第 1 章

数据库基础

20 世纪 80 年代，美国信息资源管理学家霍顿(F. W. Horton)和马钱德(D. A. Marchand)等人指出：信息资源(Information Resources)与人力、物力、财力和自然资源一样，都是企业的重要资源，因此，应该像管理其他资源那样管理信息资源。

数据是信息时代的重要资源之一。商业的自动化和智能化，使得企业收集到了大量的数据，积累下来重要资源。人们需要对大量的数据进行管理，从数据中获取信息和知识，从而帮助人们进行决策，于是就有了数据库蓬勃发展的今天。数据库技术是计算机科学中一门重要的技术，数据库技术在政府、企业等机构得到广泛的应用。特别是 Internet 技术的发展，为数据库技术开辟了更广泛的应用舞台。

本章的知识体系：

- 数据库系统
- 数据库设计的基本步骤
- 实体—联系模型
- 关系数据库
- E-R 模型转换为关系模型
- 关系数据库操作基础

学习目标：

- 了解数据库的基本概念
- 了解数据库系统设计的基本步骤
- 掌握关系数据库相关概念
- 掌握 E-R 模型向关系模型的转换
- 掌握关系数据库操作基础

1.1 数据库系统

首先，通过几个事例，介绍为什么需要数据库。

A 公司的业务之一是销售一种科技含量较高的日常生活用品，为适应不同客户群的需求，这种商品有多个型号；产品通过分布在全市的 3000 多个各种类型的零售商销售

(如各类超市、便利店等);同时,公司在全国各主要城市都设有办事处,通过当地的代理商销售这种商品。

如果读者正在管理这家公司,需要什么信息?

A 公司的管理层需要随时掌握各代理商和零售商的进货情况、销货情况和库存情况,需要掌握各销售渠道的销售情况,需要了解不同型号产品在不同地域的销售情况等,以便及时调整销售策略。A 公司的工作人员定期对代理商和零售商进行回访,解决销售过程中的各种问题,并对自己的客户(代理商和零售商)进行维护。在此过程中,公司还需要对自己的市场部门工作业绩进行考核。这个例子,涉及了产品、客户、员工和订单。

随着市场范围的不断扩大,业务量迅速增长,A 公司需要有效地管理自己的产品、客户和员工等数据。

这样大量的数据,靠人工管理已经不再可能,比较好的方法之一是用数据库系统来管理其数据。那么,应该如何去抽象数据、组织数据并能够有效地使用数据,从中得到有价值的信息呢?这正是本书要介绍的内容。

解决上述问题的最佳方案之一就是使用数据库。产生数据库的动因和使用数据库的目的是及时地采集数据、合理地存储数据、有效地使用数据,从而保证数据的准确性、一致性和安全性,在需要的时间和地点获得有价值的信息。

数据库指的是以一定方式储存在一起、能为多个用户共享、具有尽可能小的冗余度、与应用程序彼此独立的数据集合。

1.1.1 数据库的基本概念

数据库所要解决的基本问题如下。

(1) 如何抽象现实世界中的对象,如何表达数据以及数据之间的联系。

(2) 如何方便、有效地维护和利用数据。

通常意义下,数据库是数据的集合。一个数据库系统的主要组成部分是数据、数据库、数据库管理系统、应用程序以及用户。数据存储在数据库中,用户和用户应用程序通过数据库管理系统对数据库中的数据进行管理和操作。

1. 数据

数据(Data)是对客观事物的抽象描述。数据是信息的具体表现形式,信息包含在数据之中。数据的形式或者说数据的载体是多种多样的,它们可以是数值、文字、图形、图像、声音等。例如,用会计分录描述企业的经济业务,会计分录反映了经济业务的来龙去脉。会计分录就是其所描述的经济业务的抽象,并且是以文字和数值的形式表现的。

数据的形式还不能完全表达数据的内容,数据是有含义的,即数据的语义或数据解释。所以数据和数据的解释是不可分的。例如,(103501011,张捷,女,1992,北京,信息学院)就仅仅是一组数据,如果没有数据解释,读者就无法知道这是一名学生还是一名教师的数据,1992 应该是一个年份,但它是出生年份还是参加工作或入学的年份,就无法了解。

