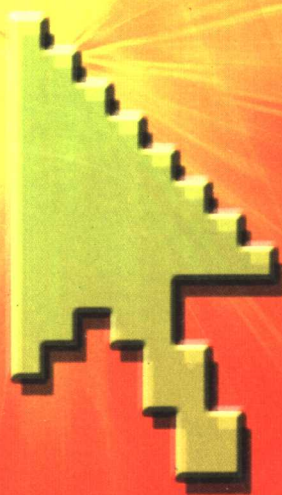


计算机 信息存储与检索

陈次白 蔡 骅 李晓鹏 郭建华 编著



國防工業出版社

<http://www.ndip.com.cn>

计算机信息存储与检索

陈次白 蔡骅 李晓鹏 郭建华 编著

国防工业出版社

·北京·

内 容 简 介

本书在讨论信息技术与计算机信息系统概念的基础上,较为详细全面地介绍了信息的获取技术与存储技术、信息编码(包括压缩编码)技术、信息处理自动化技术、信息系统数据库的特征、计算机文档的组织结构、计算机信息检索技术(包括因特网检索技术和光盘与联机检索技术)。全书涉及的内容较为全面,对当前新的信息技术有较多的体现。通过阅读本书,可以对计算机信息存储与检索的相关技术有较全面的了解和认识。书中各章附有习题,便于教学。

本书可作为大本大专科院校信息管理与信息系统专业、电子商务专业及计算机信息管理等相关专业的教科书或参考书,也可供渴望对计算机信息存储与检索技术有较多认识的读者阅读和参考。

图书在版编目(CIP)数据

计算机信息存储与检索/陈次白等编著. —北京:国防工业出版社,2003.4

ISBN 7-118-03139-9

I. 计... II. 陈... III. ① 信息存贮 ② 计算机应用—情报检索 IV. ① TP333 ② G354.4

中国版本图书馆 CIP 数据核字(2003)第 027957 号

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号)

(邮政编码 100044)

新艺印刷厂印刷

新华书店经售

*

开本 787×1092 1/16 印张 14½ 329 千字

2003 年 4 月第 1 版 2003 年 4 月北京第 1 次印刷

印数:1—4000 册 定价:20.00 元

(本书如有印装错误,我社负责调换)

前 言

人类社会在经历了四次信息革命之后，现在正经历着第五次信息革命。当今世界信息化浪潮汹涌澎湃、席卷全球。纵观 IT 发展史，数字化技术是一项划时代的成就。而数字化技术的发展在经历了以处理器为核心的 CPU 时代和以信息传输技术为中心的网络时代之后，现在以存储技术为重心的数字化技术又掀起了第三次浪潮，我们已进入了以信息存储为标志的 IT 新时代。

尽管电子计算机研制的最初目的是用于科学计算，但目光敏锐的人们已洞察到了它在信息处理领域的应用前景。第一台电子计算机诞生不久，IBM 的工程师就开始研究将计算机应用于信息的自动化处理；CD-ROM 的出现使图书情报界欢呼雀跃；网络的构建，特别是互联网的飞跃发展，更使人们欣喜若狂。计算机的广泛应用使信息不断地增殖，网络的开通使信息日益共享。人们在惊呼：信息爆炸！知识爆炸！大量的信息铺天盖地向我们涌来，它需要人们存储、维护，更期待着人们去检索、开发和利用。信息（内容）的重要性已被越来越多的事实所证明，被越来越多的人所认识。

我们都十分清楚地看到，互联网的发展，特别是在 Internet 的应用中，一个核心的问题就是如何组织、加工、存储信息，特别是如何快速有效地检索到人们所需要的信息。不论采用何种手段，目标就是一个：让信息共享，使信息增殖；让信息使我们获得财富，让信息推动社会的发展和时代的进步。为此，我们编写了此书。

本书注重内容的新颖性，同时又考虑到系统性和继承性，并尽量反映最新的相关技术。全书共分为 8 章，分别就信息的获取技术、存储技术、编码技术、自动化处理技术、信息系统的数据库、信息组织的文档结构、信息的检索特别是网络信息检索技术进行了介绍。但在阅读本书时应时刻关注当时信息技术的最新发展，讲授本书时更应补充最新实例，以便与时俱进、常讲常新。使读者对信息存储和检索的问题有一个较为全面、更为深刻的了解和认识。

全书由陈次白主编（包括大纲的制定、统稿和审定）。其中陈次白参与编写了第 1~8 章，蔡骅参与编写了第 3 章、4 章、7 章、8 章，李晓鹏参与编写了第 5 章、7 章。

本书的编写实际上历经数年的积累和准备，也是多年教学工作的结晶，更是集体努力的成果。在本书撰写过程中得到中国人民解放军工程兵工程学院孙永盛老师的鼎力相助，还得到了宗士增老师的大力支持，研究生刘亚清、刘广海也参与了很多工作，在此一并表示诚挚的谢意。本书撰写过程中参考了大量的书籍、资料、网页信息。由于历时较长，且几易书稿，有些参考文献来源已无法一一录于书后，对此我们除十分歉疚并向他们表示感谢外，还希望得到谅解。

