第一章　数据库基本知识

数据库是计算机软件的一个重要分支，是近二三十年迅速发展起来的一门新的学科。数据库主要解决的问题是数据的存储、管理、检索和利用。现在几乎所有的计算机应用系统都离不开数据库的支持。

本章介绍数据库一些最基本的概念。在介绍这些概念时力求深入浅出，容易为非计算机专业的广大用户所接受。数据库作为一门学科，有其理论体系和方法体系，概念术语也很多，但作为一个非专业的个人用户来说，目的是将数据库作为工具为自己工作服务。因此，许多专业理论和术语可以不去涉及，而把着眼点放在"如何使用它"上面。

第一节　数据库基本概念

一、什么叫数据库

数据库一词在英文里是"data base"。"data"是"数据"的意思，"base"是"基、基地"的意思，"data base"翻译过来就是"数据库"。

计算机中的"数据"一词是一个比较广泛的概念，既指数值型的数据，又指非数值型的数据。数值型的数据，例如："62127"、"41117"等。数值型数据表示数量的概念，因此可以用来进行数学运算，如求和、求积、求平均等等。非数值型的数据，指文字、符号等，例如"张三"(姓名)，"电视机"(产品名称)。非数值型的数据不能做数学运算，不能求和、求积等，但可以做非数值运算，如位置运算、逻辑运算等。注意，用数字表示的数据不一定都是数值型的数据，例如电话号码"6015948"，是用数字表示的，但不是数值型数据，因为电话号码相加、相乘是没有意义的。这里的数字代表一种"代码"。代码常常用来标识不同事物。例如电话号码"7015948"不同于"6015948"，它们之间不存在谁比谁大、谁比谁多的数量关系。日期数据，例如"1993 年 12 月 29 日"，或"1993.12.19"，一般也作为非数值型数据。虽然它可以比较大小，例如，相减得出多少天，但日期本身并不具有数量的概念。计算机能处理非数值数据，极大地扩充了它的应用范围，从而渗透到社会的各行各业。

什么叫数据库呢？数据库可以简单地理解为"在计算机中存储起来的工作数据的集合"。这里所说的"工作数据"随应用方面的不同而不同。一个企业的"工作数据"，就是这个企业在生产、销售、管理活动中使用的各种数据，如产量、销售量、利润、成本(数值型数据)以及组织机构、职工姓名、规章制度、总结报告、文件档案(非数值型数据)等；而一个机关的"数据"就是办公方面的各种数据，如请示、报告、批示、决定、文件、档案、函电、通知等。每个单位、每项工作或每位个人在日常工作中都有自己的工作数据。将这些工作数据按照一定的要求组织存储在计算机中，就是本单位、本项工作或本人的数据库。

数据库中的数据不是乱堆乱放的。要尽量减少重复(冗余)，一个职工的姓名，可能在

财务部门(发工资)、人事部门(档案)、行政部门(住房子)等许多部门使用,但在数据库中最好只存有他的一个姓名。这样做不但节约存储空间,更重要的是可以避免造成数据之间的不一致。数据库的数据还应有一定的结构,这种结构反映数据之间自身的逻辑关系,按照这种关系组织存放数据,可以不因为具体的应用而改变不同的存储方法。这好像图书馆中的图书一样,是按照图书自身的分类法,独立于任何具体的读者存放的,不因为不同读者的不同需要,而改变图书的存放结构。

二、什么叫数据库管理系统

图书馆的图书靠一套分类编目方法和图书馆管理人员进行管理。那么存储在数据库中的成千上万的数据,又靠什么来管理呢?是靠一个计算机软件来管理,这个软件叫做"数据库管理系统"(DBMS)。正是通过这个数据库管理系统完成对数据库中数据的组织、存放、检索等各种操作。数据库管理系统是数据库的大管家,是数据库的核心和灵魂。我们学习数据库,就是要学习数据库管理系统的使用方法。数据库管理系统本身是个复杂的计算机程序,但呈现在用户面前,却并不那么复杂,人们都可以学会使用它。

按照历史发展进程,数据库管理系统已经历了四代,分别是文件、层次、网状和关系数据库管理系统。目前微机上安装使用的主要是关系数据库管理系统。像本书后面章节要介绍的 dBASE Ⅲ,以及 Foxbase,Oracle,Sybase 等都是关系型的数据库管理系统。层次、网状数据库管理系统是较早的产品,除在大型计算机上还有安装使用外,现在已很少见到了。

三、什么叫数据库应用系统

什么叫"数据库应用系统"?并没有一个严格的定义,我们可以理解为:一个计算机的应用系统中,如果使用了数据库,就称为一个数据库应用系统。比如一个工资管理系统,使用了数据库,就是一个数据库应用系统。由于现在很多应用系统都使用数据库,所以就有各种各样的"管理系统"、"信息管理系统"等名称。当然,这些管理系统不仅使用数据库,还使用程序设计语言编写的程序。

关于"数据库"、"数据库管理系统"和"数据库应用系统"三个概念,是有所不同的。"数据库"强调的是"存储的数据",所以关心的是数据和存放数据的介质。"数据库管理系统"强调的是"对数据库的管理"、"对数据的操作",所关心的是它的功能。数据库必须在数据库管理系统的管理下,才能实现建库、装载数据、对数据进行查询、修改等各种操作。"数据库应用系统"强调的是"应用",关心的是应用效果。一般说来,一个数据库应用系统由三部分组成:数据库、数据库管理系统和应用程序。各自的作用是:数据库存放工作数据,数据库管理系统管理数据库中的数据,应用程序用来完成对数据的自动化处理和向用户提供更加友好的界面。三者结合一起,形成一个数据库应用系统。虽然我们力图讲清楚这三个概念的不同,但在用户口头上,常常笼统称之为"数据库",这时需要根据具体情况去理解。例如,常听到"你用的是什么数据库?","我用的是 dBASE Ⅲ"。这里的"数据库"指的就是"数据库管理系统"。"你的数据库名字是什么?""是 SBGL.DBF"。这里的"数据库"指的是一个数据的集合,即在"SBGL.DBF"中存放着工作数据。本书以后章节的叙述中,在不造

