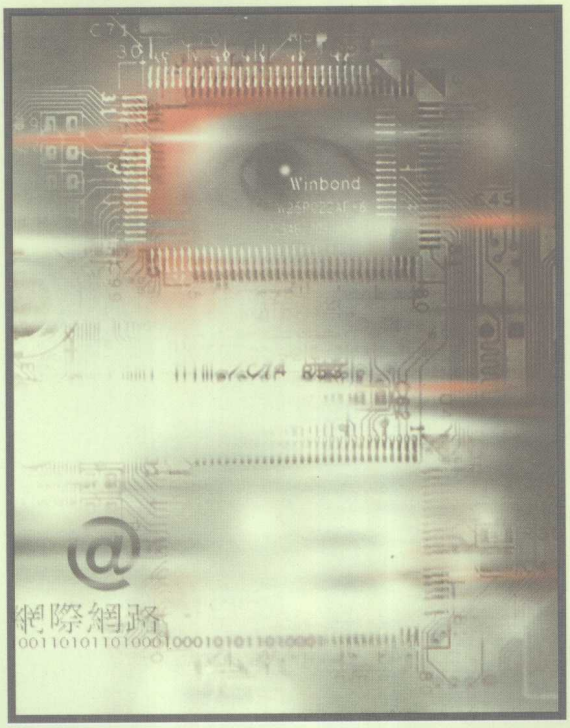
智能信息检索

ZHINENG XINXI JIANSUO

凌云 章志勇 欧阳毅 刘军◎著



中国科学技术出版社
CHINA SCIENCE AND TECHNOLOGY PRESS

智能信息检索

凌　云　章志勇
欧阳毅　刘　军　著

中国科学技术出版社
CHINA SCIENCE AND TECHNOLOGY PRESS
·北　京·
BEIJING

图书在版编目(CIP)数据

智能信息检索/凌云等著. —北京:中国科学技术出版社,2006.12
ISBN 7-5046-1524-2

Ⅰ.智… Ⅱ.凌… Ⅲ.计算机网络-情报检索 Ⅳ.G354.4

中国版本图书馆 CIP 数据核字(2007)第 022703 号

自 2006 年 4 月起本社图书封面均贴有防伪标志,未贴防伪标志的为盗版图书。

内 容 提 要

本书总结了作者和研究小组多年来在智能信息检索领域的科研成果、学术研究成果和教学经验,吸收了许多学者近些年来所发表的最新理论与科技成果,跟踪该领域发展前沿,具有较强的系统性、逻辑性和准确性。

全书的内容一共分为七章,涵盖了智能信息检索的主要研究方向,针对智能信息检索的各个主要技术研究方向进行了详细和深入浅出的阐述,其包含的内容主要有:智能信息检索的各种相关概念、特点以及发展趋势、信息采集、数据挖掘和知识发现、网页净化、互联网信息理解、Web 网页分类和聚类、面向特定领域的智能信息搜索原型等。此外,本书的内容还包含了 Web 分析、分类和检索的应用实例。本书力求理论与技术统一、技术结构合理,使得本书在具有较高的学术价值的同时,又具有很好的易读性。

本书即可以作为高等院校自动化、计算机、电子和通信等专业研究生和高年级本科生的教材和教学参考书,也可以作为计算机信息处理、信息检索等相关研究领域工程技术人员的入门阅读书籍和参考阅读书籍。

中国科学技术出版社出版
北京市海淀区中关村南大街 16 号 邮政编码:100081

策划编辑 林 培 孙卫华 **责任校对** 林 华
责任编辑 孙卫华 程安琦 **责任印制** 安利平

电话:010-62103210 传真:010-62183872
http://www.kjpbooks.com.cn
科学普及出版社发行部发行
北京长宁印刷有限公司印刷

*

开本:787 毫米×1092 毫米 1/16 印张:11.375 字数:262 千字
2006 年 12 月第 1 版 2006 年 12 月第 1 次印刷 定价:25.00 元
书号 ISBN 7-5046-1524-2/G·442

前　言

在人类步入信息社会的时代，信息同物质、能量构成人类社会的三大资源。物质提供材料，能量提供动力，信息提供知识与智慧。因而，信息已成为促进科技、经济和社会发展的新型资源，它不仅有助于人们不断地揭示客观世界，深化人们对客观世界的科学认识，消除人们在认识上的某种不定性，而且还源源不断地向人类提供生产知识的原料。随着近些年来计算机的普及，尤其是20世纪90年代互联网蓬勃兴起之后，人们摆脱了信息贫乏的桎梏，进入了一个信息极度丰富的社会。由于信息的种类非常丰富，数量非常庞大，因此当信息的来源不再是问题时，如何快捷准确地获取感兴趣的信息就成为人们关注的主要问题。但信息的异构、分散以及海量等特性对检索技术提出了非常高的要求，各种信息检索、过滤、提取技术逐渐成为研究的重点，并且在近些年来获得了极大的成功。例如，以Web搜索引擎为代表的信息检索技术已经取得了很大成功，Google、百度、Yahoo等搜索引擎已深入到大家的日常工作和生活之中，成为获取信息不可或缺的工具。

从社会信息化建设以及国民经济发展的需要出发，促使我们对智能信息检索理论与技术进行了长期研究，并且在为研究生和本科生讲授智能信息检索相关课程以及指导研究生对智能信息检索进行研究的过程中积累了一些研究成果和经验，这些都促使我们动笔撰写了《智能信息检索》一书。