尽管我们已有许多准备和积累，但临到提笔仍有力不从心之感。加之篇幅所限特别是信息技术日新月异发展迅猛，书中论及不到和欠妥之处恐在所难免，诚望读者多多赐教。

作者

2003 年 4 月

目 录

第 1 章 概论	1
1.1 信息系统概述	1
1.1.1 系统的概念	1
1.1.2 信息系统	4
1.2 计算机信息检索系统	5
1.2.1 计算机信息系统概述	5
1.2.2 信息网络	8
1.3 现代信息技术	9
1.3.1 信息技术的定义	9
1.3.2 现代信息技术简介	10
习题 1	15
第 2 章 信息获取与存储技术	16
2.1 信息获取技术	16
2.1.1 图像扫描技术	16
2.1.2 数字照相技术	22
2.1.3 条型码技术	25
2.1.4 触摸屏技术	31
2.1.5 手写输入技术	33
2.2 信息存储技术	36
2.2.1 存储技术概述	36
2.2.2 信息的印制存储	38
2.2.3 信息的缩微存储	42
2.2.4 信息的磁介质存储	47
2.2.5 信息的半导体存储技术	52
2.2.6 信息的激光存储技术	54
2.2.7 电子纸与电子书存储技术	57
2.2.8 存储技术新进展	61
2.3 计算机信息存储结构	62
2.3.1 信息的逻辑结构与存储结构	62
2.3.2 计算机存储系统	63
2.3.3 存储管理与数据备份	64
2.3.4 磁带信息的存储结构与格式	68

2.3.5 磁盘的信息存储结构·····	71
2.3.6 光盘信息存储结构与光盘刻录·····	72
习题 2 ·····	75
第 3 章 信息编码技术 ·····	77
3.1 信息编码·····	77
3.1.1 信息编码的原则·····	77
3.1.2 字符编码·····	78
3.1.3 汉字输入输出编码·····	82
3.1.4 变长码·····	84
3.2 信息压缩技术·····	86
3.2.1 信息压缩概念·····	86
3.2.2 文本信息压缩技术·····	87
3.2.3 图像信息压缩技术概述·····	91
3.2.4 彩色静止图像压缩编码标准·····	93
3.2.5 运动图像压缩编码标准·····	95
3.2.6 音频信息压缩技术·····	99
3.2.7 文件压缩软件 ·····	101
3.3 信息分类编码 ·····	102
3.3.1 信息代码的特点和信息分类原则 ·····	102
3.3.2 信息分类方法 ·····	102
习题 3 ·····	104
第 4 章 信息处理的自动化技术 ·····	105
4.1 信息处理的自动化发展历程 ·····	105
4.2 自动标引技术 ·····	106
4.2.1 自动标引的概念与原理 ·····	106
4.2.2 西文自动标引技术 ·····	108
4.2.3 汉语文献的自动标引 ·····	110
4.2.4 自动标引技术的应用 ·····	116
4.3 自动文摘技术 ·····	117
4.3.1 自动文摘技术发展历程 ·····	117
4.3.2 自动文摘方法 ·····	118
4.4 文献自动分类技术 ·····	121
4.4.1 文本自动分类技术概述 ·····	121
4.4.2 文本自动分类的关键技术 ·····	121
4.4.3 文献聚类技术 ·····	125
4.4.4 分类系统性能衡量 ·····	126
习题 4 ·····	126
第 5 章 信息管理的数据库 ·····	128
5.1 数据库技术概述 ·····	128

5.1.1 数据管理的发展	128
5.1.2 数据库的分类	129
5.2 传统的信息资源数据库	130
5.2.1 文献型数据库	130
5.2.2 非文献型数据库	134
5.3 多媒体数据库与非结构化数据库	137
5.3.1 多媒体数据库	138
5.3.2 iBASE 非结构化数据库	144
5.4 数据库技术的新进展	148
5.4.1 面向对象数据库	148
5.4.2 分布式数据库系统	149
5.4.3 并行数据库系统	150
5.4.4 智能数据库	150
5.4.5 数据仓库	151
习题 5	153
第 6 章 文档结构	155
6.1 两类外存数据的组织	155
6.1.1 两类外存数据	155
6.1.2 记录式文件的基本属性	156
6.2 常用文件的组织	157
6.2.1 顺序文件与逻辑有序的顺序文件	157
6.2.2 索引文件与倒排文件	159
6.2.3 散列文件和相对文件	159
6.3 超文本与流媒体	161
6.3.1 超文本方式	161
6.3.2 流媒体技术	163
6.4 图形文件与其他文件格式	166
6.4.1 网页图形格式	166
6.4.2 电子图书格式	171
6.4.3 其他文件格式	173
习题 6	176
第 7 章 计算机信息检索技术	178
7.1 信息检索技术概述及其发展	178
7.2 布尔逻辑检索与加权检索	180
7.2.1 逻辑检索	180
7.2.2 加权检索	182
7.3 截词检索、限定检索与词表检索	183
7.3.1 截词检索	183
7.3.2 限定检索	184

