

计算机信息存储与检索

高 星 编著



中国科学技术大学出版社

计算机信息存储与检索

高 星 编著

中国科学技术大学出版社

1995·合肥

内 容 简 介

本书系统地概述了计算机科学在信息存储与检索中的应用、信息存储媒介的发展变化、国内外计算机信息检索系统的发展与应用、机读数据库的类型以及机检系统的开发与评价,并详细介绍了著名的国际联机检索系统 DIALOG,给出了检索实例。全书共分 10 章,内容新颖、理实交融、适用面广、便于普及。

本书可作为高等学校信息管理专业、科技信息专业、计算机专业、图书馆学专业、档案学专业、专利学专业、工商管理专业、市场营销专业及其他与信息相关专业的教材或教学参考书,也可作为科技人员、工商管理人員、决策科学人員以及图书管理人員等信息工作者学习检索知识和技能的自学参考书。

前 言

在科技、经济高速发展的今天,面对浩瀚的、飞速更新的信息,如何有效地存储有价值的信息,如何充分利用最新信息,以提高个人、企事业单位、国家的竞争能力,已经引起人们的普遍关心和重视。存储技术、计算机技术和通信技术的结合,大大缩短了时间和空间的距离,为存储和利用信息提供了物质可能。从经济发达国家正着手筹建信息高速公路、发展中国家正积极疏通信息渠道等情况来看,信息的存储、检索及利用已成为迫切需要解决的重要问题之一。

为此,在高等学校培养学生掌握建立各类信息库和开发计算机检索系统的技术,熟悉获取信息的途径和技能,学会正确建立检索策略及准确分析信息的能力,完善知识结构,增强信息意识,已成为必然。这也是作者编著本书的目的。同时,作者还希望本书能有助于提高政治、经济等领域各级决策人员的领导素质水平、更新知识的能力和熟悉业务的能力,以便作出正确的、科学的、具有创造性的决策。

全书共分 10 章。第一章简单阐述了信息的概念;第二章详细介绍了存储技术;第三、四、五、六章集中论述了计算机在信息检索中的应用、机读数据库的类型、文档结构及检索技术、联机检索系统的概况;第七章详细介绍了几种典型的国际联机检索系统以及模拟软件;第八章涉及了最新存储技术光盘的应用;第九章简要介绍了机检系统的开发过程和评价标准;第十章提供了五个检索实例;附录提供了我国现有的部分数据库目录、DIALOG 系统文档专业主题词分类索引、DIALOG 系统常用检索指令和 DIALOG 系统数据库目录。全书各章均附有思考题,适合教学与自学。

0408911

在本书编写过程中,得到了美国 Hawaii 大学李曦、UMKC 大学 Harley 教授、California 大学盛红燕、Texas 州 A&M 大学潘汇源、广电部科学研究院朱智勇、中国科学技术大学研究生罗玉群、王轶男等对本书的大力支持和帮助,在此深表感谢。中国科学技术大学图书馆馆长黎难秋教授在百忙之中仔细审阅了全部书稿,就本书的框架结构和内容编排提出了中肯的意见,在此也一并表示衷心的感谢。

由于本人水平和时间有限,难免有疏漏之处,敬请专家和读者批评指正。

高 星

1995 年 2 月于中国科学技术大学

图书在版编目(CIP)数据

计算机信息存储与检索/高星 编著. —合肥:中国科学技术大学出版社,1995年8月

ISBN 7-312-00680-9

I 计算机……

II 高星 编著

III ①计算机 ②信息 ③存储 ④检索

IV TP

凡购买中国科大版图书,如有白页、缺页、倒页者,
由印刷厂负责调换

中国科学技术大学出版社出版发行

(安徽省合肥市金寨路96号,230026)

安徽省金寨县印刷厂印刷

全国新华书店经销

开本:850×1168/32 印张:11.25 字数:292千

1995年8月第1版 1995年8月第1次印刷

印数:1—4000册

ISBN 7-312-00680-9/TP·96 定价:10.50元

62

目 次

前 言	(1)
第一章 信息概述	(1)
1.1 信息引论	(1)
1.1.1 信息的概念	(1)
1.1.2 信息源	(3)
1.1.3 信息的特点	(3)
1.4 信息处理的演变	(5)
1.2 信息技术和信息社会	(6)
1.2.1 计算机技术	(6)
1.2.2 数据库技术	(7)
1.2.3 网络技术和新型通信技术	(8)
1.2.4 走向无纸社会	(9)
1.3 计算机信息系统	(11)
1.3.1 计算机信息系统的理论基础	(11)
1.3.2 计算机信息系统的开发	(13)
1.3.3 计算机信息系统的基本结构	(15)
思考题	(16)
第二章 信息的存储	(17)
2.1 存储技术	(17)
2.1.1 信息的印刷存储	(17)
2.1.2 信息的缩微存储	(18)
2.1.3 信息的磁存储	(21)
2.1.4 信息的半导体存储	(23)
2.1.5 信息的光存储	(24)