智能信息检索是一门理论性很强、应用非常广泛、有几十年发展历史的科学技术，要在一本书里对它的各个研究方向进行详细的介绍是非常困难的。本书的撰写原则是坚持基础与深度并重，即在广泛地、详细地介绍智能信息检索的各种应用的同时，也深入地介绍了智能信息检索各个研究方向的基础理论知识，使内容具有易读性、深入性。在这个基本原则指导下，对智能信息检索的相关概念、特点和发展趋势、信息采集、数据挖掘和知识发现、网页净化、互联网信息理解、Web网页分类和聚类、面向特定领域的智能信息搜索原型等相关研究方向的最新技术进行深入地分析和介绍。同时本书也给出了Web分析、分类、检索等具体应用实例。

本书的内容除了总结了作者和研究小组的科研成果、学术研究成果和教学经验外，还参考了许多国外以及国内的有关资料，如每章后面所列举的参考文献。最后，本书还得到了其他一些学者非常有益的帮助，这里对他们表示衷心的感谢。

凌　云

于浙江工商大学

2006年10月20日

目 录

第一章　信息检索的基础知识

第一节　信息检索概述

一、信息检索是时代发展的必然产物

在人类步入信息社会的时代，信息同物质、能量构成人类社会的三大资源。物质提供材料，能量提供动力，信息提供知识与智慧。因而，信息已成为促进科技、经济和社会发展的新型资源，它不仅有助于人们不断地揭示客观世界，深化人们对客观世界的科学认识，消除人们在认识上的某种不定性，而且还源源不断地向人类提供生产知识的原料。随着科技的进步和飞速发展，现代社会的信息（Information）可以说是浩如烟海，而且正以爆炸式的速度不断增长。它的内容包括了生活和工作的各个领域，如农业、生物、气象、天文学、地理、物理、化学、数学、计算机、医疗、历史、法律、政治、环境保护、文学、音乐和电影等几乎所有领域，它是知识、信息的巨大集合，是人类的资源宝库。可以说，我们所处的21世纪是信息的世界。那么什么是信息，人们又是怎么理解信息的呢？

信息是什么？半个世纪以来，科学界一直在对其定义进行积极地探讨。信息的定义在不同的领域人们赋予它不同的定义，但是，迄今为止还没有一个公认的定义能被社会各界一致接受。现在，关于信息的定义多达数十种，它们都从不同的角度反映了信息的某些特征。英文中的“信息”一词（Information）的含义是情报、资料、消息、报导、知识的意思，所以长期以来人们就把信息看做是消息的同义语，把信息定义为能够带来新内容、新知识的消息。《辞海》对信息的解释是：音讯、消息和信号系统传输和处理的对象，泛指消息和信号的具体内容和意义。此外，还有些观点认为信息是消息、情报、信号、数据和知识的总和；信息是通过文字、数据和各种信号来传递、处理和表现客观事物特性的知识流。但是，随着人类知识面以及接触的事物种类的扩大，发现信息的含义要比消息、情报的含义广泛得多，不仅消息、情报是信息，指令、代码、符号语言、文字等，一切含有内容的信号都是信息。从信息表现形式来看，信息是用文字、数据或符号等形式通过一定的传递和处理来表现各种相互联系客观事物在运动变化中所具有功能、作用、效用、效应、功用等各种特征属性的总称。信息是对客观世界中各种事物的变化和特征的反映，是客观事物之间相互作用和联系的表征，是客观事物经过感知或认识后的再现。目前，有的学者提出了比较具有普遍意义的定义：“信息是指应用文字、数据或信号等形式通过一定的传递和处理，来表现各种相互联系的客观事物在运动中所具有的特征性内容的总称。”简单地说，信息是物质、事物、现象的属性、状态、关系标记的集合。

信息普遍存在于整个宇宙之中，它无处不在，无时不有，是人们认识世界，了解世

界、改造世界，取之不尽，用之不竭的宝贵资源。信息是事物存在的方式、形态和运动规律的表征，是事物具有的一种普遍属性，它与事物同在，存在于整个自然界和人类社会。信息是物质相互作用的一种属性，它涉及主客体双方，信息表征信源客体存在方式和运动状态的特性，所以它具有客体性、绝对性。但接收者所获得的信息量和价值的大小，与信宿主体的背景有关，表现了信息的主体性和相对性。信息的增长速度和利用程度，已成为现代社会文明和科技进步的重要标志之一，信息具有如下一些基本特点。

（1）信息具有普遍性和客观性

世间一切事物都在运动中，都有一定的运动状态和状态方式的改变，因而一切事物随时都在产生信息，即信息的产生源于事物，是事物的普遍属性，是客观存在的，它可以被感知、被处理和存贮、被传递和利用。信息具有永恒性，信息的永恒表现在一条信息产生后，其载体可以变换，例如，我们可以毁掉一本书，但是书里所包含的信息本身并不能被消灭。

（2）信息具有中介性和共享性

信息源于事物，但不是事物本身，它是人们用来认识事物的媒介。能够共享是区别信息不同于物质和能量的最主要特征。信息所包含的同一内容可以在同一时间、同一地域被两个以上的用户分享，其分享的信息量不会因分享用户的多少而受影响，原有的信息量也不会因之而损失或减少。信息可以廉价复制，可以广泛传播，尽管信息的创造可能需要很大的投入，但复制只需要载体的成本。