7.3.3 词表检索	184
7.4 全文检索技术	185
7.4.1 全文检索技术概述	185
7.4.2 全文检索系统中的检索技术	189
7.4.3 全文检索系统软件	189
7.4.4 全文检索技术的应用	190
7.4.5 自然语言检索	191
7.5 搜索引擎技术	194
7.5.1 搜索引擎技术的发展过程	194
7.5.2 搜索引擎技术概述	195
7.5.3 搜索引擎的评价	197
7.5.4 百度网事通搜索引擎技术简介	197
7.5.5 搜索引擎的类型	199
7.5.6 搜索引擎技术的发展趋势	200
7.6 多媒体信息检索技术	203
7.7 信息挖掘技术	206
7.7.1 数据挖掘技术	206
7.7.2 文本信息挖掘	208
习题7	210
第8章 光盘数据库与联机检索系统	211
8.1 光盘数据库概述	211
8.1.1 光盘检索与传统手工检索的比较	211
8.1.2 光盘检索与国际联机检索的比较	211
8.1.3 光盘检索与网络检索比较	212
8.1.4 光盘数据库检索的作用与检索原理	212
8.2 常见的光盘数据库检索系统	213
8.2.1 中文光盘数据库系统	213
8.2.2 国外光盘数据库系统	214
8.3 联机检索系统	216
8.3.1 联机检索系统的服务方式	216
8.3.2 网络检索与联机检索的比较	216
8.3.3 联机检索的基本步骤	217
8.3.4 常用的联机检索系统	218
习题8	221

第1章 概 论

人类社会以其持续增长的加速度，由既往奔向未来。百万年蒙昧，数万年游牧，几千年农耕，几百年工商；现在它正在经历一场前所未有的巨变，由工业时代迈向信息时代！

物质、能量、信息是人类可资利用的三项战略资源。物质可以被加工成材料，为工具构造形体；能量可以被转换成动力，为工具注入活力；信息可以被提炼成知识，为工具提供智慧。美国哈佛大学信息政策研究中心主任欧廷格（A. G. Oettinger）对三者作了如下描述：“没有物质，就什么东西也不存在；没有能量，就什么事情也不发生；没有信息，就什么东西都无意义。”

信息科学技术是 20 世纪科学技术宝库中最为辉煌的领域之一。它和材料科学技术、能源科学技术共同构成了现代科学技术的三大支柱，它们的发展和广泛应用大大地推进了人类文明的进程。

1.1 信息系统概述

1.1.1 系统的概念

1. 系统的定义

系统概念来源于人类的长期实践活动和科学总结。“系统”这个词早在古希腊时代就已使用，亚里士多德关于整体性、目的性、组织性的观点，以及关于事物相互联系的思想，是古代关于系统的一种朴素概念。我国古代思想家老子用朴素的唯物主义哲学思想，阐述了自然界的统一性和整体性。19 世纪以来，自然科学取得伟大成就，使人类对自然界的相互联系的认识有了很大提高。马克思、恩格斯的辩证唯物主义认为，物质世界是由无数相互联系、相互依赖、相互制约、相互作用的事物和过程所形成的统一整体。这就是系统概念的实质。钱学森指出：“系统思想是进行分析和综合的辩证思维工具，它在辩证唯物主义那里取得了哲学的表达形式，在运筹学和其他系统科学那里取得了定量的表达形式，在系统工程那里获得了丰富的实践内容。”他还指出：“20 世纪中期现代科学技术的成就，为系统思维提供了定量方法和计算工具，这就是系统思想如何从经验到哲学到科学，从思辩到定性到定量的大致发展情况。”

“系统”一词来自拉丁语，是“群”与“集合”的意思。在韦氏大辞典中，“系统”一词定义为“有组织的或被组织化的整体”，是“形成集合整体的各种概念、原理的综合”，是“以有规律的相互作用或相互依存形式结合起来的对象的集合”。因此，“系统”可以定义为具有一定功能的、相互间具有有机联系的、由许多要素组成的整体。

依据系统思想建立起来的完整科学体系称为系统科学。它的基础理论是系统学；它

的技术基础是运筹学、控制论、信息论等；它的应用技术是系统工程。

系统学所研究的是一般系统，尤其是复杂系统所具有的概念、性质和演化规律。它所反映的是自然界中各门科学、各个领域系统中共同的东西。系统工程是处理系统的工程技术，其目的是使系统达到整体最优或满意。

2. 系统的特性

系统和其他事物一样，具有本身固有的区别于其他事物的属性或性质，一般可归纳为以下几个方面：

(1) 目的性 系统工作者进行系统的构思、设计、分析与控制。运转时，必须事先弄清其目的性，否则是无法构成一个良好和有序的现实系统的。换句话说，系统工程学就是研究使系统顺利达到某种目的的一门学科。如军事作战系统就应按最易保存自己战胜敌人的目的来配备兵力兵器等各种资源和进行作战的组织指挥等军事活动；经济管理系统就应按如何获得最佳经济效益的目的来优化配置人力、设备和各种资源等。

