的语义.

针对SSDB的特点,我们提出了一种新的语义数据模型MICSUM[3].MICSUM由九种语义成分和两种语义成分上的代数操作构成,支持时间序列等多种复杂数据类型,以统计表和C-关系为数据操作的基本单位,可定义具有行与列描述属性的二维统计表,并同时支持宏数据与微数据的描述,提供了模拟统计与科学数据的高级描述机构.这里的统计表完全是统计学家熟悉的统计表,C-关系则是关系模型中关系的扩充.

本文以MICSUM为基础,研究SSDB上的代数操作.

目前,已有一些人从事SSDB上代数操作的研究,并取得了一些成果.Su考虑到SSDB的特点,把属性划为标识(identifing)属性和统计(summary)属性,提出一种非第一范式关系,称为G-关系,并定义了G-关系上的代数操作[5].Ozsoyoglus和Matos针对SSDB应用的要求,扩展了Klug的关系代数和演算[4],允许关系的属性可以是集合值属性(set-valued attribute),并定义了新的操作[6].[5]和[6]定义的代数和演算操作以关系概念为基础,仍然受着关系方法的束缚,不能很好的适应SSDB的特点和SSDB的应用要求.

本文以语义数据模型MICSUM[3]为基础,以C-关系、原子统计表和复合统计表为操作对象[3],定义了SSDB上的操作,这些操作构成了C-关系、原子统计表和统计表集合上的代数,简称MS代数.本文从两个方面扩充了关系代数.一是MS代数操作具有更丰富的语义和更广泛的操作对象,摆脱了关系方法的束缚;二是MS代数包括很多支持统计分析查询的新代数操作.MS代数是构造对用户友好的统计与科学数据库查询语言的理论基础.

由于C-关系、原子统计表和复合统计表在形式和语义两方面都有所不同,我们将在第三、四节分别定义C-关系、原子统计表和复合统计表上的代数操作.

§2. 预备知识

SSDB中的数据集合可分为两类.一类是微数据集合,用来存储描述个体或事件的数据.另一类是宏数据集合,用来存储对微数据进行综合分析而得到的结果数据.每个微数据集合对应一个下边定义的C-关系.每个宏数据集合对应一个下边定义的一维或二维统计表.SSDB中的数据按其性质可以分为两种.一是直接用于统计分析的数据.这类数据被抽象地称为SSDB的统计属性或统计数据.二是说明性数据,用来描述统计数据的语义.这种数据被称为SSDB的描述属性或描述数据.描述属性又分为行描述属性和列描述属性.所有描述属性函数地确定统计属性.

C-关系r是一个n元组的集合,这里n>0.n元组的每个分量有一个名字,称为r的属性.C-关系模式R是一个属性集合,记为$R(A_1,\cdots,A_n)$,其中,A_i是R的属性.Deg(R)是R的属性个数.C-关系的每个属性都有一个值域.值域可以具有整数、实数、字符串、时间、日期、相对时间、时间序列、集合、向量、矩阵等数据类型[3].Att(r)表示C-关系r的属性集合.

设$n\geqslant 0, m\geqslant 0, p\geqslant 0$,但n和m不能同时为0.统计记录是一个三元组$((r_1,\cdots,r_n),(c_1,\cdots,c_m),(s_1,\cdots,s_p))$.统计记录的第一分量是一个n元组,其每个分量的名字称为行描述属性;

第二分量是一个 m 元组,其每个分量的名字称为列描述属性;第三分量是一个 p 元组,其每个分量的名字称为统计属性.原子统计表是一个统计记录的集合.原子统计表模式 PT 由行描述属性集合、列描述属性集合和统计属性集合组成,记作

$PT(ROW:A_1,\cdots,A_n;COL:B_1,\cdots,B_m;SUMM:S_1,\cdots,S_p)$,

其中 ROW 表示$\{A_1,\cdots,A_n\}$是行描述属性集合;COL 表示$\{B_1,\cdots,B_m\}$是列描述属性集合;SUMM 表示$\{S_1,\cdots,S_p\}$是统计属性集合.如果 $n=0$ 或 $m=0$,则称 PT 为一维统计表.原子统计表的每个属性有一个值域.值域的数据类型也可以是整数、实数、字符串、时间、日期、相对时间、时间序列、集合、向量、矩阵等[3].Rdeg(PT)、Cdeg(PT)和 Sdeg(PT)分别表示 PT 的行描述属性个数、列描述属性个数和统计属性个数.Ratt(PT)、Catt(PT)和 Satt(PT)分别表示 PT 的行描述属性集合、列描述属性集合和统计属性集合.图 1 的(a)给出了一个原子统计表的例子,其模式为

PT(ROW:国家,性别;COL:年龄,种族;SUMM:心血管疾病人数,癌病人数).

从图易见,MICSUM 模型中的一个统计表可以方便地表示应用领域中的多个具有相同行与列描述属性的统计表.复合统计表是 $k(k\geqslant 1)$个具有相同行或列描述属性的原子统计表的并集.下面定义由具有相同行描述属性的原子统计表构成的复合统计表的模式.可以类似地定义由具有相同列描述属性的原子统计表构成的复合统计表的模式.对于 $1\leqslant i\leqslant k$,设原子统计表 PT_i 的模式为

$PT_i(ROW:A_1,\cdots,A_n;COL:B_{i1},\cdots,B_{im_i};SUMM:S_{i1},\cdots,S_{ip_i})$,

则由 $PT_1,\cdots,PT_k$ 构成的复合统计表 T 的模式为

$$T(ROW:A_1,\cdots,A_n;$$
$$COL:B_{11},\cdots,B_{1m_1};\cdots;B_{k1},\cdots,B_{km_k};$$
$$SUMM:S_{11},\cdots,S_{1p_1};\cdots;S_{k1},\cdots,S_{kp_k}),$$

图 1 的(b)给出了一个由两个原子统计表构成的具有相同列描述属性的复合统计表的例子,其模式为

T(ROW:地区,年代;产地,年代; COL:石油化工产品;SUMM:产值;产量).

$Rdeg_i(T)$、$Cdeg_i(T)$和 $Sdeg_i(T)$表示 T 的第 i 个原子统计表的行描述属性个数、列描述属性个数和统计个数.$Ratt_i(T)$、$Catt_i(T)$和 $Satt_i(T)$分别表示 T 的第 i 个原子统计表的行描述属性集合、列描述属性集合和统计属性集合.

