国家自然科学基金资助项目专著

统计数据库保密原理

魏仲山 编著

学出版社

国家自然科学基金资助项目专著

统计数据库保密原理

魏仲山 编著

天津大学出版社

内 容 提 要

本书系统介绍了统计数据库模型、对统计数据库的攻击方式和统计数据库的安全措施等。书中介绍的安全方法简单实用，并配有不少例题，便于读者掌握和理解。本书是使用统计数据库人员的良师益友。

本书可供统计数据库研究和应用人员、计算机应用专业人员阅读，也可供大专院校有关专业师生参考。

统计数据库保密原理

魏仲山 编著

*

天津大学出版社出版
（天津大学内）
河北省邮电印刷厂印刷
新华书店天津发行所发行

*

开本：787×1092毫米1/32　　印张：6 1/4　字数：140千字
1989年10月第一版　　1989年10月第一次印刷
印数：1-4500
ISBN 7-5618-0166-1

TP·22

定价：2.90元

前　言

目前，计算机应用已深入到社会各个领域。计算机系统存贮了大量经济、国防、生产经营等信息和科技知识，是现代化社会的基础和支柱。然而，由于计算机信息处理、存储、传输、使用和管理上的脆弱性，使它容易滥用、溢漏、丢失、泄露，也容易受到窃取、窜改、冒充和破坏的威胁。计算机信息可分为应用信息和系统信息两种。对这些信息的威胁大致可分为三类：

（1）对计算机应用信息机密性的威胁；

（2）对计算机应用信息真实性和完整性的威胁；

（3）对计算机系统信息的破坏性威胁，包括计算机病毒在内。

本书只涉及上述第一类威胁。数据库内信息和数据的安全性，一般受存取控制机构保护。但是，对统计数据库而言，有一种特殊安全性问题存在。一个统计数据库既要发布统计信息为用户共享，又要保护机密信息不被泄露。这是存取控制机构力所不及的。通过多次相关查询获得统计数据，就有可能从中推断出有价值的数据，从而使统计数据库泄密。

信息的共享和保护、破坏和安全、扩散和保密互相对立，互为制约。计算机信息保护学就是研究计算机系统内信息保护方法的一个总课题。它的基础理论包括存取控制、加密控制、信息流向控制、隔离控制和推断控制。存取控制是对存取信息的授权理论，它与存取信息的内容不相干。反之，加密控制则

使信息内容更加难懂，以使不为对手识破。它与授权并无直接关系。信息流向控制涉及信息传递并非是权的转让。若对手借机滥用授权，则就有使信息流向失控而达到顺手牵羊的作案目的。基于互斥原理的隔离控制以牺牲信息共享为代价，使信息互相隔离而达到保护信息安全的目的。推断控制所面临信息泄密威胁是，在满足授权条件下，利用可获取信息内容而推断出机密信息。实质上，这是因“滥用信息内容”而构成的失密威胁，即统计数据库面临的安全威胁，这也是难以防止的。不言而喻，统计数据库安全性对经济建设有重要意义。

本书由国家科委信息技术政策顾问、机械电子工业部高级顾问王运丰教授审阅，在此表示感谢。由于作者水平所限，本书会有若干不足和错误，望读者批评指正。

魏仲山

1989年4月

目　　录

第一章 引 言

统计数据库是指这样一种数据库，即从中取得的信息是关于一实体集子集的汇总统计信息。在统计数据库里，除了禁止对数据库非法存取和修改的一般问题外，还有特殊的安全问题存在。保护统计数据库的目的是，由该数据库发布统计信息时，保证不会使其中受保护的信息泄密。

有许多数据库只为提供统计数据所用；如人口普查数据库就是这样。通用数据库既可为统计存取用，又可为非统计存取信息用。例如，在医疗数据库内，医生可以直接存取病人的医疗记录；而研究人员只许存取这些记录的汇总统计信息。我们关注的主要是通用数据库内对保密信息的保护机制，但也同时关注专用统计数据库的保护。