(2) 整体性 系统应由两个以上的要素或部分组成，各要素或部分之间存在着联系，从而构成一个有机的整体，以实现其目的和功能。系统科学家贝塔朗菲指出：机械论的错误观点之一，就是简单分解和简单相加。他认为应该以整体的观点来纠正过去那种错误地分解的观点。从而提出了关于系统组成的著名定律：整体恒大于各孤立部分的简单之和。

(3) 相关性 科学发展的全部成就证明了现实世界普遍联系的观点。系统中相互关联的要素或部件形成了“部件集”、“要素集”。它集中了各部件或要素的特性和行为相互制约与相互影响的关系，正是这种相关性确定了系统性特有整体的形态与功能。

(4) 复杂性 现代系统一般是多结构、多目标、多功能、多参数、多层次、多输入、多变化的系统。系统通常处在一个多变的环境约束之中，其输入具有多个参数，且表现在时间空间或数值上的随机性和不确定性，系统本身往往具有多结构层次演变，只有进行一系列运算分析和比较，才能权衡出较优的方案。

(5) 适应性 系统与周围环境之间通常都有物质、能量和信息交换。环境的变化会引起系统特性的改变。相应地引起系统内部各要素或部分之间相互关系与功能的变化。因此，一般结构良好的系统必须具有反馈系统、自适应和自学习系统，以保持对客观环境的适应能力。

(6) 动态性 系统的动态性是指其状态与时间的关系。由于物质与运动的不可分离性。各种物质的特性、结构、形态、功能及其规律都是通过运动表现出来的，要认识系统必须要研究系统的运动。开放系统与外界有物质、能量和信息交换，而系统内部结构也可随时变化，因而系统的发展是一个有方向性、周期性的动态反馈过程。

在当今科学技术高度发展的现代化社会里，事物间的联系日趋复杂，出现了各式各样的多种大系统。且这些系统通常都是开放性系统，它们与所处的环境即更大的系统发生着物质、能量和信息等的交换关系。系统同环境的不适应将对系统的存在产生不利影响，这是系统的外部条件要求。从系统内部看，它们通常由许多层次的分系统组成。系统与分系统之间有着纵横交叉的复杂关系。但是不管这些关系如何复杂，有一条基本原则是不变的，那就是下层系统以达到上层系统的目标为任务，横向各分系统必须用系统总目标协调行动，各附属分系统要为实现系统整体目标而存在。因此，任何分系统的

适应或不健全，都将对系统的整体功能和目标产生不利影响。

3. 一般系统及一般系统论

人们在日常生活、工作中所置身的都是具体的系统，如：交通系统、商业系统、金融系统、工业系统、农业系统、教育系统、经济系统、文艺系统、军事系统、社会系统等等。可以这样说，人们不能脱离系统而存在。那么什么是一般系统呢？一般系统指的是，研究一切系统所具有的共同的一般性特征及规律，并不是去研究某个具体系统的组成、具体结构、具体关系及具体运行过程，这种抽象的系统就称其为“一般系统”。研究一般系统的目的是更深刻地认识系统演化的一般规律，以使人们从整体上更好地把握系统的发展。

20 世纪初生物学家贝塔朗菲针对当时流行的只着眼于将自然现象分解的机械论观点与方法，指出“不能只是孤立地研究部分和过程，还必须研究各部分的相互作用，应把生物作为一个整体或系统来考虑。”从而提出了适用于一般化系统的原理及规律，即“它不是属于专门系统的理论，而是适用于一般系统的通用原理”，并称其为“一般系统论”。它的主题表述可导出对一般系统都有效的原理。

一般系统具有哪些普遍适用的一般规律呢？比如系统具有“整体性”，就是说，不能把系统割裂成要素孤立地去研究，应该注意研究要素及要素间的相互作用与相互影响。从系统整体性出发，就可引出一个基本的系统观，即“系统整体功能大于组织系统要素的功能和”。例如，人们常说“三个臭皮匠等于一个诸葛亮”，其意义就在于此。当然，相互作用及相互影响也可能是消极的，数学上称其为“负”的，也会导致“三个和尚没水吃”，其原因是他们的能量消耗在内耗上。再有，一般系统多具有“层次性”，即从系统结构上看是分层的，如行政组织系统分为科、处、局、部、委……；军事组织系统分为排、连、营、团、师、军……都是系统表现出的层次性。还有，一般系统都具有“动态性”，就是任何系统都不是静止的，而是时时刻刻在运动着、发展着。一般系统还具有其他一般性质，如“同形性”等等。