（3）信息具有相对性和特殊性

世间一切不同的事物都具有不同的运动状态和方式，会以不同的特征展现出来，事物的属性也随着时间在不断变化。信息是事物属性的表现，不同的事物以及事物不同的属性都给人们带来不同的信息。由于事物在不停地变化，所带来的信息也在不停地变化，变化的信息带来的效益也随之不断地变化。信息在某一时刻价值非常高，但过了这一时刻，可能没有任何价值。例如，现在的金融信息，在需要知道的时候，会非常有价值，但过了这一时刻，这一信息就会毫无价值。又如战争时的信息，敌方的信息在某一时刻有非常重要的价值，可以决定战争或战役的胜负，但过了这一时刻，这一信息就变得毫无用处。所以说，大部分信息具有非常强的时效性。正是因为信息具有很强烈的时效性，我们对信息的及时了解和把握就非常重要。

（4）信息具有实质性和传递性

事物在运动过程中和形态改变上所展现出的表征是事物属性的再现，这些事物属性被人们认知后，它们就构成了信息的实质内容。信息依附于一定的载体传递后，才能被人们接受和运用。

这些年来，计算机技术、通信技术、高密度存储技术以及互联网技术发展非常迅猛，信息的采集、传播、利用从规模和速度上都达到了空前的水平，实现了全球范围的信息共享。随着计算机技术、通信技术、网络技术的迅速发展，人类接触到的信息量也随之爆炸性的增加。由于这些信息并没有一个传统的信息过滤机制，信息质量也参差不齐，因此人们很难快速而准确地从海量的信息中准确地获取自己最需要的信息，这使得人们在海量信息面前无所适从。同时，现代社会是一个讲求效率的社会，由于信息随着

事物的变化而在不停的运动变化，信息的价值时效性要求我们尽可能地迅速掌握信息，这时信息检索系统就成为人们生活中的一个重要的研究领域。信息检索起源于20世纪50年代，早在1951年，Calvin Moores提出了“信息检索”（Information Retrieval）这一术语。

“Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, or machines that are employed to carry out the operation.” —By Calvin Moores

“信息检索”这个术语用来描述如下过程：客户找寻信息的请求被转换成相关资料的汇集。当时的信息检索仅仅是指对文本文件的检索，后来随着网络和多媒体技术的发展，信息检索经历了很大的发展与变化，出现了各种多媒体信息检索系统。如图1-1所示，随着社会的发展，信息的种类也在不断地扩大，人们接触的信息从早先的文本信息发展到三维信息，中间包括了音频信息、图像信息以及视频信息。由于人类接触到的信息种类在不断地增加，信息检索的对象也在不断的增加，先后出现了图像、视频和音频以及三维检索系统等，后来人们统一把这些检索系统称为信息检索系统。到目前为止，信息检索技术已经成为人们研究的重要领域，它具有广泛的使用意义。

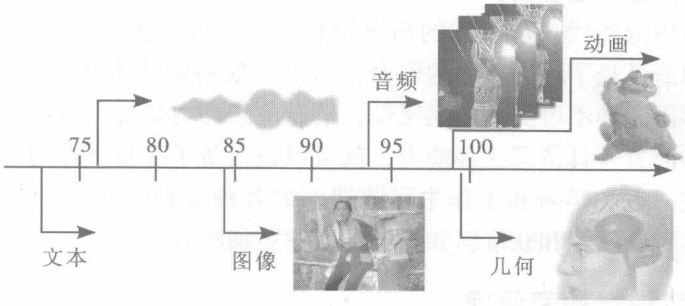


图1-1　信息的发展

海量的信息使得人们要把这些信息进行分类，进行总结和归纳，而采用人工的方式处理信息已经难以跟上时代的步伐。计算机的出现极大地提高了人类处理信息的速度和信息量，把计算机相关技术引入信息检索已经成为信息检索的必然趋势。人们可以利用计算机信息检索系统跨越不同的地域，在短时间内查阅各种大型数据库，能快速地对几十年前到现在的各种文献资料进行回溯检索，及时获得自己需要的信息。计算机检索系统的数据库由于采用了计算机处理，信息数据库中的信息更新速度很快，检索者随时可以检索到所需的最新信息资源。从目前看来，计算机信息检索的重要意义和作用主要体现在以下两方面。

1）紧跟时代步伐。在当代社会，人们需要不断学习，不断更新知识，才能适应社会发展的需求。正如柏林图书馆大门上所刻：“这里是人类知识的宝库，如果你掌握了它的钥匙，这里的全部知识就是你的”。信息检索就像一把开启知识宝库的钥匙，掌握并有效利用它，便能获得和利用人类的精神财富，并使其转化为社会物质财富，并创造出更多的精神财富，推动社会的进步和发展。美国工程教育协会曾估计，学校教育只能赋予人们所需知识的20%～28%，而72%～80%的知识是走出学校后，在研究实践和

生产实践中根据需要，不断再学习而获得的。美国未来学家托夫勒曾说过："科学越来越发展，人们按照自己需要创造资源的能力就越来越大，到那时，唯一重要的资源就只剩信息和知识，信息和知识就成为未来的中心贸易"。由此可见，信息和知识在经济发展中的重要性。在信息海量增长的今天，只有学会和掌握信息检索的方法与技能，才能从复杂多样、质量参差、污染日趋严重的信息中迅速、准确地查获自己所需的信息和知识，从而充分开发利用人类的知识宝库。因此，我们掌握信息检索的方法与技能，是形成合理知识和更新知识的重要手段，是做到无师自通、不断进取的主要途径，是适应时代步伐的重要保证。

