文献信息数据库建库技术

戴维民　葛　敏　韩建新
王兰诚　王绍平　刘永丹　著

北京图书馆出版社

文献信息数据库建库技术

戴维民　葛　敏　韩建新
王兰诚　王绍平　刘永丹　著

北京图书馆出版社

图书在版编目(CIP)数据

文献信息数据库建库技术/戴维民等著.—北京:北京图书馆出版社,2001.6

ISBN 7-5013-1777-1

Ⅰ.文… Ⅱ.戴… Ⅲ.文献数据库-基本知识 Ⅳ.G356.1

中国版本图书馆 CIP 数据核字(2001)第 027180 号

书名 文献信息数据库建库技术

著者 戴维民等著

出版 北京图书馆出版社 (100034 北京西城区文津街 7 号)
发行 (010)66126153 传真 (010)66174391
E-mail Btsfxb@publicf.nlc.gov.cn

经销 新华书店

印刷 北京双桥咸宁侯印刷厂

开本 850×1168 毫米 1/32

印张 11.3125

字数 280(千字)

版次 2001 年 6 月第 1 版 2001 年 6 月第 1 次印刷

印数 1—3000

书号 ISBN 7-5013-1777-1/G·486

定价 22.00 元

目　录

第1章 绪 论

1.1 数据库的概念

1.1.1 数据库的定义

现实世界充满着信息。所谓信息泛指反映特定事物的消息、情报或知识,这种消息、情报和知识以可被感受的声音、文字、图像、符号等为表征,并可通过各种方式传播。信息是对客观事物的反映。

以特定的符号表达的信息称为数据。数据是信息的一种量化表示,数据反映信息,而信息依靠数据来表达。表达信息的符号可以是数字、文字或图形。计算机所能接受、处理和存放的信息须是数字化的信息,因此,必须人为地把信息转换成可以被计算机接受的数据,也就是以二进制形式存储在计算机内并被计算机加工处理的数据。

一定数量的数据存储于特定的存储介质之上,以特定的结构相互联结于一体,形成了数据库。数据库是以一定的组织方式存储在一起,相互关联而独立于应用程序之外,并能为多个用户所共享的数据集合。

1.1.2 数据库的特点

1.有序性

数据库不仅存储数据，而且连同数据之间的逻辑关系一起存储，也就是说数据库中的所有信息不是无序的堆积，而是相互之间存在着一定的逻辑关系，构成一个有序的信息系统。正是这种有序性，使得用户可以方便地查检和利用数据库中的数据。

2.共享性

数据库的数据一经输入，即可多次多重输出，可以为多个用户所共同利用，甚至是同时利用。有效地克服了信息传递的时空障碍。

3.独立性

数据的逻辑组织和数据的物理存储方式与用户的应用程序无关，一旦数据结构改变，与这些数据有关的程序不需要重新编写和调试，可节省大量的开支。这样，也使得同一个数据库可以被不同的应用程序所使用。

4.完整性

数据库所聚集的数据应全面地反映特定主题范围内的相关信息，形成完整的覆盖面，而且通过不间断的数据维护活动，始终保持数据的这种完整和正确。

5.最小冗余性

数据库数据之间多维的逻辑结构使得每条数据可以通过不同的途径得以反映，从而避免了数据的重复存储，最大限度地减少了数据的冗余度，节省空间和数据更新的开支，同时保证了数据的一致性，提高了信息存储利用的效益。

1.2 数据库的类型

1.2.1 按数据库性质划分

按其提供数据的性质，数据库可以分为文献数据库、数值数据

库、事实数据库和多媒体数据库。

1. 文献数据库

文献数据库是指报道各类文献及文献信息的数据库，包括书目数据库和全文数据库。

(1) 书目数据库

书目数据库是只存储有关主题领域各类文献资料的书目信息，以二次文献的形式报道文献的数据库，如题录数据库、文摘数据库、引文数据库、期刊目次数据库以及图书馆馆藏目录数据库等。书目数据库以简略的形式向用户报道文献的信息，提供查找、获取文献的线索。这类数据库信息量大，信息密度高，文献报道范围广，数据连续性、累积性强，是用户快速查找文献的有效工具。

书目数据库中的数据来源于期刊论文、会议论文、研究报告、专利文献、学位论文、图书、政府出版物、报纸等各种不同的一次文献，是经过加工、压缩的派生性数据。

书目数据库的数据结构一般比较简单，记录格式也较为固定。因此，其建库速度往往比较快，建设费用相对比较低。

文摘索引数据库和图书馆馆藏目录数据库是最常用的书目数据库。

文摘索引数据库的内容性质与书本式文摘索引相同，主要是简要地通报有关领域特定时期内发表的文章，供人们查阅与检索。它提供确定的文献来源信息，即能准确鉴别相对应的原始文献。但是，它一般不提供原始文献的收藏地点。

图书馆馆藏目录数据库又称为“机读目录”，即 MARC(Machine-Readable Catalogue)。它主要报道特定图书馆实际收藏的各种文献资料的书目信息和存储地址。它既是一般用户查找图书馆资料的工具，又是图书馆业务部门的业务管理工具。它所报道的数据内容详细，除描述文献本身以外，还有许多附加信息，如业务加工信息、管理信息、馆藏信息等，记录格式也比较统一。我国