4. 系统论、控制论和信息论

20 世纪 40 年代，由于自然科学、工程技术、社会科学和思维科学的相互渗透与交融汇流，产生了具有高度抽象性和广泛综合性的系统论、控制论和信息论。

系统论是研究系统的模式、性能、行为和规律的一门科学。它为人们认识各种系统的组成、结构、性能、行为和发展规律提供了一般方法论的指导。系统论的创始人是美籍奥地利理论生物学家贝塔朗菲。系统是由若干相互联系的基本要素构成的，它是具有确定的特性和功能的有机整体。如太阳系是由太阳及其围绕它运转的行星（金星、地球、火星、木星等等）和卫星构成的。同时太阳系这个“整体”又是它所属的“更大整体”——银河系的一个组成部分。世界上的具体系统是纷繁复杂的，必须按照一定的标准，将千差万别的系统分门别类，以便分析、研究和管理。如：教育系统、医疗卫生系统、宇航系统、通信系统等等。如果系统与外界或它所处的外部环境有物质、能量和信息的交流，那么这个系统就是一个开放系统，否则就是一个封闭系统。开放系统具有很强的生命力，它可能促进社会的发展和科学的进步。

人们研究和认识系统的目的之一，就在于有效地控制和管理系统。控制论则为人们系统的管理和控制提供了一般方法论的指导，它是数学、自动控制、电子技术、数理

逻辑、生物科学等学科和技术相互渗透而形成的综合性科学。控制论的思想渊源可以追溯到遥远的古代。但是，控制论作为一个相对独立的科学学科的形成却起始于 20 世纪 20~30 年代，1948 年美国数学家维纳出版的《控制论》一书，标志着控制论的正式诞生。几十年来，控制论在纵深方向得到了很大发展，已应用到人类社会各个领域，如经济控制论、社会控制论和人口控制论等。

为了正确地认识并有效地控制系统，必须了解和掌握系统的各种信息的流动与交换，信息论为此提供了一般方法论的指导。语言是人与人之间的信息交流的工具，文字扩大了信息交流的范围，19 世纪电话和电报的发明及应用使信息交流进入了电气化时代。信息论最早产生于通信领域，现在已同物质和能源一起构成了现代文明的三大支柱。

1.1.2 信息系统

信息系统是与信息加工、信息传递、信息存储以及信息利用等有关的系统。信息系统一般包括数据处理系统、管理信息系统、决策支持系统、专家系统以及办公自动化系统和多媒体信息系统。

1. 电子数据处理系统

电子数据处理系统（Electronic Data Processing System, EDPS）是用计算机系统代替人工进行事务性数据处理的系统，所以也称之为事务处理系统（Transaction Processing System, TPS）。电子数据处理系统是由设备、方法、过程，以及人所组成并完成特定的数据处理功能的系统。它包括对数据进行收集、存储、传输或变换等过程。例如，在数据变换这一过程内就有一系列操作都属于数据处理，像数据的识别、复制、比较、分类、压缩、变形及计算活动等。数据处理系统的系统目标与决策无直接关系，是一种纯数据处理系统。

2. 管理信息系统

管理信息系统（Management Information System, MIS）是在事务处理系统基础上发展起来的第二代信息系统。最初的管理信息系统（始于 1961 年）的特点是有了公共数据文件或数据库，初步实现了数据的统一管理和资源共享。到了 20 世纪 70 年代，以数据通信为基础，以方法库和模型库应用为特点的管理信息系统得到了很大的发展。

3. 决策支持系统和专家系统

决策支持系统（Decision Support System, DSS）的概念是美国学者 S.莫顿（S.Morton）于 20 世纪 70 年代初首次明确提出的。决策支持系统是把数据处理的功能和各种模型等决策工具结合起来，以帮助决策的电子计算机信息处理系统。它能够在复杂的迅速变化的外部环境中，给各级管理人员或决策者提供有关的信息资料，并协助决策者制定和分析决策。决策支持系统使用的计算技术是数据库、模型库以及可能进行实时处理的计算机网络系统。其基本特征是能够协助管理者或决策者，特别是协助高层管理者制定决策。系统的重点在于易变性、适应性以及快速的响应和回答。系统允许用户自己启动和控制。决策支持系统面对的是决策过程，它的核心部分是模型体系的建立，它提供了方便用户使用的接口。广泛地建立和应用决策支持系统，将极大地提高决策的科学水平。

专家系统（Expert System, ES）作为人工智能的一个重要分支，初创于 20 世纪 60 年代中期，期间研制的 DENDRAL 和 MACSYMA 系统称为第一代专家系统。70 年代专

家系统趋于成熟并产生了一批卓有成效的专家系统,在此期间专家系统的一个重大进展是 R.Davis 提出的元知识的概念。所谓元知识就是关于知识的知识,或者说元知识就是关于怎样组织、管理、利用、解释和获得系统中知识的知识。Davis 认为,一个好的专家系统不仅需要大量的专门知识去解决问题,而且应该使用元知识来管理、解释、使用和获得系统的专门知识。70 年代中期前的专家系统多属于信号解释型和故障诊断型,70 年代后期开始出现了一些其他类型的专家系统,如设计型、规划型、教育型和预测型等。80 年代则进入了专家系统的黄金时代,专家系统有了突飞猛进的发展,大批新系统不断涌现,应用领域不断扩大并渗透到了政治、经济、军事等重大决策部门。对美国政府各种决策有较大影响的兰德(Rand)公司研制的专家系统主要集中在这几个方面。在传统的信息资源管理机构中,专家系统的应用领域也在不断扩大,情报检索、文献分类、馆藏开发和用户培训等方面都有成功的专家系统原型。