成歧义的情况下，笼统使用“数据库”一词。

四、数据格式

数据库是存储数据的，数据格式即数据的外在形式对数据存储有着重要的意义。

我们来看看下面两种数据在格式上有什么不同。

1. 学生成绩表

学　号	姓　名	数　学	语　文	历　史	地　理
001	张三	70	80	90	95
002	李四	95	85	75	65
003	王五	80	80	80	90

在表中列出的是三个学生的各科成绩。这个成绩表的外在格式是“表格”形的。表格形数据的基本特点是横行竖列。表的一行表示一个学生各科成绩的情况，表的一列则表示不同学生的同一科成绩的情况。例如“学号”这一列，表示三个同学的三个不同的学号。“语文”这一列，表示三个学生的语文成绩。

“表格”形数据在日常工作中大量遇到，有些看来不是“表格”形的数据，也能化为表格来表示。

2. 外交备忘录写作说明

备忘录写作说明

备忘录是一种郑重性次于照会、使用范围日趋扩大的外交文书。一般可用来阐明自己对某一问题的立场观点，说明某一问题在事实上或法律方面的细节，记述或更详尽地陈述、补充自己谈过的事实或论点、论据，答复或驳复对方照会、备忘录中的观点、意见。有时也可以作为一种提醒或催询某事的形式。备忘录可以单独使用，也可作为照会、声明、通知等的附件。此外，它还可用以记载双方会谈所达成的一致意见，此种备忘录称谅解备忘录，是一种特殊的协议形式，须经双方签字。

备忘录在格式和写法上主要有以下几个特点：

(1) 一般用第三人称写成。

(2) 一般不用标题。但较重要的可标“备忘录”或“关于××问题的备忘录”。交涉重要问题的说帖(备忘录的一种)也可以“说帖”为题交给对方。

(3) 一般不写抬头，正文开始也不必写客套话，而直接陈述实质性意见或其他内容，有时也可用“事由：××××”或“××××谨就××问题申述如下：”字样开头。文尾一般也不需致敬语。

(4) 落款处一般只写明日期、地点，有些也可完全无落款。

(5) 面交的或作为附件的备忘录一般均不编号，不盖章，单独送交的须编号、盖章。

备忘录可以事先写好面交，也可以事后整理送去。

这一写作说明的外在格式是“文章”形的。文章形式的数据的基本特点是分篇、分段、分句。一篇文章有一个标题，篇由段组成，段由句子组成，句子由字词(字串)组成。

“文章”形的数据在日常工作中也大量遇到。各种书籍、文章、文件、报告等等都是文章

形的数据。这类数据又称为“文本形的全文”数据，或“全文”形的数据。

以上两类典型的数据格式，也是日常工作中大量遇到的两种数据格式。当然，还有图形数据、图象数据等格式，但已超出本书讨论的范围，这里不再涉及了。

此外，工作中的数据格式还有卡片形式的、诗词形式的、文摘形式的等等，但几乎都可以归属在“表格”形式、“全文”形式或者它们的混合形式之中。

第二节　关系数据库

关系数据库是当前最流行的数据库管理系统，它具有简单、灵活、数据独立性高、理论严格等特点。目前市场上流行的 dBASE、FoxBASE、Oracle、Sybase 等都是关系数据库产品。

一、关系数据库的数据结构

数据结构是区分数据库类型的重要标志之一，不同类型（层次、网状、关系）的数据库有不同的数据结构。关系数据库的数据结构是“关系”。这里“关系”一词不同于我们日常生活中“关系”的概念，而是一个数学名词。我们不去探究它的数学定义，而是把它理解为“一个二维表”即可。所谓二维表，有时简称为“表”，就是由若干行、若干列组成的数据表。比如上一节中的学生成绩表，就是一个二维表。因此，关系数据库的数据结构是二维表结构，或表结构。

现在我们以上一节的学生成绩表为例介绍二维表这种数据结构的有关概念。

字段：表的列称为字段，或称为数据项，或称为属性等。本例中共有六个字段。这些字段的名字是“学号”、“姓名”、“数学”、“语文”、“历史”、“地理”。填在字段名下面的数据称为字段的“值”。例如“学号”这个字段的“值”有“001”，“002”，“003”三个。

记录：表中填有数据的行称为记录。本例中有三个记录。

第一个记录为：001，张三，70，80，90，95。

第二个记录为：002，李四，95，85，75，65。

第三个记录为：003，王五，80，80，80，90。

由上述定义可以看出，字段值组成记录。而记录组成文件，文件组成数据库。一个数据库至少有一个文件，一个文件可以有若干个记录，一个记录可以有若干个字段。

在数据结构中，字段是很重要的，因为数据是放在一个一个字段中的。字段有两个重要性质，一个是它的数据类型，一个是它的数据宽度（长度）。所谓“数据类型”就是上一节中说过的“数值型”还是“非数值型”。在关系数据库中数值型数据有时也用“N”表示；非数值型数据，即文字型的、符号型的数据，统称为字符型的数据，有时也用“C”表示。“数据宽度”也称为数据的长度。数值型数据宽度用位表示，例如“6”表示数据总长度为 6 位。小数中小数点也占一位，比如拟表示一个 3 位整数、2 位小数的数据，其长度应定义为 6 位。字符型数据长度用字符个数表示。例如“6”表示该字段最长可以为 6 个字符。汉字情况下，一个汉字占 2 个字符，所以宽度为“6”时，表示最多可以存放 3 个汉字。第一节中说过，代码（例如本例中的“学号”）和日期虽然用数字表示，但其数据类型是字符（C）型的。初学者