§1.1 数据库的安全问题

数据库的安全问题很多，处理方法各异。需要保护的数据库，应使数据不致受到非法修改或破坏，不让未经授权的用户读出数据。安全问题并不是数据库系统独有的，整个计算机系统都存在着信息保护问题。由于数据库内存有大量数据，而且为各用户共享，这就使数据安全问题很突出。

对数据库而言，主要保护机制是存取控制。这就是说，对于何范围内的数据，在何条件下，规定哪一个用户许可进行哪一种操作。

数据库存取控制决策是以若干因素为基础的。例如，对事件保密的信息而言，除非在规定时间内为一组专门用户开放，或为某专门终端使用，否则它禁止任何用户使用。对值保密的信息，其存取决策取决于数据现行值。如，用户不可读出收入超过一定数额的人员的档案记录。对状态保密信息在数据库管理系统处于动态状况下起作用。例如，只当数据库对一文件处于未锁状态时，用户方能够打开该文件。对模式保密信息，其存取决策取决于对数据用法的规定。如授权用户调用分类程序对某一种文件进行分类，此时它既不许可读程序逻辑，也不许可读该文件。在这种情况下，分配给用户的许可操作是只执行。但是，为了作分类而授予分类程序的许可操作为读。由此可见，分类程序代表用户读数据，但该程序不能把数据返到用户读出。这种对模式保密信息的存取决策，需要对数据安全上的弱点进行掩饰。另外，用户可用特定存取次序推断出对存取历史保密的信息。为此，数据库管理系统就须记录存取历史并加以保护。

综上所述，保密信息可分为二类：一是同周围环境无关，即不依赖于相关域；一是同周围环境有关，即有赖于相关域。对事件，值和状态保密的信息而言，是不依赖于相关域的，其含义是数据库管理系统只查验事件寄存器内容，值场或触发器状态，就可立即作出存取决策。为对当前的存取请求作出确认或否认，系统不必回忆过去的存取历史。在不依赖于相关域的存取控制机构中，在软件上不论是采用视图机构，还是采用修改查询机构，均易于设计和实现。

对存取模式和存取历史保密的信息，依赖于相关域。为此，数据库管理系统必须知道相关数据关系，以确定存取决策。假设有5个数据项D_1至D_5，并有下列保护属性：

（1）若项1读（D_1，R），或项2执行（D_2，E），则项3不能读。否则，项3能读（D_3，R）。

（2）若项4打印（D_4，P），则项5不打印。若项5打印（D_5，P），则项4不打印。换言之，只有一项打印。

此数据库如图1-1(a)所示。图内结点代表数据项和其保护属性，结点间关系用箭头代表。这样，第（1）项保护属性在该图内用（D_3，R）至（D_1，R）和（D_2，E）的箭头表示。

假设提交下列二个经简化的数据管理作业。作业1为

（A） 执行D_2；

（B） 打印D_5；

（C） 读D_1。

作业2为

（A） 打印D_4；

（B） 读D_3；

很明显，作业1提出的请求可予确认，因它不会损害规定的保护属性。若数据库管理系统正处理作业2，就有二项对数据项D_3和D_4的新请求。对这二个新请求，显然不能确认。因为如果对该请求确认，就会破坏数据的安全，而此种数据的安全是在数据库建立期间由保护属性规定的，如图1-1(a)所示。设用户结点至数据项的箭头，代表用户向该数据项的请求。同理，数据项至用户结点的箭头，代表对请求的确认。图1-1(b)表示作业1结束，图1-1(c)表示作业2向数据项D_3和D_4提出请求后的数据库状态。

利用图1-1(c)，可自动识别因确认作业2的二个新请求而构成对安全的破坏。如果用户结点和数据项结点间的箭头构成一个循环，就有可能使依赖于相关域的保护属性受到破坏。例如，该图内自用户结点指向数据项D_4的请求箭头，自D_4至D_5

