INTERNET 和信息检索

唐永林　葛巧珍　主编

编著者　唐永林
葛巧珍
陈　荣
侯丽英

华东理工大学出版社

图书在版编目(CIP)数据

INTERNET 和信息检索/唐永林,葛巧珍主编.
—上海:华东理工大学出版社,2000.8
ISBN 7-5628-1064-8

Ⅰ.I... Ⅱ.①唐...②葛... Ⅲ.情报检索 Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(2000)第 32661 号

INTERNET 和信息检索

唐永林
葛巧珍 主编

华东理工大学出版社出版发行
上海市梅陇路 130 号
邮政编码 200237 电话 64250306
新华书店上海发行所发行经销
常熟市印刷八厂印刷
开本 787×1092 1/16 印张 11.5 字数 278 千字
2000 年 9 月第 1 版 2000 年 9 月第 1 次印刷
印数 1—5000 册

ISBN 7-5628-1064-8/TP·108 定价 20.00 元

图书在版编目(CIP)数据

INTERNET 和信息检索/唐永林,葛巧珍主编.
—上海:华东理工大学出版社,2000.8
ISBN 7-5628-1064-8

Ⅰ.I... Ⅱ.①唐...②葛... Ⅲ.情报检索 Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(2000)第 32661 号

INTERNET 和信息检索

唐永林
葛巧珍 主编

华东理工大学出版社出版发行

上海市梅陇路 130 号

邮政编码 200237 电话 64250306

新华书店上海发行所发行经销

常熟市印刷八厂印刷

开本 787×1092 1/16 印张 11.5 字数 278 千字

2000 年 9 月第 1 版 2000 年 9 月第 1 次印刷

印数 1—5000 册

ISBN 7-5628-1064-8/TP·108 定价 20.00 元

1 概 论

走进21世纪，人们很快发现所处的信息环境正在日新月异地发生着巨大的变化。信息技术的飞速发展推动着信息社会的不断发展。一台微机、一部电话，加上一台调制解调器，顷刻之间使整个世界变得很小很小。在信息社会中，信息随时随地影响着人们的生活、学习和工作。信息与物质、能量已并列为现代社会的三大支柱之一。信息社会中，人们搜集、处理、流通、控制和利用信息的能力已达到了新的高度。今后随着光信息技术、生物信息技术以至人工智能的发展，将使信息收集、处理、传输、控制和利用的有效性更加复杂。与此同时，“信息和知识成为经济发展中超劳动力和资本的最重要的因素”的认识越来越被世人所认可。信息的重要性和利用现代化信息技术快速获取有用信息之间的矛盾日益加剧。众所周知，知识经济是建立在知识和信息的生产、分配和使用之上的经济，知识经济更注重信息和知识的扩散与使用。一个企业乃至一个国家，经济成功的决定因素在很大程度上取决于对知识和信息的搜集利用的效率，可以说对知识和信息的搜集利用是否有效，直接关系着一个国家的综合国力和国际竞争力。

1.1 信息社会中信息的特点

1.1.1 类型多、数量大

由于科学技术的迅猛发展，各种信息载体发展很快，特别是电子信息载体的发展大有将传统印刷型信息载体取而代之的趋势。信息的增长几乎是每隔七八年就要在原有的基础上翻一番。这使发展大容量的信息载体有了迫切需要，而信息技术的快速发展，又使这一要求的实现有了可能。随着磁、光存储技术的发展，电子化信息载体达到了存储的大容量化和高密度化。

信息源除了传统的印刷型刊物之外，因特网信息达到了空前的“爆炸”。据统计，大概有186个国家接入因特网的网络有24万个。预计到2000年将接入各种网络100万个。网上资源包罗万象，信息资源的优势前所未有。另外，随着各种光盘数据库网络信息资源的迅速发展，估计，现有出版的光盘数据库逾万种，包含的信息量超过100多亿条。而且，类型多，信息涵盖的科学技术领域广泛。

