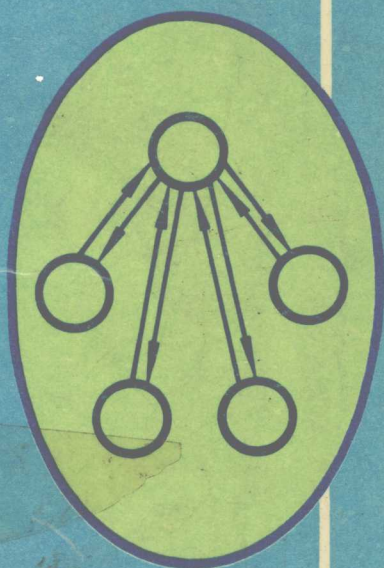


数据库原理与设计



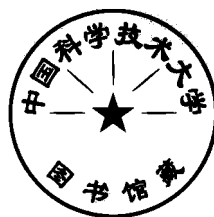
张	龙	祥
黄	正	瑞
陶		影

67

学出版社

数据库原理与设计

张龙祥 黄正瑞 陶 影 编著



西南交通大学出版社

(川)新登字 018 号

内 容 提 要

本书较全面地介绍了数据库的基本构造、关系数据模型及语义模型、SQL 语言、关系数据库设计理论基础、数据库管理系统、数据库设计、数据库的完整性与安全性、网状数据库系统,以及当前数据库的新发展简况。本书的内容以当前流行的关系数据库技术为中心,以标准数据库语言 SQL 为重点,充分反映当前数据库技术的主流,同时,兼顾已广泛使用的网状数据库系统。本书强调理论与实践紧密结合,数据库原理与设计并重,不但包含了必要的数据库的理论基础知识,而且较详细地叙述了数据库设计的方法和过程。本书叙述严谨,循序渐进,并备有适量的例题和习题。

本书可供高等院校计算机专业、管理信息专业高年级学生及研究生作为教材,也可供从事数据库设计与应用及数据处理等方面工作的人员使用。

数据库原理与设计

张龙祥 黄正瑞 陶影 编著

*

西南交通大学出版社出版发行
(成都 二环路北一段)

新华书店经销

成都飞机工业公司印刷厂印刷

*

开本: 787×1092 1/16 印张: 15.8125

字数: 369 千字 印数: 1—2000 册

1995 年 4 月第 1 版 1995 年 4 月第 1 次印刷

ISBN 7—81022—785—8/T·138

定价: 12.65 元

前 言

数据库技术已在当代的社会生活中获得了广泛的应用。为了适应高校数据库教学的需要,铁道部高教司教学指导委员会于1990年决定编写本教材。

本书内容按教学大纲要求组织,以当前流行的关系数据库技术为中心,以标准数据库语言SQL为重点,充分反映当前数据库技术的主流,同时兼顾铁路部门已广泛使用的网状数据库系统。本书强调理论与实践紧密结合,数据库原理与设计并重,学以致用。

本书共十章,材料的组织上逻辑隐含为三大部分,第一至四章构成数据库系统的基础知识与应用篇,第五至八章构成数据库设计篇,第九、十章介绍非关系数据库系统。在教学中可以根据需要灵活取舍本书的内容。

本书可供高等院校计算机专业、信息管理专业的高年级学生及研究生作教材,也可供从事数据库设计与应用及数据处理等方面工作的人员使用和参考。

本书由张龙祥任主编。本书初稿的第一、二、四章由陶影编写,第三、六、八、十章由张龙祥编写,第五、七、九章由黄正瑞编写。在内部使用的基础上,根据审稿专家的意见,由张龙祥对第一、二、三、六、八、十章,黄正瑞对第四、五、七、九章做了仔细的修订。

本书由西南交通大学周志成教授任主审。周志成教授和上海铁道学院胡谋教授、国防科技大学郑若忠教授、中南工业大学刘劭天副教授、长沙铁道学院胡湘陵教授、华东交通大学李巍讲师仔细地审阅了书稿,提出了许多宝贵的改进意见。北方交通大学刘建国教授曾提供一些有关SQL的参考资料。在此特向他们表示衷心的感谢。

由于编者的水平有限,书中谬误在所难免,望读者不吝指正。

作者

1994年5月

目 录

第一章 基础知识	1
第一节 数据库系统	1
第二节 数据模型	6
第三节 数据库系统的构造	13
第四节 数据库管理系统	17
第二章 关系数据模型	23
第一节 概念与术语	23
第二节 关系代数	26
第三节 关系演算	31
第四节 关系运算的安全限制	33
第五节 关系代数表达式的优化	36
第三章 关系数据库语言 SQL	43
第一节 SQL 语言组成	43
第二节 数据库定义	47
第三节 数据库查询	51
第四节 数据操纵	61
第五节 数据控制	63
第六节 嵌入式 SQL	65
第四章 数据库管理系统	73
第一节 数据库管理系统和操作系统的相互关系	73
第二节 数据库管理系统的层次结构	74
第三节 数据存贮管理	78
第四节 数据存取管理	81
第五节 并发控制	90
第六节 关系运算的实现算法	99
第七节 关系数据库管理系统 ORACLE	101
第五章 关系数据库设计理论	111
第一节 关系模式及其评价	111
第二节 函数依赖	113
第三节 函数依赖公理体系	115
第四节 关系模式的分解	120
第五节 关系模式的规范化	125
第六节 多值依赖和第四范式	132
第七节 联结依赖和第五范式	136
第六章 语义模型	141
第一节 概述	141