在关系数据库中,上述数据是一组属性值,属性是它们的语义。例如,这组数据描述

的是学生，描述学生的属性包括学号、姓名、性别、出生日期、籍贯、所属学院，则上述数据就是这一组属性的值。

通过对数据进行加工和处理，从数据中获取信息。数据处理通常包括：数据采集、数据存储、数据加工、数据检索和数据传输/输出等环节。

数据的 3 个范畴分为：现实世界、信息世界和计算机世界。数据库设计的过程，就是将数据的表示从现实世界抽象到信息世界（概念模型），再从信息世界转换到计算机世界（数据世界）。

2. 数据库

数据库（DataBase，DB）是存储数据的容器。通常，数据库中存储的是一组逻辑相关的数据的集合，并且是企业或组织经过长期积累保存下来的数据集合，是组织的重要资源之一。数据库中的数据按一定的数据模型描述、组织和存储。人们从数据中提取有用信息，信息的积累成为知识，丰富的知识创造出智慧。

3. 数据库管理系统

数据库管理系统（DataBase Management System，DBMS）是一类系统软件，提供能够科学地组织和存储数据，高效地获取和维护数据的环境。其主要功能包括数据定义、数据查询、数据操纵、数据控制、数据库运行管理、数据库的建立和维护等。DBMS 一般由软件厂商提供，例如，Microsoft 的 SQL Server、Access 等。

4. 数据库系统

一个完整的数据库系统（DataBase System，DBS）由保存数据的数据库、数据库管理系统、用户应用程序和用户组成。DBMS 是数据库系统的核心，其关系如图 1.1 所示。用户以及应用程序都是通过数据库管理系统对数据库中的数据进行访问的。

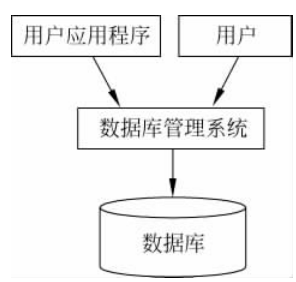

图 1.1　数据库系统的组成

通常一个数据库系统应该具备如下功能。

（1）提供数据定义语言，允许使用者建立新的数据库并建立数据的逻辑结构（Logical Structure）。

（2）提供数据查询语言。

（3）提供数据操纵语言。

（4）支持大量数据存储。

（5）控制并发访问。

1.1.2　数据库系统的特点

数据库系统是进行数据存储和管理应用的系统，它的特点如下。

1. 数据结构化

数据库中的数据是结构化的。这种结构化就是数据库管理系统所支持的数据模型。使用数据模型描述数据时，不仅描述了数据本身，同时描述了数据之间的联系。关系数据库管理系统支持关系数据模型，关系模型的数据结构是关系满足一定条件的二维表格。

2. 数据高度共享、低冗余

数据的共享度直接关系到数据的冗余度。数据库系统从整体角度看待和描述数据,数据不再面向某个应用而是面向整个系统。因此,数据库中的数据可以高度共享。数据的高度共享本身就减少了数据的冗余,同时确保了数据的一致性,同一数据在系统中的多处引用是一致的。

3. 数据的独立

数据的独立性是指数据库系统中的数据与应用程序之间是互不依赖的。数据的独立性包括逻辑独立性(数据库的逻辑结构和应用程序相互独立)和物理独立性(数据物理结构的变化不影响数据的逻辑结构)。

4. 数据实现集中管理和控制

文件管理方式中,数据处于一种分散的状态,不同的用户或同一用户在不同处理中其文件之间毫无关系。利用数据库可对数据进行集中控制和管理,并通过数据模型表示各种数据的组织以及数据间的联系。

数据库管理系统的管理和控制功能主要包括以下几点。

(1) 安全性控制:防止数据丢失、错误更新和越权使用。

(2) 完整性控制:保证数据的正确性、有效性和相容性。

(3) 并发控制:在同一时间周期内,允许对数据实现多路存取,又能防止用户之间的不正常交互作用。

(4) 故障恢复:由数据库管理系统提供一套方法,可及时发现故障和修复故障,从而防止数据被破坏。数据库系统能尽快恢复其运行时出现的故障,这些故障可能是物理上或是逻辑上的错误。