心血管疾病人数(万人)			年龄 {12,…,20} 种族 黑人	年龄 {12,…,20} 种族 白人	年龄 {21,…,64} 种族 黑人	年龄 {21,…,64} 种族 白人
国家	美国	性别 男	10	10	20	20
		性别 女	10	10	20	20
	苏联	性别 男	15	15	25	25
		性别 女	15	15	25	25

癌病人数（万人）			年龄 {12,…,20} 种族 黑人	白人	{21,…,64} 种族 黑人	白人
国家	美国	性别 男	10	10	15	15
		性别 女	10	10	15	15
	苏联	性别 男	20	20	25	25
		性别 女	20	20	25	25

(a)一个原子统计表实例

产量或产值			石油化工产品 石油	天然气
地区	黑龙江	年代 1980	19000000	13000000
		年代 1985	21000000	15000000
		年代 1990	30000000	17000000
	辽宁	年代 1980	20000000	13500000
		年代 1985	23000000	14600000
		年代 1990	31000000	18000000
产地	大庆	年代 1982	38900000	27000000
		年代 1986	43700000	31000000
		年代 1990	61900000	35000000
	辽河	年代 1982	41000000	28000000
		年代 1986	23000000	29000000
		年代 1990	63000000	39000000

(b)一个复合统计表实例

图1　原子统计表和复合统计表实例

显然,原子统计表是复合统计表的特例,即只有一个原子统计表的复合统计表.以后我们统称原子或复合统计表为统计表.

SSDB模式是一个序列$<MS_1,\cdots,MS_n>$,其中,MS_i是一个C-关系或统计表模式.SSDB模式$<MS_1,\cdots,MS_n>$的一个实例I是序列$<ms_1,\cdots,ms_n>$,其中,ms_i是C-关系或统计表,称为模式MS_i的实例.如果ms_i是C-关系,则$Att(ms_i)=Att(MS_i)$.如果ms_i是统计表,则$Ratt(ms_i)=Ratt(MS_i)$、$Catt(ms_i)=Catt(MS_i)$、$Satt(ms_i)=Satt(MS_i)$.

设ms是一个C-关系或统计表,$X=\{X_1,\cdots,X_k\}\subseteq Att(ms)$或$X\subseteq Ratt(ms)\cup Catt(ms)\cup Satt(ms)$.如果$t\in ms$,则$t[X]$是元组t在属性组X上的值$(t[X_1],\cdots,t[X_k])$.令$t_1$和$t_2$是ms的两个元组或统计记录.令$\theta=\{\theta_1,\cdots,\theta_k\}$,$\theta_i$是属性$X_i$的值域上的二元关系.$t_1[X]\theta t_2[X]$表示对于所有$1\leqslant i\leqslant k$,$t_1[X_i]\theta t_2[X_i]$.

设$\sum$是元组或统计记录的集合,D_i表示$\sum$中所有元组或统计记录的第i分量或第i统

计属性值的集合，$S=\{s_1,\cdots,s_n\}$是任意集合. 聚集函数 $f_i(x)$是 $P(D_i)$(D_i 的幂集合)上的单位函数. 逆聚集函数 $f_i(S,x)$是 D_i 上关于 S 的多值函数. 对于任意 $x\in D_i$，

$$f_i(S,x)=\{(s_1,y_1),\cdots,(s_n,y_n)\}$$

其中，y_i 与 s_i 相对应.

在本文以后的讨论中，我们用 EXP 表示 SSDB 上的 MS 代数表达式集合，I(E)表示 MSA 表达式 E 的实例，即 C—关系或统计表.

§3. C—关系和原子统计表上的代数操作

我们首先定义 C—关系和原子统计表上的基本代数操作，然后讨论可由基本代数操作构成的常用宏代数操作. 本部分定义的代数操作的操作分量皆为 C—关系或原子统计表.

3.1 基本代数操作

1. 常量

对于任意常量 $C\in D$(D 是由整数、实数、字符串、时间、日期、相对时间、时间序列、集合、向量或矩阵组成的集合)，$C\in EXP$，而且 C 的实例是 I (C)=C.

2. C—关系和统计表

C—关系、原子统计表或复合统计表 ms 是 MS 代数表达式，而且 I (ms)=ms.

3. 集合的并与差

设 $E_1,E_2\in EXP$，E_1 和 E_2 同为 C—关系或原子统计表. 如果 E_1 和 E_2 是 C—关系，令 $Deg(E_1)=Deg(E_2)$. 如果 E_1 和 E_2 是原子统计表，令 $Rdeg(E_1)=Rdeg(E_2)$、$Cdeg(E_1)=Cdeg(E_2)$、$Sdeg(E_1)=Sdeg(E_2)$. 无论哪种情况，E_1 和 E_2 的对应属性必须具有相同的数据类型. E_1 和 E_2 的并与差定义为 $E_1\cup E_2$ 与 E_1-E_2，且 $E_1\cup E_2\in EXP$、$E_1-E_2\in EXP$，其语义分别为

$$I(E_1\cup E_2)=\{t\mid t\in I(E_1)\text{或 }t\in I(E_2)\},$$

$$I(E_1-E_2)=\{t\mid t\in I(E_1)\text{且 }t\notin I(E_2)\}.$$

图 2 给出了求原子统计表并的例子. 结果表中出现了一些空值.

4. 笛卡尔积

设 $E_1,E_2\in EXP$.

如果 E_1 和 E_2 同为 C—关系. E_1 和 E_2 的笛卡尔积是 $E_1\times E_2\in EXP$. $Att(E_1\times E_2)=Att(E_1)\times Att(E_2)$，$Deg(E_1\times E_2)=Deg(E_1)+Deg(E_2)$，而且

$$I(E_1\times E_2)=\{t_1\cdot t_2\mid t_1\in I(E_1)\text{且 }t_2\in I(E_2)\}.$$

其中·表示连接.

如果 E_1 或 E_2 是原子统计表，不妨令 E_1 是原子统计表，则 $E_1\times E_2\in EXP$ 是 C—关系，$Deg(E_1\times E_2)=Rdeg(E_1)+Cdeg(E_1)+Sdeg(E_1)+Deg(E_2)$，而且 $I(E_1\times E_1)=\{t_r\cdot t_c\cdot t_s\cdot t_2\mid((t_r),(t_c),(t_s))\in I(E_1)\text{且 }t_2\in I(E_2)\}$. 如果 E_1 和 E_2 都是原子统计表，则 $E_1\times E_2\in EXP$ 是 C—关系.