一定要将这个概念搞清楚。

本例中六个字段，其字段名、数据类型、数据宽度可以说明如下：

字段名	类型	宽度	小数
学号	字符型	3	
姓名	字符型	6	
数学	数值型	5	1
语文	数值型	5	1
历史	数值型	5	1
地理	数值型	5	1

这就是学生成绩数据库的数据结构表。其中，学号、姓名是字符型字段。“学号”宽度为3个字符，以数字表示时，从001到999，可以表示不大于1000名学生的不同学号。如果学生总数超过1000，学号的宽度应定义得再宽一些。“姓名”的宽度为“6”，即可以放3个汉字。一般情况下“姓名”有三个字就够了。但如果学生中有像“欧阳慧敏”、“司徒小聪”这样四个汉字或更多汉字（如少数民族人名）的，那就要把“姓名”的宽度定义得宽一些。“数学”、“语文”等四个字段为数值型字段，宽度为5位，小数点前3位，小数点后1位。数值型数据将来是可以进行运算的，比如求总分、平均分等。

由以上讨论可以看到，关系数据库特别适合处理表格形式的数据。

二、关系数据库的功能

关系数据库的功能，可以分为四个部分：建库功能、查询功能、维护功能、程序设计功能。

建库功能，实际上就是定义数据库数据结构的功能。将一个数据库的数据结构定义好了，该库的存储框架就形成了。以后将一个一个数据装入到这个框架中，就成了一个数据库。

查询功能，就是按一定条件从数据库中查找出符合条件的数据来，查询是最基本的数据库操作。关系数据库可以有多种查询方式，例如：

（1）按某一字段查。如查姓名为张三的各科成绩。

（2）按多字段组合查。如查出语文成绩大于85分，并且数学成绩也大于85分的学生来。这里涉及到两个字段“语文”和“数学”。

（3）按逻辑运算查。逻辑运算指“与”、“或”、“非”的运算。如查出“或数学大于等于85分，或语文大于等于85分的学生”，这就是一个逻辑“或”的运算。

（4）按比较运算查。比较运算指大于、小于、等于、不等于、大于等于、小于等于等运算。如查出历史成绩小于60分的学生。

（5）按数学运算查。算术运算包括加减乘除以及其他一些函数运算。如查出四科总成绩大于320分的学生，这里需对四科成绩求和。

（6）精确查。所谓精确查就是必须给出精确的查询条件才能查。比如“姓名＝张三”。“张三”两个字不能多也不能少，也不能错（例如中间增加一个空格）。上面几项其实都属于

此为试读，需要完整PDF请访问：www.ertongbook.com

精确查询。

(7) 模糊查。所谓"模糊"查,是指记不清字段的全部值,而只记得其中部分值的查询。例如:记不清"张三"的具体名字,只记得姓"张",这种情况就要模糊查:"查出所有姓张的学生成绩"来。当姓张的记录都查出来时,根据查出结果,选出所需的记录。这种模糊查,实际上是一种"批量查",即凡姓张的一批都查出来。

关系数据库的查询功能是丰富的,这是因为它支持逻辑运算、比较运算和算术运算,而由这些运算构成的查询条件是很丰富的。

维护功能,就是对数据库中的数据进行维护操作的功能。一般包括对字段、记录的修改、增加、删除,对数据结构的修改以及其它一些操作。

程序设计功能。一般情况下从数据库中查询出来的结果还要经过处理,例如进行计算或按一定格式打印出来,或形成一个报表等等。做这些处理时就需要编写程序。编写程序一般是使用专门的程序设计语言进行的。但现在的关系库产品中也常常提供灵活方便的程序设计的功能。

三、关系数据库的开发方式

关系数据库有三种开发方式可供选择:

1. 手工编写程序的开发方式

目前大多数数据库应用系统是由计算机专业人员手工编写程序开发出来的。计算机专业人员根据用户需求,使用程序设计语言(如C语言)或数据库管理系统本身自含的命令编写应用程序,开发出完整的数据库应用系统。这种方式的优点是界面漂亮,提供有菜单、窗口等功能,用户有时只要点点菜单,就可以轻松地对数据库进行操作。这种方式适于进行复杂数据处理的情况,或输出复杂报表的情况。但这种方式需要由专业人员编写程序。如果用户自己不会程序设计,而请人来编写,则需要支付一笔对大多数个人用户来说相当可观的费用。如从市场购买类似程序,除花钱之外,以后程序出了故障维修也是个问题。还有,如果应用情况发生了变化,这样的应用程序又需要重新修改,重新投资。

2. 直接使用数据库命令的开发方式

用户不编写程序,直接利用数据库管理系统提供的建库功能、查询功能、维护功能开发出一个应用系统来。这种方式因为不需要编写程序,只需要学习一些有关命令即可进行开发,所以容易些。这种方式灵活、直接,效率也高。但因为没有程序自动控制,所以自动化程度差,用户界面也比较差。如果只是为了个人使用,不追求商品化,也不完成复杂的数据处理工作,这种方式还是一种好的方式,既实用掌握起来又不太困难。本书第六章至第十章介绍这种方法。

