21世纪 21世纪高等院校规划教材

陈次白 丁晟春 颜端武
李晓鹏 王 瑛 刘广海 编著

信息存储与检索技术

（第2版）

国防工业出版社
National Defense Industry Press

21 世纪高等院校规划教材

信息存储与检索技术

（第2版）

陈次白　丁晟春　颜端武
李晓鹏　王　瑛　刘广海　编著

国防工业出版社
·北京·

图书在版编目(CIP)数据

信息存储与检索技术/陈次白等编著. —2 版. —北京：
国防工业出版社,2008.8
21 世纪高等院校规划教材
ISBN 978-7-118-05862-8

Ⅰ.信… Ⅱ.陈… Ⅲ.①信息存贮-高等学校-教材
②情报检索-高等学校-教材 Ⅳ. TP333 G252.7

中国版本图书馆 CIP 数据核字(2008)第 104550 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 17½ 字数 398 千字
2008 年 8 月第 2 版第 1 次印刷 印数 1—4000 册 定价 30.00 元

(本书如有印装错误,我社负责调换)

国防书店：(010)68428422 发行邮购：(010)68414474
发行传真：(010)68411535 发行业务：(010)68472764

前 言

江河行地，日月经天；生命轮回，万物争辉；时空变幻，气象万千。我们生活在一个事物千差万别的世界，我们生活在一个充满矛盾的宇宙，这就是信息的世界，这就是“赛伯”空间。随着事物在时间和空间中运动，信息在产生和扩散。古往今来，人们采用各种可能的手段保存信息，使人类历史得以不断传承。有时，我们觉得信息实实在在，俯拾皆是，触手可得；有时，又感到虚无缥缈，无所适从，遥不可及。有时发现信息十分稀缺，有时又变得应接不暇。信息的数量是如此巨大，信息的形式（格式）是如此纷繁复杂，信息的重要程度又如此之高，必须要存储保护好；信息如此杂乱，应当进行序化优化。

有人说过，知识的一半就是知道如何去获取它。面对信息的世界、知识的宝库，我们应如何去找寻、如何去发掘、如何去利用，这就是我们面临的课题和任务。在数字化的空间，在网络化的环境下，信息如何存储、如何设计和构建检索系统以便顺利地获取，这就是本书要讨论的问题。

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。纵观 IT 发展史，数字技术是一项划时代的成就。而数字技术的发展在经历以处理器为核心的 CPU 时代和以信息传输技术为中心的网络时代之后庄严地宣告：以存储技术为重心的数字技术又掀起了第三次浪潮，我们已进入了信息存储为标志的新时代。

电子计算机的出现，特别是计算机网络和通信技术的发展，使人们处理信息的能力大大提高。人们发现：网络，特别是互联网络（因特网），实际上就是一个信息的海洋，一个知识的宝库。互联网的发展，特别是在因特网的应用中，一个核心的问题就是如何组织、处理、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，使信息增值；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。

本书是在 2006 年出版的《计算机信息存储与检索》一书的基础上，经过进一步精简、更新，加强技术性，充实新内容而成。它既可以看做是原书的再版，又可以认为是重新编著。目前，关于信息存储和信息检索方面的书已有多种。其中大多侧重于检索系统的操作使用；有些虽然以技术为主，但又不够全面。本书以技术，特别是计算机技术为主，兼顾传统的技术（如印刷、缩微、条码等技术），从存储设备、存储原理、数据模型、检索算法、检索原理到文件构造等多方面的技术问题均有论述。

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 9 章，分别就信息获取技术、信息存储技术、信息编码技术、文本信

息处理的自动化技术、文件组织与文件格式、信息检索模型、信息检索技术和信息检索系统及其应用等进行了讨论。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，以跟上信息技术发展的步伐。

全书由陈次白主编、统稿和审定。其中丁晟春参与编写了第4章、第6章，颜端武和李晓鹏参与编写了第5章、第7章，王瑛参与编写了第9章，刘广海参与编写了第8章，陈次白编写了第1章、第2章、第3章、第9章并参与编写了其它各章。

本书的编写实际上历经数年的研究和积累，也是多年教学工作的结晶，更是集体努力的成果。在本书撰写中，研究生刘亚清、李毅博、刘丽娟也参与了很多工作，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、杂志、网页信息等各种资料。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，但临到提笔仍觉力不从心。加之篇幅所限，特别是信息技术日新月异发展迅猛，书中对很多问题，特别是一些新课题和理论性较强较深的问题没有展开讨论。此外，还有许多论及不到和欠妥之处，诚望读者诸君多多赐教，提出宝贵意见。