5. 数据库发展过程

美国学者詹姆斯·马丁在其《信息工程与总体数据规划》一书中,将数据环境分为四种类型,阐述了数据管理即数据库的发展过程。

(1) 数据文件

在数据库管理系统出现之前,程序员根据应用的需要,用程序语言分散地设计应用所需要的各种数据文件。数据组织技术相对简单,但是随着应用程序的增加,数据文件的数量也在不断增加,最终会导致很高的维护成本。数据文件阶段,会为每一个应用程序建立各自的数据文件,数据是分离的、孤立的,并且随着应用的增加,数据被不断地重复,数据不能被应用程序所共享。

(2) 应用数据库

意识到数据文件带来的各种各样的问题,于是就有了数据库管理系统。但是各个应用系统的建立依然是“各自为政”,每个应用系统建立自己的数据库文件。随着应用系统的建立,孤立的数据库文件也在增加,“数据孤岛”产生,数据仍然在被不断地重复,数据不能共享,并且导致了数据的不一致和不准确。

(3) 主题数据库

主题数据库是面向业务主题的数据组织存储方式,即按照业务主题重组有关数据,而

不是按照原来的各种登记表和统计报表来建立数据库；强调信息共享(不是信息私有或部门所有)。主题数据库是对各个应用系统“自建自用”数据库的彻底否定，强调各个应用系统“共建共用”的共享数据库；所有源数据一次一处输入系统(不是多次多处输入系统)。同一数据必须一次一处进入系统，保证其准确性、及时性和完整性，经由网络—计算机—数据库系统，可以多次、多处使用。主题数据库由基础表组成，基础表具有如下特性：原子性(表中的数据项是数据元素)、演绎性(可由表中的数据生成全部输出数据)和规范性(表中数据结构满足三范式要求)。

(4) 数据仓库

数据仓库是将从多个数据源收集的信息进行存储，存放在一个一致的模式下。数据仓库通过数据清理、数据变换、数据集成、数据装入和定期数据刷新来构造。建立数据仓库的目的是进行数据挖掘。

数据挖掘是从海量数据中提取出知识。数据挖掘是以数据仓库中的数据为对象，以数据挖掘算法为手段，最终以获得的模式或规则为结果，并通过展示环节表示出来。

1.1.3　数据管理技术的发展

随着计算机应用范围的不断扩大，以及各领域对数据处理的需求不断增强，数据管理技术在不断地发展。

计算机数据管理随着计算机硬件、软件技术和计算机应用范围的发展而不断发展，经历了如下三个阶段：人工管理阶段、文件系统阶段、数据库技术阶段。对数据有效地管理，是为了对数据进行处理，数据处理的过程包括数据收集、存储、加工和检索等。

1. 人工管理

20 世纪 50 年代中期以前，计算机主要用于数值计算。从硬件系统看，当时的外存储设备只有纸带、卡片、磁带，没有直接存取设备；从软件系统看，没有操作系统以及管理数据的软件；从数据看，数据量小，数据无结构，由用户直接管理，且数据间缺乏逻辑组织，数据依赖于特定的应用程序，缺乏独立性。

人工管理数据阶段的特点如下。

(1) 数据不保存：一个目标计算完成后，程序和数据都不被保留。

(2) 应用程序管理数据：应用程序与所要处理的数据集是一一对应的，应用程序与数据之间缺少独立性。

(3) 数据不能共享：数据是面向应用的，一组数据只能对应一个程序。

(4) 数据不具有独立性：数据结构改变后，应用程序必须修改。

2. 文件系统阶段

20 世纪 50 年代后期到 60 年代中后期，计算机应用从科学计算发展到了科学计算和数据处理。1954 年出现了第一台用于商业数据处理的电子计算机 UNIVACI，标志着计算机开始应用于以加工数据为主的事务处理阶段。这种基于计算机的数据处理系统从此迅速发展起来。这个阶段，硬件系统出现了磁鼓、磁盘等直接存取数据的存储设备；软件系统有了文件系统，处理方式也从批处理发展到了联机实时处理。文件系统阶段的数据管理特点如下。

(1) 数据可以长期保存。数据能够被保存在存储设备上,可以对数据进行各种数据处理操作,包括查询、修改、增加、删除等操作。

(2) 由文件系统管理数据。数据以文件形式存储在存储设备上,有专门的文件系统软件对数据文件进行管理,应用程序按文件名访问数据文件,按记录进行存取,可以对数据文件进行数据操作。