3. 使用工具软件的开发方式

有没有一种兼有上述两种方式优点的开发方式呢?有,这就是使用工具软件进行开发。

所谓工具软件,实际上是一个具有开发功能的软件,用户使用这样的软件可以大大减轻编写程序的工作或者根本不用编写程序,就能开发出一个数据库应用系统。好的工具软件开发出的应用系统既有友好、漂亮的菜单、窗口界面,又可以进行复杂的处理,还可以输

出复杂的报表，而开发者本人可以不懂任何程序设计，只要熟悉本专业的业务，回答系统提出的问题就可以了。本书第五章就介绍这种产品。但是这样的开发工具是要花钱购买的。如果经济条件允许，购买一个这样的软件工具，换来方便和开发效率，仍是可取的。另外，工具软件也有不足之处，对于一些特殊的应用开发，也许它不能完成。

四、关系数据库的开发过程

不管采用哪种开发方式开发关系数据库应用系统，其开发过程大体是相同的，其步骤如下：

数据分析：这是开发的第一步。对所要开发项目的数据格式、数据要求、使用要求等进行调查、了解，并进行分析，得出分析结果。有关数据分析在本书第二章中详细介绍。

建库：根据分析结果，选定某一种数据库管理系统，并根据系统提供的操作，建立数据库的框架（数据结构）。按照数据结构要求准备好数据，将数据加载（放入）到数据库中。

查询：数据库建好之后，即可投入应用。根据数据库管理系统提供的查询命令或专门为应用系统编写的应用程序，实现数据库的各种查询。

输出：根据需要将查询的结果按照一定格式显示或打印输出。

维护：数据库建成后，随着不断地使用，可能还要增加新的数据，或修改已有数据，或删去不用的数据。这些操作称为数据库的日常维护工作，它们可以通过数据库的维护操作命令来完成，也可以通过专门的应用程序来完成。

第三节　全文数据库

前面已经介绍过，日常工作中的数据格式主要有两种，一种为表格形式，一种为全文形式。关系数据库的数据结构是二维表，因此特别适合管理表格式的数据。但是，如果用关系数据库管理全文数据，就不方便了。因为篇、段、句子结构的全文数据格式，很难转为字段、记录的表状结构。全文数据适于存储在全文数据库中。

一、什么叫全文数据库

什么叫全文数据库？顾名思义，就是存储全文信息的数据库。全文数据库是相对于前面介绍的表格结构的数据库而言的。

近些年来，随着计算机广泛、深入地应用，产生了大量的电子全文数据。例如，随着办公自动化的开展，机关中的大量公文、报告都在计算机上产生。随着电子印刷技术的推广，印刷出版业大量图书、期刊、报纸改为计算机激光照排，还有电子出版物逐渐上市，电子信箱逐渐开通，电子全文数据资源越来越丰富。另一方面，由于信息化步伐的加快，人们对全文数据的检索要求越来越强烈。人们不再满足于目录文摘的检索，而是希望能够直接获得全文。这样全文数据库迅速发展起来。据有关统计资料表明，1985 年至 1989 年五年间，全世界全文数据库增长了近 50%，在所有数据库中增长率最高。著名的国际联网的美国 Dialog 情报检索系统 300 个数据库中，1990 年就有 100 多个库提供全文检索服务，占数据总数的三分之一。

对于个人用户来说，不管是在工作单位，还是在家庭里，也会越来越多地遇到全文数据库的问题。例如，科技文献、新闻资料、公文报告、法律条文、合同协议、文稿手记、个人文集等等都可能使用全文数据库。

全文数据库的管理软件称为全文检索系统，有的也称为全文文本检索系统，或称为全文信息管理系统。一般说来，全文检索系统至少应该具有以下几个特点：

(1) 对全文数据库进行管理。

(2) 对文章中任何词都可以进行检索，即每个词、甚至每个字都可以作为检索的根据。

(3) 能对检索词之间的关系进行位置运算和逻辑运算。

全文数据库及其检索系统出现较晚，汉字全文数据库及其全文检索系统出现得更晚一些，但发展很快，现已涌现出一批功能很强的产品。本书第三章和第四章将介绍两个产品，这两个产品各有特色，适合广大个人用户使用。

二、全文数据库的结构

全文数据库的结构不像关系数据库那样有明确的定义，而是随全文检索软件不同而不尽相同。这里提出全文数据库结构的问题，是从便于用户使用全文数据库出发的。一般说来，全文数据库分为文献、篇、段、句、字的结构层次。全文数据库由若干文献组成，每篇文献由若干段落组成，每个段落由若干句子组成，而每个句子由若干字组成。但什么叫做篇，什么叫做段落，什么叫做句子，往往随全文检索系统软件的研制者不同理解而不同。加上日常生活、工作中全文的具体式样千姿百态并无严格的规定，所以全文数据库的结构并无明确的定义。例如，有些软件中规定“凡起始缩进 2 个字的就是一个新段的开始”，这似乎符合书本、文章关于自然段落划分的概念。但是一首诗怎么办？诗的每一句都可能缩进 2 个字，能说诗的一句就是一段吗？因此在这里我们不笼统谈论全文数据库的数据结构，而是在使用具体的全文检索系统软件时，再根据具体情况加以说明。

显然，全文数据库适合管理全文格式的数据。

三、全文检索系统的功能

全文检索系统一般提供建库、检索、维护等功能。

建库功能。软件不同建库功能也不尽相同。一般说来建库功能应该包括定义数据库、加载数据(单篇加载或批量加载)、检查不合法数据等功能。

检索功能。全文检索系统一般提供下面的检索功能：

1. 任意字、词查找

即查找出正文中包含有检索字、词的篇、段、句子来。这个检索字、词可以是文章中任意的字、词。

这种查找有时也称为模糊查找。检索者只模模糊糊记得正文中的片言只语，甚至只是几个不连续的字，也可以查找出有关的文献来。任意字、词查找是全文检索基本的、主要的检索操作。