第二节	实体—联系模型	142
第三节	实体—联系模型转换为关系模式	149
第四节	实体—联系模型的扩充	151
第五节	函数语义模型	152
第六节	SMITH 图解法	153
第七章	数据库设计	158
第一节	信息系统及其与数据库的关系	158
第二节	数据库设计及数据库系统的生命周期	160
第三节	数据库设计方法	163
第四节	关系数据库的设计方法	170
第五节	数据库设计评价	185
第六节	数据库设计辅助工具	190
第八章	数据库的安全性及完整性	196
第一节	概述	196
第二节	数据库的安全性	197
第三节	数据库的一致性与完整性	200
第四节	数据库恢复	202
第九章	网状数据库系统	206
第一节	DBTG 系统及其结构	206
第二节	DBTG 的数据结构	207
第三节	数据组织和存取策略	211
第四节	数据描述语言	215
第五节	数据操作语言	221
第六节	网状数据库管理系统简介	227
第十章	新型数据库系统	230
第一节	分布式数据库系统	230
第二节	演绎数据库系统	236
第三节	面向对象数据库系统	239
第四节	工程数据库系统	241
第五节	多媒体数据库系统	244
参考文献		247

第一章 基础知识

随着科学技术的发展,人们掌握和处理的信息越来越多,然而要想充分地开发与利用这些信息资源,就必须对大量的信息进行识别、存贮、处理与传递。众所周知,人脑在信息的识别、信息的分析与综合、推理和联想等方面有很强的能力,但在记忆信息、快速处理信息方面的能力较弱,因此要从大量而又经常变化的信息中提取一些关键信息并不是很容易的。以电子计算机为基础的数据库技术由于具有信息存贮量大、处理速度快、传输速度快、逻辑推理严密、重复性高且不会疲劳、能够有效合理地存贮各种信息、能为有关应用准确快速地提供有用信息等特点,正好弥补了人类加工处理信息等方面的不足,使其很快成为了信息处理的强有力工具。

自从 60 年代末出现数据库技术以来,它已经渗透了工农业生产、商业、行政管理、科学研究、工程技术和国防军事等各个部门,它是管理信息系统(MIS),办公室自动化(OA)系统和决策支持系统(DSS)等应用系统的核心部分。可以说哪里有信息处理,哪里就有数据库技术的应用。故在计算机领域中,70 年代被称为是数据库的年代。

到了 70 年代末与 80 年代为适应各种需要又产生出很多数据库分支,它们是:

- 为适应分布式系统的需要而产生的分布式数据库。
- 为适应推理需要而产生的演绎数据库。
- 为适应人工智能需要而产生的知识库。
- 为图形处理的需要而产生的图形数据库。
- 为图象处理的需要而产生的图象数据库。
- 为数值统计的需要而产生的统计数据库。
- 为计算机辅助设计的需要而产生的 CAD 数据库。
- 为综合管理图文音数据的多媒体数据库。
- 为提高数据库效率而产生的数据库机器。

这些数据库大都建立在基本数据库的基础上,因此我们有必要掌握和学习基本数据库系统的原理和技术,用以解决各种计算机应用中的实际问题。

第一节 数据库系统

一、信息世界

人类社会错综复杂;其中存在着各种各样的事物、事物的属性、事物之间的联系、以及人们对各种事物的管理活动。然而所有的管理活动都离不开信息。农民从事农业生产需要土质、产量、肥料价格、气候情况等信息;工厂制定生产计划需要市场需求、材料来源、生产能力等有关信息;在日常生活中需要天气、物价、亲友的通信地址及电话号码、日程安排等信息。总之,我们的一切活动时时刻都和大量的各种信息打交道,所有的这些信息就构成了信息世界。人类社会的

一切活动都离不开对信息的处理,由于信息世界中的绝大部分信息都可以用数据来表示,因此,信息处理实际上也可以说是数据处理。从其发展的历史来看,数据处理大致分为手工处理、机械处理和电子处理三个阶段。

从原始社会到十九世纪末,由于社会生产力和科学技术的限制,数据处理处于低级的手工处理阶段。这一阶段主要使用的计算工具有算盘、基于齿轮结构的六位加法器、计算尺和微分机等。该阶段的特点是计算工具低级、精确度差、处理能力低,且离不开手工操作。

1890年,美国中央统计局的 H. Hollorith 为了编制人口统计表的需要,发明了卡片制表机,使数据处理跨入了机械处理阶段。该机能以半自动方式进行卡片的穿孔、校验、分类、整理和制表等工作。与手工阶段相比,一部分手工操作由机械所代替,因此,无论在数据处理能力和效率上都有很大提高,但其效率受到机械设备性能的限制。

1946年第一台电子计算机 ENIAC 的诞生,标志着一个崭新的数据处理阶段——电子处理阶段的开始。电子计算机以它自动、快速的处理,大容量的数据存贮,灵活的输入/输出,彻底改变了以往数据处理效率低、手工操作多、可靠性差、与当时社会生产力的发展不相适应的落后状态。随着计算机硬件和软件的发展,特别是大容量磁盘存贮器的应用,使计算机不仅能进行各类数值计算,同时还能进行文字和图象等各种数据处理,使数据处理技术获得突飞猛进的发展。今后我们所谈的数据处理均指电子数据处理。