1.1.2 传递快

电子化传输和信息技术的快速发展，使原始状态的信息从无序变为有序的工作大大加

快，并经过像蛛网一样高速运行的网络将整个世界的空间范围缩小在方寸之间。真可谓“秀才不出门，能知天下事”。

1.1.3　共享程度高

信息资源具有共享性，这是信息资源优于物质资源和能源资源的重要特征。在网络环境下，时、空范围得到了最大程度的延伸和扩展。一旦某一信息资源上网后，全世界每一个终端用户都可同时共享同一份信息资源，使有限的信息资源得到最大限度的利用。

1.2　通信技术、网络和现代信息传输

1.2.1　通信技术

我们知道，任何信息只有经过一定的载体以及一定方式的传递，才能被他人利用。远古时的举火为号到纸张印刷术的出现已大大扩展了信息的流通范围。自从人类学会利用电和电磁波以来，信息技术的变革大大加快。电报、电话、收音机、电视机的发明使人类的信息交流与传递快速而有效。20 世纪中期开始，半导体、集成电路、计算机的发明，数字通信、卫星通信的发展形成了新兴的电子信息技术，使人类利用信息的手段发生了飞跃。计算机网络就是通信技术和计算机技术二者高度发展和密切结合而形成的。

采用某种方法，通过某种介质或传输线将信息从一个地方送到另一个地方称之为通信。现代的通信技术主要分为两种：模拟通信和数据通信。模拟通信是指利用模拟信号传输信息的通信方式。模拟信号的频率、振幅或相位随着信息（如声、光等）的变化而变化。在发送端将信息变换成模拟信号加以传输，在接收端将收到的模拟信号还原成信息，可以传输声音、文字和图像，占用频带窄，但抗干扰能力弱。所谓的数据通信就是通过通信系统完成二进制编码的字母、数字、符号及数字化的声音、图像，信息的传输、交换、存储和处理。在这基础上产生的数据通信系统，可把分散在各地的各种终端连接起来，实现远程终端和计算机之间的数据传输。远程终端所产生的数据能及时传送到中央处理器进行处理，其结果可以马上送回到远程终端，这样人们就可以在远离计算机的地方直接使用计算机的各种资源。而用通信线路将分散在不同地方并具有独立功能的多个计算机系统互相连接起来的计算机网络可以共享资源。

简单地说，数据通信是计算机之间或者终端与计算机之间的通信。计算机、终端设备及数据传输结合起来的系统称之为数据通信系统。它具有输入、输出数据的功能，且具有差错控制、传输控制等通信控制功能，以实现正确的数据传输。

1.2.2　网络

网络也就是人们所称的计算机网络。是指将分散在各处，且具有独立功能的多台计算机终端及其附属设备，通过通信设备和线路连接起来，运用功能完善的通信软件按照网络协议进行数据通信，以实现资源共享的系统。

在实际工作中，人们与计算机和计算机网络的联系越来越密切，而因特网的出现可以说是计算机网络系统优越性的最充分的体现。计算机网络一般可分为广域网（WAN）和局域

网(LAN)。

广域网又称远程计算机网络,一般不受地区的限制。范围可延伸到全国或全球。它是利用公共电话网、电报网、租用线路或专用线路,把远程计算机或终端设备连接起来,实现远程计算机间的通信。

局域网一般在一个相对较小范围的特定区域内部建立起来的通信网络。

互联网络也就是我们平时所讲的互联网(Internet)。这是一种遍布全球的网终,是各个计算机信息网络平台的总网络,是成千上万信息资源的总称,简言之,是网间之网。从技术角度看 Internet 是一个互相衔接的 IP 网,由成千上万的局域网、企业网及全球性计算机网络的实时互联,且所有的互联都是通过 TCP/IP 来实现。

正因为计算机网络可使网络中的用户实现各计算机或终端之间的数据库、应用程序、软件以及包括打印机等设备的资源共享,因此网络中如有某台计算机出现故障,可由其他计算机替代其工作,从而提高系统工作的可靠性,并可通过网络实现集中管理。可以这么说,没有联网的计算机就达不到真正的信息交换与共享,先进的计算机也不能充分地发挥其功能。一句话,没有计算机网络也就不存在今天的信息时代。