各类图书馆的馆藏目录数据库都按照《中国机读目录通讯格式》(CNMARC)著录。

(2) 全文数据库

全文数据库是存储文献全文或其中主要部分,以一次文献的形式直接提供文献的源数据库。通常将经典著作、法律条文及案例记录、重要的科技期刊、新闻报道和百科全书、手册、年鉴,以及图书馆所藏的其他重要文献的全部文字或主要文字转换成计算机可读形式,建成数据库。用户可以从中直接检出所需的原始文献。

与书目数据库相比,全文数据库有许多特点和优点,主要表现在以下方面:

①检索直接。能直接检索出原始文献或解决问题所需的文献资料,不必进行二次检索(即根据检出的书目信息再去查找原文)。

②报道详尽。文献的正文部分或附属部分都可以检索获得和显示,用户可以直接查看到文献正文中的每一段、每一句和每一个词,甚至还可以看到某些边缘性信息。

③传递快速。用户可以通过检索系统迅速地浏览、检索和获得文献原文,不受地理位置的限制。

④标引全面,使用简单。绝大多数全文数据库多利用计算机进行全文自动抽词标引,生成倒排档。能为用户提供标题、著者、关键词、摘要等多重检索途径,使用方便。

⑤用户接口多为菜单驱动型,或采用较简单的检索命令,易学易用。

⑥检索语言多用自然语言,少数用受控语言。检索方法除使用布尔检索以外,位置检索占有相当突出的地位。

而且,随着信息处理技术和储存技术的发展和突破,原先困扰全文数据库发展的存储空间、传递速度、文本转换和信息内容稳定性、可靠性等技术问题已得到基本解决,全文数据库获得了迅速的发展和普及,成为数据库技术和产业的重要发展方向。

2. 数值数据库

除了以文献单元为存储和处理单元的文献数据库外，还有一些以数值、事实、术语或图像为存储和处理单元的数据库，即非文献数据库。非文献数据库包括数值数据库、事实数据库、术语数据库和多媒体数据库等。

数值数据库是一种以自然数值形式表示、计算机可读的数据集合，主要记录和提供特定事物的性能、数量特征等信息。其信息报道范围常覆盖某一大类的专业领域。如在军事上，部队的实力、装备、武器的性能等信息；在商业和经济领域中，特定产品的性能特征、价格趋势、国家经济增长率等数值信息；在科技领域，物质的物理化学性质、结构、频谱及实验数据、计算公式等。

数值数据库所记录的数值数据是人们从文献资料中分析提炼出来的，或是从实验、观测和统计工作中直接得到的。数据库生产者将这些数据收集起来，经过核实、检验和加工整理，按一定的方式组织起来，利用计算机进行存储和检索，就成了数值数据库。如果数据库中还含有定义数值或说明这些数据项所必需的文字(文本数据)，则又称为文本—数值数据库。与文献数据库相比，数值数据库是人们对信息进行深度加工的产物，它可以直接提供解决问题时所需的数据，是进行各种统计分析、定量研究、管理决策和预测的重要工具。

数值数据库对所收录的数据的可靠性要求比较高，有时还需要列出数据的误差估计、数据来源和实验条件等，以减少用户使用中的误差。数值数据库存储的数据通常成组排列，他们本身并不被检索，但它们与字母或数字形式表达的可检索的关键词或叙词相联结，成为可检索的数据。数值数据库检索的结果是一个或一组特定的数值。

为了满足用户的不同需要，许多数值数据库还常附有特定的检索软件包，以提供数值的运算、统计分析、分类、排序和重组等功

能。这就使数值数据库系统除了具备一般检索功能外，还常能提供如下一些功能：

(1)数据运算能力。可以根据现有数据建立函数关系，进行数值的计算，例如可以要求系统以某一年为基准年，重新确定时序的指数；要求系统进行不同计量单位的换算；要求计算时达到某种精确度，等等。

(2)图形处理能力。例如，有的系统能提供分子结构式或晶体结构图、光谱图等，有的可以处理非线性座标等。

(3)报表生成能力。能根据用户的需要自动生成各种报表，或对输出数据进行排序和重新组织，或允许转存数据，建立新的文档。

(4)数据分析能力。例如，为复杂的现象构造仿真模型；或利用各种数值分析技术进行预测、可靠性计算和假设检验；甚至可利用对数据的观察和统计分析产生新的思想等。

(5)试验模拟能力。能够模拟具有数学模式的试验；模拟化学反应过程等。

3.事实数据库

事实数据库是存储有关事物(如人物、机构、事件等)的一般指示性描述的参考数据库。因此又称之为“指示性数据库”或“指南数据库”。其主要用途是供用户查询特定事物的发生时间、地点、过程等简要情况。

事实数据库的类型很多。按信息类型划分，常有收录各种人物传记信息的人物传记数据库，收录各种公司的生产与经营活动信息的公司名录数据库，存储各种基金信息的基金指南数据库，存储各种技术标准或规程信息的技术标准指南数据库，存储各种计算机软件目录信息的软件数据库，存储各种产品和商品信息的产品指南数据库以及面向各行业管理的管理信息数据库，存储各种武器装备的技术性能、制造技术、使用方法的武器装备数据库等