4. 办公自动化系统和多媒体信息系统

办公自动化系统(Office Automation System, OAS)和多媒体信息系统(Multimedia Information System, MMIS)是电子数据处理系统、管理信息系统和决策支持系统等的综合应用。它可为组织决策提供准确及时的信息支持,还可提供方法、手段和模型等多种辅助决策支持。

多媒体技术的发展提供了高速传递大容量数据的重要手段,使地理信息系统(GIS)、计算机辅助设计与制造(CAD/CAM)、医学辅助诊断(MAD)以及远程医疗、远程会议的实现成为可能。可以说,分布式多媒体信息系统是未来信息系统的发展方向。

1.2 计算机信息检索系统

1.2.1 计算机信息系统概述

1844 年,恩格斯明确指出:“科学发展的速度至少也是和人口增长的速度一样的;人口增长同前一代人的人数成比例,而科学的发展则同前一代人遗留下来的知识量成比例。因此,在最普通的情况下,科学也是按几何级数发展的。”科学发展的事实表明:科学认识的发展是一种加速运动;科学认识发展的形态愈高,进化愈快;科学认识的发展是永无止境的。曾有人估算,截止 1980 年,人类社会获得的科学知识的 90%是第二次世界大战后 30 余年获得的。科学的发展,产生了大量的知识和文献,并发展出了新学科。

作为科学知识的基本载体,科学期刊与论文的增长速度是极为迅速的。世界第一份科学期刊,据说是 1665 年问世的《伦敦皇家学会哲学论坛》,此后,期刊数目连续增长:1750 年有 10 种,1800 年有 100 种,1850 年有 1000 种,1900 年有 10000 种,1965 年突破 10 万,现在全世界每天发表论文约 6000 篇~8000 篇。发表论文的数量每隔一年半就增加一倍。而在因特网上发表的信息则更是急剧增长。

19 世纪恩格斯曾经把自然科学分为四大类:力学、物理学、化学和生物学。进入 20 世纪,学科越分越细,专业愈来愈多,边缘学科、交叉学科、横断学科、综合学科相继出现。据统计,现在学科总数已达 1000 多个群体,2500 多门类。

知识的急剧增长被人们形容为“信息爆炸”和“知识爆炸”。有人估计,即使一个技

术人员不停地阅读技术文献,一生也只能阅读本专业领域中技术资料的5%。技术人员在查找资料上所花费的时间与精力大大超过他们在考虑实际的技术问题方面所花费的时间与精力。随着现代化社会生活的节奏加快,对各级各类的管理、决策与信息传送速度要求也越来越高。信息的准确性和重要性往往是以时间为前提的,没有高速、高效的信息传送,信息化将是一句空话。此外,在信息处理中对信息加工的要求也有了质的变化,对信息保存的安全性、可靠性要求越来越高。这就导致信息处理的方法越来越复杂;信息处理中所涉及的知识越来越多,关系越来越复杂。为此,人们早就盼望有一种机器能代替人去做这些信息处理的工作。电子计算机的出现使人类的宿愿成为现实。第一台电子计算机诞生之初,还主要是用于科学计算。但随后的发展,使计算机的应用迅速扩展到了信息处理领域,并成为计算机应用的主流,计算机信息系统随之而诞生,现在的信息系统已离不开计算机了。

信息系统可以不涉及计算机等现代技术,甚至可以是纯人工的。但是,现代通信与计算技术的发展,使信息系统的处理能力得到很大的提高。目前各种信息系统中已经离不开现代通信与计算机技术,所以现在所说的信息系统一般均指人、机共存的系统。

计算机信息系统,是指包括计算机硬件资源、能完成数据采集、分析、加工处理、存储、检索和传输信息的全过程有机整体。具体包括信息资源、物理设备资源、技术处理软件和功能服务软件等等。

不同领域都对计算机信息系统从不同角度进行了研究。计算机和通信领域研究信息交换、信息处理、信息共享和信息传输所需要的技术;行为科学领域研究计算机信息系统对人产生的影响。在技术上涉及到计算机体系结构、操作系统、软件工程、数据库技术、通信与网络技术、数据交换技术、图像处理技术、多媒体处理技术、人工智能技术等等多种学科。同时,计算机信息系统是人类智慧的产物,也是一个依赖于人的系统。

1. 计算机信息检索的发展过程

计算机信息检索的发展大体经历了以下4个阶段:

(1) 脱机信息检索阶段 1954年美国海军兵器中心图书馆建成了世界上第一个计算机信息检索系统。它应用IBM-701计算机实现了军械管理。20世纪50年代至60年代中期处于脱机信息检索的阶段。用户提出检索要求后,必须由专职检索员把用户的检索要求编制成“检索提问式”,并以文档的形式存储在磁带上,提供定题信息检索服务。由于采用单机独立进行,又称为脱机信息检索。