1.2.3 现代信息传输

任何通信的目的都是把信息从一点送到另一点。要完成这种信息传送,最常用的办法是把信息先附加(调制)在一个电磁波(载波)上,然后把被调制的载波送(传播)到目的地,在目的地接到电磁波后,再把信息复原(解调)。这种系统通常由无线电射通信和光波频段通信组成,见图 1-1。

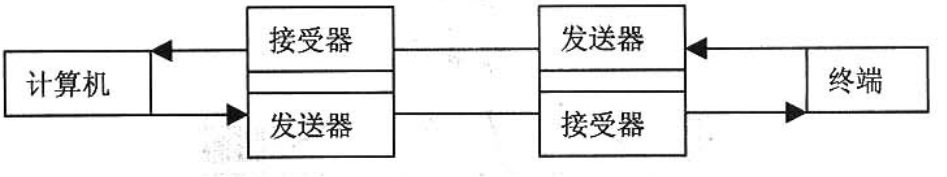

图 1-1 信息传输流程

简言之,信息技术的高速发展、通信现代化和计算机网络的有机结合为信息传输提供了快速、安全、有效的工具,在信息技术的基础上实现了信息资源的网络化,为信息使用者提供了获取信息的最优化的服务网络,使信息生产者在横向联合、资源共享、互利互惠的条件下,实现信息资源的集成化和市场化。即提高信息质量达到信息资源建设的相互协作,协调统一,同时,信息源之间也获得了相互调剂、互通有无的结果。

1.3 计算机网络的组成和类型

计算机网络大致由计算机系统、终端子系统和通信子系统 3 个部分组成。我们也可简单看作是通信子网络和资源子网的组成。其中通信子网络主要完成数据传输、交换通信和控制;资源子网络主要是提供共享的硬件、软件和数据库等资源,并进行相应的数据处理。

计算机网络的类型很多,根据不同的特性和需要,可建立不同的计算机网络形式,下面

界新的学术动向，找出新的增长点；有利于学到新的知识，发现新的问题，启发我们产生新的思维，从而找到自己研究的方向和突破口，有利于避免重复研究；有利于最大限度地集中多方面的智慧；有利于进行跨学科、跨层次、跨地区、跨国家的全球性研究；有利于创新资源的合理配置。

1.4　网上信息资源

信息技术的进步和互联网的日益完善，人类正在进行信息史上最巨大的一项工程——将世界现有的信息，诸如报纸、期刊、专刊文献等等都放到网络上去，同时也不停地在网络上产出数不胜数的新信息。整个网络正在堆积成一个前所未有的超级大型数据库。传统观念中的知识宝库——图书馆也将只是一个概念上的“图书馆”而已。数字化图书馆(Digital Library)随着计算机网络技术、数据库技术、多媒体技术的发展应运而生。

在网络环境和虚拟实境(virtual reality)技术配合下，情报用户已经不必亲自到某一图书馆去借阅书刊，而只需坐在计算机终端前利用网络进行检索与阅览就足够了。无论是上海图书馆还是中国国家图书馆，不管是美国国会图书馆还是大英图书馆，对我们终端用户来说都是等距离的，而且可以“自由”进出，随需获取。