二、数据管理

所谓数据管理是对数据的组织、存贮、检索和维护等工作,它是数据处理的核心。随着计算机软件与硬件的发展,数据管理技术也经历了三个阶段。

1. 人工管理阶段

它是计算机数据管理的初级阶段。该阶段的主要特点是数据在外存的物理结构与用户观点的逻辑结构完全一致,计算机系统仅提供基本的输入/输出操作,而无统一的数据管理软件,对数据的管理完全由各个程序员在其程序中进行管理。程序员在编制其应用程序时,必须考虑数据的逻辑定义和组织、数据存放的存贮设备、物理存贮方式和地址分配,并通过物理地址来存取数据,表示处理流程的程序与其处理的数据,相互结合成一个整体。此阶段数据的逻辑结构与物理结构之间的关系可用图 1-1 表示。

由于数据的物理结构是由应用程序员根据应用的要求设计的,因此用户要花费很大精力去考虑具体的物理细节,并且由于很难实现多个应用程序共享数据资源,造成数据大量重复;当数据的物理结构或存贮设备改变时,其应用程序必须重新编制,用户的负担很重。

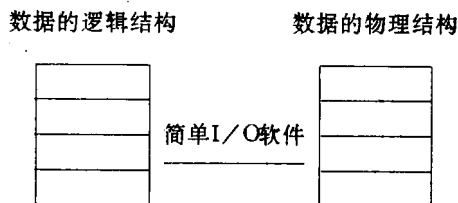


图 1-1 人工管理阶段的特征

2. 文件系统阶段

随着计算机硬件性能的改进和软件的发展,出现了文件管理系统,负责对数据进行管理。此时各种用途的应用程序仍然具有各自的数据,数据仍然具有两种形式,即用户看到的逻辑结构(称为逻辑文件)和实际存贮的物理结构(称为物理文件)。但此时逻辑文件和物理文件之间已允许有所区别,文件管理系统提供从逻辑文件到物理文件的转换。因此在进行程序设计时,用户可以不考虑具体的物理存贮结构,而直接利用文件管理系统中所提供的存取方式来设计应用程序。当物理存贮结构发

生改变时,通常无需改变用户程序,只要改变文件管理系统中的存取方式就行了,从而部分实现了逻辑数据和物理数据的独立性,在文件管理系统中,还提供了多种文件组织形式,如顺序文件组织、链接文件组织和直接存取文件组织。数据的逻辑文件和物理文件之间的关系可用图 1—2 表示。

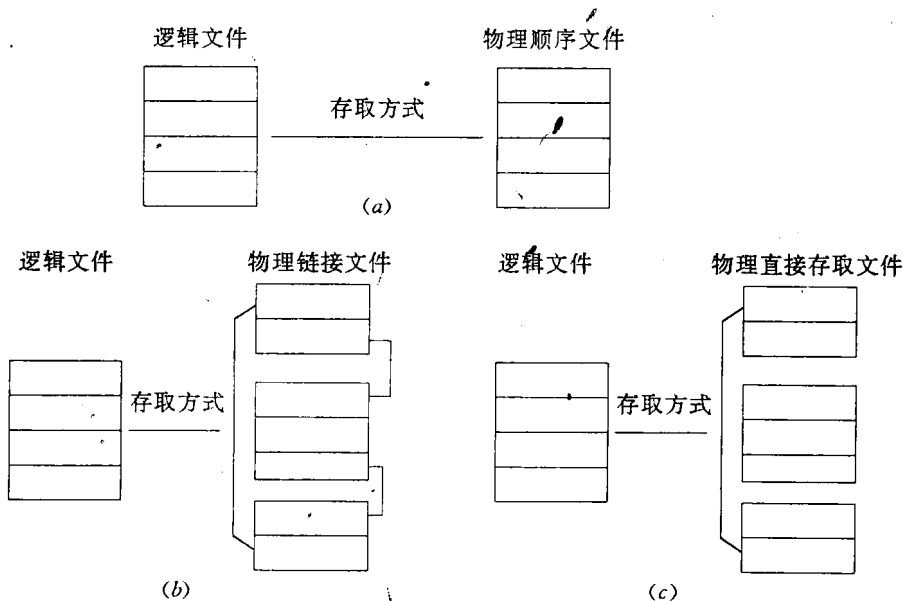


图 1—2 文件系统的数据组织

这一阶段尽管比人工管理阶段有很大的改进,但一些根本性的问题仍然没有解决,这主要表现在以下四个方面:

(1)数据冗余度大 由于一个数据文件只为某个特定的应用程序服务,不同的应用程序使用相互独立的数据文件,因此,相同的数据同时出现在几个数据文件中的情况几乎是不可避免的。这种数据大量重复的现象,称为冗余。它降低了存储空间的有效利用率。

(2)易造成数据的不一致性 由于同一数据重复存贮在由不同的应用程序(用户)使用和维护的文件中,不能保证其更新的同时性和正确性,以致造成同一数据在不同的文件中有不同的值。

(3)程序与数据的相互依赖,简称数据依赖。在文件管理方式下,数据文件由使用它的应用程序建立,一旦应用程序改变,原来的数据文件也必须相应改变,否则就无法使用。反之,如果数据文件存放的存贮设备、存贮方式和存取方法改变,则使用这些数据文件的应用程序也必须作相应修改,否则,应用程序就无法运行。这就是说,数据文件依赖于应用程序而存在,否则,就失去其存在的意义;反之亦然。