2）节省科研时间，提高工作效率。科学研究具有继承和创造两重性，整个科学发展史表明，积累、继承和借鉴前人或他人的研究成果是科学发展的重要前提。科学研究的两重性要求科研人员在探索未知或从事研究工作之前，应该尽可能地占有与之相关的信息，即利用计算机信息检索的方法，充分了解国内、国外，前人和他人对拟探索或研究的问题已做过哪些工作，发展动向如何等，这样才能做到心中有数，防止重复研究，将有限的时间和精力用于创造性的研究中。根据国内外有关材料表明，科研人员花费在查找资料上的时间是相当多的，一般占本人工作时间的一半左右。科研人员掌握信息检索的方法，能熟练地查找自己所需的信息资料，这无疑将大大缩短查询信息资料的时间，这等于增加或延长了科研人员的寿命，这是发展科学技术的一个巨大潜力。因此，信息检索是科学研究必不可少的前期工作，信息检索能力是当今科研工作者应该具备的一种重要的素质能力。具备了这种能力，就可以说具备了终身学习的能力基础，它可以帮助人们解决在一生的学习和工作中可能遇到的各种疑难问题；掌握了索取知识的门径，使自己在接受新教育和获取新知识的过程中更加顺利。

二、计算机信息检索原理

信息检索通常是指从以任何方式组成的信息集合中，查找特定用户在特定时间和条件下所需信息的方法与过程。它主要是指根据用户的查询条件，完成对记录信息的查找过程。并将查询结果以一定的方式呈现给用户，以方便用户对相关信息进行浏览或进行进一步的查询。完整的信息检索系统还包括了信息的存储以及信息检索系统的管理。例如图书馆把书籍和期刊按题目、作者、年份以及出版社组织安排，读者可以根据书籍的题目和作者等信息对书籍进行检索。由于现代社会信息量的急剧增长，把计算机相关技术引入信息检索已经成为信息检索的必然趋势。简单地说，计算机信息检索是指利用计算机存储和检索信息；具体说，计算机检索系统就是指人们利用各种计算机技术，使用特定的检索方式、各种检索关键字以及各种检索策略，从计算机检索系统的信息数据库中检索出所需的信息，继而再由终端设备显示或打印的过程。因此，计算机信息检索的实质是将描述特定用户所需信息的提问特征，与信息存储的检索标志进行异同的比较，然后从中找出与用户需要相一致或基本一致的信息，然后按一定的优先级排序呈现给用户。计算机检索系统又称现代化检索系统，其主要特点是：检索速度快、可以采用灵活的逻辑运算和基于特征的匹配方式，便于进行多元概念检索并能提供远程检索。为实现计算机信息检索，必须事先将大量的原始信息加工处理、以数据库的形式存储在计算机

中，如图1-2所示，计算机信息检索的全过程应也包括大致以下两个方面内容。

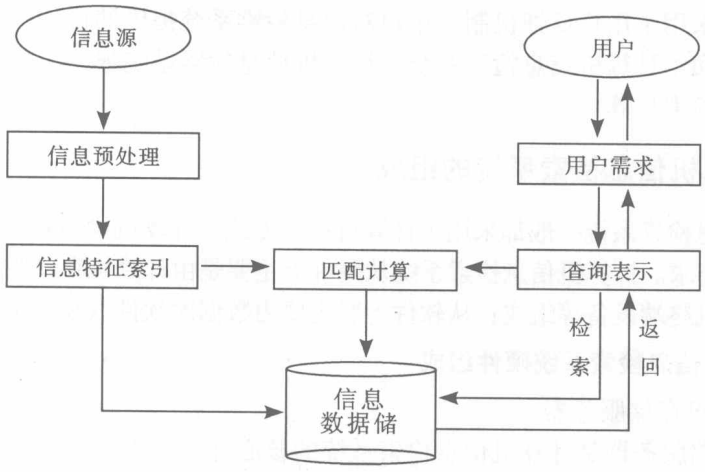


图1-2　信息检索系统原理图

（1）信息索引建立和信息存储过程

计算机检索系统的信息存储过程是用手工或者自动方式将大量的原始信息进行预处理加工，然后再把这些经过预处理加工的数据按一定格式输入计算机，存储在各种存储介质上，例如磁盘阵列或光盘阵列中，这样完成计算机信息检索系统的信息加工存储过程。对大量无序的信息资源进行特征计算，使之有序化，并按科学的方法存储，组成检索工具或检索文档，即组织检索系统的过程是计算机检索系统的一个重要过程。目前，信息索引基本可以分为两类：一类是基于目录的索引，这种方法主要是利用人工的方法对信息建立目录，然后把目录作为计算机信息检索系统的数据库索引，这种方法在海量的数据信息面前已经显得不足。一类是基于特征的索引，这是目前计算机信息检索系统采用的主要索引方式，它计算信息的各种内在特征以及外在的联系，然后把信息的特征和联系作为计算机信息检索系统的数据库索引，这种索引方式具有自动化程度高和速度快等优点。信息的特征计算又称为基于内容的信息特征计算，如何有效地提取信息的各种特征是信息检索领域里一个非常关键的研究问题，成为一个研究热点。

