

● 高等学校试用教材

数据库理论

及新领域

施伯乐 楼荣生
崔靖 蒋家福



高等教育出版社

高等学校试用教材

数据库理论及新领域

施伯乐 楼荣生

崔靖 蒋家福

高等教育出版社

内 容 提 要

本书对关系数据库的理论进行了系统地概括和总结,并介绍了当前数据库发展的几个新领域.书中汇集了大量当今数据库理论和技术的最新研究成果.主要内容有四个方面:(一)关系数据库理论,包括依赖理论、数据库设计、符号表、跟踪算法、表示理论、无环数据库等;(二)分布式数据库,包括查询优化、并发控制、可靠性;(三)知识库,包括查询语言的实现、查询求值过程的优化;(四)面向对象数据库,包括数据库新领域的需求、面向对象的概念和特点、几个典型的 OODBMS、可扩充数据库的途径及可扩充的方面、几个可扩充的 DBMS.本书内容丰富新颖,结构安排合理,阐述严密清楚,许多章节配有习题.

本书可作为高等院校计算机专业本科高年级学生、研究生的教材或参考书,也可供从事数据库的理论研究和实际应用的科技人员学习参考.

高等学校试用教材

数据库理论及新领域

施伯乐 楼荣生

崔 靖 蒋家福

*

高等教育出版社出版

高等教育出版社照排中心照排

新华书店总店北京科技发行所发行

化学工业出版社印刷厂印装

*

开本 850×1168 1/32 印张 11.875 字数 300 000

1990 年 9 月第 1 版 1990 年 9 月第 1 次印刷

印数 0001—1 610

ISBN 7-04-003135-3/TP·71

定价 3.05 元

前 言

数据库技术从 60 年代末至今已有 20 年的历史了, 20 多年来它吸引了许多专家和学者, 他们一方面从事数据库管理系统 (DBMS) 的实际研制, 一方面从事数据库的理论与算法的研究。

传统数据库的三种模型, 以关系数据库见长。由于它提供的概念简单, 使用方便, 并且建立在丰富的理论基础上, 使关系数据库模型能得以形式化地描述和处理。因此, 近年来有关关系 DBMS 的研制日益增多, 世界上的各种大型计算机系统 (如 IBM, Vax 等) 均配置了关系 DBMS。与此同时, 有关关系数据库设计的理论也日益丰富, 包括各种数据依赖的表示、关系数据库的各种规范形式、有关的算法以及数据库模型的性质, 涉及到集合、关系、超图等许多方面。本书的第二章至第九章试图介绍关系数据库的有关理论, 其主要内容有: 依赖理论、设计算法、符号表、表示理论、跟踪算法及无环数据库等。

由于实际需要和计算机网络的广泛应用, 分布式数据库的研制也随之出现。目前已研制出了较成功的分布式系统, 如 SDD-1 和 R* 等, 商用的系统也已投入使用。不少系统都是建立在关系数据库之上的。分布式数据库涉及并行控制、查询优化和可靠性等较难处理的内容。本书结合实践从理论上介绍半连接处理、事务可串行化、调度器的算法、多版本控制及分布式系统的确认协议等有关理论。书中的第十章至第十二章介绍这方面的内容。

近几年来, 由于人工智能技术和数据库技术相结合, 提出了知识库、专家系统的研制和有关理论。在日本、美国、欧洲等国家许多专家都认为, 新一代计算机的核心之一是知识库, 因此出现了研究知识库的热潮。本书将对逻辑数据库查询语言的实现、查询求值过程的优化等核心问题加以介绍, 内容包括: 规则/目标图、

基本俘获规则、递归俘获规则、边侧信息传递、魔集法、计数法等。目的是使读者对当前知识库中主要研究的问题有所了解,这些内容将在第十三章和第十四章中介绍。

随着计算机技术和计算机应用的发展,对数据库技术也提出了新的要求,如 CIMS, CASE, OIS 等领域提出了传统数据库难以表示的要求,因此专家们希望数据库技术有所突破,这就出现了面向对象的数据库研究热潮。尽管这方面的研究还很不成熟,也缺乏形式的描述,但看来是一个新的重要方向。本书将介绍对象的概念、定义、操纵和特点、几个面向对象的数据库系统以及面向对象数据库的研究领域;与此同时还介绍了可扩充数据库系统,包括它的扩充功能、方式、可扩充的方面以及几个可扩充的 DBMS。由于这方面的理论研究还不成熟,大部分是内容罗列,因此篇幅较长。本书的第十五章和第十六章介绍这些内容。