2.2 计算机存储概述	(26)
2.2.1 信息的逻辑结构	(26)
2.2.2 信息的存储结构	(27)
2.2.3 计算机存储系统	(28)
思考题	(35)
第三章 计算机在信息检索中的应用	(36)
3.1 计算机信息系统的组成	(36)
3.2 信息检索自动化发展简史	(36)
3.3 文献存储自动化	(40)
3.3.1 自动标引	(40)
3.3.2 自动文摘	(44)
3.3.3 自动排序	(46)
3.4 信息检索自动化	(47)
3.4.1 计算机检索系统的构成	(47)
3.4.2 脱机批处理	(49)
3.4.3 联机分时处理	(50)
思考题	(52)
第四章 检索系统的数据库	(53)
4.1 数据库系统概述	(53)
4.1.1 数据处理的发展	(53)
4.1.2 数据库系统的特征	(55)
4.1.3 数据模型	(57)
4.1.4 数据库的分类	(58)
4.2 文字型数据库	(59)
4.2.1 书目型数据库	(59)
4.2.2 指南型数据库	(61)
4.2.3 全文型数据库	(71)
4.2.4 新闻报道型数据库	(74)
4.3 数值型数据库	(75)

4.3.1	数值型数据库概述	(76)
4.3.2	数值型数据库的结构与特点	(78)
4.4	图象数据库	(80)
4.4.1	图象处理发展简况	(81)
4.4.2	图象处理	(82)
4.4.3	图象数据库的类型	(84)
	思考题	(85)
第五章	文档结构与检索技术	(86)
5.1	文档结构及检索原理	(86)
5.1.1	顺排文档及检索原理	(86)
5.1.2	倒排文档及检索原理	(88)
5.1.3	索引文档及检索实现过程	(89)
5.2	检索方法	(91)
5.2.1	定性检索	(91)
5.2.2	定量检索和检索条件	(95)
5.3	文档检索技术	(97)
5.3.1	表展开法与检索处理	(97)
5.3.2	顺排文档的检索处理	(101)
5.3.3	用逆波兰表示逻辑检索式	(103)
5.3.4	倒排文档的检索处理	(108)
	思考题	(110)
第六章	联机检索系统	(111)
6.1	联机检索系统的功能与环境	(111)
6.1.1	联机检索系统的功能	(111)
6.1.2	联机检索系统的运行环境	(113)
6.1.3	基本通信设施	(113)
6.2	联机检索系统的服务方式和服务对象	(120)
6.2.1	联机检索系统的服务方式	(120)
6.2.2	联机检索系统的服务对象	(121)

6.3 联机检索的基本步骤与途径	(124)
6.3.1 联机检索的基本步骤	(124)
6.3.2 联机检索的基本原理及途径	(133)
思考题	(135)
第七章 国际联机检索系统	(136)
7.1 DIALOG 系统概述	(136)
7.1.1 DIALOG 系统发展简介	(136)
7.1.2 我国与 DIALOG 系统联机概况	(139)
7.1.3 DIALOG 系统的服务方式	(141)
7.1.4 DIALOG 系统的文档结构	(144)
7.1.5 DIALOG 系统的用户界面	(146)
7.2 DIALOG 系统的基本检索指令	(151)
7.2.1 DIALOG 系统的基本指令	(151)
7.2.2 DIALOG 系统的基本运算符和保留检索功能	(162)
7.3 DIALOG 系统的联机步骤及服务时间	(171)
7.3.1 联机具体步骤	(171)
7.3.2 DIALOG 系统的服务时间和费用	(173)
7.4 DIALOG 系统的数据库	(175)
7.4.1 DIALOG 数据库的来源和类型	(175)
7.4.2 选择数据库的辅助工具	(178)
7.4.3 DIALOG 系统的商情服务	(186)
7.5 DIALOG 系统模拟软件 DIALTWIG	(194)
7.5.1 DIALTWIG 的环境和结构	(194)
7.5.2 DIALTWIG 的检索原理与操作步骤	(195)
7.6 ORBIT 系统简介	(204)
7.6.1 ORBIT 系统的基本检索指令	(206)
7.6.2 ORBIT 系统中的限定检索	(207)
7.6.3 ORBIT 系统的文档结构及基本检索方法	(209)