作 者

2008年7月

目 录

第1章　信息检索概论

物质、能量和信息是人类可以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，什么东西也就不存在；没有能量，什么事情也就不能发生；没有信息，什么东西也就无意义。"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　信息与信息系统

1.1.1　信息概述

虽然人类自古以来就在利用信息，但是，人类认识和研究信息的概念和内涵却是近百年内的事情。直到20世纪40年代，在美国数学家克劳特·香农（C. E. Shannon）创立了信息论以后，"信息"一词才成为一个科学的概念。但对于信息的含义，至今仍是众说纷纭，莫衷一是，各种信息定义都从不同侧面反映了信息的某些特征。

在日常生活中，人们所说的"信息"，是指音信、消息和情况，是人们在相互交流中要告诉对方的某种内容。在西方国家的文字中，信息一词来源于拉丁文"Information"，大致解释为消息、情报、知识、见闻、通知、事实、数据等。这些解释基本上都是从字面上来理解的。

信息论创始人C. E. Shannon从研究通信理论出发，认为信息是在通信的任何可逆重新编码或翻译中保持不变的东西。控制论创始人科学家维纳（N. Wiener）提出，信息是在人们适应外部世界，并且使这种适应为外部世界感觉到的过程中，同外部世界进行交换的内容的名称。控制就是复杂的、有组织的系统在外界环境发生变化时，能够根据"变化"进行调整。在控制的过程中，控制系统必须及时得到外部环境的信息、系统自身各组成部分的状态信息以及控制效果的反馈信息，并对所得到的信息进行加工和处理，不断发出指令信息，保证控制系统的正常运行。因此，可以说控制的过程就是信息输入、加工处理和输出的过程。维纳在1948年发表的名著《控制论——动物和机器中的通信与控制问题》一书中曾经指出"信息就是信息，不是物质，也不是能量。"

从概率的角度看，信息是用以消除不确定性的东西，即人们把关于事物的某种东西传给对方，使之消除知识上的不确定性。信息是系统的组织程度或有序程度的标记。该定义是通过与热力学中的概念"熵"进行类比推理而来的。人们常用熵来表示系统的无组

织状态或无序状态，信息作为与“熵”相对的概念提出来，也即“负熵”。

信息是数据处理的结果。这个定义是从信息处理的角度讲的。它把未经过加工的原始资料，无论是数字、文字，还是符号、图像、信号，都称为数据，而把信息理解为加工原始资料后得到的、便于使用的结果。

“信息”概念的广泛应用，引起许多哲学工作者对信息本质的探讨，使“信息”从一个科学概念上升到一个哲学范畴。他们认为，信息是以物质能量在时空中某一不均匀分布的整体形式所表达的物质运动状态和关于运动状态所反映的属性。事物都是在时间和空间中运动的，在运动过程中会发生时间和空间的变化，变化即是信息。

1.1.2 系统与信息系统

1. 系统的定义

系统(system)一词最早出现在古希腊语中，希腊文“sys-tema”指的是由部分组成的整体。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古代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，在运筹学和其它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”一词在拉丁语中是“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

系统和其它事物一样，具有本身固有的、区别于其它事物的属性或性质，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，否则无法构成一个良好和有序的现实系统。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。

(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、多输入和多变化的系统。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、自适应系统和自学习系统，以保持对客观环境的适应能力。

（6）动态性。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、能量和信息的交换，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。

1.1.3 信息系统

1. 信息系统的构成

输入原始信息，经过加工处理后，输出各种信息的系统就是信息系统。信息系统一般由信息搜集子系统、信息加工子系统、信息存储子系统、信息传播/通信子系统和提供信息子系统构成。

1）搜集信息子系统

信息搜集就是通过各种渠道广泛搜集，用一定的方法、鉴别、分析、选择和获取信息的活动。搜集信息应遵循的原则是：① 准确性原则，保证搜集信息的准确性；② 时效性原则，保证以最短的时间，最快的速度及时搜集信息；③ 连贯性原则，保证所搜集信息全面完整，主次分明；④ 开拓性原则，在信息搜集过程中要具有开拓精神，善于捕捉信息、获取信息和开发信息的价值。

信息搜集的方法主要有行政手段、经济手段、法律手段、技术手段等。

2）信息加工子系统

由于信息资源数量非常巨大，且来源广泛，要进行有效的匹配与选择，首先要对信息集合进行组织，使之按一定顺序组织起来，即有序化问题。有序就是信息按一定的规则排列起来。使信息按照一定的规则排列起来的方法通常称为情报检索语言。

情报检索语言在信息加工过程中用来描述信息的内容特征（或外表特征），从而形成检索标志；在检索过程中用来描述用户的检索提问，从而形成提问标志；当检索标志和提问标志完全匹配或部分匹配时，即为检索到的所需信息，信息检索标志的集合就是检索工具。