$Deg(E_1\times E_2)=Rdeg(E_1)+Cdeg(E_1)+Sdeg(E_1)+Rdeg(E_2)+Cdeg(E_2)+Sdeg(E_2)$，而且 $I(E_1\times E_2)=\{t_{1r}\cdot t_{1c}\cdot t_{1s}\cdot t_{2r}\cdot t_{2c}\cdot t_{2s}\mid((t_{1r}),(t_{1c}),(t_{1s}))\in I(E_1)\text{且}((t_{2r}),(t_{2c}),t_{2s}))\in I(E_2)\}$.

以后我们将看到，组合笛卡尔积与其他操作可以由两个相关的统计表构造新的统计表.

E1:

人口数(万人)		年		
		1983	1984	1985
城市	哈尔滨	247	252	269
	长　春	172	176	171

E2:

人口数(万人)		年		
		1981	1982	1983
城市	长春	269	189	172
	吉林	251	254	265
	沈阳	201	208	297

E1∪E2:

人口数(万人)		年				
		1981	1982	1983	1984	1985
城市	哈尔滨	---	---	247	252	269
	长　春	269	189	172	176	171
	吉　林	251	254	265	---	---
	沈　阳	201	208	297	---	---

图 2　一个求原子统计表并的例子

5. 选择

设 $E\in EXP$,E 的选择定义为 $Sele(E,F)\in EXP$. 如果 E 是 C－关系,则 Sele(E,F)也是 C－关系,$Deg(Sele(E))=Deg(E)$. 如果 E 是原子统计表,则 Sele(E,F)也是原子统计表,$Rdeg(Sele(E,F))=Rdeg(E)$,$Cdeg(Sele(E,F))=Cdeg(E)$,$Sdeg(Sele(E,F))=Sdeg(E)$. Sele(E,F)的语义是

$$I(Sele(E,F))=\{t\mid t\in I(E)\text{而且 }F(t)\text{为真}\},$$

其中 F(t)表示 t 满足命题 F. F 是条件布尔表达式,由多个命题通过逻辑运算符 OR、AND、NOT 连接而成. 命题的形式是 AθX,其中 A 是 E 的属性名,X 是常量或 E 的属性名,$\theta\in\{=,\neq,\leqslant,<,\geqslant,>,\subseteq,\supseteq,\supset,\subset,\notin,\in,\equiv,<>,->,<-,<=,=>\}$. 下边对于不同的数据类型定义 θ 的语义. 以下,我们用 Val(Y)表示 Y 中存储的值(如果 Y 是属性名)或 Y 本身(如果 Y 是常量).

如果 Val(A)和 Val(X)是整数或实数,则 $\theta\in\{=,\neq,<,\leqslant,>,\geqslant\}$,即 θ 是通常的算术比较符.

如果 Val(A)和 Val(X)是字符串或正文,则

(1)A=(或≠)X 为真表示 Val(A)与 Val(X)相同(或不同).

(2)$A\theta X(\theta\in\{<,\leqslant,>,\geqslant\})$为真表示按字典顺序 $Val(A)\theta Val(X)$.

(3)$A\theta X(\theta\in\{\subset,\subseteq,\supset,\supseteq\})$为真表示 Val(A)是 Val(X)的(真)子串或反之.

如果 Val(A)和 Val(X)是日期、时间或相对时间,则 $A\theta X(\theta\in\{=,\neq,<,\leqslant,>,\geqslant\})$为真表示 Val(A)等于、不等于、先于、先于等于、后于或后于等于 Val(X).

如果 Val(A)和 Val(X)是集合,$A\theta X(\theta\in\{=,\neq,\subset,\subseteq,\supseteq,\supset\})$为真表示 $Val(A)\theta Val(X)$.

如果 Val(A)和 Val(X)是有序集合,则

(1)AθX($\theta\in\{=,\neq,\subset,\subseteq,\supset,\supseteq\}$)为真表示 Val(A)$\theta$Val(X).

(2)A<(或>)X 为真表示 Val(A)中所有元素都先(或后)于 Val(X)中所有元素. A≤(或≥)X 为真表示 Val(A)中所有元素都先于等于(或后于等于)Val(X)中所有元素. A<-(或->)X 为真表示 Val(A)中某些元素先(或后)于 Val(X)中所有元素. A<=(或=>)X 为真表示 Val(A)中某些元素先于等于(或后于等于)Val(X)中所有元素.

如果 Val(A)和 Val(X)是向量,则

(1)A=(或≠)X 为真表示向量 Val(A)与 Val(X)相等(或不等).

(2)A<(或>)X 为真表示 Val(A)中所有分量小于(或大于)Val(X)中相应分量. A≤(或≥)X 为真表示 Val(A)中所有分量小于等于(或大于等于)Val(X)中相应分量.

(3)A<-(或->)X 为真表示 Val(A)中某些分量小于(或大于)Val(X)中对应的分量. A<=(或=>)X 为真表示 Val(A)中某些分量小于等于(或大于等于)Val(X)中相应的分量.

(4)A$\not<$(或$\not>$)X 为真表示 Val(A)按字典序先于(或后于)Val(X).

(5)A$\subset$(或$\supset$)X 为真表示 Val(A)是 Val(X)(或 Val(X)是 Val(A))的真子向量. A$\subseteq$(或$\supseteq$)X 为真表示 Val(A)是 Val(X)(或 Val(X)是 Val(A))的子向量.

如果 Val(A)和 Val(X)是矩阵,则

(1)A=(或≠)X 为真表示矩阵 Val(A)与 Val(X)相同(或不同).

(2)A<(或>)X 为真表示 Val(A)中所有元素小于(或大于)Val(X)中对应的元素. A≤(或≥)X 为真表示 Val(A)中所有元素小于等于(或大于等于)Val(X)中对应元素.

(3)A<-(或->)X 为真表示 Val(A)中部分元素小(或大)于 Val(X)中对应元素. A<=(或>=)X 为真表示 Val(A)中部分元素小于等于(或大于等于)Val(X)中对应元素.