1.4.1　网络信息资源的类型

网络信息资源主要分为以下3种：

(1) 数字化图书馆提供的各种信息服务。

(2) 因特网上的各种信息源。

(3) 光盘数据库形式的各种科技、经济和商业性的电子化信息数据库。

对于广大科研人员来说，这3类信息源是获取信息的主要来源。

1.4.2　数字化图书馆和虚拟图书馆

什么是数字化图书馆？广义言之，一个数字化的图书馆应是计算机可处理信息的集成或此类信息的储存处。按美国研究图书馆协会(ARL)提出，数字化图书馆的要素为：

(1) 数字化图书馆不是一个简单的实体。

(2) 数字化图书馆需要用技术来连接众多资源。

(3) 许多数字化图书馆和信息服务之间的连接对终端用户是透明的。

(4) 广泛地存取数字化图书馆信息和服务是同一个目标。

(5) 数字化图书馆馆藏并不限于文献替代品，已延展至不能从印刷格式表达或传递的数字式人工制品。

目前，在对于数字图书馆、虚拟图书馆问题的认识上，大多数图书情报学专家认为两者应是相同的，但严格分析起来，两者还是有一定的区别。

数字图书馆主要是指利用计算机技术对文字等符号进行二进制编码转化为可视文图的图书馆。它可能还是某一图书馆的馆藏。

虚拟图书馆主要是指由许多分支网络通过网络构成的一个网络中心，它是无形、虚拟的图书馆，我们只感到它的存在而不能真正看到这个全球性“大图书馆”的庐山真面目。

但不管是数字图书馆,还是虚拟图书馆,都必须是网络化的图书馆。也就是说没有了网络化这个基本特点,两者都无从谈起。建设数字化图书馆是新世纪各国图书馆工作努力开发的重点。我国已着手建立的"全国高校文献资源保障体系"(CALIS),可以说是这一努力的创始,其国家级中心设于北京大学。地区级中心有华中的武汉大学、华南的中山大学、华东北的南京大学、华东南的上海交通大学和复旦大学等。除了由全国中心引进 OCLC 的 First Search Uncover 等大型数据库,以及 Science 等全文数据库外,以各地区中心馆为主共建具有中国高校特色的数据库,目的是通过整个高校图书馆的共同努力改变目前我国信息资源个体重复与总体贫乏交互存在的现象。

目前,世界上典型的虚拟图书馆首推美国的联机计算机图书馆中心(OCLC),其特点是采取了相对集中式的馆际协作模式。OCLC 拥有成员馆多达 25 000 多个,遍及世界上 63 个国家,多种语言的书目记录有 3 700 万条,反映 6 亿条馆藏记录。

OCLC 除自己的通信网络外,First Search(FS)系统还采用 Internet 传输信息、数据与图像,为用户提供服务。凡购买了 FS 检索卡,得到 OCLC 授权号与密码号的用户,都可作为最终用户以自己的 Internet 终端机直接联机访问 FS 系统完成检索。FS 的另一特点是实现了文献书目到文献以至全文的一体化服务。在 FS 数据库中有 ASC Ⅱ 全文,印刷文献的扫描图像和 Internet 文件等多种形式的全文文献。

1.4.3 因特网信息源

互联网中最杰出的代表——因特网(Internet)。据统计,因特网已覆盖全球 186 个国家和地区,涉及到 24 万左右网络的资源,入网主机超过 1 288 万台,全球网民 1 亿多人,预计到 2005 年将会超过 10 亿。又据我国互联网信息中心(CNNIC)调查:到 1999 年 6 月底,我国上网用户就达 890 万,上网计算机 350 万台,CN 域名有 29 045 个,WWW 站点数约 9 906 个。因特网是人类信息活动史上的一次革命,它的产生和发展给信息资源有效配置带来了深刻的影响。

因特网更是各种数据库的汇集处,网上的数据库不下万种,其他信息更是难以记数。可以说,这些资源是其他任何环境下的信息资源所无法比拟的。

1.4.4 光盘数据库信息资源

光盘是一种存储量大、价格低廉的计算机信息存储设备。它是利用激光、计算机、数字通信和光电集成等现代化高科技技术的产物。1 张 CD-ROM 光盘其容量在 650MB ~ 900MB。目前,记录有文献信息数据库的光盘在图书馆中得以广泛应用。据统计,已出版的光盘数据库逾万种,包含的信息量超过 100 多万亿条。光盘数据库类型多、品种全,信息涵盖的科学技术领域广,且覆盖的时间跨度大,地域广,检索方便。

光盘数据库检索利用方便、快捷。从 20 世纪 90 年代起,由于光盘塔和光盘网络软件的不断发展,光盘数据库网络化应用发展飞快。我国各单位建立的光盘数据库检索网络系统已逐渐占据计算机信息检索系统的主导地位,成为人们最常用的信息检索系统。光盘数据库检索网络的出现,一方面解决了光盘数据库的大量发展和单机阅览需频繁换盘给用户带来的不便;另一方面能供大量用户同时共享同一种数据库,充分发挥信息资源的效益,为计算机信息检索的普及发挥了很大的作用。