情报检索语言按其词形可分为分类型检索语言、语词型检索语言、代码型检索语言和

引文型检索语言。按情报检索语言的组配程序可分为先组式语言和后组式语言。分类法是按文献内容的知识属性来描述信息的一种信息处理方法，比较常见的是体系分类法。

语词型检索语言不考虑学科门类，按词去组织信息集合和检索工具。语词型检索工具最早是标题语言，而后经过元词法发展成为叙词法。除此以外，还有主要由计算机实现的关键词法，保持上下文主题法等。例如，中国人民大学图书馆图书分类法（人大法）、中国科学院图书馆图书分类法（科图法）、中国图书馆图书分类法（中图法）和中国图书资料分类法（资料法）。国内主要叙词表有中国科技情报所和北京图书馆于1980年编制的“汉语主题词表”，原国防科工委情报所于1985年编制的“国防科技叙词表”，化工部、机械部、电子部等都编制过本领域的叙词表。

对存在于一定载体的信息外表特征进行加工的过程称为著录。对信息的内容进行分析，根据主题法或分类法给出主题标志或分类标志的加工过程称为标引。计算机信息加工就是利用计算机编制目录，建立数据库的过程。

3）信息存储子系统

信息存储有着悠久的历史。自古以来，人们一直探索着记录和保存信息的方法与载体。结绳记事可以说是人类最早的存储信息的方法。信息大量存储的实现可能还是在文字、纸和印刷技术发明之后，但信息存储技术真正的飞跃还是近百年以来的事。

就人体来说，收集信息是感觉器官的功能，传递信息是神经系统的功能，存储信息是大脑功能的一部分。大脑存储信息的功能称为记忆。直到今天，纸印刷仍是信息存储的主要方式。纸印刷存储按存储方式属于机械存储，即以物质的机械形变（或涂覆）来存储信息。人们不仅以文字的形式来记录信息，还以图像的形式直接拍摄各种活动，这种存储方式为光存储，即利用光学手段对信息进行存储，如光盘存储、全息存储、缩微存储等。此外，还有利用磁性物质的磁性来存储信息的磁存储，利用电荷来存储信息的半导体存储。

总之，按存储的方式，信息存储技术可分为机械存储、光存储、磁存储和半导体存储。按信息存储的内容可分为文字存储、代码存储、数字存储、声音存储、图像存储、景象存储。数字技术的发展使得各类信息均可利用“0”、“1”进行综合存储和处理。

4）信息传播/通信子系统

信息传播子系统就是把信息从一个地方传到另一个地方的系统，又称为通信系统。从广义来说，各种信息的传递均可称为通信。绝大多数的通信是以电流或电磁波为载体而传递信息的，因此通信也称为电信，现在把光作为载体的光通信也得到了广泛应用。现代通信的发展，尤其是计算机网络技术的发展，使人们可以突破时空的局限，便捷地获取各种各样有用的和及时的信息，通信系统已经是集信息应用和计算机网络于一体的现代通信网络。

5）信息提供子系统

信息系统的最终目的是为用户提供信息服务工作，信息提供子系统作为信息系统与用户的接口，担负着信息搜集、加工处理、存储和传输各子系统功能的最终集成与实现的任务。信息提供子系统包括信息系统为用户所提供的服务和信息系统为用户输出信息的方式。

2. 信息系统的演进

信息系统由人、设备、信息、规则等要素组成，实现信息的搜集、整理、加工、处理、传

递，提供利用的综合体。其本身也具有不同的类型：

（1）按信息系统的规模划分，可分为小型信息系统、中型信息系统和大型信息系统。

（2）按信息系统的分布范围划分，可分为局域网、城域网、广域网、国际互联网和国家信息基础结构。

（3）按信息系统所属的领域划分，可分为工业信息系统、经济信息系统、科技信息系统等。

（4）按信息系统的使用范围划分，可分为专用信息系统和公共信息系统。

系统的性质是多方面的，因此信息系统的划分也是多方面的，并且不同类型的信息系统可以相互转化。信息系统不仅存在着不同的类型，而且也存在着产生、成长、衰老和更新的变化过程。新型信息系统的不断问世，反映了信息系统的演化过程。

从各类信息系统产生的时间来分析，最早是在20世纪50年代产生的数据传输加工系统（TPS），在60年代产生了管理信息系统（MIS），在70年代出现了情报检索系统和办公自动化系统，在80年代出现了决策支持系统和专家系统。

