第6卷第2期
1995年2月

软 件 学 报
JOURNAL OF SOFTWARE

Vol. 6, No. 2
Feb. 1995

# 统计数据库管理系统的设计与实现*

曾红卫　陈永年

(上海科技大学计算机系,上海 201800)

**摘要**　统计与科学数据库与常规商用数据库有很大差异,利用常规的商用数据库管理系统建立统计与科学数据库是不合适的.本文以我们开发的统计数据库管理系统(SSDBMS)为基础,讨论SSDBMS的数据模型数据压缩、安全保密、统计查询等一系列技术.

**关键词**　统计数据库(SDB)、分类属性、汇总属性、汇总表、混合模型、数据压缩.

统计数据库主要用于统计与科学数据的存储、操纵、统计分析及归纳等.如气象、水文、人口普查、科学实验等领域.它与常规商用数据库相比有许多特性:

·数据可分成分类属性与汇总属性.分类属性用于描述汇总属性的特性,分类属性的联合构成整个元组的组合键.而汇总属性则用于统计分析,如水文中的"年代"、"水系"、"河名"、"站名"等属于分类属性.而相应的,为上述属性所唯一地确定的水文要素则属于汇总属性.由于在统计数据库中只需将分类属性集合中任何一个属性的实例值进行改变,而其它的实例值保持不变,就可以得到一个新元组,所以采用常规的记录形式进行存贮就会造成很大的重复;

·统计数据库中的分类属性与科学试验中测量参数属性很相似,而它的汇总属性又与科学数据中的测量值属性很相似.因此,科学数据库也可借用分类属性及汇总属性的概念.而且科学数据的整理也需要借助于整套统计分析算法.所以,我们将这两种数据库用同一种数据库管理系统来支持是合适的;

·汇总属性中常常有大量相同的实例值(如零),应进行压缩以减少冗余;

·统计分析往往只涉及一个或少数几个属性上的数据.如果用记录为单位对数据进行存贮、检索,将耗费许多不必要的I/O时间.因此,应当在存贮格式上采用与常规记录格式完全不同的,而以同一属性的实例值存放在一起的转置文件格式;

·统计数据库所支持的运算与常规数据库也不一样,它除了一般检索外,还需支持大量的统计运算和其它特有的操纵;

·允许对数据库的统计查询,又要防止个体数据的泄露,包括统计推理;

·数据量庞大且含大量元数据.元数据可包含数据生成者、生成时间、生成原因等信息.由于统计科学数据库很庞大,数据很复杂,用户很容易把这方面的信息遗忘而影响使用,因

* 本文1992-12-11收到,1993-04-06定稿

作者曾红卫,1966年生,讲师,主要研究领域为数据库.陈永年,1927年生,教授,主要研究领域为软件,体系结构.
本文通讯联系人:曾红卫,上海201800,上海科技大学计算机系

此元数据的管理是很重要的.元数据库一般可以用统计数据库管理系统实现;

·用户可根据自己的需要抽取较小的数据集合构成子库和汇总表,子库和汇总表通常可以采用类似选择操作的运算以选出有用的元组集合,或采用类似投影的运算以选出有用的属性集合,并通过统计查询进行汇总.

因此,需要开发一种面向建立统计与科学数据库的新型数据库管理系统——统计数据库管理系统(SSDBMS).下面几节总结和分析SSDBMS的设计与实现技术.

## 1 体系结构

SSDBMS采用图1所示体系结构.

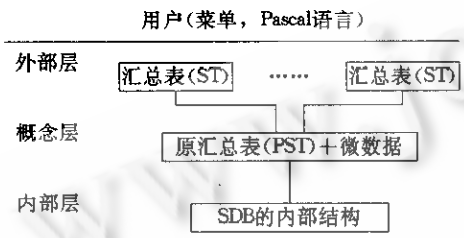


图1 SSDBMS体系结构

其中:汇总表(ST),用作SSDBMS的外部结构,构成亲统计用户接口.统计用户对其可进行下列操作:定义ST的模式;生成ST值;对ST作统计查询;绘图(直方图、折线图、饼图);制表.