(2) 联机信息检索阶段 随着现代通讯技术的不断发展,到了20世纪60年代中后期至70年代初期出现了利用大、中型计算机,通过通信网络连接远程多个计算机终端的联机信息检索系统,这种系统由于具有分时操作能力,因此多个相互独立的远程终端用户能同时与主机进行“对话”并进行实时联机检索。从而使联机检索逐步进入世界范围,出现了一批著名的国际联机检索系统,使计算机信息检索步入联机信息检索阶段。

(3) 光盘信息检索阶段 从1972年荷兰飞利浦公司最早研制的激光唱盘,到1983年日本首张CD-ROM的问世,光盘作为计算机的外部存储设备引起了世界信息界的极大兴趣。随着1985年第一张CD-ROM数据库产品——BIBLIOFILE《美国国会图书馆机读目录》的诞生,光盘就成为大型脱机式数据库的主要载体。光盘信息检索因此得到迅猛的发展。

(4) 网络信息检索阶段 Internet (因特网) 的诞生、信息高速公路的建设, 使因特网逐渐成为人们进行信息交流的新场所。伴随着公共数据传输技术在科技信息传递领域的应用、联机信息检索的逐步网络化, 使信息检索步入了网络化信息检索的新阶段。

我国的计算机信息检索起步较晚, 20 世纪 70 年代中期, 我国才开始从事计算机信息检索的研究, 1980 年初与 DIALOG、ORBIT 连通并开始提供国际联机信息检索服务。到目前为止, 全国已有 200 多个联机检索终端, 提供与 DIALOG、ORBIT 等 20 多个国际联机信息检索系统的连接与信息检索服务。90 年代光盘信息检索在全国各地科技信息部门得以普遍应用。随着互联网的开通, 特别是因特网的诞生、信息高速公路的建设, 信息检索步入网络化信息检索的新阶段。网络信息检索成为计算机信息检索的一个重要发展方向, 计算机信息检索越来越受到社会各界的广泛重视与关注。

2. 计算机信息检索系统的组成和工作原理

1) 计算机信息检索系统的组成

计算机信息检索系统是以磁带、磁盘、光盘等为存储介质的数据库系统, 检索人员根据某种目的, 在一定的时间内通过计算机或联机网络的终端机, 使用特定的检索指令和检索语言, 利用计算机从数据库中检索所需要信息或文献的系统。它主要由计算机、数据库、通信网络、检索终端和辅助设备等组成。

① 计算机 它是计算机信息检索系统的核心部分, 包括硬件和软件。系统的检索速度和存储容量主要取决于机器的硬件部分。软件负责管理数据库、操作系统和处理检索请求等, 其中检索软件决定整个系统的检索能力。

② 数据库 它是由一系列相关记录组成的信息集合, 是检索系统中的信息源, 可存储在磁盘、磁带或光盘等载体上, 通过检索软件提供数据库内满足用户检索要求的信息资源。

③ 通信网络 它是联系检索终端与计算机的桥梁, 主要起到确保信息传递畅通的作用。

④ 检索终端 它是用户与检索系统传递信息进行“人机对话”的装置, 有电传终端、数传终端和微机终端几种, 现在普遍使用的是微机终端。

⑤ 辅助设备 辅助设备主要有打印机、光盘驱动器、磁盘驱动器、磁带机、调制解调器、扫描仪、缩微机等。

2) 计算机信息检索系统的工作原理

计算机信息检索的实质一般是“匹配运算”, 与手工检索的基本原理相同。当然手工检索时的检索提问无须“提出”, 只在自己的意识之中, 用手翻阅检索工具, 利用眼看和大脑思维来对检索工具中的文献款目进行“扫描”、“匹配”, 找出命中文献。而在计算机检索时则是把提问变成机器能识别的检索提问式并输入计算机中, 由计算机自动对数据库中各文档的记录进行扫描、匹配。实质上, 就是把检索要求变成检索提问输入计算机, 与数据库中信息记录的特征进行类比、组配, 把能够匹配的信息记录检索出来的自动化过程。随着人工智能技术的发展, 计算机检索技术开始突破所谓“匹配运算”, 从认知和理解的角度进行检索。

计算机信息检索与手工检索相比, 检索点较多、检索速度快、查全率较高、需要懂得机检知识、检索综合效率较高。

1.2.2 信息网络

在现代术语中,网络可定义为一组个体或组织为实现某种特定目标相互联系而形成的一个系统。如图书馆之间为更好地实现资源共享和更好地为用户服务而组成的图书馆互借网络。又如通过通信线路把分散的计算机资源和信息资源连接起来的网络。后者是一个狭义的基于计算机系统的网络。

一般认为,信息网络主要是指各种信息系统所构成的信息网络体系,以及为了支持这个网络体系有效运行所需要的网络标准、通信协议、操作规程、传输编码等。信息网络用于信息资源高速地传递、交换和共享。因此,信息网络既包括通信网络的建设,又包括信息资源网的建设。计算技术、通信技术、网络技术、微电子技术等现代信息技术的结合,对人类的信息处理方式产生了全面的影响。例如:

- 信息传递时效加快,远程登录是对信息的实时操作,电子邮件虽为异步传输,但可在很短时间内向世界任何一地完成内容的发送;