1.5　计算机信息检索

计算机技术、通信技术及网络技术的飞速发展，使信息的检索手段发生了质的飞跃。由于信息生产和传递的无序扩展，造成信息混乱和信息污染。不管采用何种方式去寻找所需的信息，都会碰到信息混杂、淤积、拥挤、无序和信息过剩等问题，使科研人员在获取所需的信息时，产生大量人力、物力的浪费，造成了无限增长的信息与人们有限接受、处理信息能力之间的矛盾愈来愈尖锐。

人类社会经过三次大的革命，而20世纪出现的计算机和电信互联网络则揭开了信息革命的序幕，将给人类社会带来更为剧烈的变化。如何应付信息时代的挑战，关系到国家的前途和民族的命运。

21世纪将是多种载体、多种形式的文献共存互补的世纪。其中以电子格式和通过各类计算机工作站在荧屏上提供给用户阅读或打印输出的文献将会越来越多，而这些电子文献大体上以——检索类、报道类、研究类三种类型出现，又以题录、索引、文摘与全文同时存在，而本书着重讲述的是如何利用检索类电子文献，通过计算机及网络系统，从各种不同类型的信息数据库中获取我们所需的信息，并用检索到的信息来满足日常的生活、学习和科研工作中对文献、数据和事实的迫切需要。

1.5.1　数据库检索系统的基础知识

数据库的检索系统是人们从“无序”的“海洋般”信息源中方便、快速地获取有用信息的有效工具。所谓的信息检索系统，即是信息存贮与检索的全过程。一个好的检索系统必须经过整序后有价值的信息集合，并按一定的方式进行合理的存放，最后能使用户方便地从中提取3个要素组成。其中关键是存贮和检索。简言之，存贮是将无序的信息变为有序。而只有达到了信息的有序化，才能供用户有效地获取信息。用户不管是查找有用的文献也好，还是查找数据或者事实，没有快速方便的检索系统，就很难使“有用的信息”发挥应有的作用。而在计算机信息检索系统中，数据库是系统赖以生存的核心。20世纪七八十年代，数据库技术趋向成熟，到80年代末，据报道开放型的数据库数量已达到4 000多个。90年代更是有了快速增长，而且，数据库类型有了大的发展。多媒体技术在数据库中的广泛使用，光盘数据库日益普及，同时，计算机软件技术的发展使信息数据库服务更为方便。而现代化通信技术使数据库的服务不断延伸。“信息高速公路”更是使数据库的服务达到了前所未有的水准。应该说，任何数据库的建立都是为某种信息目的而建立的，并为特定用户提供数据库的检索服务。

1.5.2　数据库的一般结构

数据库的种类很多，其中对科研人员来说，文献型数据库使用频率最高。不同的数据库其结构各异，提供的检索途径也有差异，但无论什么类型的文献数据库，其基本特征是相同的。

(1) 记录、字段、文档：

记录　是数据库中的基本文献单元。一个记录是对某一实体的完整描述，如一本书、一

篇专利文献等等，每一记录都有一个独特的记录号。记录越多，数据库的容量就越大。

字段　是组成记录的基本信息单元。每一字段都描述文献的某一方面的特征，如著者(AU)字段、出版项(PR)字段、标题(TI)字段、文摘(AB)字段等等，各个字段的组合就成了记录。字段的检索功能对提高文献的查找效率有很大的作用。