（2）信息的需求分析和检索过程

计算机信息检索系统的检索过程是：用户对自己需要检索的内容加以分析，明确检索范围，弄清主题概念，然后利用系统检索提供的检索接口来表示进行计算机检索。计算机按照用户的要求使用一系列的检索策略，选出符合要求的信息，并且将信息按一定的优先级排序输出。这个过程是分析用户的信息需求，利用已组织好的检索系统，按照系统提供的方法与途径检索有关信息，即检索系统的应用过程。本质上看，计算机检索的过程实际上是一个比较、匹配的过程，检索提问只要与数据库中的信息的特征标志及其逻辑组配关系相一致，则属“命中”，即找到了符合要求的信息。因此，信息检索的实质是将描述特定用户所需信息的提问特征，与信息存储的检索标志。此外，由于在信息索引建立和信息存储的过程中，由于信息的特征计算相对于人的理解来说，是一个低层次的计算，这种低层次的计算与人的高层次的理解之间有一个差距，这个差距限制了

计算机信息检索系统的准确性。目前，为了克服特征计算与人的理解之间的差别，很多信息检索系统采用了用户反馈机制，用户对信息检索系统提供的检索结果进行评价，然后把评价再返回给计算机信息检索系统，计算机信息检索系统根据用户的反馈调整检索策略，进行再检索计算。

三、计算机信息检索系统的组成

目前，信息检索系统一般都采用了计算机检索方式，计算机检索系统已经成为信息检索发展的必然要求。计算机信息检索系统从硬件上主要是由计算机存储服务器、通信网络以及个人计算机终端设备等组成；从软件上看主要由数据库软件以及检索系统软件组成。

1. 计算机信息检索系统硬件组成

（1）计算机存储服务器

计算机存储服务器是计算机信息检索系统的核心部分。当前，我们正处在一个信息爆炸的时代，数据的存储量已经不仅仅是用 KB、MB、GB 甚至 TB 来计算，在不远的将来，人们所谈论的将是 PB（1petabyte = 1000terabytes）甚至 EB（1exabyte = 1000petabytes）。根据 IDC 公司的统计报告，企业数据的增长速度是每 9 个月增长 100%。计算机信息检索系统中的作业系统和数据采掘中，大量的、频繁的数据计算以及移动将会对系统所在的区域网或者广域网造成巨大的影响。此外，计算机信息检索系统的信息存储设备的分布对计算机信息检索系统也有巨大的影响，这种影响使得计算机信息检索系统的存储服务器成为提高检索效率的一个关键因素。

目前，出现了多种新的存储技术和存储结构，这些技术如雨后春笋般给计算机存储模式带来了巨大的变化。而今，引人注目的计算机存储技术，主要有网络附加存储（NAS）技术和存储区域网络（SAN）技术。他们的优势就是在于能够为网络上的应用系统提供多样、快捷、简便的存储资源，另一方面有能共享存储资源并对其实施集中式的管理，成为现今理想的存储管理和应用模式。NAS 系统拥有一个独立的存储服务器，类似于一个专用的文件服务器，不过这种专用文件服务器去掉了通用服务器原有的大多数计算功能，只提供文件系统功能存储服务。SAN 系统是一种面向网络的存储结构，与 NAS 最不同的是，它是以数据存储为中心的，也就是说，存储设备不再限制于服务器系统，而表现为服务器节点上的“网络磁盘”，在服务器操作系统看来，就像网络盘与本地盘一样。这样扩展性能得到极大的改善，每台主机扩大更多的可控存储的容量，还可以通过级联交换机来连接多个存储设备以扩展容量。

采用存储区域网，可以通过快速的、专用的光纤网络，将上百个甚至几千个存储设备连接起来，组成低成本的、易于管理的存储区域网络。存储区域网不仅可以减少数据移动对现有的网络系统的压力，从而降低存储的成本，而且可以通过将存储设备的集中，方便地进行监视和调整，从而实现灵活方便的管理。目前，大型的计算机信息检索系统都采用了存储服务器，存储服务器的品牌主要有 IBM、DELL、惠普、联想等。对于海量信息存储来说，存储服务器的主要性能参考指标有：

1）存储服务器的存储性能。存储服务器的存储性能是存储服务器系统设计的重要原则之一，存储服务器的存储性能应该能够满足应用系统峰值的需求，并有进一步扩展

的空间，包括容量和性能的扩展。从计算机的发展历史来看，计算机的芯片发展速度按照摩尔定律，已经提高了成千上万倍，而计算机 I/O 速度，即磁盘系统接口速度，则从 SCSI 的每秒 5MB 到目前业界最快的 FC－2 协议每秒 200MB，只提高了 40 倍。因此选择性能最佳的磁盘系统，可以有效地提高计算机系统的 I/O 性能，从而提高计算机信息检索系统的整体性能。

2）存储服务器的扩展性。计算机信息检索系统的数据量增长已经呈几何级数的增长，例如美国著名的 EI 检索系统每周要更新一次数据库。检索系统每年数据成倍地增长早已经不是新闻了，为了保证磁盘存储系统的增长满足企业今后发展的需要，对磁盘存储系统的扩展性应从磁盘系统的容量扩展性、磁盘系统的扩展兼容性以及磁盘驱动器的兼容性几个方面考虑。由于磁盘系统本身设计会有一定的局限，磁盘系统的容量最大可扩展性能力是否满足企业今后数据发展的需要，是企业选择磁盘系统时应当考虑的一个方面；由于磁盘系统的发展也是日新月异，用户在对存储服务器进行存储扩容时还要考虑新磁盘系统与旧设备之间的兼容性，即产品系列有连续性；由于磁盘驱动器的技术近年来飞速发展，磁盘的存储容量已经从 9GB、18GB 发展到 36GB、73GB、160GB 磁盘，转速也已经从 1 万转发展到 1.5 万转。只有能够支持不同容量、不同转速的磁盘驱动器，才能最大限度地保护用户的投资。