(4)A$\subset$(或$\supset$)X 为真表示 Val(A)是 Val(X)(或 Val(X)是 Val(A))的真子矩阵. A$\subseteq$(或$\supseteq$)X 为真表示 Val(A)是 Val(X)(或 Val(X)是 Val(A))的子矩阵.

如果 Val(A)和 Val(X)具有时间序列数据类型,则

(1)A=(或≠)X 为真表示 Val(A)与 Val(X)定义在相同(或不同)的时间间隔上. 当 Val(X)为常量时,上式为真表示 Val(A)的时间间隔等于(或不等于)Val(X).

(2)A==(或<>)X 为真表示 Val(A)与 Val(X)相同(或不同).

(3)AθX($\theta\in\{\leqslant,\geqslant,<,>,<=,<-,=>,->\}$)为真的语义于 Val(A)和 Val(X)为矩阵时相同.

如果 Val(X)是由与 Val(A)类型相同的元素构成的集合、向量或矩阵,A$\in$(或$\ni$)X 为真表示 Val(A)属于或不属于 Val(X).

6. 投影

设 E$\in$EXP. 如果 E 是 C-关系,令 X$\subseteq$Att(E),则 E 在 X 上的投影 Proj(E,X)$\in$EXP 是一个 C-关系,Deg(Proj(E))=|X|(X 中属性个数),而且

$$I(Proj(E))=\{t[X]\mid t\in I(E)\}.$$

如果 E 是原子统计表,令 X=R$\cup$C$\cup$S,R$\subseteq$Ratt(E),C$\subseteq$Catt(E),S$\subseteq$Satt(E),S、R、C 满足:S$\neq$0(空集)则 R=Ratt(E)、C=Catt(E). E 在 X 上的投影 Proj(E,X)$\in$EXP 是一个原子统计表,Ratt(Proj(E))=R,Catt(Proj(E))=C,Satt(Proj(E))=S,而且

$$I(Proj(E,X))=\{((t_r),(t_c),(t_s))\mid \text{存在 } t\in I(E),\text{使 } t_r=t[R],t_c=t[C],t_s=t[S]\}.$$

7. 聚集

设 $E\in EXP, F=\{f_{i_1},\cdots,f_{i_k}\}$是聚集函数族.

如果 E 是 C—关系,令 $X\subseteq Att(E), Y\subseteq Att(E), X\cap Y=0, |Y|=k, f_j$ 用于属性 $A_j\in Y$. E 的聚集 $Agg(E,X,Y,F)\in EXP$ 是一个 C—关系,$Deg(Agg(E,X,Y,F))=|X|+k$,而且

$$I(Agg(E,X,Y,F))=\{(t[X],y_{i_1},\cdots,y_{i_k})\mid t\in I(E)\text{而且}$$
$$y_{i_j}=f_{i_j}(\{u\mid u\in I(E)\text{且 }u[X]=t[X]\})\}.$$

例如,我们有 C—关系 E 如下:

B1	B2	B3	A1	A2	C1
1	1	1	1	10.1	N
1	1	1	2	10.5	F
2	2	2	3	9.8	T
2	2	2	4	6.1	M

令 X={B1,B2,B3},Y={A1,A2},F={f1,f2},f1=SUM 与 f2=MAX 是求和与求最大值函数,则:Agg(E,X,Y,F)为

B1	B2	B3	A1	A2
1	1	1	3	10.5
2	2	2	7	9.8

如果 E 是原子统计表,令 $X\subseteq Ratt(E), Y\subseteq Catt(E), Z\subseteq Satt(E), |Z|=k, f_j$ 用于统计属性 $A_j\in Z$. E 的聚集 $Agg(E,X,Y,Z,F)\in EXP$ 是一个统计表,$Rdeg(Agg(E,X,Y,Z,F))=|X|, Cdeg(Agg(E,X,Y,Z,F))=|Y|, Sdeg(Agg(E,X,Y,Z,F))=k$,而且

$$I(Agg(E,X,Y,Z,F))=\{((t[X]),(t[Y]),(y_{i_1},\cdots,y_{i_k}))\mid t\in I(E)\text{而且}$$
$$y_{i_j}=f_{i_j}(\{u\mid u\in I(E)\text{且 }u[X]=t[X],u[Y]=t[Y]\})\}.$$

设 E 表示图 1(a)所示统计表. 令 X={国家},Y={年龄},Z={心血管疾病人数,癌病人数},F={SUM,MAX},则图 3 给出了 Agg(E,X,Y,Z,F).

心血管疾病人数(万人)	年龄 {12,…,20}	年龄 {21,…,64}
国家 美国	40	80
国家 苏联	60	100

癌病人数(万人)	年龄 {12,…,20}	年龄 {21,…,64}
国家 美国	40	60
国家 苏联	80	100

图 3　统计表聚集结果

8. 逆聚集

设 $E\in EXP, F=\{f_{i_1},\cdots,f_{i_k}\}$是逆聚集函数族,$Y=\{A_{i_1},\cdots,A_{i_k}\}$是 E 的属性集合(如果 E 是 C—关系)或 E 的统计属性集合(如果 E 是统计表),f_{i_j}用于 $A_{i_j}\in Y$.

如果 E 是 C—关系,令 X 是一个 C—关系,$Z=Att(E)-Y$,E 的逆聚集 Dagg(E,X,Y,F)

∈EXP 是一个 C－关系，Deg(Dagg(E,X,Y,F))=Deg(E)+Deg(X)，而且

$$I(Dagg(E,X,Y,F))=\{(t_1[Z],t_2[Att(X)],y_{i_1},\cdots,y_{i_k})\mid t_1\in I(E),$$
$$t_2\in X \text{ 而且对 } 1\leqslant j\leqslant k,(t_2,y_{i_j})\in f_{i_j}(X,t_1[A_{i_j}])\}.$$

如果 E 是原子统计表，令 X 是一个无统计属性的特殊原子统计表，Z=Satt(E)－Y，E 的逆聚集 Dagg(E,X,Y,F)∈EXP 是一个原子统计表，

$$Rdeg(Dagg(E,X,Y,F))=Rdeg(E)+Rdeg(X),$$
$$Cdeg(Dagg(E,X,Y,F))=Cdeg(E)+Cdeg(X),$$
$$Sdeg(Dagg(E,X,Y,F))=k,$$