文档　是由各个记录构成的集合。

字段、记录和文档的构成应为树状结构，见图 1－6。

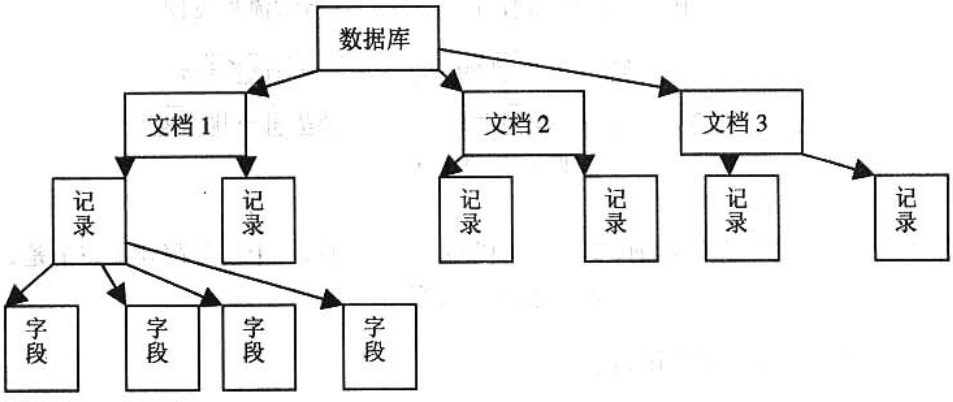

图 1－6　树状结构的数据库构成

(2) 顺排档、倒排档：

顺排档　是按其记录的编号顺序线性排列的文档。在按顺排档编制的检索方式中，每查找一条记录都是从第一条记录开始，按顺序依此往下查，直到找到记录为止。顺序文档结构简单，管理方便，节省存贮空间，但检索速度较慢。

倒排档　对于要求多途径检索的信息数据库来说，主要采用的还是倒排档。这是因为，从信息用户角度出发，用户提出的问题是以用户熟悉的检索项交于计算机处理的，如主题、分类、著者、文献名一直到时间等内容。只有号码组成的顺排档显然是不能胜任的。而倒排档是将数据库中具有某些属性的字段值和具有该字段值的记录号构成的文档。在检索过程中，不是由记录号查属性，而是由属性查记录号，相对顺排档检索功能正好相反，故称为倒排档。

在数据库中，并不是所有的字段都要建立倒排档，一般只是对“可检词”建立倒排文档。简言之，倒排档是满足文献检索用户的检索习惯而建立起来的，它大大提高了信息检索的效率。

(3) 基本索引、辅助索引：

基本索引　各种数据库提供的字段检索默认值。其特征为：当某一检索词进入数据库而未对检索词进行字段限制检索时，数据库默认该词在某些特定字段中进行检索。如 Dialog 系统，其基本索引的构成字段为四个：DE(叙词字段)、ID(题内关键词字段)、TI(标题字段)和 AB(文摘字段)。一般情况下，数据库大都采用基本索引这一方法，只是所限定的字段项有所不同。

辅助索引　与基本索引的不同在于，检索词进入数据库时，一定得标明用何种字段进行

检索。如要查某著者写的文献，其输入方法应为：AU（著者字段名缩写）= SMITH，BLACK B. 如未将著者放入指定字段，系统将其自动放入基本索引中检索。

1.5.3　计算机检索的基本原理

通过计算机获取文献信息，是以“电子化”的手段达到手工检索所无法相比的效率。计算机检索的方法和各种数据库所用的检索指令有所差异，但其本质都是一样的，无非是以用户提出的条件如何充分运用匹配运算达到目的的问题，其检索流程见图 1－7。

信息检索的需求 → 进行课题分析 → 选择适用的数据库 → 确定检索词 →
编制检索式 → 进数据库实行检索词匹配运算 → 显示检索结果 → 输出文献

图 1－7　检索流程图

用户在上机进行信息检索前，要对整个机检原理中的每一个环节都要了解清楚，而其中又以“匹配”过程最为关键，即计算机检索的检索算法。

1.5.4　计算机检索的匹配运算

计算机不具备人脑的思维能力，它只能把用户提出的复杂的需求变成计算机“机械性”“匹配”所需的一组检索词的组合。即将文献的特征标识（索引词）与检索提问标识（检索词）在字面上达到一致就算完成“匹配”，然后将作为命中的文献输出。在执行匹配运算时，检索效果好坏对于选什么检索词，采用何种逻辑运算，如何截词等关系重大。

（1）确定检索词：

由于检索词挑选得正确与否对于提高机检的效率有很大的关系，因此在选定检索用的数据库后，确定检索词有两个要求：① 必须能正确反映你的检索要求；② 必须符合数据库对输入词语的要求。对于第二个要求，由于各数据库对于输入词语的规定并不一样，但我们可用一定的方法来满足其对词语的要求。如是叙词（DE）字段，所用的检索词往往是经规范化的词语，这类词语的运用比较严格，而分类（Classification）字段则受数据库本身所采用的那种特定的分类法的限制，如国际十进分类法、中国图书馆图书分类法、世界专利分类法等等。而诸如 TI、AB、ID 字段则往往是自由词，对检索词没有特殊要求。另外，还要确定数据库是运用多元词（两个或两个以上的词组成的概念）检索，还是非要用单元词才行；或者可以单元词、多元词同时使用。