1.2 计算机信息检索

1.2.1 信息检索简述

1. 信息检索

信息检索的概念有狭义和广义之分。广义的信息检索包括信息的存储和检索两个过程（Storage and Retrieval），全称又叫做“信息存储与检索”（Information Storage and Retrieval）。信息存储是指工作人员将大量无序的信息集中起来，根据信息源的外表特征和内容特征，经过整理、分类、浓缩、标引等处理，使其系统化、有序化，并按一定的技术要求建成一个具有检索功能的工具或检索系统，供人们检索和利用；而信息检索是指运用编制好的检索工具或检索系统，查找出满足用户要求的特定信息。

狭义的信息检索则仅指该过程的后半部分，即从某一信息集合中找出所需信息的过程，相当于人们通常所说的信息查询（Information Search）。

2. 文本信息检索

作为检索对象的文本信息具有不同的形式，有的以文献的形式出现，有的以数据或事实的形式出现。根据检索对象的形式不同，信息检索又分为文本文献检索和数据检索。凡以文本文献（包括文摘、题录或全文）为检索对象的，就叫文本信息检索（Document Retrieval）。凡以数据或事实为检索对象的，则是数据检索（Data Retrieval 或 Fact Retrieval）。

文本信息检索是信息检索的一部分，是其中最重要的一部分。从性质上说，文本信息检索是一种相关性检索，系统不直接解答用户所提出的技术问题本身，只提供与之相关的文献供用户参考。数据检索则是一种确定性检索，系统可直接回答用户提出的问题，即直接提供用户所需要的、确切的数据或事实；而且检索的结果一般也是确定性的。

依据检索对象的不同，文本信息检索可分为：以查找文献线索为对象的文献检索；以查找数值与非数值混合情报为对象的事实检索；以查找数据、公式或图表为对象的数据检索；以查找文献全文为对象的全文检索。

3. 信息检索的基本原理

信息检索的基本原理是：通过对大量的、分散无序的文献信息和多媒体信息进行搜集、加工、组织、存储，建立各种各样的检索系统，并通过一定的方法和手段使存储与检索这两个过程所采用的特征标志达到一致，以便有效地获得和利用信息资源。其中存储是为了有效的检索，而检索又必须先进行合理的存储。

存储是检索的基础，检索信息是存储信息的相反过程。了解检索系统的结构和组成，有助于人们对各种检索系统、检索工具特征的认识，从而正确选择检索系统与工具，改善信息检索的效果。

文献信息的存储和检索的一般过程如图 1－1 所示。

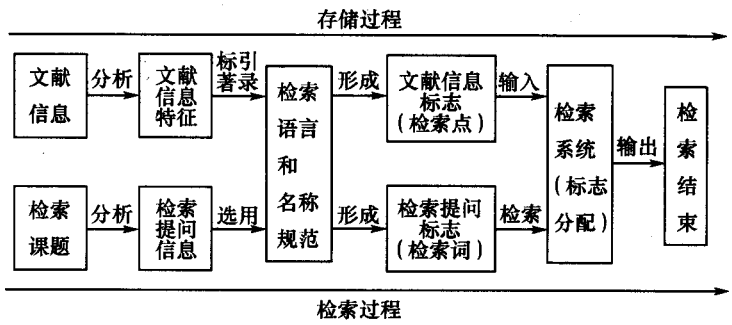

图 1－1　文本信息的存储和检索的全过程

4. 多媒体信息检索

多媒体信息包括文本、图形、图像、声音、视频等多种不同类型的信息，对多媒体信息的检索可以采用关键字标引后，像文本信息那样去进行检索；但更多的是采用基于内容的检索方式，也就是首先要对多媒体信息流进行分析，提取多媒体的特征，然后进行多媒体数据的分割和识别分类，最后对识别出来的语义建立索引，提供检索。多媒体信息分析检索流程如图 1－2 所示。

图 1－2　多媒体信息分析检索流程

在基于内容的检索中，由于特征值为高维向量，不具有直观性，因此必须为其提供一个可视化的输入手段。例如，在图像检索中，目前采用的特征包括主颜色、颜色直方图、纹理分布、草图等，根据特征的种类不同，可采用两种特征输入手段，即操纵交互输入方式和模板选择输入方式。同时还可应用浏览检索和样本检索的特征输入手段。由于多媒体检索是近似检索，所以查询结果一般都不止一个，一般将满足条件的图片按其相似度从大到小排序，返回前面的若干个。用户可通过选择与检索特征最接近的结果，逐步逼近来完成检索过程。

基于音频内容的检索有多种方法，如直喻（属于某一类的声音）、感知特性（用可理解的物理特性来描述声音，如基频）、主观特性（用描述语言来说明声音）和拟声（在某些音质上类似于要找的声音）。在应用中，先利用各种分析技术把声音变成一组参数，然后对