原汇总表(PST)和微数据,用作SSDBMS的概念层结构,是构成ST的基本块.PST能自然地映射到内部层.用户对ST的操作将映射为对PST的操作:定义PST的模式;根据微数据生成PST;PST值的存贮和存取;查询一组PST以支持对ST的统计查询;对微数据和PST的查询及添、删、改;安全保密.

SDB的内部表示,SSDBMS的内部层.采用转置文件组织和数据压缩技术,以及元数据管理技术.除了存贮微数据外,还存贮PST中的宏数据.

## 2 数据模型

为了适应统计与科学数据的特点,SSDBMS采用与常规数据库不同的数据模型.模型的要点是用两种语义节点(叉乘节点,简称X节点;聚集节点,简称C节点)将分类属性组成层次树形结构;汇总属性实例值则可存贮在根据相应的分类属性实例值所确定的位置中,这样就构成了转置文件.

叉乘节点:表示属性的多维性质.

聚集结点:表示属性数据间的从属关系或组成关系.

图2给出了简单人口统计中分类属性组成的树及汇总倒置文件——即人口统计模型.

引言中谈到,分类属性联合构成统计库的组合键,通过展开分类属性树,即可求得分类属性实例值与汇总属性实例值的对应关系.换言之,汇总属性倒置文件是按分类属性树展开的顺序存放的.显然,这种模型既能加快检索速度,又节省了大量存储空间.

然而,在实际应用中所遇到的统计与科学数据往往比较复杂,它们在某些分类属性集合上是十分接近上述统计数据库模型的,但对于其它的分类属性集合则不能有效地用上述模型支持,也就是它们的属性实例值要比模型所展开的要稀疏得多.以水文数据为例,降水量数据的分类属性应当是"测站"、"年"、"月"、"日"、"时"、"分",汇总属性是"降水量"等.由于

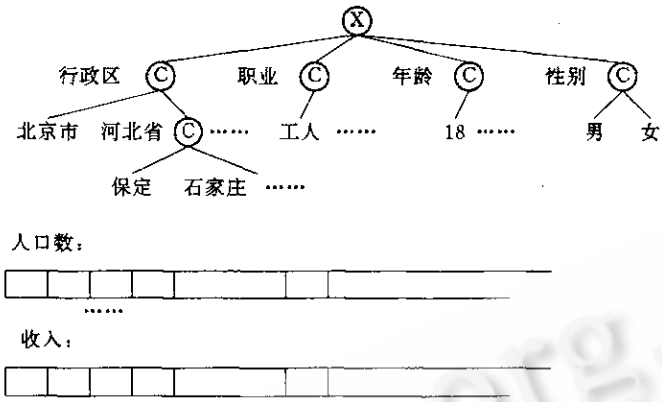


图2 数据模型实例

降水时间是很不规则,有时很长时间如数 10 天甚至几个月不降雨,有时则出现大雨或暴雨.此时测量时间间隔需要短到每 5 分钟一次.如果将它的分类属性的时间间隔定为 5 分钟,则可以满足大、暴雨测量的需要,但对于不降雨或降小雨时,数据过于稀疏.虽然可以用压缩方法解决,但花费太大,使用不方便.由此,提出了一种新的混合模型以支持这类数据库.

混合模型将分类属性集合分为两部分:适合上述模型的部分用树描述,其余部分用一般关系模型描述.如水文库中,"测站"、"年"、"月"构成分类属性树;"日"、"时"、"分"则用关系描述.如此,分类属性树的一个展开实例值不再是对应汇总属性的一个实例值,而且对应分类属性关系的多个记录或汇总属性的多个实例值,但关系记录与汇总实例值是一对一的.为此,在树与关系记录、汇总属性间建立指针文件,以支持这种一对多的联系.如图 3 所示水文库模型.