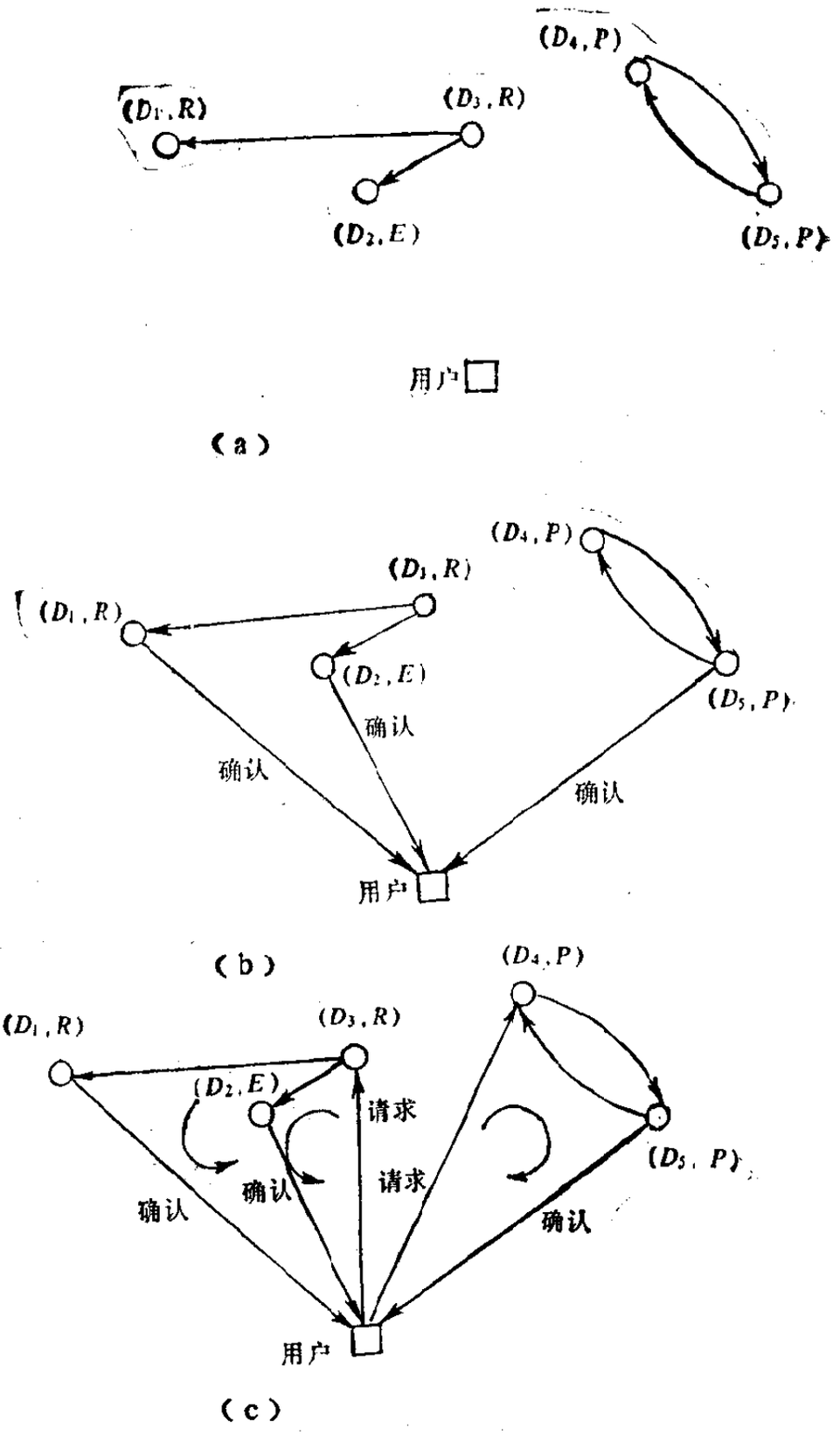


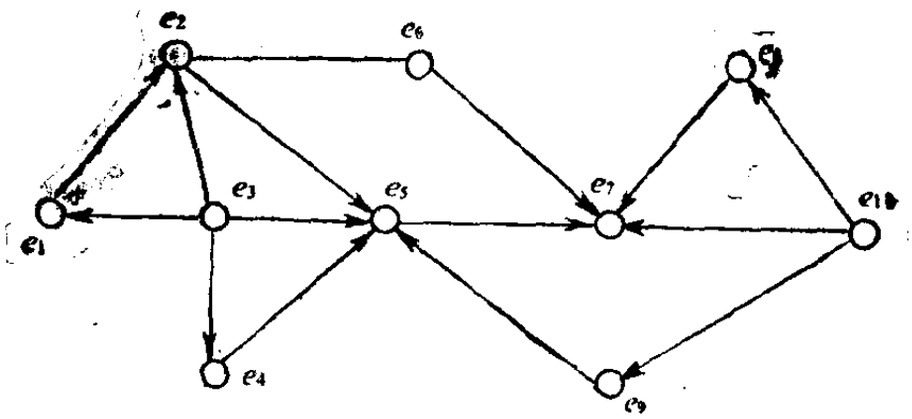
图 1-1　有相关域保护属性的数据库存取控制

的关系箭头和自D_5至用户结点的确认箭头，就构成了一个环。在该图内有三个环存在，说明受到袭击有三种可能性。数据库管理系统识别闭环的能力，就使系统能够发现可能受到的威胁和破坏。

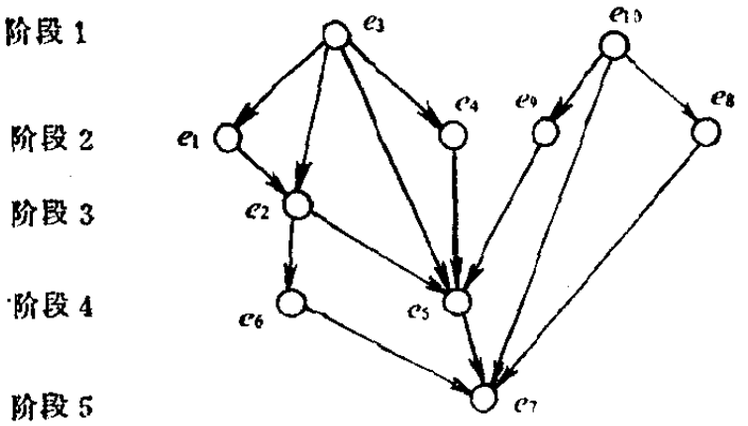
综上所述，只要表示有相关性信息保护要求的是利用嵌入到数据库内各数据项的有向图，则数据库存取控制机构就可使用上述对信息安全定义实现有相关性信息的保护和发现可能出现的对安全的威胁。自然，用户(蓄谋或无意地)非法存取受保护数据时，这种威胁是存在的。但是，合法用户使用规定的保护属性是允许存取受保护数据的。因为各种信息保护要求间的关系会很复杂，所以提出实现可行目标是为保证信息不受任何威胁而满足最大限度地存取数据项并确定相应存取顺序。对此，用图1-2(a)说明有相关性信息保护要求的存取控制问题。该图由10个结点（成对数据项和保护属性）和15个箭头（嵌入保护要求）组成。用不同级（箭头只由顶向下）重新绘制出有向图，如图1-2(b)所示。由此可见，如果确定出正确的存取请求顺序图1-2(c)，即第一次请求为e_3，e_{10}，最后一次请求为e_7，则可对整个数据库作安全存取。