由于本书涉及的面较广,新领域的内容也较多,因此参加协助编写的同志也较多,有复旦大学王雄、陈昉、柳守中、李心宇、马学强、梁兼栋等同志,华东师范大学李玉茜教授、上海交通大学白英彩教授对书稿进行了认真的审阅,提出了不少宝贵的意见。对以上同志的大力支持和帮助在此表示衷心的感谢。由于编者水平有限,书中一定有很多不足之处,望读者指正。

作 者

1990.2

目 录

第一章 数据库概述	(1)
第二章 关系和关系模式	(6)
§2.1 实例	(6)
§2.2 关系的形式定义	(8)
§2.3 键(Key)	(9)
§2.4 关系的更新	(10)
习题	(12)
第三章 关系操作	(15)
§3.1 布尔操作	(15)
§3.2 选择、投影和连接操作	(18)
§3.3 商操作	(26)
§3.4 常关系和更名属性	(28)
§3.5 等值连接操作和扩充操作功能	(30)
§3.6 关系代数	(34)
习题	(36)
第四章 无损分解	(38)
§4.1 问题提出——关系分解	(38)
§4.2 投影与自然连接的关系	(40)
§4.3 函数依赖 (Functional Dependence)	(42)
§4.4 多值依赖 (Multivalued Dependence)	(43)
§4.5 连接依赖 (Join Dependence)	(45)
习题	(49)
第五章 依赖理论	(52)
§5.1 带依赖集的关系模式	(52)
§5.2 函数依赖公理	(53)
§5.3 多值依赖公理	(57)
§5.4 依赖基	(60)

§5.5 多值依赖公理的完备性定理	(63)
§5.6 嵌入连接依赖	(65)
习题	(66)
第六章 关系数据库范式	(68)
§6.1 只与函数依赖有关的关系数据库范式	(68)
§6.2 分解式规范化算法	(72)
§6.3 保持依赖的无损分解	(74)
§6.4 依赖集的等价变换 (一):无冗余化和既约化	(76)
§6.5 依赖集的等价变换 (二):最小化	(78)
§6.6 保持依赖的无损分解的完整算法	(83)
§6.7 第四范式	(86)
§6.8 第五范式	(87)
习题	(88)
第七章 符号表和跟踪算法	(90)
§7.1 投影连接映照	(90)
§7.2 符号表	(92)
§7.3 符号表等价性与模式等价性	(94)
§7.4 带限制的符号表等价性	(98)
§7.5 跟踪算法	(102)
§7.6 检验隐含依赖	(104)
§7.7 跟踪算法的计算复杂度	(111)
习题	(112)
第八章 表示理论	(115)
§8.1 合理表示的概念	(115)
§8.2 数据库模式的数据等价性	(123)
§8.3 测试合理表示	(126)
§8.4 测试数据等价性	(132)
习题	(135)
第九章 无环数据库模式	(137)
§9.1 数据库模式的性质	(137)
9.1.1 完全归约的存在性	(137)

9.1.2	连接依赖与多值依赖间的等价性	(140)
9.1.3	唯一4NF分解	(141)
9.1.4	成对一致性蕴涵完全一致性	(142)
9.1.5	较小的中间连接	(143)
§9.2	数据库模式的语法条件	(145)
9.2.1	无环超图	(145)
9.2.2	连接树	(150)
9.2.3	连续交集性质	(152)
§9.3	各种条件的等价性	(152)
9.3.1	Graham 归约	(152)
9.3.2	找出连接树	(154)
9.3.3	无环数据库模式的等价性定理	(156)
	参考文献	(157)
第十章	分布式数据库的查询处理	(158)
§10.1	记号和定义	(158)
§10.2	相等连接查询	(160)
§10.3	半连接归约对查询的解	(161)
10.3.1	归约 (Reduction)	(161)
10.3.2	树查询	(162)
10.3.3	环查询	(168)
§10.4	树类查询的成员测试算法	(169)
§10.5	半连接在分布式数据库中应用的基本概念	(174)
10.5.1	数据库略图 (Profile) 和对半连接运算的估计	(175)
10.5.2	把半连接用于连接查询	(176)
§10.6	确定半连接程序的SDD-1算法	(179)
10.6.1	基本SDD-1算法	(179)
10.6.2	事后优化	(184)
§10.7	确定半连接程序的AHY (Apers, Hevner, Yao) 算法	(185)
10.7.1	简单查询	(187)
10.7.2	调度的综合	(191)
	参考文献	(193)