- 信息传播范围可大可小,信息传播的双向性和交互性增强;
- 同一信息可采用多种格式保存;
- 同一信息几乎同时可出现在不同地点,供多人共享。

总之,人们正在朝任何时间任何地点与任何人交换任何形式的信息的目标迈进。

1. 计算机网络

计算技术与通信技术结合,构造了信息处理与信息通信融为一体的信息网络。经过半个世纪的开拓,特别是近几年 Internet 的快速发展,信息网络已定位为一个国家/地区的信息基础设施,其规模和技术水平是一个国家现代化水平和综合国力的重要标志。网络技术包括网络系统技术、网络传输与交换技术、网络终端与接口技术、网络业务与网络应用技术。

根据国际电信联盟(ITU)提出的全球信息基础结构(GII)体系结构的框架性建议,可将网络结构进行垂直和水平的描述。垂直描述是从功能上将网络分解为应用层、业务网和传送网等,简称网络的功能结构。水平描述则从用户如何接入网络的实际物理连接来划分网络,可将网络分解为用户驻地网(CPN)、接入网(AN)和核心网(CN)。

网络传输可以分为有线传输和无线传输两大类。信息网络最基本的任务是传输和交换信息。为了交换信息发展了电路交换和存储转发两类信息交换方式。传统的网络终端有电话机、传真机、电视机、计算机等。它们是伴随着电信网、广播电视网和计算机网各自独立发展而发展的。数字化浪潮推动着上述网络融合为数字化信息网络。网络终端也向功能高度综合、具有更强的渗透力、更适合人的需要方向发展。PC机既是信息检索、网络计算的工具,同时又可作为IP电话、IP电视。以信息家电为基础的终端设备将比PC机更为普及。移动式和佩戴式终端创造出无论何处都可以上网、都可以计算的环境。信息网络的终端将越来越普遍地利用多媒体技术,实现人机接口的多媒体化。人工智能的应用将产生更多新的网络业务和网络应用。

2. 信息资源子网

信息网络建设,最终是信息资源开发的过程。没有信息流通的网络将是无用的网络。人类信息资源的开发方式是计算机信息处理系统的发展交织在一起的,并大致经历了

以下阶段。

(1) 独立的信息处理方式 这即为最初的单机处理阶段。这种方式可解决信息资源有序化、规范化处理和检索效率问题，但存在着信息交换、资源共享和统一管理等方面的困难。

(2) 集中的信息处理方式 亦即主机/终端方式。这种方式形成了最早的联机网络。其优点是信息组织效率高、资源共享方便，但价格昂贵且数据更新难度大、时效性差。

(3) 自上而下的信息处理方式 它采用数据库技术和广域网技术，根据行业和管理需要建立专门的信息网络。这种方式的传输手段有所进步，数据可及时更新，但从长远看不易保证数据的完整性、时效性和正确性。

(4) 分布处理与统一共享方式 这种方式采用了 Internet 有关技术，在不同节点、不同用户开发独立的信息资源。资源分散存放和维护，采用统一的技术标准，从而实现了信息资源在更大范围内共享。其优点是信息传播快、范围广，但信息资源的规范性和安全性都比较差，易受攻击和破坏，难以在较大范围内进行有效的组织和管理。

总之，基于 Internet 和 Intranet（内部网）技术的发展为信息资源的开发提供了新的手段，推动了信息网络的开发和建设。

1.3 现代信息技术

1.3.1 信息技术的定义

一切科学技术都是人类社会生产实践发展到一定阶段的产物。人类通过自己的各种器官同自然界打交道，而当人类各种器官功能的实际水平同所要求的水平之间存在某种差距时，这种差距就成为一种动力，使得人类一定要找出各种办法来消除这种差距。人类在实践中摸索的结果，会导致对自然规律认识的深化，并找到扩展人类相应器官功能的实际办法。前者就是科学的发展，后者则是技术的进步。

辅人律是科学技术发展的规律之一，即技术的功能是辅助人的各种器官的功能的，这称为技术功能论。按照技术功能论的思想，可以给出信息技术的定义：一切能够扩展人的信息器官功能的技术就是信息技术。

联合国教科文组织（UNESCO）给信息技术的定义是：应用在信息加工和处理中的科学、技术与工程的训练方法和管理技巧；上述方法和技巧的应用；计算机及其与人、机的相互作用；与之相应的社会、经济和文化等诸种事物。

从以上关于信息技术的定义可做如下理解：

(1) 信息技术是人类信息器官的功能扩展，是人类神经系统的延伸。工业革命带来的机械动力设备扩展了人的肢体的力量；信息革命带来的计算机及其他信息设备则延伸了神经系统和扩充开发了人类的智慧。

(2) 信息技术一般所指的是“一系列与计算机相关的技术”。

(3) 这些技术能够对数量巨大的、格式变化的、分布在不同地点的各种信息进行记忆、处理、展示、发送和使用。

(4) 信息技术越来越多地同文本、图形、声音和动态视频等多种媒体的变换相联系。