2. 逻辑运算

全文检索系统一般都支持字、词之间的逻辑运算。例如查出既包含有“信息”又包含有“计算机”的文献(“与”运算),查出包含有“信息”或包含有“文化”的文献(“或”运算)。

3. 比较运算

在全文数据库每篇正文之处设置一定的检索项目(有的软件称为属性,有的软件称为表征项),可以实现这些检索项目的比较运算。例如,在正文处设置有“作者”“发表日期”等检索项的,可以实现“作者=祁红”,“发表日期>1990.12.31”的检索操作。

4. 位置运算

所谓位置运算可以理解为:按照被检索字、词之间的位置关系所进行的检索操作。例如:查出文献中包含有“北京”和“公司”两个词的文献,但要求“北京”一词在前,“公司”一词在后,两个词之间最多可以有若干个其它字词。

全文检索系统提供的检索功能是丰富的。熟练掌握各种运算方法,可以构造出各种各样的检索表达式,从而实现各种各样的检索。但是,有一种情况需要注意,一般的全文检索系统不适宜进行数学运算(例如统计、汇总),这是全文检索系统和传统数据库如关系数据库的一个比较明显的区别。

维护功能。全文检索系统的维护功能一般包括增加文件、删除文件、修改文件、优化数据库等。

四、全文数据库的开发过程

全文数据库特别适合于个人用户建立个人文库,因为它的数据准备很简单,向数据库加载数据也很方便,检索方法通俗直观,用户很容易掌握。而且可以适应各种各样的数据格式,不但适合全文文本格式的数据,也同样适合于文摘、目录、卡片甚至表格形式的各种数据。只要不作数学运算,一切格式的数据都可以被看成是一种文本数据而进入全文数据库。

全文数据库的开发过程也很简单。一般包括数据准备、数据加载、数据检索、数据维护几个环节。

1. 数据准备

数据准备,是指对计划加载到全文数据库中的数据进行收集、整理、归类等预先处理的过程。

数据准备的首要工作是数据收集。加载到全文数据库中的数据可以从多种途经获得,有时并不需要花费大量人力去录入,而是对身边的电子数据注意加以收集即可获得。下面是一些常见的数据来源:

(1) 电脑打字产生的文件　在各企事业单位、公司、各类办公室中以及个人电脑打字所产生的文件是一个很宝贵的数据资源。因为这些文件往往是本部门、本单位或个人的工作文件,有很强的使用价值。平时加以收集,零存整取,积少成多,便可以建成一个全文数据库。使用文字处理软件 WS(Word Star)编辑的文件是文本数据,可以直接加载到库。使用 Edline 编辑的文件也属于文本数据。使用 WPS 编辑的文件需先经过转换后加载到库,WPS 本身就有这种转换功能。

(2) 电子印刷产生的文稿　目前很多印刷单位已经屏弃了铅字印刷,实现了计算机

激光照排，或采用了其它电子排版系统。大一点的单位一般也有自己的计算机轻印刷系统或桌面排版系统，其印刷的文件、书稿等都是从电脑上录入的，这也是全文数据库的重要数据来源。采集这些数据时，同样需要先确认是不是文本文件，如不是，应先加以转换。例如国内流行的华光、北大方正的排版系统产生的二扫(.S2)文件，就不是文本数据，需先转换成文本数据(已有这样的转换软件)才能加载入库。

(3) 计算机网上传送的文件　联网的计算机用户可以从网上提供的数据库中获得所需数据，或通过电子信箱获得各种数据。这些数据一般也是文本数据，因此也可以直接加载入库或稍加整理加载入库。

(4) 电子出版物　随着信息化进程的不断提高，社会上将会有越来越多的电子出版物出现。例如年鉴、百科全书、法规、辞典以及重要文献全集等都可能制成电子出版物发行，国外这方面已很发达。直接购买电子出版物建立自己的全文数据库，能更快更直接地获得数据和资料。

(5) 专门组织人力录入建库　对于单位和个人所独有的数据，无法从外界获得，则需要专门组织人力录入。对于单位和个人来说，如果这些全文资料和文献是很必要的、常要查询的，则应该投入人力和财力建立全文数据库。

数据搜集起来之后，要进行一些简单的分类。一般是按照数据内容进行分类。同一类内容将来加载到同一库中，这样便于查找。如果同一类内容的文献存放在不同的库中，将来检索时，就需要跨库进行。有些全文检索系统可能不支持跨库检索，即使支持的，跨库检索也常需要花费更长的机器时间。如果数据总量不大，比如不超过一百万字，也可不进行分类。分类对于数据量大的情况效果比较明显。

数据经过分类之后，还要对每篇的数据格式加以整理。全文数据库原则上可以容纳各种各样的格式，甚至包括表格形式。但不管是什么样的格式，在全文数据库中都是以文本数据存在的。当格式多种多样时，应加以整理，使文献的格式规范化。如标题与正文之间空几行，段落开始行缩进几个字，如果有属性值，应该放在什么位置，如此等等，各篇都应该一致，这样检索的结果才能显示整齐、易于观看。另外，对于目录式文件、文摘式文件、卡片式文件等不同格式的文件，其“段”和“句”如何定义才适于检索需要，才能达到理想的显示效果，也要事先考虑好，确定下来，然后按照确定的规范对不符合格式要求的每篇文献进 行整理。当然，如果平时在产生数据文件时，已经是规范的格式，这些工作也就不必再做了。

