中国计算机学会
学术著作丛书

马垣 著

非经典关系数据库理论

清华大学出版社

中国计算机学会

CCF

中国计算机学会

学术著作丛书

马 垣著

非经典关系数据库理论

清华大学出版社

北京

内容简介

本书系统介绍了非经典关系数据库理论的国际前沿成果及作者本人的研究成果，内容包括约束关系模型、时态关系模型、空值关系模型、偏序关系模型、概率关系模型、对象关系模型、粗糙关系模型、关系中的说明性更新、关系中的相容与蕴含及关系中的说明性扩充等。

本书是数据库领域的学术专著，可作为高等院校计算机专业研究生或本科高年级学生的教材，也可作为相关专业科技工作者的参考书，对一般数据库的研究及教学有很大的参考价值。

图书在版编目(CIP)数据

非经典关系数据库理论/马垣著．—北京：清华大学出版社，2005.9
(中国计算机学会学术著作丛书)
ISBN 7-302-11181-2

Ⅰ．非…　Ⅱ．马…　Ⅲ．关系数据库　Ⅳ．TP311.138

中国版本图书馆 CIP 数据核字(2005)第 061289 号

出 版 者：清华大学出版社　　地　址：北京清华大学学研大厦
http://www.tup.com.cn　　邮　编：100084
社 总 机：010-62770175　　客户服务：010-62776969
责任编辑：薛　慧
封面设计：常雪影
印 刷 者：北京鑫丰华彩印有限公司
装 订 者：北京鑫海金澳胶印有限公司
发 行 者：新华书店总店北京发行所
开　本：175×245　印张：24.75　字数：504 千字
版　次：2005 年 9 月第 1 版　2005 年 9 月第 1 次印刷
书　号：ISBN 7-302-11181-2/TP·7407
印　数：1～3000
定　价：49.00 元

清华大学出版社计算机学术著作出版基金

评审委员会

中国计算机学会学术著作丛书

序

Preface

1998 年 3 月我曾为马垣的著作《关系数据库理论》作序，那是我在学习和思考后，对关系数据库和其他数据库技术的发展形成的一些认识。时隔 6 年，作者又有了新的著作《非经典关系数据库理论》，这是作者在数据库理论研究领域辛勤耕耘的又一成果。

关系数据库从 1970 年诞生到现在已有 30 多年了。30 多年来，它的应用领域越来越广，影响的范围越来越大。2003 年全世界关系数据库年销售额已达 70 亿美元，根据 IDC 市场研究公司预测，2008 年全世界关系数据库年销售额将达到 200 多亿美元。

关系数据库之所以能如此快速的发展，究其原因就是它从诞生就得到了严格的理论指导与支持。1970 年 6 月，IBM 公司的高级研究员 Edgar Frank Codd 发表的论文“A relational model for large shared data banks”开创了关系数据库的新时代。该论文于 1983 年被 ACM 列为 1958 年以来具有里程碑意义的 25 篇论文之一。1981 年 Codd 获图灵奖。

关系数据库和许多事物一样也在与时俱进。20 世纪 90 年代以来，关系数据库的研究领域及应用领域（例如从

数据库发现知识领域)更加扩大,提出了很多新思想,引出了很多新方向,取得了很多新成果。这些新思想、新方向、新成果全都远远超出了原来的经典研究范围,它们形成了前景非常广阔的“非经典关系数据库”。