(4)应用程序设计困难 由于在应用程序中含有实现数据操作的程序设计(如查找、插入、删除功能的设计),这就要求程序员必须具备较高的算法和程序设计能力。

所有这些问题,文件系统本身无法解决。为此,计算机软件工作者经过长期不懈的努力,提出了数据库的概念。

3. 数据库系统阶段

以数据库为中心的数据库系统,是当代数据管理的主要方式。它克服了文件系统的弊病,

是一种较完善的高级数据管理方式,获得了广泛的应用。

三、什么是数据库系统

通俗地可以把数据库比作仓库。仓库是保存和管理物资的,并能根据其服务对象的要求随时提供所需物资。数据库则是存贮和管理数据并负责向用户提供所需数据的机构。就象仓库不能简单地同库房等同起来一样,也不能把数据库仅仅理解为存贮数据的集合,而应视为一个系统,即数据库系统。一个仓库系统不管其规模大小,都有五个基本部分:物资、库房、管理机构、管理人员和服务对象。数据库系统的情况与仓库系统的情况相类似,也包括有五个主要组成部分:数据库、运行环境、数据库管理系统、数据库管理员和用户。

1. 数据库 数据库是某一组织中各种应用所需要保存和处理的数据的集合,在该集合中完全地或部分地消除了数据的冗余。例如,一个数据库可能同时包括职工文件(职工号、姓名、地址、部门、工资等)和业务档案文件(职工号、姓名、部门、完成项目、评价等)。在这两个文件中存在有一定的数据冗余,而构造数据库时,就可以从其中一个文件中消除姓名、部门这两项数据,以减少这种冗余。

2. 运行环境 在数据库系统中有存放数据文件的大容量存贮器,还有相应的输入设备、输出设备、中央处理机、系统软件等。它们构成了数据库系统的运行环境。数据库系统要求的运行环境与数值计算的运行环境的侧重点有所不同。在数值计算中,计算机的运行速度决定着整个程序的运行速度;而在数据处理中,数据处理的速度除了与计算机本身的运算速度有关外,更主要的因素就是 I/O 所占的时间。由于数据库系统要处理大量的数据,所以它对内存、外存容量要求也比数值计算大得多。在为数据库系统选择运行环境时,要着重考虑 I/O 的速度和存贮容量。

3. 数据库管理系统(DBMS) 它是实际存贮的数据和用户之间的一个接口,负责处理应用程序存取和管理数据库的各种请求。DBMS 使用户在使用数据库时无需考虑数据库的物理存取结构,它向用户提供数据库的高层次视图,DBMS 还负责数据库完整性检查、安全性检查、并发控制、故障恢复等工作。

4. 数据库管理员(DBA) 主要负责设计、建立、管理和维护数据库,协调各用户对数据库的要求等,他们对于程序语言和系统软件如 OS,DBMS 等都要熟悉,还要熟悉各部门的所有业务工作。DBA 的职责将在第三节中较详细地讨论。

5. 用户 它是数据库系统的服务对象。一般而言,一个数据库系统有两类用户:应用程序员和终端用户。应用程序员用高级程序设计语言编写使用数据库的应用程序。在应用程序中根据需要向 DBMS 发出适当的请求,由 DBMS 对数据库执行相应的操作。这类用户通常称为批处理用户。

终端用户从联机终端上以交互方式向系统提出各种操作请求,使用数据库中的数据。终端用户除了在联机终端上通过系统提供的人—机接口交互式地使用数据库外,还可以通过应用程序提供的人—机接口在联机终端上交互地使用数据库。

四、数据库系统的作用及优越性

现在,从微型机到大型机都配置有数据库系统用于数据处理。与文件系统相比,数据库系统的作用及优越性有以下几个方面。

1. 数据的集中管理 在数据库中,各个应用程序所使用的数据是由数据库系统统一规划,按照一定的数据模型组织和建立的,由系统统一管理和集中控制。它能较好地反映数据之间的自然联系,减少数据冗余。由于是从一个组织的全部应用来全盘考虑统一规划数据的结构,所以尽管这种结构对某个应用可能不是最好的,但从整体来说是较优的。

2. 具有数据独立性 数据库系统提供了两级数据独立性。其一是物理独立性,即当数据的物理结构发生改变时,不影响数据库的逻辑结构,从而无需修改原来的应用程序;其二是逻辑独立性,即当数据库的总体逻辑结构发生改变时,无需修改原来的应用程序。

3. 实现数据共享 数据库中的数据可供多个应用程序用于不同的目的。它不但能被现有的各种应用所共享,而且也能为未来新开发的应用所共享。

4. 数据的冗余度小 文件系统是根据应用程序的需要而各自建立的,当不同的应用程序所使用的数据有许多部分相同时,也必须建立各自的文件,从而存贮数据大量冗余,产生大量存贮空间浪费、更新冗余数据困难、更新时间增加等弊病。而数据库则从整体观点来组织和存贮数据,使重复的数据仅需存贮一份,并为多个应用所共享,从而大大减少了数据冗余。

5. 避免了数据的不一致性 当本应相同的数据项在不同的应用中出现不同的值时,就出现了数据的不一致性。数据的不一致性主要是由数据冗余引起的。数据库系统通过消除冗余和增加一致性保护机制保证了数据的一致性。

6. 可以实施安全性保护 数据的安全性主要指数据保密,防止数据的非法使用。在数据库系统中,通过给用户规定一定的存取权限等安全措施,可以保证当非法用户试图存取敏感数据时,数据库就进行安全性检查,对于不符合相应存取权限的用户,系统拒绝执行存取操作。对于数据库中每部分数据的各种类型的操作(检索、插入、删除、修改等),数据库系统都可以实施不同的安全检查。

7. 有利于保证数据的完整性 数据的完整性是指数据的正确性、准确性和有效性。数据的不一致性是失去完整性的例子。此外,还有其它一些因素使数据库失去数据的完整性,例如,在一个数值型数据中出现了字母、特殊符号等,或一个工人一周的工作时间超过 200 小时等都是失去完整性的例子。数据库系统通过由 DBA 定义的完整性约束规则,对每一次更新操作实施完整性检查,可以保证数据的完整性。