2. 数据加载

数据准备好之后，便可以加载到数据库中去了。加载之前，当然需要先定义一个数据库的名字。全文数据库的定义很简单，往往只给一个名字即可。加载数据可有单篇方式或批量方式。单篇方式一次加载一篇，适于平时随有文献随时加载的情况。批量方式一次加载多篇，适于集中大量加载的情况。批量加载时有时对文件名的命名方法有某些要求，原来文件名如不符合要求，需加以修改，这些修改是很简单的。

数据加载后，数据库就算建立起来了。

3. 数据检索

数据库建立起来之后，便可以投入服务了。根据全文检索系统提供的检索功能对数据

库进行检索。具有编辑器功能的全文检索系统还可以一边编辑一边检索，将检索到的结果剪贴过来。数据检索是数据库的具体应用，不同的用户有不同的应用方式，如简单查询、复杂查询、汇编资料等。

4. 数据维护

数据库建立之后，还需要经常对数据库的内容进行更新、追加和清理，以保证数据库的实用性、有效性和时新性。数据维护操作根据全文检索系统提供的维护功能进行。

小　结

本章首先介绍了数据库的一些基本概念：什么叫数据库，什么叫数据库管理系统，什么叫数据库应用系统。接着介绍了日常生活中常见的两种典型的数据格式。一种是表格形式的，一种是文本形式的。表格形式的数据适合建立关系型数据库，因为这和关系数据库的二维表的数据结构相容。紧接着第二节介绍了关系数据库。文本形式的数据适合建立全文数据库，因为这和全文数据库的篇、段、句的结构相容。紧接着在第三节介绍了全文数据库和全文检索系统。但是有一点需要说明的是，建立什么样的数据库有时并不是根据数据外在格式来确定。实际上，用于管理（如生产管理、设备管理、图书管理、学籍管理、病历管理、客户管理等等各式各样的管理）的信息更适于建立关系型数据库，关系数据库提供丰富的字段查询功能和统计、计算功能都是管理工作中常用的功能。而用于文献检索的数据更适于建立全文数据库。全文检索系统关于字、词、句的模糊查询、位置运算等在文献检索中常常使用。也许这才是两类数据库的实质区别。

本章从用户的应用角度出发，将数据库分为两类，表格式的数据库和全文的数据库。这样分类容易为用户所接受，容易根据用户身边的工作数据的外在格式进行数据库的选择。但这种分类不是严格意义上的分类。

本章的基本概念和基本知识将在第二章及其以后章节中得到应用。因此，仔细阅读本章是必要的。

习　题

1. 什么叫数据库？
2. 什么叫数据库管理系统？
3. 什么叫数据库应用系统？它主要由几部分组成？
4. 关系数据库的数据结构是什么样的？哪类数据适合建立关系型数据库？
5. 全文数据库的数据结构是什么样的？哪类数据适合建立全文数据库？
6. 全文数据库中也可以有表格形式的数据，这和关系数据库中的表格形式的数据有什么区别？

此为试读，需要完整PDF请访问：www.ertongbook.com

第二章　数据分析

第一章中我们介绍了数据库的基本概念，这些概念也许还是比较抽象的，我们还不会把身边的工作转化为数据库问题，当然也就无法进一步去求解。这一章就要解决这个问题，这就是数据分析。数据分析的目的就是通过对身边的工作数据进行调查、了解和分析，将工作中的数据管理问题转化为数据库问题，以便进一步去建立数据库、操作数据库。数据分析是初学者入门的关键，是通往建立数据库应用系统的桥梁。

本章第一节概述数据分析的内容和方法，第二节、第三节列举十二个事例，说明数据分析的具体方法。

第一节　数据分析

数据分析包括数据调查、数据分析和书写分析结论等三个方面。

一、数据调查

数据库是对数据进行管理的，因此首先要了解清楚被存放的数据的情况，这就是数据调查。数据调查先从需求开始。需要对什么工作进行管理？这项工作中使用的数据有哪些？数据格式如何？数据量有多大？对这些数据如何使用？常用的查询是什么？常用的报表是什么？如此等等，这些都是数据调查的内容。

数据调查中有一个关键，就是收集、调查原有手工系统下的数据报表。未计算机化之前的手工系统往往是通过手工数据报表方式实现信息管理的。这些报表集中了这项工作的工作数据。因此获得这些报表就获得了数据。而且，从报表中还可以看出数据格式和数据之间的相互关系。数据调查的另外一个关键就是了解清楚经常从哪些方面使用这些数据、查询这些数据。这两点抓住了，数据调查的主体就抓住了。

二、数据分析

在调查的基础上对数据进行分析。数据分析要解决以下几个问题。1. 根据调查的数据格式决定采用什么类型的数据库，即是表格式的数据库？还是全文型的数据库？2. 根据调查的报表种类和数量确定建立一个库还是多个库。3. 将调查的数据格式转化为数据库的数据结构。对于关系型数据库，就是转化为二维表。对于全文数据库，就是确定出篇、段、句的格式。4. 计算出数据量。完成了这几步，数据分析的主要内容就完成了。

分析数据时有两点要注意。一是要弄清原始数据和再生数据的概念。原始数据是指需要原始录入的无法由其它数据生成的、必不可少的数据。这些数据需要录入计算机，计算机才能进行管理。再生数据是指可以由原始数据通过计算产生的数据。再生数据不必由人工录入，可以由数据库产生。例如，学生成绩表中的单科成绩都是原始数据，而总成