第十一章 并行控制 (Concurrency Control)	(194)
§11.1 概念与记号	(195)
§11.2 正确性理论	(197)
11.2.1 终态可串行性 (Final-State Serializability)	(197)
11.2.2 视图可串行性 (View Serializability)	(199)
11.2.3 冲突可串行性 (Conflict Serializability)	(200)
§11.3 调度器算法	(202)
11.3.1 封锁方法 (Locking)	(202)
11.3.2 时间印方法 (Timestamp)	(204)
11.3.3 冲突图方法	(205)
11.3.4 分布式调度器	(206)
§11.4 多版本并行控制 (Multiversion Concurrency Control)	(207)
11.4.1 多版本系统的正确性	(208)
11.4.2 多版本调度器的能力	(210)
11.4.3 多版本调度器算法	(213)
参考文献	(214)
第十二章 可靠性 (Reliability)	(216)
§12.1 集中式系统的可靠性理论	(216)
12.1.1 强化的调度器	(217)
12.1.2 数据管理器 (Data Manager)	(222)
§12.2 分布式系统的确认协议 (Commit Protocol)	(226)
12.2.1 原子性确认协议 (Atomic Commitment Protocol)	(227)
12.2.2 两段确认协议 (Two-phase Commit Protocol)	(229)
12.2.3 三段确认协议 (Three-phase Commit Protocol)	(232)
参考文献	(236)
第十三章 逻辑数据库查询语言的实现	(238)
§13.1 逻辑数据库的查询	(239)
§13.2 规则/目标图和基本俘获规则	(249)
§13.3 基于递归的俘获规则	(255)
参考文献	(259)
第十四章 逻辑数据库查询求值过程的优化	(260)
§14.1 引言	(260)

§ 14.2 边侧信息传递	(262)
§ 14.3 装饰规则集	(267)
§ 14.4 魔集法 (Magic Set Method)	(270)
§ 14.5 计数方法 (Counting Method)	(277)
参考文献	(281)
第十五章 面向对象数据库系统	(282)
§ 15.1 新的应用领域及其对数据库技术的要求	(282)
15.1.1 新的应用领域	(282)
15.1.2 新的应用领域对数据库技术的要求	(284)
§ 15.2 面向对象技术与数据库技术的结合	(286)
15.2.1 面向对象的概念与特点	(287)
15.2.2 面向对象的数据库管理系统 (OODBMS)	(297)
§ 15.3 几个典型的OODBMS	(300)
15.3.1 VBASE	(300)
15.3.2 GemStone	(306)
15.3.3 Iris	(311)
15.3.4 ORION	(317)
§ 15.4 与OODB有关的研究领域	(326)
15.4.1 关系数据库系统	(326)
15.4.2 语义数据模型	(327)
15.4.3 嵌套关系 (Nested Relations)	(327)
15.4.4 数据库程序设计语言与持久性程序设计语言	(327)
15.4.5 可扩充的数据库系统	(327)
参考文献	(328)
第十六章 可扩充的数据库系统	(329)
§ 16.1 实现可扩充性的途径	(330)
16.1.1 扩充DBMS功能的三种方式	(330)
16.1.2 构造可扩充的DBMS的两种方式	(331)
§ 16.2 可扩充的各个方面	(333)
16.2.1 面向应用的扩充	(333)
16.2.2 面向系统的扩充	(336)
§ 16.3 几个可扩充的DBMS	(339)

16.3.1	POSTGRES	(340)
16.3.2	Exodus	(351)
16.3.3	GENESIS	(357)
16.3.4	其他可扩充的DBMS 原型系统	(361)
§16.4	可扩充的数据库与面向对象的数据库的比较.....	(366)
参考文献		(367)

第一章 数据库概述

在计算机数据管理的历史上,曾出现过三次飞跃,第一次是数据库技术的出现,它使数据管理技术步入了一个新的时代,第二次是关系数据模型的诞生,它标志着数据库技术走向成熟,随着计算机科学技术的进一步发展,数据库技术和网络、人工智能、软件工程、自然语言等相结合,不断扩展形成新的发展方向,出现第三次飞跃,特别在知识库、面向对象数据库等方面,在未来的几年将有所突破。