8. 可以发现故障和恢复正常状态 硬/软件的故障及用户操作的失误,都有可能使数据库遭到局部性地或全局性地破坏。数据库系统有一套及时发现故障,并迅速地把数据库恢复到故障发生以前正确状态的措施,如转贮、日志、检查点等方法。

9. 可以平衡矛盾的要求 DBA 与任何个别用户不同,它了解一个组织总的要求。所以,它能够有效地构造数据库,以便对该组织提供最好的服务。例如,它可以对最重要的应用来说存取速度最快,而对某些别的应用的性能影响又相当小这一角度来选择数据库的构造和在存贮器中的表示方法。

10. 有利于实施标准化 由于数据库的集中控制,DBA 能够保证在数据表示中遵守所有该应用的标准。这些标准可以是:团体、设备、部门、工业、国家和国际标准中的任一个。尤其存贮数据格式的标准化,因为它有助于数据交换或系统之间的流通。

11. 为用户提供了方便的用户接口 用户可使用查询语言操作数据库,也可以用程序方式(用高级语言和数据库操纵语言编制的程序)操作数据库。

12. 允许并发使用数据库 为了防止由于多个用户并发操作数据库时,相互干扰引起数

数据库发生数据不一致性问题,在数据库系统中提供了并发控制功能,以便对并发操作采取必要的控制措施。

第二节 数据模型

在数据库中存贮和管理的数据,都来自客观事物,那么怎样把现实世界中的客观事物抽象为能用计算机存贮和处理的数据呢?这有一个逐步转化的过程,一般讲,它分为三个阶段,又称为三个世界,即:现实世界、信息世界及机器世界。下面首先介绍怎样把现实世界中的客观事物逐步抽象为能用计算机存贮和处理的数据,然后介绍在数据库中怎样来表示和组织这些数据。

一、信息的三个世界

1. 现实世界

现实世界即客观存在的世界。在现实世界中客观存在着各种运动着的物质,即各种事物及事物之间的联系。客观世界中的事物都有一些特征,我们正是利用这些特征来区分事物。一个事物可以有许多特征,通常都是选用我们感兴趣的以及最能表征该事物的若干特征来描述该事物。以人为例,常选用姓名、性别、年龄、籍贯等作为描述一个人的特征。有了这些特征,就能很容易地把不同的人区分开来。

世界上各种事物虽然千差万别,但都息息相关,也就是说,它们之间都是互相关联的。事物间的关联也是多方面的,我们仅选择那些感兴趣的关联,而没有必要选择所有的关联,如在教学管理系统中仅选择“教师教学生”这种感兴趣的关联。

2. 信息世界

现实世界中的事物及其联系由人们感官感知,经过人们头脑的分析、归纳,抽象形成成为信息。对这些信息进行记录、整理、归类 and 格式化后,它们就构成了信息世界。为了正确直观地反映客观事物及其联系,有必要对所研究的信息世界建立一个抽象的模型,称之为信息模型。目前较为流行的一种信息模型是实体联系模型,它用实体—联系图来刻画。在实体联系模型中,事物用实体(Entity)表示,事物的特征用属性(Attributes)表示。教学管理的一个简单的实体联系模型见图 1—3,图中矩形表示实体、椭圆表示属性、菱形表示实体间的联系。有关实体联系模型的更详细情况将在第六章讨论。

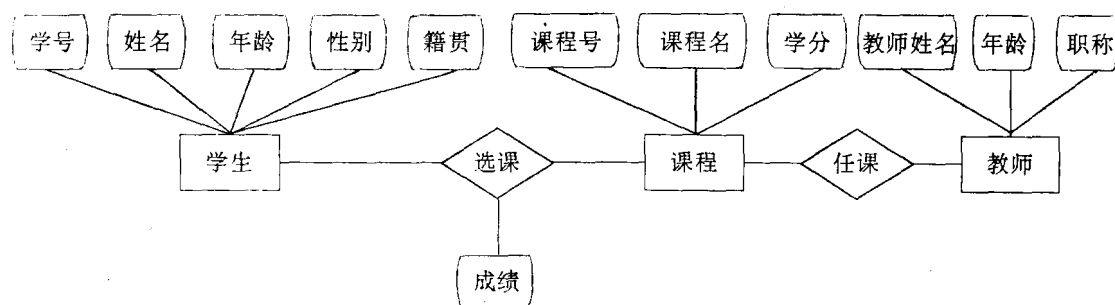


图 1—3 教学管理的实体联系模型

3. 机器世界

早期的计算机只能处理数据化的信息(即只能用字母、数字或符号表示),所以用计算机管

理信息,必须对信息进行数据化,即将信息用字符和数值表示。数据化后的信息称之为数据,数据是能够被机器识别并处理的。当前多媒体技术的发展使计算机还能直接识别和处理图形、图象、声音等数据。数据化了的信息世界称之为机器世界。通过从现实世界到机器世界的转换,为数据管理的计算机化打下了基础。

二、数据模型

得到能用计算机存贮和处理的数据是一个问题,怎样在数据库中表示和组织这些数据又是一个问题。一个部门或单位所涉及的数据很多,而且数据之间的联系又是错综复杂的,这些数据如何组织呢?正像图书馆管理图书那样,图书馆有大量的藏书,这些书籍如何保存?如何管理?组织得好可方便读者使用,组织得不好会影响读者使用。对于数据库也存在同样的问题,数据组织得好坏会影响系统效率和用户对数据库的使用。数据库技术在处理数据组织时,是从全局出发,对数据内部的联系和用户要求进行综合平衡来考虑的。在设计数据库全局逻辑结构时,现有的数据库管理系统允许使用三种方法。