目前,国内外集中而系统地论述非经典关系数据库理论的书还非常少,《非经典关系数据库理论》一书是我国的第一部,是作者在其1999年的著作《关系数据库理论》的基础上撰写的新的、全面反映非经典关系数据库理论的学术专著,其中也包括作者本人的成果。书中的内容是目前非经典关系数据库理论的世界前沿,绝大多数内容都是目前国内外相关的书籍中没有介绍过的。例如,该书研讨的“单依赖”集合就是人们在研究不完全关系中函数依赖的可加性时提出的一种全新概念,是在20世纪90年代前国际上未曾研究过的。它的提出彻底解决了函数依赖的可加性问题。而更有意义的是,它还使经典关系数据库中的一些遗留问题有了答案。再如该书介绍的一种“二元关系约束”也是经典关系数据库中函数依赖、多值依赖、连接依赖、等值产生依赖、元组产生依赖等诸多约束都未能包括的全新约束。这种约束在实际应用中的意义不亚于经典关系数据库中的约束。近些年来国内外对具有复杂对象的时态关系进行了很多的研究,各种成果也都超出了经典关系理论的范畴,该书中均做了详细的论述。粗糙关系、偏序关系、概率关系、对象关系等新的关系模型更是层出不穷,该书全面研究了这些关系模型,并对研究成果一一进行了论述。该书内容翔实,视野开阔。显然,它的出版将会与国内外学者的研究成果相辅相成,成为国内同行研究工作中非常有用的武器,书中的最新成果对大家的工作都有很好的参考价值,并具有很好的启发和帮助作用。不论在理论上,还是在实践上,该书的出版都对我国关系数据库的研究向更广、更深的方向发展起到非常重要的作用。

本书作者于20世纪70年代初开始从事编制计算机程序的工作,80年代初开始了对数据库理论的研究,1989年出版了《VAX-11 网状数据库》一书,1999年出版了《关系数据库理论》一书。作者原先在冶金部门工作,是在繁忙的工作之余从事数据库理论研究的。这期间他不仅在工作上获得了许多国家级及省部级奖项,而且在《计算机学报》、《计算机工程》等刊物上发表了许多关系数据库理论的论文。作为业余研究人员,他于1995年获得了国家自然科学基金资助。作者现于鞍山科技大学计算机科学与工程学院担任研究生导师。他从一个业余研究人员成为本领域卓有成就的执著的专业研究人员,这在国内是少见的,他那种锲而不舍、刻苦钻研的精神值得我们敬佩与学习。

我祝贺这本书的出版,并高兴地把这本书推荐给广大读者。祝我国的数据库理论研究人才辈出,祝我国的数据库实际应用欣欣向荣。

中国科学院研究生院　罗晓沛

2004年10月

目 录

Contents

第1章 约束关系模型

在20世纪70年代初，E. F. Codd的一系列论文开创了关系数据库理论。这些理论在关系数据库的应用中发挥了巨大的作用。20世纪后期，它的应用更是扩展到了很多其他领域。应用范围的扩大及数据库自身的发展都使这些早期的理论(一般称为经典关系数据库理论)需要进一步推广与改革。近年来，这些推广与改革取得了丰硕的成果，人们将这些新成果称为非经典关系数据库理论。在本章，我们将首先研究一种非经典的约束关系模型。

1.1 二元约束关系

众所周知，在经典关系理论中，函数依赖、多值依赖、连接依赖、元组产生依赖、等值产生依赖等起到了重要的作用，而实际上，它们都是加在数据库上的一些约束。另一方面，查询中数据应满足的条件实际也是加在查询结果上的约束。随着关系数据库使用的深入，人们发现还有一种非常广泛地出现在各种实际问题中的、既可以加在原始数据库上又可以加在查询结果上的"二元关系约束"。这种二元关系约束是在两个属性的值域间给定了一个(些)二元关系，每个元组在这两个属性上都只能取这个(些)二元关系中的值。例如，在一个毕

业生档案的数据库中，由于毕业生就业是双向选择，所以在属性“姓名”及属性“就业单位”的值域间就存在着两个二元关系约束，一个由学生填写的“志愿表”组成，另一个由各单位填写的“可接收学生表”组成。分别如以下两表所示：