(3) 应用程序通过文件系统访问数据文件,使得程序与数据之间具有一定的独立性。

(4) 数据共享差、数据冗余大。仍然是一个应用程序对应一个数据文件(集),即便是多个应用程序需要处理部分相同的数据时,也必须访问各自的数据文件,由此造成数据的冗余,并可能导致数据不一致;数据不能共享。

(5) 数据独立性不好。数据文件与应用程序一一对应,数据文件改变时,应用程序就需要改变;同样,应该程序改变时,数据文件也需要改变。

3. 数据库技术

20 世纪 70 年代开始有了专门进行数据组织和管理的软件数据库管理系统,特别是在 20 世纪 80 年代后期到 90 年代,由于金融和商业的需求,使其得到了迅猛的发展。应用数据库管理系统管理数据具有如下特点。

(1) 数据结构化。

(2) 数据共享性高,冗余度低,易扩充。

(3) 数据独立性高。

(4) 数据由 DBMS 统一管理,完备的数据管理和控制功能。

1.2 数据库设计的基本步骤

数据是一个组织机构的重要资源之一,是组织积累的宝贵财富,通过对数据的分析,可以了解组织的过去,把握今天,预测未来。但这些数据通常是大量的,甚至是杂乱无章的,如何合理、有效地组织这些数据,是数据库设计的重要任务之一。

正如前面所述,数据库是企业或组织所积累的数据的聚集,除了每一个具体数据以外,这些数据是逻辑相关的,即数据之间是有联系的。数据库是组织和管理这些数据的常用工具。

数据库设计讨论的问题是:根据业务管理和决策的需要,应该在数据库中保存什么数据?这些数据之间有什么联系?如何将所需要的数据划分到表和列,并且建立起表之间的关系。数据的三个范畴为:现实世界、信息世界和计算机世界。数据库设计的过程,就是将数据的表示从现实世界抽象到信息世界(概念模型),再从信息世界转换到计算机世界(数据世界)。

数据库设计的目的在于提供实际问题的计算机表示,在于获得支持高效存取数据的数据结构。数据库中用数据模型这个工具来抽象和描述现实世界中的对象(人或事物)。数据库设计分为四个步骤,如图 1.2 所示。

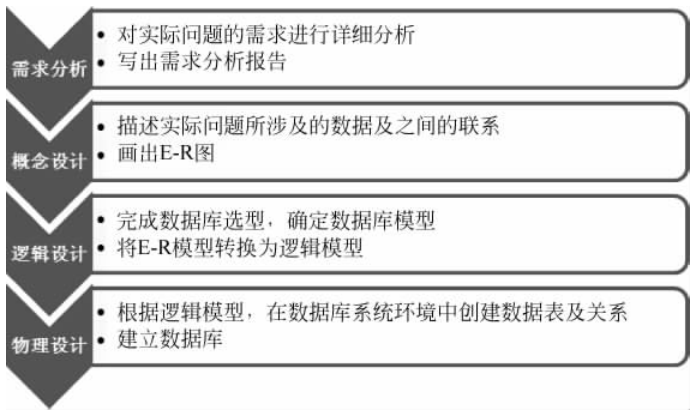

图 1.2　数据库设计的步骤

1. 数据库的需求分析

对需要使用数据库系统来进行管理的现实世界中对象(人或事物)的业务流程、业务规则和所涉及的数据进行调查、分析和研究,充分理解现实世界中的实际问题和需求。需求分析的策略一般有两种,自下向上的方法和自上向下的方法。

(1) 自下向上的方法

对事物进行了解,理解实际问题的业务规则和业务流程。在此基础上,归集出该事物处理过程中需要存放在数据库中的数据。

例如,一个产品销售数据库,需要保存客户的哪些数据? 可以做出一个二维表格,每一列是一个数据项,每一行是一个客户信息,可能包括:客户姓名、地址、邮政编码、手机号码等。

(2) 自上向下的方法

从为描述事物最终提供的各种报表和经常需要查询的信息着手,分析出应包含在数据库中的数据。

例如,上述产品销售数据库的客户信息,是否需要按客户性别进行统计分析? 如果需要,就应该增加一列“性别”数据项。

进行需求分析时,通常会同时使用上述两种方法。自下向上的方法反映了实际问题的信息需求,是对数据及其结构的需求,是一种静态需求;自上向下的方法侧重点在于对数据处理的需求,即实际问题的动态需求。