3）存储服务器的可靠性。数据是计算机信息检索系统最重要的资产，数据的可靠性很大程度上依靠存储设备的可靠性，存储设备的可靠性主要是指磁盘系统的可靠性。因此，作为磁盘系统的选择，可靠性永远是存储服务器的第一考虑。目前，磁盘系统的可靠性主要考虑的技术有冗余电源、写缓存的数据保护、RAID 数据保护以及总线（包括内部总线和外部总线）等。

（2）通信网络

从 19 世纪 40 年代到 20 世纪 30 年代，电磁技术被广泛用于通信。1844 年电报的发明以及 1876 年电话的出现，开始了近代电信事业，为人们迅速传递信息提供了方便。从 20 世纪 30 年代到 60 年代，电子技术被广泛用于通信领域。微波传输、大西洋电话电缆以及 1960 年美国海军首次使用命名为“月亮”的卫星进行远距离通信，标志着远程通信事业的开始。纵观计算机网络的发展历史可以发现，它和其他事物的发展一样，也经历了从简单到复杂，从低级到高级的过程。在这一过程中，计算机技术与通信技术紧密结合，相互促进，共同发展，最终产生了计算机网络。计算机网络的诞生直接导致了目前社会生活中信息量的激烈剧增。

因特网（Internet）是当前世界上最大的国际性计算机互联网络，而且还在不断的发展之中。在 1964 年 8 月，巴兰（Baran）在美国兰德（Rand）公司“论分布式通信”的研究报告中提到了存储转发的概念。在 1962～1965 年，美国国防部高级研究计划署（Advanced Research Projects Agency，ARPA）和英国的国家物理实验室（National Physics Laboratory，NPL）都在对新型的计算机通信技术进行研究。英国 NPL 的戴维斯（David）于 1966 年首次提出了“分组”（packet）这一概念。到 1969 年 12 月，美国的 DARPA 的计算机分组交换网 ARPANET 投入运行。ARPANET 连接了美国加州大学洛杉矶分校、加州大学圣巴巴拉分校、斯坦福大学和犹他大学四个节点的计算机，ARPA-

NET 的成功运行使计算机网络的概念发生了根本性的变化。ARPANET 的成功，标志着计算机网络的发展进入了一个新纪元，计算机网络从单节点网络向多节点网络发展。早期的面向终端的计算机网络是以单个主机为中心的星型网，各终端通过电话网共享主机的硬件和软件资源。但分组交换网则以通信子网为中心，主机和终端都处在网络的边缘，主机和终端构成了用户资源子网，用户不仅共享通信子网的资源，而且还可共享用户资源子网的丰富的硬件和软件资源。这种以资源子网为中心的计算机网络通常被称为第二代计算机网络。

在第二代计算机网络中，相互通信的计算机必须高度协调工作，而这种“协调”是相当复杂的。为了降低网络设计的复杂性，早在当初设计 ARPANET 时就有专家提出了计算机网络层次模型，分层设计方法可以将计算机网络庞大而复杂的问题转化为若干较小且易于处理的子问题。1974 年 IBM 公司宣布了它研制的系统网络体系结构 SNA（System Network Architecture），它是按照分层的方法制定的。DEC 公司也在 70 年代末开发了自己的网络体系结构——数字网络体系结构（Digital Network Architecture，DNA）。有了网络体系结构，使得一个公司所生产的各种机器和网络设备可以非常容易地被连接起来。但由于各个公司的网络体系结构是各不相同的，所以不同公司之间的网络不能互联互通。针对上述情况，国际标准化组织（International Standard Organization，ISO）于 1977 年设立专门的机构研究解决上述问题，并于不久后提出了一个使各种计算机能够互连的标准框架——开放式系统互联参考模型（Open System Interconnection / Reference Model，OSI / RM），简称 OSI。OSI 模型是一个开放体系结构，它规定将网络分为七层，并规定每层的功能。OSI 参考模型的出现，意味着计算机网络发展到了第三代。

在 OSI 参考模型推出后，网络的发展道路一直走标准化道路，而网络标准化的最大体现就是 Internet 的飞速发展。在 Internet 的飞速发展过程中，各个计算机厂商既遵守了 OSI 的七层协议，但是又不是严格的遵守。这些厂商把 OSI 的七层协议的部分进行了合并，使得协议层数减少到四层，这就是 TCP / IP 协议模型。如图 1 - 3 所示，TCP / IP 协议模型与 OSI 七层协议模型的内容基本完全一致，但是 TCP / IP 协议模型更加面向实际生产和应用，而 OSI 七层协议模型更加具有学术化气息。