姓名	单　位
张强	中心医院
张强	康复医院
张强	人民医院
李晓	人民医院
李晓	铁路医院
⋮	

单　位	姓名
中心医院	王丽
中心医院	赵晶
中心医院	张强
中心医院	孙胜
邮电医院	赵晶
⋮	

再如在一个 α 无回路数据库 $\Omega=\{R_1,R_2,R_3,R_4\}$ 中，$R_1=\{$SUPPLIER，COST，PART$\}$，$R_2=\{$PROJECT，COUNT，PART$\}$，$R_3=\{$PROJECT，DATE，SUPPLIER$\}$，$R_4=\{$PROJECT，SUPPLIER，PART$\}$。这里，在 SUPPLIER 与 PART 的数据域之间可能存在某个供应商只供应某些部件的二元约束关系（由产品手册组成）。在 PROJECT 与 PART 的数据域之间可能存在某个工程需要某些部件的二元约束关系（由工程设计清单组成）。在 SUPPLIER 与 PROJECT 的数据域之间可能存在某个供应商愿意向某些工程供应部件的二元约束关系，以及 PROJECT 与 SUPPLIER 的数据域之间可能存在某个工程希望从某些供应商那里购买部件的二元约束关系。再如有时还可能是一些抽象的二元约束关系，例如在库存管理的数据库中，在属性“生产日期”与属性“销售日期”的数据域之间就有“小于”这样一个约束关系。这个二元关系虽然不需要像前面给出的“姓名”与“单位”之间的二元关系那样全部写出来，但这个二元关系确实是存在的。在两个表示时间区段的属性的数据域之间还可能有图 1.1 所示的多种二元约束关系。

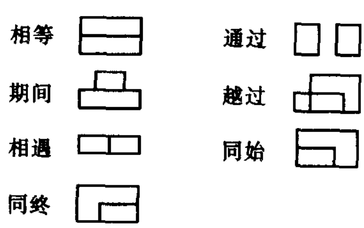

图 1.1　时间区段的二元关系

这些二元关系的约束是普遍存在的，但却是经典关系数据库理论中的各种“依赖”都未涉及的。因此为了更好、更优化地对存在着二元关系约束的数据库进行查询、输入、修改等操作，就必须对二元约束关系集合的包含问题、等价问题、极小化问题及其他各种有关问题也认真地进行研究。这里，包含问题是指如何判定一个

二元关系约束集合 s_1 是否包含另一个二元关系约束的集合 s_2，即如何判定是否适合 s_1 的元组也都适合 s_2；等价问题是指如何判定 s_1 与 s_2 等价，即判定是否适合 s_1 的元组也适合 s_2，而且适合 s_2 的元组也适合 s_1；极小化问题是指如何将给定的二元关系约束的集合 s 极小化，即如何找出包含非平凡二元约束关系最少的与 s 等价的二元关系约束的集合。只有解决了这些问题，特别是解决了极小化问题，才能优化我们的操作。而解决这些问题的基础又是解决怎样才能知道适合二元关系约束集合的全部元组的问题。下面先给出有关概念的形式化定义，然后研讨解决这一问题的数学工具，并解决这一问题，最后给出各种应用实例。

1.1.1 形式化定义

定义 1.1 设 $R = \{A_1, \cdots, A_n\}$ 是一个关系模式，属性 A_i 的数据域是 $dom(A_i)(i = 1, \cdots, n)$。$r_{ij} \subseteq dom(A_i) \times dom(A_j)$ 是属性 A_i 与属性 A_j 数据域间的二元约束关系，它所产生的约束称为二元关系约束，$1 \leqslant i, j \leqslant n$。若 C 是 R 上的二元约束关系的集合，$\langle a_1, \cdots, a_n \rangle$ 是 R 上的元组，当且仅当对 C 中的每一个二元约束关系 r_{ij} 都有 $\langle a_i, a_j \rangle \in r_{ij}$ 时称 $\langle a_1, \cdots, a_n \rangle$ 适合 C。称(R,C) 为二元约束关系模型，简称约束关系模型。属性域的并集 $U = dom(A_1) \cup \cdots \cup dom(A_n)$ 称为约束关系模型的基域。基域上所有适合 C 的元组的集合称为约束 C 的解集。若 C_1, C_2 是 R 上两个二元约束关系集合，C_1 的解集是 C_2 解集的子集，则称 C_2 包含 C_1。若 C_1 与 C_2 的解集相同，则称 C_1 与 C_2 等价。 □