所以，把图1-2(a)变换成图1-2(b)，是采用了一种算法，即使用此种存取请求顺序，用户不仅得到对所请求数据的确认，还满足相关域保护对数据的要求。

在此基础上，讨论数据库环境下对一个专利包的保护要求。专利包为许多用户使用和共享。然而，由于专利性质，对该包的存取必须受到控制。可以有两类控制。首先是必须明确对专利包本身的保护。即在此情况下，合法用户没有读，写或打印的存取权。只有专利包所有者才有这些权利。这就是说，合法用户仅有执行的存取权。其次是对该包使用必须为ii



（a）



（b）

第一次请求：$\{e_3, e_{10}\}$　　第二次请求：$\{e_1, e_4, e_8, e_9\}$
第三次请求：$\{e_2\}$　　第四次请求：$\{e_5, e_6\}$
第五次请求：$\{e_7\}$

（c）

图 1-2　确定安全的存取顺序

费和其他用途受到监控。每一次使用专利包，专利包所有者的监控器就将作用。这样对该包使用确实受到控制和保护。所以，数据库管理系统应提供用于自动实施的相关程序机构。第一类存取控制可由传统数据库系统提供。对该系统的不同用户而言，各数据项（程序或信息）带有不同保护属性。这是一种不依赖于相关域的保护机制。第二类存取控制，是在现行程序

执行之前先执行别的程序。这实际上是有赖于相关域保护更为普遍形式的一种特殊情况。这就是说，除非对别的数据项已作存取，否则不能对一数据项（程序或信息）进行存取。这样，如果这些数据项受到妥善保护，则因为系统在可能受到袭击之前拥有抢先的能力，所以不会因别的存取而受到攻击。

从存取控制理论上讲，如果存在一个有赖于相关域保护关系集的数据库，则为保证安全应能找出数据库的最大子集，使对该子集的一项许可存取不致对该数据库其他项拒绝存取造成任何损害。如果存在一个有赖于相关域保护关系集的数据库，则为保证安全应能找出一种存取顺序，不致对实现这些保护要求造成破坏。

总之，在数据库安全保护方面，要设置语义关系并使其付诸使用。说明相关域保护关系和算法特征，使数据库管理系统的存取控制机构可以实现预期的安全措施，而不管数据库内的语义关系和推断关系如何。数据库语义说明和实施算法，就构成智能存取控制机构。可以指出，此种存取控制机构真正把共享资源当作语义上相关实体的集合处理。实现安全不仅靠拒绝存取受保护实体，也靠拒绝存取可推断的实体，即语义上与受保护实体相关的实体。数据库安全的目标，应当同智能存取控制机构结合。

§1.2 统计数据库的安全

统计数据库是各数据项的集合，其中包含许多机密信息。用户对各数据项的统计信息有存取权。假设统计数据库的各个数据元是机密的，用户没有对各个数据元的存取权，仅对各个数据元的汇总统计信息有存取权，各数据元的机密性会得到保

证。但是，事实并非如此。统计数据库的安全性不能用存取控制保护机制保证，而是由推断控制保护机制决定。为了解释这个问题，先用一个实例说明。

现在讨论大学生一门课程的考分统计数据库。每项记录有五个属性，即性别（N为女，Y为男）、奖励情况（N为非三好学生，Y为是）、届别（用年代表示）、家庭经济情况（N为无助学金，Y为有）和健康情况（N为一般，Y为健康）。假设五个属性的每种可能序列至多识别一个学生。在给定序列下，也可能不存在符合该描述序列的学生。每个记录还有另外的两个场。一是含有学生的姓名，一是其考试分数（在0和100之间）。此例遵循的原则是把学生考分作为机密信息。现设拥有该数据库的大学教务处，许可用户获得存储于此数据库内数据的统计信息。实现此数据库的安全应采取这样的方式，即至少含有两个登记项的查询才是允许的。在本例中，答案应是所涉及的全部学生考分的总和，否则答案为“无可奉告”。例如，某个用户提出的询问是（*，N，1977，Y，Y），即查询全部学生考分的总和。这些学生是非三好学生、1977年听的课、有助学金并身体健康的。星号是指不问学生的性别。如果有两个学生，答案应是其考分的总和；如果少于两个学生，答案是“无可奉告”。很显然，不可能多于二个学生（一个男生和一个女生），但是完全可以一个没有。因为根据题设用户不能取出单个记录项，现在提出四个查询并得到下列四个答案：