2. 数据库的概念设计

数据库的概念设计是在需求分析的基础上,建立数据的概念模型(Conceptual Data Model);用概念模型描述实际问题所涉及的数据以及数据之间的联系;这种描述的详细程度和描述的内容取决于期望得到的信息。一种较常用的概念模型是实体—联系模型(Entity-Relationship Model,E-R 模型)。E-R 模型是一种较高级的数据模型,它不需要使用者具有计算机知识。E-R 模型用实体和实体之间的联系来表达数据以及数据之间的联系。

例如,产品销售数据库,供应商是实体,客户是另一个实体,产品是实体,订单是实体,并且它们之间是有联系的;使用 E-R 模型描述这些实体以及它们之间的联系。

3. 数据库的逻辑设计

数据库的逻辑设计是根据概念数据模型建立逻辑数据模型(Logic Data Model),逻辑数据模型是一种面向数据库系统的数据模型,本书使用目前被广泛使用的关系数据模型来描述数据库逻辑设计:根据概念模型建立数据的关系模型(Relational Model);用关系模型描述实际问题在计算机中的表示;关系模型是一种数据模型,用表的聚集来表示数据以及数据之间的联系。数据库的逻辑设计实际是把 E-R 模型转换为关系模型的过程。

E-R 模型和关系模型分属两个不同的层次,概念模型更接近于用户,不需要计算机知识,属于现实世界范畴;而关系模型是从计算机的角度描述数据及数据之间的联系,需要使用的人具有一定的计算机知识,属于计算机范畴。

4. 数据库实现(数据库的物理设计)

依据关系模型,在数据库管理系统(如 Access)环境中建立数据库,Access 把数据组织到表格,表格由行和列组成。简单的数据库可能只包含一个表格,但是大多数数据库是包含多个表的,并且表之间有关系。

例如,产品销售数据库,就应该至少包含供应商表、客户表、产品表、订单表等,这些表通过主键建立联系。

1.3 实体—联系模型

数据库设计的过程就是利用数据模型来表达数据和数据之间联系的过程。数据模型是一种工具,用来描述数据(Data)、数据的语义(Data Semantics)、数据之间的联系(Relationship)以及数据的约束(Constraints)等。数据建模过程是一个抽象的过程,其目的是把一个现实世界中的实际问题用一种数据模型来表示,用计算机能够识别、存储和处理的数据形式进行描述。在本节中,将介绍一种用于数据库概念设计的数据模型:E-R 模型。一般地讲,任何一种数据模型都是经过严格定义的。

理解实际问题的需求之后,需要用一种方法来表达这种需求,现实世界中使用概念数据模型来描述数据以及数据之间的联系,即数据库概念设计。概念模型的表示方法之一是用 E-R 模型表达实际问题的需求。E-R 模型具有足够的表达能力且简明易懂,不需要使用者具有计算机知识。E-R 模型以图形的方式表示模型中各元素以及它们之间的联系,所以又称 E-R 图(E-R Diagram)。E-R 图便于理解且易于交流,因此,E-R 模型得到了相当广泛的应用。

1.3.1 基本概念

下面介绍 E-R 模型中使用的基本元素。

1. 实体

实际问题中客观存在并可相互区别的事物称为实体(Entity)。实体是现实世界中的对象,实体可以是具体的人、事、物。例如,实体可以是一名学生、一位教师或图书馆中的

一本书籍。

2. 属性

实体所具有的某一特性称为属性(Attribute)。在 E-R 模型中用属性来描述实体,例如,通常用“姓名”“性别”“出生日期”等属性来描述人,用“图书名称”“出版商”“出版日期”等属性描述书籍。一个实体可以由若干个属性来描述。例如,学生实体可以用学号、姓名、性别、出生日期等属性来描述。这些属性的集合(学号,姓名,性别,出生日期)表征了一个学生的部分特性。一个实体通常具有多种属性,应该使用哪些属性描述实体,取决于实际问题的需要或者说取决于最终期望得到哪些信息。例如,教务处会关心、描述学生各门功课的成绩,而学生处可能会更关心学生的各项基本情况,如学生来自哪里,监护人是谁,如何联系等问题。