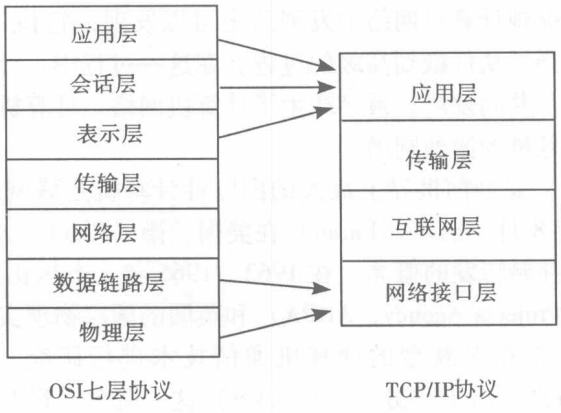


图 1 - 3　网络协议

（3）个人计算机终端设备

个人计算机终端设备是用户与检索系统相互传递信息进行“人—机对话”的装置，它包括个人电脑、打印机等设备。目前，个人计算机终端设备已经非常普及，这为信息检索系统的发展提供了有力保障。

2. 计算机信息检索系统软件

（1）数据库技术

数据处理是人类生活的一个重要研究内容，数据处理大致经历了三个阶段：手工管理阶段（20 世纪 50 年代中期）、文件系统管理阶段（50 年代后期至 60 年代中期）和数据库系统阶段（60 年代后期）。今天，研究用计算机进行数据处理的方法已成为计算机科学技术中的一个主要研究方向，其中，数据库技术是其中的一个重要研究内容。

计算机数据库系统的萌芽出现于 20 世纪 60 年代。当时计算机开始广泛地应用于数据管理，对数据的共享提出了越来越高的要求。传统的文件系统由于各种限制已经不能满足人们对大量数据进行有效管理的需要。能够统一管理和共享数据的数据库管理系统（DBMS）因此应运而生。数据库通常是指特定的信息集合，而数据库管理系统是对数据库进行管理和控制的软件。这些管理和控制功能主要包括数据的定义、数据存取和修改、数据库的运行管理、数据库的建立和维护等。除了功能方面的要求外，对于数据库系统性能方面也有一定要求，其中之一就是能够及时准确地满足多个用户的并发存取操作，另外还有能够保证数据库事务的原子性、时刻保持数据的一致性，要求在硬件和操作系统正常工作的情况下独立的并发操作互不影响、不丢失数据。在人们生活中，各种数据信息是种类繁多，为了对各种类型的数据信息进行有效的管理，必须提取出这些数据信息的共性，使用一定的数据模型来表示数据信息的共性。数据模型是数据库系统的核心和基础，各种数据库系统软件都是基于某种数据模型建立的。按照数据模型的特点，我们可把数据模型分为网状数据模型、层次数据模型以及关系数据模型和面向对象数据模型。与此对应，我们根据数据模型可以将数据库系统大致分成网状数据库、层次数据库和关系数据库以及根据专业领域内各种信息特点建立的面向对象数据库系统等。

1）网状数据库。最早出现的数据模型是网状模型，以网状模型为基础的数据库管理系统称为网状数据库管理系统。网状模型中以记录为数据的存储单位，记录包含若干数据项，并且网状数据库的数据项可以是多值的和复合的数据。在每个记录中有一个唯一的标志它的内部标识符，称为码，它在一个记录存入数据库时由数据库管理系统自动赋予。网状数据库是导航式（Navigation）数据库，数据库的记录之间彼此相互联系，组成一个关系网，一个记录的修改和移动会影响整个数据库的其他记录。对于网状数据库，用户在操作数据库时不但说明要做什么，还要说明怎么做。例如在查找语句中不但要说明查找的对象，而且要规定记录的存取路径。世界上第一个网状数据库管理系统是美国通用电气公司 Bachman 等人在 1964 年开发成功的 IDS（Integrated Data Store）。

2）层次数据库。由于网状数据库的记录之间彼此相互联系，组成一个关系网，一个记录的修改和移动会影响整个数据库的其他记录，因此网状数据库的使用有很多不便之处，后来人们对网状数据库进行了改进，提出了层次数据库。层次型数据库管理系统是紧随在网络型数据库出现之后而出现的。现实世界中很多事物是按层次组织起来的，

层次数据模型的提出，首先是为了模拟这种按层次组织起来的事物。层次数据库也是按记录来存取数据的，但是层次数据模型中最基本的数据关系是基本层次关系，它取消了网状数据模型记录之间的任意联系结构，它只允许数据库记录和自己上下层记录进行联系，这种联系代表了记录型之间一对多和一对一的关系，和网状数据库相比，它取消了多对一的关系。在层次数据库中，有且仅有一个记录型无双亲，这个节点称为根节点，其他记录型有且仅有一个双亲。在层次模型中从一个节点到其双亲的映射是唯一的，所以对每一个记录型（除根节点外）只需要指出它的双亲，就可以表示出层次模型的整体结构。层次模型是树状的，最著名最典型的层次数据库系统是 IBM 公司的 IMS（Information Management System），这是 IBM 公司研制的最早的大型数据库系统程序产品。