- 层次方法 将数据库数据按树结构的形式进行组织;
- 网状方法 将数据库数据按图结构的形式进行组织;
- 关系方法 将数据库数据按表结构的形式进行组织。

不同形式的数据组织方法,对应有不同类型的数据操作集合,以及为了保证数据的完整性规定的完整性规则集合。这三者(数据组织、数据操作集合和完整性规则集合)组成了数据库的数据模型。模型是人们对客观世界的认识和理解,是对客观世界的近似描述。当前流行的三种数据模型是层次模型、网状模型和关系模型。层次模型发展最早,适合描述层次结构的数据。关系模型虽然发展较晚,但有一系列优点,它的数据二维表视图易于理解和使用,有严密的关系数据理论指导数据库的设计,因而获得了广泛的应用。关系数据模型是当今数据库中最主要的数据模型,但它的查询效率略逊于网状模型。网状模型开发较早而且也有一定优点,当前使用仍较多。在本节将对这三种模型的数据组织作一简单介绍,在以后章节中还要对关系模型和网状模型的典型数据库系统作详细介绍。

在介绍数据模型之前,先介绍几个概念和术语。

1. 记录与数据项

在数据模型中,用记录表示实体,用数据项表示实体的属性。一个数据项包含有:数据项名、数据类型和数据长度等内容。由于一个实体具有若干属性,故记录亦由若干数据项组成。一般就采用属性名作为描述它的数据项名。

2. 型与值

实体与属性有型与值之分,所以表示它们的记录和数据项也分为“型”与“值”两级。例如数据项“年龄”的型是“名称为年龄,数据特征为三位的十进制整数”,而它的值是“1岁,2岁,...,200岁等”。记录型是数据项型的一个有序组,同理,记录值是数据项值的同一有序组。记录型是一个框架,只有给它的每个数据项取值后才得到记录。

为了便于理解,在图1-4中以学生数据为例表示了信息在三个世界中的有关术语及其联系。看图时要特别注意实体与属性、型与值的区分,以及三个世界中各术语的相应关系。

数据模型的所有术语都用在型一级,整个模型就象一个框架,给它填上具体的数据值就得

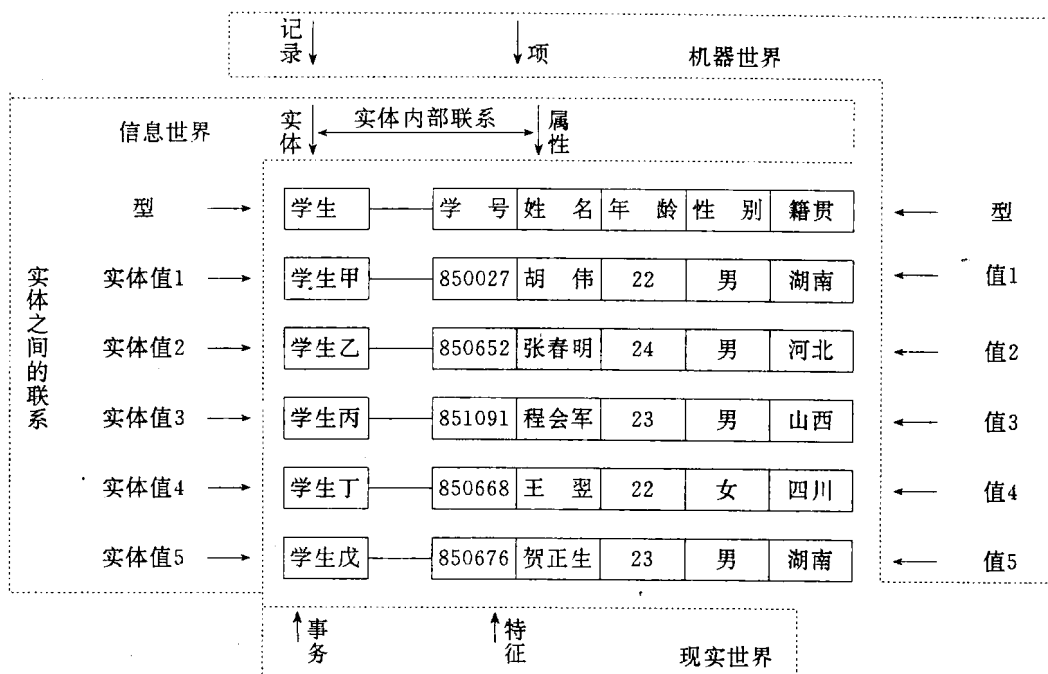


图 1—4 信息三个世界的术语联系

到数据模型的一个实例。为了简单在以后的讨论中对于实体、属性等不再加上后缀“型”或“值”了,同一术语在型与值两级中平行使用。它们具体的含义可根据上下文的含义来判断。

三、层次数据模型

层次数据模型的数据库是最早研制成功的数据库,它就是美国 IBM 公司于 1968 年开发的 IMS 系统。层次模型实际上是一个树形结构。它是以记录为结点、以联系为边的有向树。在层次模型中,最高层只有一个记录,该记录称为根记录,根记录以下的记录称为从属记录。一般说来,根记录可以有任意多个从属记录,每一从属记录又可以有任意多个低一级的从属记录等等,直到任意级。若干个树型结构形成森林。那么如何把实体联系模型转换成层次模型呢?一般需要经过以下几步:

1. 选择某一实体作为层次模型的根记录;
2. 将与父记录相联系的实体作为该记录的从属记录;
3. 将实体的属性转换成相应记录的数据项;
4. 将实体间联系的属性(不起联系作用的属性)转换成从属记录的数据项;
5. 再找从属记录的从属记录,直至构成一个层次模型。

在层次模型中联系已不再被明显地表达出来,而是被隐含在上下层关系中。图 1—3 的教学管理模型,对应的层次模型如图 1—5 所示。图中由于选择的根记录不同,得到了分别以学生实体、课程实体和教师实体作为根记录的三个层次模型。在选课联系中,由于学号、课程号两个属性起着联系学生、课程两个实体的作用,因此不转换成从属记录的数据项,而只将不起联系作用的属性分别转换成从属记录的数据项。而对于任课联系,由于没有不起联系作用的属性,所以没有属性被转换成从属记录的数据项。

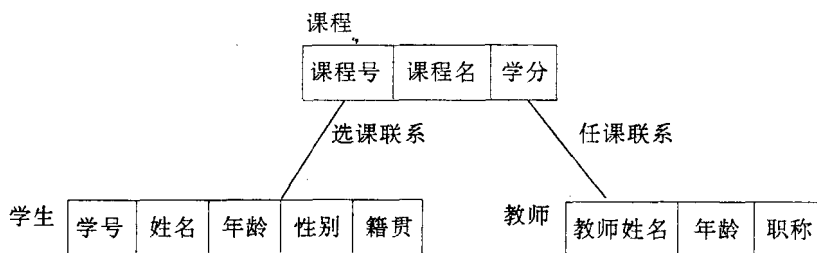


图 1—5 教学管理数据库的层次模型

从图 1—5 可以看出,层次模型具有两个较为突出的问题。首先,在层次模型中具有一定的存取路径,它仅允许自顶向下的单向查询。例如,对于图 1—5 的模型比较适合回答以下查询:查询某课程的情况;查询教授某课程的教师情况;查询选修某课程的学生情况。但不能直接回答“某教师所教的学生情况”这样的查询,因为在教师与学生之间没有自顶向下的通路。这时需要把查询折成两个子查询:先查询某教师所讲授的课程,当得到所讲授的课程名后,再查询选修这些课程的学生情况。故设计层次模型(以及网状模型)时,要仔细考虑存取路径问题,因为路径一经确定就不能改变。由于路径问题的存在,给用户带来了不必要的复杂性,尤其是要迫使用户花费时间和精力去解决那些由层次结构产生的、而不是所提查询所固有的问题。层次结构中引入的记录越多,层次变得越复杂时,则问题会变得越糟。从而使应用程序变得比问题要求的更加复杂,其结果是程序设计人员在编制、调试和维护程序时将需要花费比问题需要的更多的时间。

其次,层次模型比较适合于表示数据记录之间的一对多联系,而对于多对多、多对一的联系,会出现较多的数据冗余。例如对于图 1—5 的模型,由于一个学生可能选修多门课程,因此针对某一门具体的课程,该学生的情况就要被存贮一次,从而造成学生情况的大量重复。此外,层次模型还有以下一些问题:数据依赖性强,当上层记录不存在时,下层记录无法存贮;语义完整性差,某些数据项只有从上下层关系查看时,才能显示出它的全部含义,例如对于图 1—5 的模型,学生记录中数据项分数只有从根记录查看时才能明白它是选修哪门课程的分数;从同一实体联系模型可以构造出许多层次模型,而对于不同的层次模型,同一查询的表达方式就不同,因此要求用户必须了解模型的结构;查询的非对称性,在实体联系模型中为对称的结构,在层次模型中所用的查询语句格式可以不同,例如对于图 1—5,根据学生查课程和根据课程查学生所用的格式是不同的。

当然,不能否认层次模型是模拟现实世界中具有层次结构数据的一种很自然的方法。

四、网状数据模型

美国负责开发 COBOL 语言委员会 CODASYL 的一个小组 DBTG,于 1971 年发表的一个报告中首先提出了网状模型。在网状模型中用结点表示实体,用系(set)表示两个实体之间的联系。网状模型是一种较为通用的模型,从图论的观点看,它是一个不加任何条件的无向图。网状模型与层次模型的根本区别是:

- (1) 一个子结点可以有多个父结点;
- (2) 在两个结点之间可以有两种或多种联系。

显然,层次模型是网状模型的特殊形式,网状模型是层次模型的一般形式。把实体联系模

型转换成网状模型,一般需要经过以下几步:

(1) 将实体联系模型中的实体转换成为网状模型中的结点,实体的属性就是结点的数据项。

(2) 将实体间联系转换为结点间的系,联系的属性就是系的数据项。

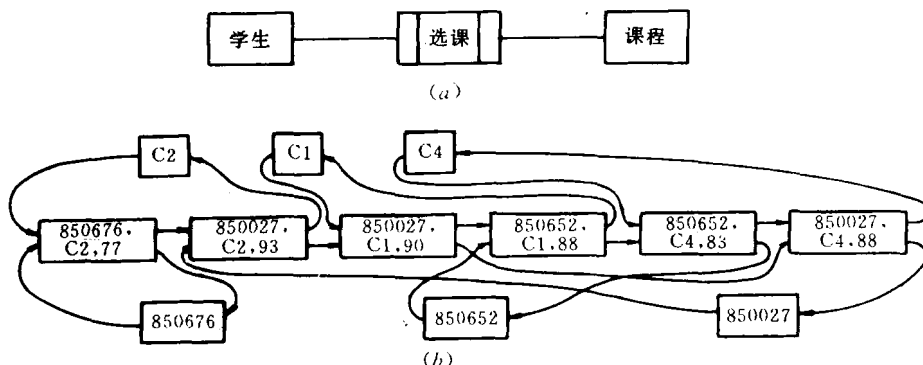


图 1—6 选课数据库的网状模型

图 1—6 表示学生选课的数据模型及其实例。为了简化,图中只取 850027、850652、850676 三个学生和 C1、C2、C4 三门课程。从图 1—6(b)可以看出,所有的实体记录都具有一个以其为始点和终点的循环链表,而每一个系都处于两个链表中,一个是课程链,一个是学生链。从而根据学生查找课程和根据课程查找学生都很方便。这种以两个结点和一个系构成的结构是网状模型中的基本结构,一个结点可以处于几个基本结构之中,这样就形成了网状结构。

网状模型在结构上比层次模型复杂,因而它在查询方式上要比层次模型较为优越。在网状模型中,对数据的查询可以用两种方式:

- (1) 从网络中任一结点开始查询;
- (2) 沿着网络中的路径按任意方向查询。

从网状结构中可以看出,这是一种对称结构,对于根据学生查课程和根据课程查学生这种对称的查询,在网状模型中所使用的查询语句格式是相同的。但是尽管网状模型要比层次模型具有对称性,也并不能使其查询变得简单,因为它支持的数据结构种类较多,这就势必造成其操作的复杂性。因此,网状模型的主要缺点是在数据结构本身及其相应的数据操作语言都极为复杂。一般说来,结构越复杂,则其功能就越强,然而所要处理的操作也越多,因此相应的数据操作语言也就变得越加复杂。而且由于其结构的复杂性,给数据库设计带来困难。

五、关系数据模型

关系模型的发展较晚,70年代初由 IBM 公司的 E. F. Codd 首先提出。关系模型是完全不同于前两种模型的一种新的模型,它的基础不是图而是表格。

大家知道,表格在日常生活中应用很广,许多数据都是以表格形式表示的。可以说,任何一个信息模型均可用二维表的形式表示出来。若把实体联系模型中的实体和联系均用二维表表示,其数据模型就称为关系模型。

在关系模型中,通常把二维表称为关系,数据的关系模型是由若干个关系模式(相当于记录型)组成的集合。表中的每一行称为一个元组,相当于通常的记录值。每一列称为一个属性,

相当于记录中的一个数据项。一个关系若有 k 个属性则称为 k 度关系。对于一个关系它有如下性质：

- (1) 没有两个元组在各个分量(或属性)上是完全相同的；
- (2) 行的次序无关；
- (3) 列的次序无关。

图 1—3 相应的关系模型如图 1—7 所示，它有五个关系模式：学生、课程、教师、选课和任课。

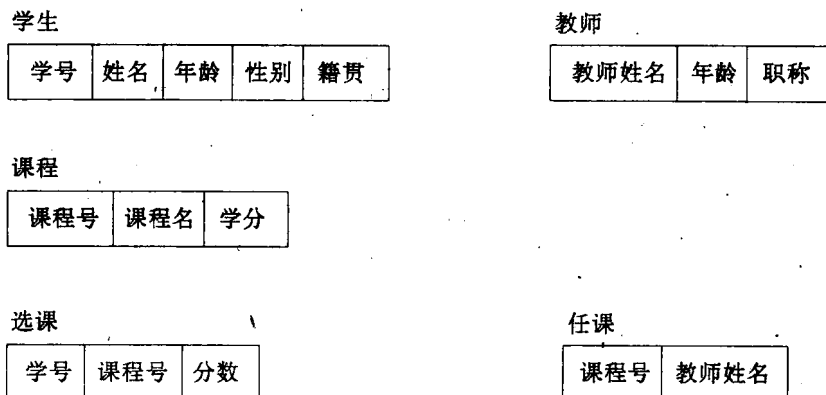


图 1—7 教学管理数据库的关系数据模型

关系模型具有以下特点：

1. 描述的一致性 即无论是信息世界中的实体还是联系都是用一个关系来描述，保证了数据操作语言的相应一致性，对于每一种基本操作功能(插入、删除、修改等)，都只需要一种操作运算即可。而前两种数据模型，由于它们支持的数据结构有几种，因而对于同一种功能需要几组不同的操作运算。

2. 利用公共属性连接 关系模型中各个关系之间都是通过公共属性发生联系的。例如学生关系和选课关系是通过公共属性学号连接在一起，而选课关系又可以通过课程号与课程关系发生联系。

3. 结构简单直观 采用表结构，用户容易理解，有利于和用户进行交互，并且在计算机中实现也极为方便。

4. 有严格的理论基础 二维表的数学基础是关系数据理论，对二维表进行的数据操作相当于在关系理论中对关系进行运算。这样，在关系模型中整个模型的定义与操作均建立在严格的数学理论基础之上。

5. 语言表达简练 在进行数据库查询时，不必像前两种模型那样需要事先规定路径，而是用严密的关系运算表达式来描述查询，从而使查询语句的表达非常简单直观。

缺点是在执行查询操作时，需要执行一系列的查表、拆表、并表操作，故执行时间较长。

六、数据模型的相互转换和比较

数据模型与被描述的客观事物之间并不是一一对应的，同一客观事物可以用若干不同的数据模型描述。例如图 1—3 的教学管理模型，就可以分别用层次、网状和关系三种模型来描