E 的模式为 T(ROW:Ratt(E),Ratt(X);COL:Catt(E),Catt(X);SUMM:Y)，而且

$$I(Dagg(E,X,Y,F))=\{((t_1[Ratt(E)]\cdot t_2[Ratt(X)]),(t_1[Catt(E)]\cdot t_2[Catt(X)]),(y_{i_1},\cdots,y_{i_k}))\mid t_1\in I(E),t_2\in X \text{ 而且对于 } 1\leqslant j\leqslant k,(t_2,y_{i_j})\in f_{i_j}(X,t_1[A_{i_j}])\}.$$

例如，令 E 表示图 3 所示统计表，X 是一个无统计属性的原子统计表 T(ROW:性别;COL:种族)，I(X)={((男),(黑人)),((男),(白人)),((女),黑人)),((女),(白人))}，Y=(癌病人数)，F=(f_1)，$f_1(X,y)$={(((男),(黑人)),y * 25/100),(((男),(白人)),y * 25/100),(((女,(白人)),y * 25/100),(((女),(黑人)),y * 25/100)}，则 Dagg(E,X,Y,F)即为图 1(a)的第二图所示统计表.

9. 构造统计表

构造统计表操作实现 C－关系到统计表的转换. 设 E∈EXP，E 是 C－关系，X,Y,Z⊆Att(E)，函数依赖 XY→Z 成立. 由 E 构造统计表操作定义为 Cst(E,X,Y,Z). Cst(E,X,Y,Z)∈EXP 是一个统计表，Ratt(Cst(E,X,Y,Z))=X，Catt(Cst(E,X,Y,Z))=Y，Satt(Cst(E,X,Y,Z))=Z，而且

$$I(Cst(E,X,Y,Z))=\{((t[X]),(t[Y]),(t[Z]))\mid t\in I(E)\}.$$

10. 构造 C－关系

构造 C－关系操作实现统计表到 C－关系的转换. 设 E∈EXP，E 是统计表，由 E 构造 C－关系的操作定义为 Ccr(E). Ccr(E)∈EXP 是一个 C－关系，Att(Ccr(E))=Ratt(E)+Catt(E)+Satt(E)，而且

$$I(Ccr(E)=\{t_r\cdot t_c\cdot t_s\mid((t_r),(t_c),(t_s))\in I(E)\}.$$

11. 随机抽样

随机抽样操作是按指定的抽样方法，在 C－关系或统计表中提取样本. 设 E∈EXP，M 是抽样方法，P 是方法参数. E 上的随机抽样 Sample(E,M,P)∈EXP 是一个 C－关系(如果 E 是 C－关系)或统计表(如果 E 是统计表). 若 E 是 C－关系，Att(Sample(E,M,P))=Att(E). 若 E 是统计表，Ratt(Sample(E,M,P))=Ratt(E)，Catt(Sample(E,M,P))=Catt(E)，Satt(Sample(E,M,P))=Satt(E).

12. 统计分析操作

设 E∈EXP，A∈Att(E)(如果 E 是 C－关系)或 A∈Satt(E)(如果 E 是统计表)，SN 是用户给出的统计分析过成名，如均值、方差、相关系数、回归系数等. E 在 A 上的统计分析 Stati(E,A,SN)∈EXP 是一个或一组数(可视为常量或仅有一个元组的 C－关系).

上述十二种代数操作是 C－关系和统计表上的基本操作，也是 SSDB 上的基本代数操作.

下面讨论的宏代数操作和第四节定义的复合统计表上的代数操作都可以由这些基本代数操作构成.

3.2 宏代数操作

本小节讨论的宏代数操作是一些常用的操作.这些操作由上述基本代数操作复合而成,故称为宏代数操作.

1. 连接

设 $E_1,E_2\in EXP,\theta=\{\theta_1,\cdots,\theta_k\}$ 是关系运算符集,其语义与选择操作中 θ 的定义相同.

如果 E_1 和 E_2 都是C—关系,令 $A=\{A_1,\cdots,A_k\}\subseteq Att(E_1)$,$B=\{B_1,\cdots,B_k\}\subseteq Att(E_2)$,$A_i$ 和 B_i 中对应属性是关于 θ_i 相容的.E_1 与 E_2 在属性 A 和 B 上的连接 $Join(E_1,E_2,A\theta B)\in EXP$ 是一个C—关系,$Deg(Join(E_1,E_2,A\theta B))=Deg(E_1)+Deg(E_2)$,而且

$$I(Join(E_1,E_2,A\theta B))=\{t_1\cdot t_2\mid t_1\in I(E_1),t_2\in I(E_2),t_1[A]\theta t_2[B]\}.$$

如果 E_1 或 E_2 是原子统计表,不妨令 E_1 是原子统计表,$A\subseteq Ratt(E_1)\cup Catt(E_1)$,$B\subseteq Att(E_2)$,则 $Join(E_1,E_2,A\theta B)\in EXP$ 是C—关系,

$$Deg(Join(E_1,E_2,A\theta B))=Rdeg(E_1)+Cdeg(E_1)+Sdeg(E_1)+Deg(E_2),$$

而且

$$I(Join(E_1,E_2,A\theta B))=\{t_{1r}\cdot t_{1c}\cdot t_{1s}\cdot t_2\mid t_1=((t_{1r}),(t_{1c}),(t_{1s}))\in I(E_1),t_2\in I(E_2),t_1[A]\theta t_2[B]\}.$$

如果 E_1 和 E_2 都是原子统计表,令 $A\subseteq Ratt(E_1)\cup Catt(E_1)$,$B\subseteq Ratt(E_2)\cup Catt(E_2)$,$Join(E_1,E_2,A\theta B)\in EXP$ 也是C—关系,

$$Deg(Join(E_1,E_2,A\theta B))$$
$$=Rdeg(E_1)+Cdeg(E_1)+Sdeg(E_1)+Rdeg(E_2)+Cdeg(E_2)+Sdeg(E_2),$$

而且

$$I(Join(E_1,E_2,A\theta B)=\{t_{1r}\cdot t_{1c}\cdot t_{1s}\cdot t_{2r}\cdot t_{2c}\cdot t_{2s})\mid t_1=((t_{1r}),$$
$$(t_{1c}),(t_{1s}))\in I(E_1),t_2=((t_{2r}),(t_{2c}),(t_{2s}))\in I(E_2),t_1[A]\theta t_2[B]\}.$$

显然,$Join(E_1,E_2,A\theta B)=Sele(E_1\times E_2,A\theta B)$.