绩、平均成绩等就是再生数据，因为它们可以由单科成绩相加或相加求平均而得到。分清原始数据和再生数据的概念，有助于弄清哪些是必须输入的数据，哪些是不必输入的数据。

另一个要注意的是要弄清楚数据库的二维表和输出报表之间的关系。二维表是数据库的数据结构表，由字段、字段类型、字段宽度等组成，必须正确、准确地定义出二维表的各个项目。而输出报表则是用户需要的报表，实际上是把从数据库中取出的数据，按照用户要求加以排列的报表，在格式上和内容上要满足用户的要求。输出报表的格式不一定和数据结构表相一致。对于一个数据库而言，其数据结构表(二维表)是唯一的，不能改变的，改变了就成了另外一个数据库。但输出报表可以是多种多样的，可根据用户的需要将这些数据做各种各样的安排。弄清楚这些概念，有助于在数据分析时，如何从手工报表转化为数据库结构表。

三、书写分析结论

经过数据分析，将分析结果书面化，按照一定的格式书写下来，形成一个报告书，这就是分析的结论。这个报告书主要内容就是根据分析得出的二维表，写出数据库的数据结构表的定义格式来。另外，报告书中还要包括数据库名字、数据量、查询要求等。这个书面报告，就是以后建立数据库的依据。

以上我们从个人用户数据库的角度出发，对数据分析的三个步骤进行了说明。一般说来，掌握了数据分析方法，基本上就可以顺利地“把工作中的数据管理问题”转化为“数据库问题”。需要说明的是，这些分析方法和内容对于大型数据库应用系统的数据分析是不够的，大型数据库应用系统的数据分析比较复杂，已超出了本书讨论的范围。

第二节　事例分析(一)

本节和下一节将运用上一节介绍的数据分析方法，对十二个事例进行分析。这些事例具有典型性。通过对这些事例的分析，初学者可以基本掌握数据分析的方法。

一、简单的学生成绩管理

事例：某中学已开设计算机课，教师要求学生用计算机对学生成绩进行管理。学生们还没有学习过数据库的知识，教师简单地讲了本书第一章和第二章的内容，学生们已经可以作出学生成绩管理的数据分析。

数据调查：全班35名学生，本学期学习七门课程，有：数学、语文、政治、物理、化学、英语、体育，各科成绩最高分为100，考分可以有一位小数。要求计算机能对每个学生的单科成绩、总成绩、平均成绩等进行计算和管理。

查询要求：

(1) 按学生姓名查。例如，查某某同学的各科成绩。

(2) 按学号查。例如查某某学号同学的成绩。

(3) 按单科查所有同学的成绩。例如，查全班每名同学的数学成绩。

(4) 按分数线查。例如，查出全班同学中平均分在 75 以上的学生。

(5) 求出全班各科平均分。例如，求出全班英语成绩平均分。

数据分析：本例要求对学生成绩进行管理，实际上就是对学生成绩表的管理。管理内容包括成绩录入、存储、查询和成绩计算、统计等。根据第一节的介绍，这种数据格式适宜建立关系数据库。全班学生建立一个库，每一个学生是一个记录，每一门学科成绩是一个字段。这样得到下面的二维表。

学号	姓名	数学	语文	政治	物理	化学	英语	体育	总分	平均分

在这个表中，“姓名”、“学号”字段为字符型数据，其余为数值型数据。最后两项“总分”和“平均分”为再生数据，不要求输入。只输入前 9 项(从“学号”到“体育”)数据即可。

数据量：全班 35 人，即 35 个记录。

分析结论：

数据库名称：XSCJK(学生成绩库)

数据库类型：关系数据库

数据结构：

字段语义	字段名称	类　型	宽　度	小　数	注　解
学号	XH	C	2		
姓名	XM	C	6		3 个汉字
数学	SX	N	5	1	3 位整数 1 位小数
语文	YW	N	5	1	
政治	ZZ	N	5	1	
物理	WL	N	5	1	
化学	HX	N	5	1	
英语	YY	N	5	1	
体育	TY	N	5	1	
总分	ZF	N	5	1	
平均分	PJF	N	5	1	

数据量：35 个记录，每个记录 53 个字节，35×53=1855 约 2000 字节

查询要求：1. 按学生姓名查　2. 按学号查　3. 按单科查　4. 按分数线查　5. 求平均分查

说明：这是一个简单的事例，但包括了数据分析的全部过程。由此例可以看出，数据调查就是根据要求找出有关数据。数据分析就是做出一个二维表。而分析结构主要是根据二维表写出数据结构表来，这个结构表是以后建立数据库的依据。

当然，本例只是学生们做的一个练习，还不实用。一个实用的例子见事例十。那里是根据学校成绩管理的具体要求做出的，表格比较复杂，考虑了学年、学期，但数据分析的方

法是相同的。

二、名片的管理

事例：某公司公关部由于业务往来，接受许多名片。过去这些名片存放在名片夹中。现欲由计算机管理，请作出数据分析。

数据调查：本例要求计算机对名片进行管理，即使用计算机中的数据库技术实现名片管理。尽管名片式样不尽相同，但所包含内容基本上大同小异。下面是一个例样：

北京鸿运公司

闻雷人 高级工程师

总经理

中国·北京　　电话：9121485

长安街　　电挂：北京

邮政编码 100032　　传真：9421485

一张名片一般包括有以下项目：单位或公司名称、姓名、职称、职务、地址、邮政编码、电话、电报挂号、传真号。

查询要求：

(1) 按公司(单位)名称查。记不清公司全称，只记得其中几个字，也可以模糊查。

(2) 按姓名查。

(3) 按地址查。

数据分析：本例要求用计算机对名片管理。名片的数据格式是卡片形式的，卡片形式可以转化为表格形式，因而适宜建立关系型数据库。

将名片转化成关系数据库的二维表如下：

单位	姓名	职务	职称	地址	邮编	电话	电挂	传真	BP 机	备注

这个二维表中有 11 项数据，比名片例样中的数据多了 2 项，即“BP 机”、“备注”两个字段。因为例样只是一个例子，不全面，在确定二维表项目时，应该考虑得更全面一点，可以增加一些项目。其中“备注”字段，可填写固定项目未能包括的内容。