$q1$ （*，Y，1978，N，Y）：142

$q2$ （*，Y，1978，*，Y）：206

$q3$ （Y，Y，1978，*，Y）：134

$q4$ （*，Y，1978，Y，Y）：“无可奉告”

可以看到，这四个询问足以确定几个学生各自的考分，先

看到记录项（N，Y，1978，Y，Y）不存在，因为（Y，Y，1978，*，Y）许可有答案，故（Y，Y，1978，N，Y）和（Y，Y，1978，Y，Y）必然存在，但（*，Y，1978，Y，Y）则无可奉告。为了简化起见，用

$x1$代（N，Y，1978，N，Y）

$x2$代（Y，Y，1978，N，Y）

$x3$代（Y，Y，1978，Y，Y）

此三个记录项均存在。这样，查询项$q1$、$q2$、$q3$可改写成线性方程组：

$$\begin{aligned} x1+x2 &= 142 \quad (\text{由}q1\text{得}) \\ x2+x3 &= 134 \quad (\text{由}q2\text{得}) \\ x1+x2+x3 &= 206 \quad (\text{由}q3\text{得}) \end{aligned}$$

此方程组等效于

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = \begin{bmatrix} 142 \\ 134 \\ 206 \end{bmatrix}$$

因为系数矩阵非奇异，可求得此方程组的解为

$$x1=72$$

$$x2=70$$

$$x3=64$$

由此可知，用（N，Y，1978，N，Y）标识的学生考分为72，（Y，Y，1978，N，Y）的考分为70，（Y，Y，1978，Y，Y）的考分为64。显然，如果不是用已有属性序列决定单值元，自然是不会知道数据库的学生考分的。直观上要求的是少于两个元的查询不予回答。现把此例内1977和1978年学生听课的数据库复制如图1-3，以便验证。

这个简例表明，在统计数据库内利用合法的查询存取权，

姓名	性别	奖励	届别	经济	健康	考分
A	N	Y	77	Y	N	74
B	Y	Y	78	N	Y	70
C	N	N	77	N	Y	86
D	Y	N	77	Y	Y	78
E	N	N	78	Y	N	67
F	Y	Y	78	Y	Y	64
G	N	Y	78	N	Y	72
H	Y	N	77	Y	N	69
I	N	N	77	Y	Y	73
J	Y	Y	77	Y	Y	60

图 1-3　简单统计数据库

可以推断出机密信息．而这些机密信息不能从直接存取中获得。

一般而言，统计数据库的信息状态由两个成分组成。一是存储于该数据库内的数据，一是外部知识。该数据库含有N个事业、企业或个体的属性。如果有M个属性（也可称变量）存在，则每个属性A_j（$1\leqslant j\leqslant M$）有$|A_j|$可能有的值。例如，上例中所说的性别为一个属性，可能有的值为两种（男或女）。设x_{ij}为个体i的属性j的值。如果下标j不起作用，则可以简单地写为x_i，它表示个体i的一个属性A的值。

为了突出统计数据库的特点，把它看成N个记录的集合。每个记录有M个场，而x_{ij}存储于记录i场j内，如图1-4所示。显然，这个图等效于关系数据库内的关系，其中各记录项对应于关系数据库的M个元组。如果存储于数据库内的信息分散到几个关系中，则图1-4所示的关系就对应于这些关系的自然联结。这里假定对全部个体定义场。每个个体有单独的记录。

外部知识涉及到各用户对该数据库占有的知识。外部知识