使用连接、聚集和选择操作可以把相关的统计表组合起来形成一个新的统计表.例如,我们有两个统计表

T_1(ROW:年,地区;COL:人口数,性别;SUMM:工业总产值),

T_2(ROW:年,地区;COL:县,人口数;SUMM:农业总产值).

从这两个统计表,我们可构造一个有关年、地区的工农业总产值统计表

$$T_3(ROW:年,地区;COL:人口数;SUMM:工业总产值,农业总产值)$$
$$=Cst(Agg(Join(T_1,T_2,A=B),X,Y,F),Z,W,Y),$$

其中,$A=\{年,地区\}$,$B=\{年,地区\}$,$X=\{年,地区,人口数\}$,$Y=\{工业总产值,农业总产值\}$,$F=\{SUM,SUM\}$,$Z=\{年,地区\}$,$W=\{人口数\}$.

2. 样本聚集

设 $E_1,E_2\in EXP$,$F=\{f_{i_1},\cdots,f_{i_k}\}$ 是聚集函数族,$A=\{A_{i_1},\cdots,A_{i_k}\}\subseteq Att(E_1)$(如果 E_1 是C—关系)或 $Satt(E_1)$(如果 E_1 是统计表),f_{i_j} 是属性 A_{i_j} 上的聚集函数.

如果 E_1 为C—关系,令 E_2 是C—关系,$Y\subseteq Att(E_1)$,$Z=Att(E_2)$,$X\subseteq Att(E_1)$,$|Y|=|Z|$.Z 中每个属性具有集合数据类型.设 Y_n 是 Y 中非集合数据类型的属性集合,$Y_s=Y-Y_n$,Z_n

和 Z_s 是 Z 中与 Y_n 和 Y_s 对应的属性集合. 样本聚集 $Tagg(E_1,E_2,X,Y,A,F)\in EXP$ 是一个 C－关系,$Deg(Tagg(E_1,E_2,X,Y,A,F))=|X|+|Z|+k$,而且

$$I(Tagg(E_1,E_2,X,Y,A,F))=\{(t_1[X],t_2[Z],y_{i_1},\cdots,y_{i_k})\mid t_1\in I(E_1),$$
$$t_2\in I(E_2),y_{i_j}=f_{i_j}(\{u\mid u\in E_1,u[X]=t_1[X],u[Y_n]\in t_2[Z_n],u[Y_s]\subseteq t_2[Z_s]\})\}.$$

如果 E_1 是原子统计表,令 E_2 是一个仅有行和列描述属性的统计表,$X=(X_1,X_2)$,$Y=(Y_1,Y_2)$,$Y_1\subseteq Ratt(E_1)$,$Y_2\subseteq Catt(E_1)$,$Z_1=Ratt(E_2)$,$Z_2=Catt(E_2)$,$X_1\subseteq Ratt(E_1)$,$X_2\subseteq Catt(E_1)$,$|Y_1|=|Z_1|$,$|Y_2|=|Z_2|$. E_2 的每个属性都具有集合数据类型. 对于 $i\in\{1,2\}$,设 Y_{in} 是 Y_i 中非集合数据类型的属性集合,$Y_{is}=Y_i-Y_{in}$,Z_{in} 和 Z_{is} 是 Z_i 中与 Y_{in} 和 Y_{is} 对应的属性集合. 样本聚集 $Tagg(E_1,E_2,X,Y,A,F)\in EXP$ 是一个统计表,

$$Rdeg(Tagg(E_1,E_2,X,Y,A,F))=|X_1|+|Y_1|,$$
$$Cdeg(Tagg(E_1,E_2,X,Y,A,F))=|X_2|+|Y_2|,$$
$$Sdeg(Tagg(E_1,E_2,X,Y,A,F))=k,$$

而且

$$I(Tagg(E_1,E_2,X,Y,F))=\{((t_1[X_1]\cdot t_2[Z_1]),(t_1[X_2]\cdot t_2[Z_2]),$$
$$(y_{i_1},\cdots,y_{i_k}))\mid t_1\in I(E_1),t_2\in I(E_2),y_{i_j}=f_{i_j}(\{u\mid u\in E_1,u[X_1]=t_1[X_1],$$
$$u[X_2]=t_1[X_2],u[Y_{1n}]\in t_2[Z_{1n}],u[Y_{1s}]\subseteq t_2[Z_{1s}],u[Y_{2n}]\in t_2[Z_{2n}],u[Y_{2s}]\subseteq t_2[Z_{2s}]\})\}.$$

例如,我们有如下两个 C－关系:

R1:

Country	Sex	Age	W1	W2
USA	M	{11,…,40}	100	5
USA	M	{41,…,70}	200	6
CAN	M	{11,…,40}	500	2
CAN	M	{41,…,70}	200	8
USA	F	{11,…,40}	100	7
USA	F	{41,…,70}	300	3
CAN	F	{11,…,40}	300	9
CAN	F	{41,…,70}	400	10

R2:

Country1	Age1
{USA}	{11,…,70}
{USA,CAN}	{11,…,70}

令 $X=\{SEX\}$,$Y=\{Country,Age\}$,$Z=\{Country1,Age1\}$,$A=\{W1,W2\}$,$F=\{SUM,MAX\}$,R1 关于 R2 的样本聚集是:

Country1	Sex	Age1	y1	y2
{USA}	M	{11,…,70}	300	6
{USA}	F	{11,…,70}	400	7
{USA,CAN}	M	{11,…,70}	1000	8
{USA,CAN}	F	{11,…,70}	1100	10

3. 描述属性置换

设 $E\in EXP$,E 表示原子统计表 $RP\subseteq Ratt(E)$,$CP\subseteq Catt(E)$,$RP'=Ratt(E)-RP$,$CP'=Catt(E)-CP$. E 的描述属性置换 $Atran(E,RP,CP)\in EXP$ 是一个原子统计表.

$$Ratt(Atran(E,RP,CP))=RP\cup CP,$$
$$Catt(Atran(E,RP,CP))=RP'\cup CP'$$

$$Satt(Atran(E,RP,CP))=Satt(E),$$

而且

$$I(Atran(E,RP,CP))=\{((t[RP]\cdot t[CP]),(t[RP']\cdot t[CP']),(t[Satt(E)]))\mid t\in I(E)\}.$$

§4. 复合统计表上的代数操作

本节讨论复合统计表上的代数操作.我们的方法是:针对复合统计表由原子统计表构成的特点,把复合统计表上的代数操作转化为一组原子统计表上的代数操作,继而利用原子统计表上的代数操作来完成复合统计表上代数操作的定义.以下用 $T(PT_1,\cdots,PT_n)$ 表示 T 是由原子统计表 PT_1、…、PT_n 构成的复合统计表,简称为统计表.$Prta(T)$ 表示 T 的原子统计表集合.$Prtn(T)$ 表示 T 中原子统计表的个数.如果不特殊说明,本节的代数表达式皆为复合统计表.这里介绍的代数操作也都是宏代数操作,可以由 3.1 节定义的基本代数操作构成.