数据量：一张名片是一个记录，名片的数量就是记录的数量。

分析结论：经过上面分析之后，可以写出分析结果如下：

数据库名称：MPK(名片库)

数据库类型：关系数据库

数据结构：

字段语义	字段名称	类　型	宽　度	注　释
单位	DW	C	100	国内外用户
姓名	XM	C	40	
职务	ZW	C	20	
职称	ZC	C	20	
地址	DZ	C	100	
邮编	YB	C	10	
电话	DH	C	20	
电挂	DG	C	10	
传真	CZ	C	20	
BP 机	BP	C	12	
备注	BZ	C	30	

数据量：略

查询要求：1. 按单位模糊查　　2. 按姓名查　　3. 按地址查

说明：这也是一个简单的数据库应用的例子。对此例有四点说明：

(1) 看似不是表格形式的数据，其实是可以转化为表格形式的，只要转化成表格形式，就可以使用关系数据库。为什么关系数据库有那么广泛的应用，就是因为日常工作中很多很多的数据都可以转化为表格形式。例如各种卡片、各种目录、各种索引、各种文摘等。实际上，表格形式的数据的特点不在于它的横横竖竖的表格线，而在于它的内容是不是呈现为一个一个字段的形式。只要是字段形式，就可以做出它的二维表，就适宜使用关系数据库。

(2) 数据分析在于得到一个数据结构，而不是具体的数据。所以要学会从一个具体的数据例样(如本例的数据例样是闻雷人的名片)中抽象出它的结构来。如从“闻雷人”抽象出“姓名”，从“北京鸿运公司”抽象出“单位”等等。

(3) 定义字段和字段宽度时，应该考虑全面。数据库结构一旦定义之后，便不再改变(不是不可以改变，而是改变一次比较麻烦)，因此定义之前，尽量考虑周到一些。比如字段个数够不够，不够就增加一些；字段宽度够不够，不够也可以加宽一些。像本例中“单位”字段的长度定为“100”，即 50 个汉字，考虑的是公司可能接受国外客户的名片，如果从国家名称、州的名称再到公司名称、部门名称可能有几十个西文字母，所以定义了 100 个字节。其它几个字段的长度也是考虑到国际情况。但是，也不要随便增加字段、增加宽度。如果增加了又不用，就浪费了存储空间。

(4) 像在第一节中讲过的，要注意数据库的数据结构和数据外在格式这两个概念的不同。数据结构一定要表示成二维表的形式。而外在形式可以是卡片或其它形式。

三、报刊目录的管理

事例：某资料室订有200种报刊，每月编辑一期《报刊目录选编》专辑，汇集某一专题方面的报刊文章目录，提供给读者，颇受欢迎。但因人手少，一年只能编12期、12个专题，同一专题出版相隔时间长，不能满足读者需要，希望采用电脑数据库对报刊目录进行管理，改善手工编选效率，增加专题门类，缩短同一专题出版间隔。请作出数据分析。

数据调查：本事例要求利用电脑数据库技术实现报刊目录管理，并在管理基础上做好《报刊目录选编》，增加期数、增加专题数。现有手工方式编辑的《报刊目录选编》的例样如下：

邮电部部长吴基传畅谈中国建立国家经济信息网的宏图大略
　　罗兰，志明，　　改革与理论，　1994年第4期
关于全球信息基础设施浪潮的几点思考
　　乌家培，科技日报，1994年4月16日

对查询的要求有：

(1) 按作者查。例如，查出某作者近来在各期刊发表的文章目录。

(2) 按标题查。记不清标题全称，只记得其中几个字，需要模糊查。

(3) 按主题词查。即按某一主题内容查出所有文章目录。例如：查出有关“信息高速公路”方面的文章目录来。

(4) 按报刊名称、期号、日期查。例如，查出某某期刊1994年第10期的全部目录。

数据分析：由现有手工方式编辑的《报刊目录选编》式样可以看到，它所包含的数据，如“标题”、“作者”、“报刊名称”等等都是一个一个的字段。在第一节中和事例二中我们都讲到，只要是字段特性的数据就可以看成表格形式的数据。这样由例样可以抽象出下面的数据结构：

标　题	作　者	报刊名称	日　期	期　号	主题词

显然这是一个二维表，因此适宜建立关系型数据库。对于这个关系数据库，如果想开发成有菜单的完整的应用系统，可以选择工具软件(第五章)来完成。如果想使用命令直接操作，可选择dBASE Ⅲ命令方式(第六至十章)来实现。

注意，为了满足前面查询要求中“按主题查”，二维表中增加了一项“主题词”。所谓主题词，就是用来表征文章主题内容的规范词语，检索时输入主题词，即可检索出包含此主题内容的文献。

根据数据分析要求，还需要计算出数据库的数据量来。由前面调查可知，此图书馆有200种报刊，平均每期30个目录，每月约有6000个目录，每年约有7万多个目录。

实现数据库管理之后，平时只要求工作人员将目录录入计算机即可。过去花费大量时

此为试读，需要完整PDF请访问：www.ertongbook.com