我们注意到，在定义 1.1 中并没有要求每一对 i, j 在属性 A_i 与 A_j 的数据域间都有约束 r_{ij}。不过，为了形式上的统一，可以认为在没有约束的属性 A_i 与 A_j 的数据域间存在着二元约束关系 $r_{ij} = dom(A_i) \times dom(A_j)$，并称它是平凡的。

另外还注意到，定义 1.1 中也没有明确限制属性 A_i, A_j 数据域间二元约束关系的个数。然而如果在属性 A_i, A_j 的数据域间存在着多个二元约束关系，例如，$r_{ij}^{(1)}, \cdots, r_{ij}^{(k)}$，那么 C 的解集中的任何一个解 $\langle a_1, \cdots, a_n \rangle$ 就都会满足：$\langle a_i, a_j \rangle \in r_{ij}^{(1)}$，$\langle a_i, a_j \rangle \in r_{ij}^{(2)}, \cdots, \langle a_i, a_j \rangle \in r_{ij}^{(k)}$，于是可用 $r_{ij} = r_{ij}^{(1)} \cap \cdots \cap r_{ij}^{(k)}$ 来代替 $r_{ij}^{(1)}, \cdots, r_{ij}^{(k)}$。显然，代替后的约束集合与原来的约束集合是等价的。

最后还可以认为，在属性 A_i 与 A_i 的数据域间存在着恒等二元关系的约束，即存在着：$r_{ii} = \{\langle a, a \rangle \mid a \in dom(A_i)\}$。这样，从现在起我们就规定在关系模式 $R = \{A_1, \cdots, A_n\}$ 上的二元约束关系的集合 C 中，对每一对 $i, j, 1 \leqslant i, j \leqslant n$，都有而且只有一个二元约束关系 r_{ij}。

下面是一个简单的约束关系模型的例子。

$R = \{A_1, A_2, A_3\}, dom(A_1) = dom(A_2) = dom(A_3) = \{a, b, c, d\}, C = \{r_{11}, r_{12}, r_{13}, r_{21}, r_{22}, r_{23}, r_{31}, r_{32}, r_{33}\}$，

其中：$r_{11} = r_{22} = r_{33} = \{\langle a, a \rangle, \langle b, b \rangle, \langle c, c \rangle, \langle d, d \rangle\}$，

$r_{12} = \{\langle a,b\rangle,\langle b,c\rangle,\langle c,d\rangle\}$,

$r_{13} = \{\langle a,b\rangle,\langle b,a\rangle,\langle b,d\rangle,\langle d,b\rangle\}$,

$r_{21} = \{\langle a,a\rangle,\langle c,b\rangle\}$,

$r_{23} = \{\langle c,d\rangle,\langle d,c\rangle\}$,

$r_{31} = r_{13}$,

$r_{32} = r_{23}$。

这里,C 的解集是$\{\langle b,c,d\rangle\}$。显然,在这种简单的情况下,通过经典关系理论中的自然连接

$$r_{11} \bowtie r_{12} \bowtie r_{13} \bowtie r_{21} \bowtie r_{22} \bowtie r_{23} \bowtie r_{31} \bowtie r_{32} \bowtie r_{33}$$

即可求出 C 的解集,但当数据域中的元素是无限的,而二元约束关系中包括大于、小于、期间等抽象关系时单纯自然连接就不够了,就需要对经典关系理论进行推广与改革。

1.1.2 约束关系运算

二元约束关系作为关系显然也可以对它进行关系运算。本章所关心的运算有4个。

设 r,s 是两个二元约束关系,则

(1) $r \cap s = \{\langle a,b\rangle \mid \langle a,b\rangle \in r$ 且$\langle a,b\rangle \in s\}$;

(2) $r \cup s = \{\langle a,b\rangle \mid \langle a,b\rangle \in r$ 或$\langle a,b\rangle \in s\}$;

(3) $r : s = \{\langle a,b\rangle \mid (\exists c)(\langle a,c\rangle \in r$ 且$\langle c,b\rangle \in s)\}$;