确定属性的两条原则如下。

(1) 属性必须是不可分的最小数据项,属性中不能包含其他属性,不能再具有需要描述的性质。

(2) 属性不能与其他实体具有联系,E-R 图中所表示的联系是实体集之间的联系。

属性的取值范围称为该属性的域(Domain)。例如,“学号”域可以是由 9 位数字组成的字符串,“性别”域是“男”或“女”,“工资”的域是大于零的数值等。但域不是 E-R 模型中的概念,E-R 模型不需要描述属性的取值范围。

3. 实体集

具有相同属性的实体的集合称为实体集(Entity Set/Entity Class)。例如,全体学生就是一个实体集。实体属性的每一组取值代表一个具体的实体。例如,(103501011,张捷,女,1992 年 12 月)是学生实体集中的一个实体,而(113520200,李纲,男,1993 年 8 月)是另一个实体。在 E-R 模型中,一个实体集中的所有实体有相同的属性。

4. 键

在描述实体集的所有属性中,可以唯一地标识每个实体的属性称为键(Key,或标识 Identifier)。首先,键是实体的属性;其次,这个属性可以唯一地标识实体集中每个实体。因此,作为键的属性取值必须唯一且不能“空置”。例如,在学生实体集中,用学号属性唯一地标识每个学生实体。在学生实体集中,学号属性取值唯一而且每一位学生一定有一个学号(不存在没有学号的学生)。因此,学号是学生实体集的键。

5. 实体型

具有相同的特征和性质的实体一定具有相同属性。用实体名及其属性名集合来抽象和描述同类实体,称为实体型(Entity Type)。表示实体型的格式是:

实体名(属性 1,属性 2,…,属性 n)

例如,学生(学号,姓名,性别,出生日期,所属院系,专业,入学时间)就是一个实体型,其中带有下划线的属性是键。

用图形表示这个实体集的方法,如图 1.3 所示。用矩形表示实体集,矩形框中写入实体集名称,用椭圆表示实体的属性。作为键的属性,用加下划线的方式表示。

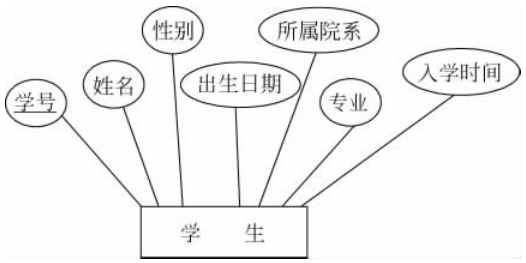

图 1.3　学生实体图形表示

在建立实体集时,应遵循的原则如下。

(1) 每个实体集只表现一个主题。例如,学生实体集中不能包含教师,它们所要描述的内容是有差异的,属性可能会有所不同。

(2) 每个实体集有一个键属性,其他属性只依赖键属性而存在。并且除键属性以外的其他属性之间没有相互依赖关系。例如,学生实体中,学号属性值决定了姓名、性别、出生日期等属性的取值(记为:学号→姓名　性别　出生日期),但反之不行。

6. 联系

世界上任何事物都不是孤立存在的,事物内部和事物之间是有联系(Relationship)的。实体集内部的联系体现在描述实体的属性之间的联系;实体集外部的联系是指实体集之间的联系,并且这种联系可以拥有属性。

实体集之间的联系通常有三种类型:一对一联系(1∶1)、一对多联系(1∶n)和多对多联系(m∶n)。

1.3.2　实体集之间的联系

1. 一对一联系

对于实体集 A 中的每一个实体,实体集 B 中至多有一个实体与之联系,反之亦然,则称实体集 A 与实体集 B 具有一对一联系(1∶1)。记为 1∶1。

【例 1.1】 某科技园要对入驻其中的公司及其总经理信息进行管理。如果给定的需求分析如下,建立此问题的概念模型。

(1) 需求分析

① 每个公司有一名总经理,每位总经理只在一个公司任职。

② 公司的数据包括:公司名称,地址,电话。

③ 总经理的数据包括:姓名,性别,出生日期,民族。

这个问题中有两个实体对象,即公司实体集和总经理实体集。描述公司实体集的属性是公司名称、地址和电话;描述总经理实体集的属性是姓名、性别、出生日期和民族。但两个实体集中没有适合作为键的属性,因此为每一个公司编号,使编号能唯一地标识每一个公司;为每一位总经理编号,使编号能唯一地标识每一个总经理。并且在两个实体集中增加“编号”属性作为实体的键。

(2) E-R 模型

① 实体型。

公司(公司编号,公司名称,地址,电话)