3）关系数据库。网状数据库和层次数据库已经很好地解决了数据的集中和共享问题，但是在数据独立性和抽象级别上仍有很大欠缺。用户在对这两种数据库进行存取时，仍然需要明确数据的存储结构，指出存取路径。而后来出现的关系数据库较好地解决了这些问题。关系数据库理论出现于 20 世纪 60 年代末到 70 年代初。1970 年，IBM 的研究员 E. F. Codd 博士发表《大型共享数据银行的关系模型》一文提出了关系模型的概念。Codd 在 70 年代初期的论文论述了范式理论和衡量关系系统的 12 条标准，用数学理论奠定了关系数据库的基础。Codd 博士也以其对关系数据库的卓越贡献在 1983 年获得了计算机领域的最高级别奖励 ACM 图灵奖，后来 Codd 又陆续发表多篇文章，奠定了关系数据库的基础。关系模型有严格的数学基础，抽象级别比较高，而且简单清晰，便于理解和使用。关系数据模型提供了关系操作的特点和功能要求，但不对数据库关系的语言给出具体的语法要求。关系数据库的操作是高度非过程化的，用户不需要指出特殊的存取路径，路径的选择由数据库管理系统的优化机制来完成。目前，我们应用的关系数据库软件主要有 IBM 的 DB2 数据库、甲骨文公司的 Oracle 数据库、微软的 SQL Server 数据库等。

4）面向对象数据库。随着像 CAD、CASE、图像处理、GIS 等新的应用领域的发展，以及传统应用领域中应用的深化，要求数据管理软件管理复杂对象，模拟复杂对象的复杂行为。于是，在 80 年代中后期产生了面向对象数据库系统，把面向对象技术与数据库技术结合起来，利用类的设施来描述复杂对象，利用类中封装的方法来模拟对象的复杂行为，利用继承性来实现对象的结构和方法的重用。目前，和关系数据库一样，市场上有许多面向对象的数据库（Object-Oriented Data Base，OODB）可供选择。

面向对象数据库系统对一些特定应用领域（例如 CAD 等），较好地满足了其应用的各种需求。但是，这种由于纯粹的面向对象数据库系统并不支持 SQL，因此在通用性方面失去了优势，其应用领域受到很大的局限。目前，在面向对象技术与数据库技术相结合的过程中，基本上是沿着两种途径发展的：一种是建立纯粹的面向对象数据库管理系统（即 OODBMS），这种途径往往是以一种面向对象语言为基础，增加数据库的功能，主要是支持持久对象和实现数据共享。另一种途径是从传统的关系数据库加以扩展，增加面向对象特性，把面向对象技术与关系数据库相结合，建立对象—关系数据库管理系统（即 ORDBMS）。ORDBMS 既支持已被广泛使用的 SQL，具有良好的通用性，又具有面向对象特性，支持复杂对象和复杂对象的复杂行为。正在制定的新的 SQL 国

际标准体现了 ORDBMS 的特征。ORDBMS 适应了某些新应用领域的需要和传统应用领域深化发展的需要，因而近几年来，ORDBMS 获得了快速的发展，成为数据库领域的一个令人关注的技术。

（2）信息检索系统软件

通过一定的检索系统软件，人们可以快速地查询各种需要的信息。此外，信息检索系统软件能够对信息进行收集、预处理、特征提取、存储以及整个信息检索系统的运行和管理。相对而言，计算机信息检索系统的硬件部分决定了系统的检索速度和存储容量，而软件部分则是充分发挥硬件的功能，确定检索方法。目前，有很多公司和研究机构开发了各种信息检索系统软件，这些信息检索软件的发展为用户提供了方便的信息查询平台。

四、计算机信息检索系统的发展

目前，信息检索在人们生活领域中获得了广泛的应用，随着计算机硬件和软件技术的发展，计算机信息检索系统也在不断的更新和发展。计算机信息检索系统根据其系统的硬性发展变化经历了如下的不同的发展阶段：脱机批处理检索阶段、联机检索阶段、光盘检索阶段以及计算机网络检索阶段。

（1）脱机批处理检索阶段

随着 1946 年世界上第一台电子计算机问世，计算机技术逐步走进信息检索领域，并与信息检索理论紧密结合起来。在利用计算机进行信息检索的早期，人们只是使用单台计算机的输入和输出装置进行信息检索，用磁带作存储介质，信息检索一般采用连续的顺序检索方式。检索部门把许多用户的检索提问汇总到一起，进行批量检索，然后把检索结果通知各个用户，用户不直接接触计算机。这种方法更适合大批量的预定信息检索，所以这种检索方式也叫做脱机批处理检索或定题情报服务。由于这种检索方式需要把许多用户的检索提问汇集在一起，因此这种信息检索方式缺乏用户交互，检索灵活性较差，并且单台电脑的计算能力和存储能力都有局限性，因此，这种信息检索系统的信息容量较小。

（2）联机检索阶段

到了 20 世纪 60 年代末期，由于计算机软硬件技术的不断提高，出现了一台主机带多个终端的联机信息检索系统。这种系统采用分时操作技术，所以用户可以使用终端设备直接与计算机进行“人—机对话”，计算机对用户的提问能及时处理并显示出结果。联机信息检索系统具有分时的检索操作能力，能够使许多相互独立的终端同时进行检索。美国在 20 世纪 60 年代末期进行了联机检索的研究，并在 1969 年研制出第一个大型的联机检索系统。在 20 世纪 70 年代末至 80 年代中期是国外数据库发展速度最快的时期，80 年代发达国家的一些计算机信息联机检索系统，通过卫星通信网络和计算机专用终端，在世界范围内提供联机信息检索服务，形成国际联机检索服务业。联机检索服务是计算机检索走向实用化、规模化、产业化的重要标志，信息载体向小型化、高密度化转化，数据库生产实现了产业化，联机检索系统的创建实现了国际化。由于国际联机检索系统具有速度快、查准率高等优点，目前已成为世界各国各个行业获得重要信息