(4) $\tilde{r} = \{\langle a,b\rangle \mid \langle b,a\rangle \in r\}$。

这里(:)是复合运算,(~)是逆运算。

在同一层小括号中,先算~,再算 :,然后算 $\cap$,最后算 $\cup$。同层小括号中的同种运算是先左后右,例如:

$$s : r : t = (s : r) : t$$

3个特殊的二元约束关系分别用以下符号代表:① 恒等关系:$\{\langle a,b\rangle \mid a=b\}$ 用 Id 表示,② 相异关系:$\{\langle a,b\rangle \mid a \neq b\}$ 用 Di 表示,③ 空关系用 $\varnothing$ 表示。另外定义二元约束关系 r_{ij} 的左域是$\{a \mid (\exists b)(\langle a,b\rangle \in r_{ij})\}$,右域是$\{b \mid (\exists a)(\langle a,b\rangle \in r_{ij})\}$。二元约束关系 r_{ij} 的场是其左域与右域的并集。

1.1.3 约束矩阵

定义 1.2 设(R,C)是一个约束关系模型,其中 $C=\{r_{ij} \mid 0 \leqslant i,j \leqslant n-1\}$,则由 $M_{ij} = r_{ij}, 0 \leqslant i,j \leqslant n-1$,定义的 $n \times n$ 矩阵M 为该模型所对应的约束矩阵,并称 C 的解集也为 M 的解集。 □

显然每个约束矩阵的主对角线元素全是恒等关系,反之,每个主对角线元素全

是恒等关系(即 $M_{ii} \subseteq \mathrm{Id}$)的以二元关系为元素的 $n \times n$ 矩阵 M,一定都对应一个约束关系模型 (R,C),其中 R 包含 n 个属性。

为了与很多文献中的记号相统一,这里统一记 R 为 $\{A_0,\cdots,A_{n-1}\}$,M 的元素为 M_{ij},$0 \leqslant i,j \leqslant n-1$。

如果 M 中有一个元素 M_{ij} 是 $\varnothing$,则称 M 无解。显然,这是 M 无解的充分条件,但不是必要条件,因为有时每个 M_{ij} 全不空,但也不存在 $\langle a_0,\cdots,a_{n-1}\rangle$ 能使所有 $0 \leqslant i,j \leqslant n$ 均有 $\langle a_i,a_j\rangle \in M_{ij}$。

1.1.4 约束矩阵的运算

设 M 与 N 是两个 $n \times n$ 约束矩阵,定义约束矩阵的 3 种运算如下:

(1) 逆运算:$\widetilde{M}$

$$(\widetilde{M})_{ij} = (M_{ij})^{\sim}$$

(2) 交运算:$N \cdot M$

$$(N \cdot M)_{ij} = N_{ij} \cap M_{ij}$$

(3) 积运算:$N : M$

$$\begin{aligned}(N : M)_{ij} &= N_{i0} : M_{0j} \cap \cdots \cap N_{i(n-1)} : M_{(n-1)j} \\ &= \bigcap_{k<n} N_{ik} : M_{kj}\end{aligned}$$

积运算也称复合运算。

应注意到约束矩阵的积运算不满足结合律,请看下例:

$$\text{设}\quad M = \begin{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{pmatrix} & (0 \quad 1) \\ \varnothing & \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{pmatrix} \end{pmatrix}, \qquad N = \begin{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{pmatrix} & (0 \quad 2) \\ \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{pmatrix} \end{pmatrix}$$

$$\text{于是}(M : N)_{00} = \bigcap_{k<2} (M_{0k} : N_{k0}) = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{pmatrix} \cap (0 \quad 0) = (0 \quad 0)$$

$$(M : N)_{01} = \bigcap_{k<2} (M_{0k} : N_{k1}) = (0 \quad 2) \cap (0 \quad 1) = \varnothing$$

$$(M : N)_{10} = \bigcap_{k<2} (M_{1k} : N_{k0}) = \varnothing \cap \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix} = \varnothing$$

$$(M : N)_{11} = \bigcap_{k<2} (M_{1k} : N_{k1}) = \varnothing \cap \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{pmatrix} = \varnothing$$