9.2 系统运行与维护	(268)
9.2.1 日常性维护	(268)
9.2.2 紧急性维护	(269)
9.2.3 完善性维护	(269)
9.3 系统评价	(269)
9.3.1 评价标准	(269)
9.3.2 评价的实施步骤	(273)
思考题	(274)
第十章 联机检索实例	(276)
实例一 检索《中国专利文摘》	(276)
实例二 检索《世界专利索引》	(280)
实例三 检索《美国大型企业名录》	(286)
实例四 检索《欧洲公司名录》	(290)
实例五 检索《化学工业文摘》与《化工厂商》	(302)
思考题	(313)
附录一 国内部分数据库一览表	(314)
附录二 DIALOG 系统文档专业主题词分类索引	(319)
附录三 DIALOG 系统常用指令一览表	(323)
附录四 DIALOG 系统数据库一览表	(325)
主要参考文献	(349)

第一章 信息概述

1.1 信息引论

人类生活的三个基本要素是材料、能源和信息,由这三个基本要素分别形成材料科学、能源科学和信息科学,被人们确认为现代科学技术的三大支柱。它们的发展大大地推动了人类文明进步的进程。

1.1.1 信息的概念

传说上古时期,洪水泛滥,诺亚自造方舟,避免了灭顶之灾。方舟在洪水中飘荡许久,诺亚想知道洪水是否已退,就放出飞鸽。待飞鸽衔回一橄榄枝,诺亚根据飞鸽的返回以及新折下的绿枝这样的信息,推断洪水已退,和平来临了。

自古以来,人们随时都在自觉不自觉地接收、传递、存储和利用信息。

“信息”这一概念最早是由奈奎斯特(Nyquist)和哈特莱(R. V. Hartley)提出来的,当时并没有引起重视。

1948年,美国科学家克劳德·仙农(Claude E. Shannon)在其著名论文《通信的数学理论》中,把“信息”解释为“两次不定性之差”,即通信的意义在于消除某种不定性。该论文成为信息论诞生的标志。

同年,美国另一名科学家维纳(N. Wiener)发表了《时间序列的内插、外推和平滑化》一文和《控制论》一书。他说:“接收信息和使用信息的过程,就是我们适应外界环境发生的一切偶然事件的

过程,也是我们在这个环境中有效地生活的过程。……要有效地生活,就要有足够的信息。”从此,“信息”的概念才被广泛应用。

如今,没有人否认现在的社会是信息的社会。信息活动已渗透到社会生活的各个领域。究竟什么是“信息”?直到现在都没有一个统一的标准定义,各种表述很多。

哈特莱最早认为:信息是指新内容、新知识的消息。

仙农则认为:信息是使不确定性消除的某种东西。

贝克特(John A. Beckett)认为:“信息”就是影响或可能影响用户的决策的有关知识。

唐纳德·桑德斯(Donald H. Sanders)认为:信息是已经整理并有用的数据。是有用处的知识,是经过加工的输出。人们占有这种知识,就可以加深对事物的理解和达到特定的目的,等等。

但有一点是肯定的,即通常所说的原始数据并不能称之为信息。数据和信息的主要区别就在于:一切信息虽然都是由数据组成的,但并非一切数据都能产生消除不确定性而导致有用的信息。

原始数据需要经过加工处理而转变为信息,经传递媒介实现信息应用。变换过程如图 1.1 所示。

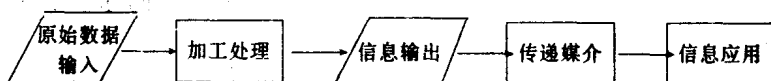


图1.1 变换过程

“Information is power”,这类口号蕴含着深刻的意义。世界上每个人都需要有关信息来指导衣食住行。买主需要比较不同卖主的产品价格和质量的信息;雇员必须获知关于雇主所提供的劳动条件和工资等信息。每一个组织的管理机构要想使该组织获得成功,取决于是否有效地组织了各分支部门之间的信息交流。每一个国家的决策部门同样需要利用内外信息作润滑剂,以使国家机器正常运转。无论是科学技术的研究,还是生产力的发展,都要利用蕴含了新知识和新内容的信息,避免重复无谓的劳动,才能跟得上时代的步伐。

1.1.2 信息源

产生信息的数据源可以依据两个标准进行划分,即地点和时间。根据地点的不同,信息源可分为内源和外源。

内源数据产生于各种组织,如医院、学校、企业、政府机构本身的活动。这些内源可以在常规的和合法的基础上提供事实,即对于内源数据的采集,各组织可以自由地规定采集方法。属于内源的数据有:病床数据、师资设备数据、经济效益数据、生产数据、销售数据、人事数据,等等。