数据库技术起源于计算机主要应用领域转变的时期,当计算机的主要应用领域从科学计算逐步转变到事务数据处理时,数据库技术便应运而生,经过一大批数据库专家十余年的不懈努力,数据库领域在理论和实践上都取得了令人瞩目的成就,成为计算机科学的主要领域之一。

计算机科学研究的一个主要问题就是怎样有效地利用计算机来解决实际问题,一个有效的计算机系统,应当能辅助我们更迅速、准确地分析和改造现实世界(图1.1)。

如果说人工智能是研究问题求解技术的话,那么数据库技术就是研究怎样模拟现实世界。

如图1.2,数据库系统有两大主要功能,一是用一个数据世界(即数据库)去合理、有效地表达现实世界的某一部分,这就是所谓的数据库设计问题,数据库系统的另一个功能是,将数据世界映照成一个意识世界(即用户界面),意识世界是数据世界以一定的模型提供给用户的界面,在层次型、网型、关系型和逻辑数据模型的数据库中,用户的意识世界分别是以树、图、表和规则的形式体现出来的。

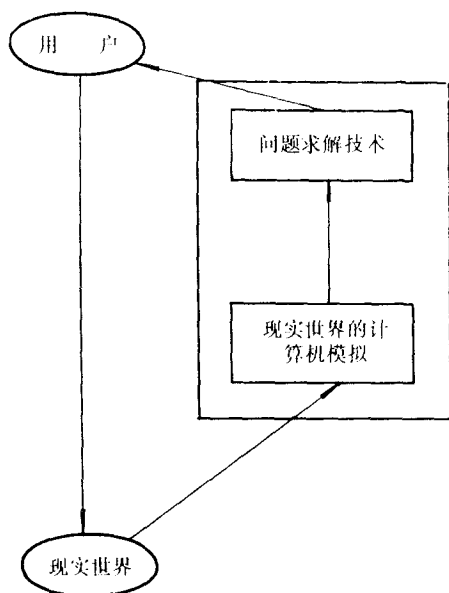


图1.1

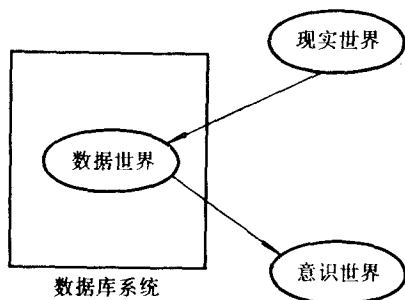


图1.2

意识世界是现实世界的再现,是一个虚拟的世界,它必须准确、如实地反映现实世界中的信息,数据世界是现实世界的浓缩,它被存储在计算机内,作为从现实世界到意识世界映象过程中的中

间媒体。

现实世界是复杂的,它可看成由许多有着相互联系的不同实体组成。实体间的联系大致可分为一对一、一对多和多对多三种。同时,现实世界又是在不断变化的,因此,数据库系统除了要准确模拟现实世界外,还必须要模拟现实世界中的变化。

数据库技术的研究主要分成下面四个不同侧面:

(1) 开发能表达丰富语义且能高效处理的新的数据模型(如逻辑数据模型、面向对象数据模型等)。

(2) 提供友好的、非过程性的高级用户接口。

(3) 复杂的多介质信息(如正文、图象、声音等)的存储;

(4) 事务的并发控制技术。

数据模型是表示实体间相互联系的模式。常见的数据模型有三种:层次模型、网络模型和关系模型。由图 1.2 可以看出,数据模型有两方面的涵义:一是表示在数据库中以什么形式来存储数据,称为存储模型;二是表示用户以什么形式来看待数据库中的数据,称为用户模型。这是两个不同的问题。例如,采用 Horn 子句模型的知识库系统,其内部存储可以采用关系模型。

关系模型自 1970 年由美国 IBM 公司的 E. F. Codd 提出后,即以其简明的结构(在关系模型中,只有唯一的一种数据结构——关系)和严密的理论基础,吸引了一大批优秀的计算机专家(如 Ullman, Maier, Berstein, Beerl 和 Armstrong 等)投身于关系理论的研究,并取得了惊人的成就。