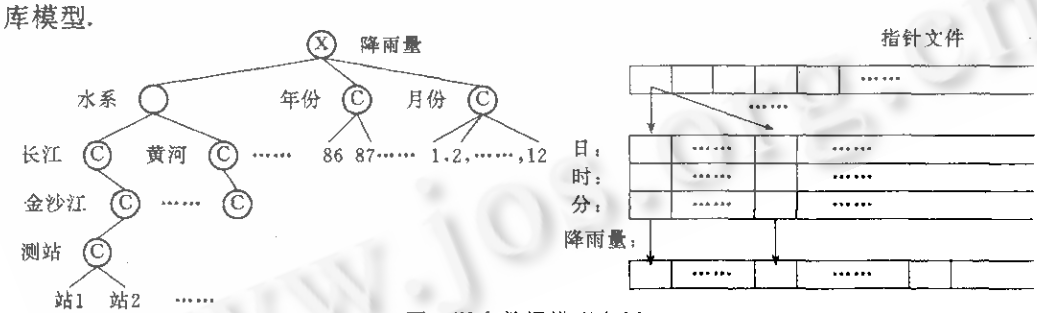


图3 混合数据模型实例

## 3 数据压缩

统计与科学数据库中往往有大量相同的数据(如零值)相邻地集结在一起,对于这种数据采用标头压缩技术最为合适.标头压缩技术的基本思想是采用标头标明所指数据串的类别(是集结在一起的某个常量,还是一个变量数据串),它还标明在压缩前的数据序数以及压缩后数据串末尾所处的字节数等.这样对于相邻地集结在一起的许多相同的常量在压缩后就可以只存放一个数据,从而达到压缩的目的.标头压缩技术还可分为用于多相邻常量和变长度数据的双计数压缩方案,多相邻常量和定长度数据及单相邻常量及定长度数据的单计数压缩方案等.

在统计数据库中,一般来讲,数据都是定长的,但重复出现的数据除零外可能还会出现少数几个常量.因此我们采用了多常量和单长度(MCSL)数据压缩方案.其具体实现是标头的前部是指示数据.如用"*"指示数据为变量,1指示重复常量"1",2指示重复常量"2"等.标头的后部分别指示该变量串或重复常量串最后一位分别在压缩前存贮序列中所处的位置.

例如:

压缩前逻辑数据库

V1 V2 V3 1 1 V4 V5 0 0 V6 1 1 0 0 V7 V8

压缩后实际存贮的数据库

V1 V2 V3 V4 V5 V6 V7 V8

标头

*3 1.2 *5 0.4 *6 1.6 1.8 *8

图4 压缩前后的数据及标头

## 4 安全保密措施

统计数据库的另一特点是既要为各类用户方便地提供有关群体的各种统计数据,又不能因此泄露个体保密信息,包括防止用户采用推导的办法而获得个体的信息,从而达到保密的目的.例如,全国或某省的某种产品的产量是不保密的,但涉及某个具体工厂的这种产量是不应泄露的,在统计数据库中虽然不允许一般用户直接访问个体机密信息,但在每个统计结果中都会有残留的原始个体数据的痕迹,如不采取适当的措施,用户就可以利用合法的统计数据进行推导而获得所需个体信息.为了防止泄露,统计数据库的安全措施一般有:限制查询集合的大小、限制查询集合交的大小、随机取样查询、对数据值进行微扰、数据库划分等方法.结合统计数据库采用的数据模型,处于各层的树节点事实上已把数据库进行了划分.越是接近根节点的上层节点,它所包含的个体越多,所以它的保密级别应当越低.相反,越接近叶节点其保密级别应越高.然而,在查询中大都会涉及多个节点,这时,计算多个节点的联合密级作为查询密级.同时,对不同类型的用户赋以不同的权限.只有当用户的权限高于查询密级查询时才被允许,否则被拒绝.这种方法称为层次保密法.作为一种选择,在SSDBMS中还支持数据微扰,即用微小的数值对查询结果进行打扰.层次与微扰两种措施可单独使用,也可联合使用.