1. 集合并与差

设 $E_1,E_2\in EXP$,$V=((i_1,j_1),\cdots,(i_k,j_k))$,对于 $1\leqslant l\leqslant k$,(i_l,j_l) 表示 E_1 的第 i_l 个原子统计表与 E_2 的第 j_l 个原子统计表相关,或可进行原子统计表级的二元代数操作.记

$$Relat(E_1,V)=\{PT_{i_1},\cdots,PT_{i_k}\}\subseteq Prta(E_1),$$

$$Relat(E_2,V)=\{PT_{j_1},\cdots,PT_{j_k}\}\subseteq Prta(E_2).$$

令 E_1 和 E_2 中由 V 指定的相关原子统计表的对应属性具有相同的数据类型,而且,对于 $1\leqslant l\leqslant k$,

$$Rdeg_{i_l}(E_1)=Rdeg_{j_l}(E_2),$$

$$Cdeg_{i_l}(E_1)=Cdeg_{j_l}(E_2),$$

$$Sdeg_{i_l}(E_1)=Sdeg_{j_l}(E_2).$$

E_1 和 E_2 的并 $Sum(E_1,E_2,V)\in EXP$ 是一个统计表,而且

$$Prtn(Sum(E_1,E_2,V))=Prtn(E_1)+Prtn(E_2)-k,\text{令其为 }n,$$

$$Prta(Sum(E_1,E_2,V))=Prta(E_1)\cup(Prta(E_2)-Relat(E_2,V)),$$

$$I(Sum(E_1,E_2,V))=T(PT_1,\cdots,PT_{n-k},PT_{n-k+1},\cdots,PT_n),$$

其中,对于 $1\leqslant i\leqslant n-k$,$PT_i\in(Prta(E_1)-Relat(E_1,V))\cup(Prta(E_2)-Relat(E_2,V))$,对于 $n-k+1\leqslant l\leqslant n$,$PT_l=PT_{i_{l-n+k}}\cup PT_{j_{l-n+k}}$.

E_1 和 E_2 的差 $Dif(E_1,E_2,V)\in EXP$ 也是一个统计表,而且

$$Prtn(Dif(E_1,E_2,V))=Prtn(E_1)-Prtn(E_2)+k,\text{令其为 }m,$$

$$Prta(Dif(E_1,E_2,V))=(Prta(E_1)-Prta(E_2))\cup Relat(E_1,V),$$

$$I(Dif(E_1,E_2,V))=T(PT_1,\cdots,PT_{m-k},PT_{m-k+1},\cdots,PT_m),$$

其中,对于 $1\leqslant i\leqslant m-k$,$PT_i\in Prta(E_1)-Prta(E_2)$,对于 $m-k+1\leqslant l\leqslant m$,$PT_l=PT_{i_{l-m+k}}-PT_{j_{l-m+k}}$.

2. 笛卡尔积

设 $E_1,E_2\in EXP$,E_1 表示统计表 $T_1(PT_1,\cdots,PT_n)$,E_2 表示统计表 $T_2(PT'_1,\cdots,PT'_m)$.E_1 和 E_2 的笛卡尔积 $Card(E_1,E_2)$ 是 $n\times m$ 个 C-关系 CR_{11}、…、CR_{nm}.对于 $1\leqslant i\leqslant n$ 和 $1\leqslant j\leqslant m$,$CR_{ij}=PT_i\times PT'_j$.

3. 选择

设 $E\in EXP$,E 表示统计表 $T(PT_1,\cdots,PT_n)$,$X=\{PT_{i_1},\cdots,PT_{i_k}\}\subseteq Prta(E)$,$F=\{F_1,\cdots,F_k\}$是一组条件表达式,$F_j$ 的定义与原子统计表选择操作中 F 的定义相同. E 的选择 $Sele(E,X,F)\in EXP$ 是一个统计表. $Prtn(Sele(E,X,F))=Prtn(E)$,而且

$$I(Sele(E,X,F))=T'(PT'_1,\cdots,PT'_n)$$

其中,对于 $1\leqslant i\leqslant n$,如果 $PT_i\in X$,$PT'_i=Sele(PT_i,F_i)$,否则 $PT'_i=PT_i$.

4. 投影

设 $E\in EXP$,$Prta(E)=\{PT_1,\cdots,PT_n\}$,$X=(X_{i_1},\cdots,X_{i_k})$,对于 $1\leqslant j\leqslant n$,$X_j=R_j\cup C_j\cup S_j$,$R_j\subseteq Ratt(PT_j)$,$C_j\subseteq Catt(PT_j)$,$S_j\subseteq Satt(PT_j)$,S_j、R_j、C_j 满足:若 $S_j\neq 0$(空集)则 $R_j=Ratt(PT_j)$、$C_j=Catt(PT_j)$. E 在 X 上的投影 $Proj(E,X)\in EXP$ 是一个统计表,$Prta(Proj(E,X))=\{PT'_1,\cdots,PT'_n\}$,其中,对于 $1\leqslant j\leqslant n$,如果 $j\in\{i_1,\cdots,i_k\}$,则 $PT'_j=Proj(PT_j,X_j)$,否则 $PT'_j=PT_j$.