外源又称环境源。外源数据牵涉到组织之外的数据的产生者和分配者。例如,病人和医生、生产者和消费者、销售者和竞争者等等外界的一般经济情况、社会情况和其他环境条件。外源将周围有竞争性的数据提供给决策者。外源数据的采集,需要依赖外部组织。决策者也只能在可以得到的各种数据中选择自己所要求的信息。

根据时间的不同,信息源可分为一次信息源和二次信息源。一次信息源是组织内部的事件和活动。从印刷文献的角度看,是印刷版本的一次文献。二次信息源是已经加工采集的信息集合。联机检索数据库中的索引或文摘文档就是二次信息源。从二次信息源获取的数据称为二次数据或派生数据。

1.1.3 信息的特点

信息活动一般是通过搜集和加工、存储和检索、传递和利用这三大环节来完成的。决策者所得到的信息,越能消除不确定因素,它的价值就越高。

人们对信息的要求决定信息应具有如下特点:

1. 准确性

准确性定义为正确的信息和一段时期内所产生的信息总量之比。对信息准确性的要求因信息的类别而不同。医院中 100 份化

验报告中发生了5次差错,医院和病人都不能容忍;而一年中的天气预报达到0.88的准确性,目前则已令人满意了。

准确性的重要意义十分明了,正确的信息才能有助于正确的决策。

2. 及时性

及时性是信息的另一个重要特点。社会的发展、科学技术的进步已使人类进入了信息的海洋。计算机技术和通信技术为人们提供了及时传递和获取信息的能力,并保证了信息的及时性。人们只要掌握或利用这些技术,就能成为时代的弄潮儿。

例如,某厂计划引进美国生产的成套设备,但对对方提出的价格吃不准。该厂有关人员及时通过国际联机检索,得知该设备技术的专利保护期已到,从而将美方报价压低了几十万美元。

3. 完整性

信息的完整性是必要的。决策者虽然要依靠准确及时的信息,但如果缺乏完整性,仍将导致错误的决策。

例如,一段时间内,许多厂家准确及时地知道某类产品畅销,却缺乏市场总需要量的信息,一哄而上,结果造成市场饱和,产品滞销。

再如,1941年,美国当局虽曾数次截获有关日本搜集珍珠港情报的密码电文,却没有把一个个相关消息综合起来。12月7日,日本偷袭珍珠港得以成功,静谧的军港毁于一旦。

4. 关联性

人们往往可以从一篇文章的标题或文摘中得知其主要内容,而不是先阅读全文后才判断这篇文章是否需要阅读。当相关信息和无关信息混在一起时,就需要加以筛选。不排除无关信息,给决策者带来的只是累赘和负担。

5. 时效性

现代社会中,信息的淘汰周期迅速缩短,决策者不必要掌握曾一度有价值而现在已过时的信息。

1.1.4 信息处理的演变

随着人类历史的进步,信息处理的方式也在不断改变。大致可以分为三个阶段:

1. 人工处理

氏族社会形成后,人们由于需要,开始出现计算行为。例如,牧羊人在绳子上打结代表一头羊;狩猎人在树木上、岩石上刻下一道印痕以示捕获到一只猎物,就是利用绳子、树枝、石子等原始方式来记录信息、交流信息。

2000 多年前,人工计算器——算盘诞生了,这一古老的计算工具流传至今。人们用手、算盘珠、纸笔记录信息和交流信息。随着信息量增加、复杂程度提高,人工处理在准确性和及时性上都不能保证人们的需要了。

2. 半机械化处理

人类在其历史发展过程中,创造出了许多工具和机器,它们完成工作的效率比人工操作高得多,甚至能完成人本身不能完成的工作。

例如,17 世纪中期出现的机械加法器;18 世纪后期出现的打字机,就将人的手算速度和书写速度提高了几倍。

半机械化处理仍离不开人的操作,但确实大大提高了信息处理的速度和准确性。

3. 计算机处理

人类总是在不断创新的,有些创新是划时代的。

1833 年,英国剑桥大学的查尔斯·巴贝奇(Charles Babbage)教授设想制造一个分析机,这实际上是计算机的雏型。

1944 年,美国哈佛大学的霍华德·艾肯(Howard Aiken)教授与助手共同设计了第一台机电计算机 Mark I 型。

1946 年,世界上第一台电子计算机 ENIAC 在美国宾夕法尼亚大学莫尔电工学院诞生了。这一创新是划时代的,是人类进步的