## 5 数据检索和统计分析

SSDBMS采用类SQL语言,就检索语言而言,语句结构为:

SELECT 输出表
FROM 库名
[WHERE 检索条件]
[GROUP BY 分组属性表]
[HAVING 条件]

其中,[ ]表示可选,输出表可含聚集和算术运算.

从 SSDBMS 的数据模型可知,库数据分为两部分:分类属性树及汇总转置文件或分类属性关系.分类属性树展开后的实例值或一一对应于汇总属性实例值或通过指针文件对应于一般关系记录和汇总属性实例值(当为混合模型时),这实际上构成了一个索引树.由此,采取下列优化措施:

(1)将检索条件分解为树属性条件和非树属性条件,树属性条件仅包含关于分类属性树中的属性,其余的检索条件构成非树属性条件;

(2)根据树属性条件建立检索条件树,即满足检索条件的分类属性树的子树;

(3)在检索条件树的基础上,根据分组属性建立分组属性树;

(4)检索操作先后在三级索引上进行,求出符合检索的树分类属性,然后取出关系记录和汇总属性值比较非树属性条件.当完成一组后,再检查 HAVING 条件.

图 5 给出了各检索层次.

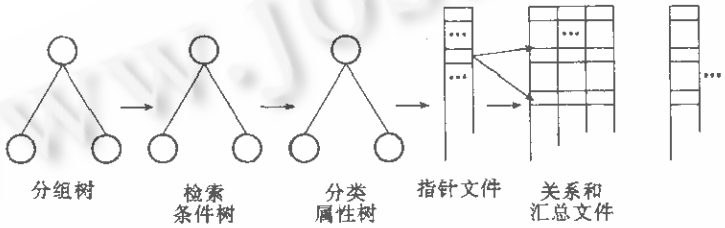


图5

SSDBMS 支持 40 多种统计分析,包括曲线拟合、矩阵运算、回归分别等,它们的数据来源于对库的检索,检索结果按规定格式存于临时中间文件中,然后对文件中数据进行统计分析,分析结果存于文本中或通过曲线、饼图、直方图输出.

## 6　结论与展望

SSDBMS 采用叉乘和聚集结点建立分类属性树,以及用转置文件存放总属性值,较好地适应了统计与科学数据的特点.压缩节省了存贮,查询优化提高了效率,特别是混合模型拓广了应用范围.然而,现在大部分应用都是以关系数据库为基础的.建立与关系数据库的良好接口十分必要.同时,为适应新的发展要求,结合面向对象技术,开发一个面向对象的统计数据库管理系统也将成为一种趋势.

## 参考文献

1　Chin F Y, Ozsoyoglu G. Statistical database design. ACM Trans. on Database System, 1981,6(1):113－139.

2　曾红卫.统计数据库中的数据压缩与有效存取.中国软件行业协会青年协会论文集(第一辑),中国软件行业协会青年协会第一届年会,北京,1988:189－194.

3　李志中.统计数据库.计算机工程与设计,1986(4).

4　陈永年,曾红卫等.统计数据库的安全性.数据库进展——中国第八届数据库学术大会论文选集,497－503.

5　Paul Chen, Arie Shoshani. A directory driven system for organizing and accessing large statistical database. A LBL Perspective on Statistical Database Management, 69－87.

# DESIGN AND IMPLEMENTATION OF SCIENTIFIC AND STATISTICAL DATABASE MANAGEMENT SYSTEM

Zeng Hongwei Chen Yongnian

(*Department of Computer Science, Shanghai University of Science and Technology, Shanghai* 201800)

**Abstract** There are many differences between scientific and statistical database and traditional DB. It is not suitable to build scientific and statistical database with traditional database management system. This paper discusses some technologies in SSDBMS, including data_model, data_compression, query, security etc.

**Key words** SDB(Statistical Database), category attribute, summary attribute, summary table, mixed data model, data compression.