5. 聚集

设 $E\in EXP$,$Prta(E)=\{PT_1,\cdots,PT_n\}$,$F=(F_{i_1},\cdots,F_{i_k})$,$F_{i_j}$是聚集函数族,同原子统计表聚集操作定义中的 F,用于 PT_{i_j},$X=(X_{i_1},\cdots,X_{i_k})$,$Y=(Y_{i_1},\cdots,Y_{i_k})$,$Z=(Z_{i_1},\cdots,Z_{i_k})$,$X_{i_j}\subseteq Ratt(PT_{i_j})$,$Y_{i_j}\subseteq Catt(PT_{i_j})$,$Z_{i_j}\subseteq Satt(PT_{i_j})$. E 的聚集 $Agg(E,X,Y,Z,F)\in EXP$ 是一个统计表,$Prta(Agg(E,X,Y,Z,F))=\{PT'_1,\cdots,PT'_n\}$,其中,对于 $1\leqslant j\leqslant n$,如果 $j\in\{i_1,\cdots,i_k\}$,则 $PT'_j=Agg(PT_j,X_j,Y_j,Z_j,F_j)$,否则 $PT'_j=PT_j$.

6. 逆聚集

设 $E\in EXP$,$Prta(E)=\{PT_1,\cdots,PT_n\}$,$F=(F_{i_1},\cdots,F_{i_k})$,$F_{i_j}$是逆聚集函数族,与原子统计表逆聚集操作定义中的 F 相同,用于 PT_{i_j},$S=(S_{i_1},\cdots,S_{i_k})$,$S_{i_j}$是无统计属性的特殊原子统计表,$Y=(Y_{i_1},\cdots,Y_{i_k})$,$Y_{i_j}\subseteq Satt(E)$. E 的逆聚集 $Dagg(E,S,Y,F)\in EXP$ 是一个统计表,$Prta(Dagg(E,S,Y,F))=\{PT'_1,\cdots,PT'_n\}$,其中,对于 $1\leqslant j\leqslant n$,如果 $j\in\{i_1,\cdots,i_k\}$,则 $PT'_j=Dagg(PT_j,S_j,Y_j,F_j)$,否则 $PT'_j=PT_j$.

7. 随机抽样

设 $E\in EXP$,$Prta(E)=\{PT_1,\cdots,PT_n\}$,$M=(M_{i_1},\cdots,M_{i_k})$,$P=(P_{i_1},\cdots,P_{i_k})$,$M_{i_j}$是抽样方法,用于 PT_{i_j},P_{i_j}是方法 M_{i_j}的参数. E 上的随机抽样 $Sample(E,M,P)\in EXP$ 是一个统计表. $Prta(Sample(E,M,P))=\{PT'_1,\cdots,PT'_n\}$,其中,对于 $1\leqslant j\leqslant n$,如果 $j\in\{i_1,\cdots,j_k\}$,则 $PT'_j=Sample(PT_j,M_j,P_j)$,否则 $PT'_j=PT_j$.

8. 统计分析操作

设 $E\in EXP$,$Prta(E)=\{PT_1,\cdots,PT_n\}$,$SN=(SN_{i_1},\cdots,SN_{i_k})$,$A=(A_{i_1},\cdots,A_{i_k})$,$A_{i_j}\subseteq Satt(PT_{i_j})$,$SN_{i_j}$是用户给出的统计分析操作,用于 PT_{i_j}的的统计属性 A_{i_j},E 在 A 上的统计分析 $Stati(E,A,SN)\in EXP$ 是一个或一组统计值(可视为常量或仅有一个元组的 C-关系).

我们可以类似地定义复合统计表上的连接、样本聚集、属性置换等操作. 下边定义复合统计表上常用的原子表置换和原子表投影操作.

9. 原子表置换

设 $E\in EXP$,$I(E)=T(PT_1,\cdots,PT_n)$,$P=(i_1,i_2,\cdots,i_n)$是(1,2,…,n)的一个置换. E 的原子表置换定义为 $Ttran(E,P)\in EXP$. $I(Ttran(E,P))=T'(PT_{i_1},\cdots,PT_{i_n})$.

10. 原子表投影

设 $E \in EXP, I(E)=T(PT_1,\cdots,PT_n), P=(i_1,i_2,\cdots,i_k)$，对于 $1 \leqslant j \leqslant k, 1 \leqslant i_j \leqslant n$. E 上的表投影 $Tproj(E,P) \in EXP$ 是一个统计表. $I(Tproj(E,P))=T'(PT_{i_1},\cdots,PT_{i_k})$.

参考文献

[1]A. Shoshani and H. K. T. Wong, Statistical and Scientific Database Issue, IEEE Transactions on Software Engineering, SE-11: 10(1985), 1040-1047.

[2]李建中，统计和科学数据库系统的要求——与面向事务处理数据库系统的比较，计算机科学，1(1990)，65-70.

[3]李建中，孙文隽，统计与科学数据库的数据模型，计算机学报，10(1991)，757-763.

[4]A. Klug, Equivalence of Relational Algebra and Relational Calculus Query Languages Having Aggregate Functions, J. ACM 29: 3(1982), 699-717.

[5]S. Y. W. Su, SAM*: A Semantic Association Model for Corprate and Scientific-Statistical Databases, Information Sciences, 29: 2, 3(1983), 151-200.

[6]G. Ozsoyoglu, Z. M. Ozsoyoglu and V. Matos, Extending Relational Algebra and Relational Calculus with Set-Valued Attributes and Aggregate Functions, ACM Transactions on Database Systems, 12: 4(1987), 566-592.

全国第十一届数据库学术会议

征　文　通　知

中国计算机学会软件专业委员会，办公自动化专业委员会数据库专业学组拟于一九九三年九月在西安联合召开全国第十一届数据库学术会议。

1. **征文内容**：①数据库模型与语言，②数据库理论与算法，③多介质数据库，④工程数据库，⑤特种数据库（主动，模糊，统计，时序），⑥面向对象数据库，⑦分布式数据库，⑧知识库系统，⑨数据库应用（CAD/CAM，MIS，OA），⑩数据库管理系统的体系结构及实现技术。

2. **截稿日期**：1993 年 4 月 15 日（邮戳日期）

3. **来稿要求**：①上述征文内容中理论研究和开发成果未公开发表者均可应征。②每篇论文字数不超过 8000 字（含图表，附录，参考文献）。③论文摘要不超过 200 字。④字迹清楚，工整，用稿纸（20×20）书写，一式两份。⑤来稿无论录用与否均不退还，故请自留底稿。⑥论文请寄“**西安西北工业大学 168 信箱　蒋泽军**（710072）”，信封请注明“**征文**”。