整个 70 年代以及 80 年代初期,关系数据库的研究得到了突飞猛进的发展。有关关系数据库的新理论与技术不断涌现,如关系代数、关系演算理论、依赖理论、规范理论、泛关系理论、无环超图理论、查询优化技术、符号表跟踪技术、空值以及视图修改问题等。J. Ullman 的《数据库系统原理》和 D. Maier 的《关系数据库理论》这两本书的出版,标志着关系数据库理论已走向成熟。70 年代被称为是“数据库的时代”,E. F. Codd 本人也于 1980 年获

得计算机科学最高奖——ACM 图灵奖。

关系数据模型的最大优点是用户接口简洁、友好,并具有牢固的理论基础,关系模型能提供非过程性的、系统导航的关系查询语言,还允许一次存取多个记录,同时,由于查询优化技术的发展,特别是对高效连接算法的研究,使得关系系统的效率不断提高,出现了许多关系模型的数据库商用产品。

当然,关系模型也有不少不足之处,如语义表达能力差,不能直接表达实体概念,没有推理能力等等,随着计算机的广泛应用,用户对 CAD、CASE、OA、AI 的数据处理提出了许多新的要求,关系数据模型显得有些力不从心,为此,近几年来数据库研究已逐渐转向知识库和面向对象数据库等方面,尽管如此,关系数据库仍活跃在数据库领域中,目前,世界上约有上百万个关系数据库系统在运行,人工智能专家也已将关系表示列为一种有效的知识表示方法,值得一提的是,现在对知识库的研究也大多建立在关系数据库系统的基础上,因此,对关系数据库理论进行归纳总结,是一件很有意义的工作,它必将有益于对数据库和面向对象数据库等技术的研究,这也就是本书的宗旨所在。

知识库是当前数据库领域的热门话题,知识库由事实和规则组成,知识库系统的一个显著特点是它具有较强的推理机制,能管理大量的事实和规则。

目前,知识库系统的研究重点主要在查询处理(包括递归、否定、集合等机制的处理)上,递归查询常采用优化方法,将选择、投影操作下推,如魔集法、计数法等,否定的经典处理方法是采用所谓的“封闭世界假设”策略,集合的处理一般采用把带“集合项的规则集”重写为一般的规则集的方法,当前,世界各国的一些研究机构,如美国的 MCC,日本的 ICOT,英国的 Alvey 和欧洲共同体的 Esprit 等,都在致力于与知识库有关的研究。

关系模型和 Horn 子句模型都是面向值的,它们有产生各种异常的隐患,因此,近年来人们正致力于面向对象的数据数据库系统的

研制。它将面向对象系统中的一些先进的思想,如对象的封闭性、带类型、继承、共享、可扩充性、延迟连接等,引入数据库系统中,以解决面向值系统中的许多问题。尽管面向对象数据库的发展目前正处于初始阶段,还缺乏一种公共模型和一套形式化的理论基础,其研究工作仍偏重于实验性,但无可否认,这是一个很有前途和生命力的数据库发展方向。

第二章 关系和关系模式

用关系来描述实体以及实体间的联系,其优点是具有一致性.在直观上把关系看成表格的形式,这是十分有益的.因为表格结构简单,能简明地表达内容.对表格中内容的插入、删除、修改以及多个表格的组合也十分方便.用户、应用程序员和数据库管理人员都易于掌握和使用.关系型数据库主要采用表格数据来表示和实现实体间的关系.尽管关系数据库实现效率(占用空间和花费时间)低于其他类型数据库,但目前仍有不少数据库专家从事关系优化理论和技术的研究,以此来提高它的效率.总的来说,关系型数据库结构简单、理论严密且使用方便,目前已得到广泛使用.本章通过实例来建立关系型数据库的基本概念:关系模式、关系、键,说明它们的形式定义和内涵,以及关系更新的一些操作.

§2.1 实 例

例 2.1.1 在一张飞机时刻表里,用编号(NUMBER)表示飞机航班,且标明起点(FROM)、终点(To)、起飞时间(DEPARTS)和到达时间(ARRIVES).这些航班(实体集)的信息可用关系来表示.这个关系可看成表 2.1.1,它是一个由行和列组成的二维表.

用来描述实体或实体间联系(简称联系)的表,应该具有如下的特点:

1. 列表示实体或联系的某个属性.
2. 每列指定一个彼此相异的名字,称为属性名,简称属性.
3. 列(行)的相对次序允许改变.