（2）布尔逻辑运算：

布尔检索是目前应用最为广泛的检索方法。常用的布尔逻辑运算有 3 种：逻辑与、逻辑或、逻辑非。

逻辑与　检索时一般输入“AND”或运算符“＊”。表示两个检索词概念之间的交叉和限定，A＊B，表示数据库中一篇文献应同时含有“A”和“B”的概念才能检出。例如，powder and coating（粉末涂料），命中的文献同时含有 powder 和 coating 两词，因而，用逻辑“与”组配可以提高检索的查准率。

逻辑或　检索时一般输入“OR”或者运算符“＋”。表示检索词与检索词之间概念的并

列组合，A + B，表示文献中只要含有 A、B 中任何一方即算命中。例如，Vitamin C（维生素 C）+ Ascorbic Acid（抗坏血酸）两词为同一种物质的不同名称，同位概念运用逻辑或能扩大检索范围，提高检索的查全率。

逻辑非　检索时一般输入“NOT”或运算符“ - ”。表示检索词概念不包含某种概念关系的一种组配，A-B，表示在数据库整体中含有 A 的文献范围中排除含有 B 概念的文献记录。例如：以 Dialog 系统为例。

```
?    SS   fuel(燃料)not gas fuel(气体燃料)
     S1      16766       fuel
     S2      8332        gas fuel
     S3      7542        1 not 2
```

由于逻辑非实际上反映的是集合的差运算，也可以说是一种排除法的检索过程，用逻辑非运算检索能提高检索的查准率。

注意：在遇到相关概念排除时，往往要对其命中结果考虑周全，并谨慎使用。

（3）截词检索：

截词检索是作为弥补布尔运算检索的不足而发展起来的，可以部分解决由于检索式中遇到的检索词为词干（词根）相同，词义相近的词，或同一词的单、复数形式，或同一词的动、名词形式，或同一词的英、美两种不同的拼写方法等由于列举不全造成的漏检现象。如果充分利用计算机数据库所特有的截词功能，对检索词作相应的技术处理，使检索词能更全面地表达检索要求，提高检索词的检索效果。截词的基本情况有：词尾截词、词头或词中间截词。由于截词检索是进行词的部分比较，不要求检索词和索引词全等，所以这是一种相似性运算，相当于用“或”逻辑扩展检索的范围，可以提高检索的查全率。

（4）位置逻辑检索：

当输入两个以上检索词时，限定词与词之间在文献中的逻辑间隔，是一种词与词之间在文献中关系密切程度的体现。主要有如下几种情况：

Same Document　表示用户输入的两个以上检索词只要是在整个记录中出现均可，不论词与词是紧靠在一起还是分散各处，都算命中。

Same Paragraph　表示用户输入的两个以上的检索词限于出现在同一段落中，不论词与词是紧靠在一起还是分散各处，都算命中。

Words Apart　表示用户输入的两个以上的检索词之间必须作出间隔限定。比如，词与词之间不能分隔，但词与词之间可以颠倒，词与词之间容许间隔 1 个词，间隔 2 个词，间隔 *n* 个词等等。

Exact Order　表示用户输入的两个以上的检索词之间不容许有间隔，且词与词之间不可以颠倒。

同样表示位置算符的还有：

Directly adjacent　直接邻近，表示词与词之间不可分隔。

Adjacent or one intervening word　表示词与词之间可有 **1** 个词的间隔。

Up to three intervening words　表示词与词之间可允许多至 **3** 个词的间隔。

Anywhere in the same paragraph　表示词与词的出现只要是在同一段落即可。

Anywhere in the same field　表示词与词的出现只要是在同一字段即可。

With　表示算符两侧的词必须紧挨着，且词序不可颠倒，但词与词之间容许插入 n 个词。

Near 表示算符两侧的词必须紧挨着，但词序可颠倒，且词与词之间容许插入 n 个词。

有的数据库中还容许词与词之间前后位置可以互换。位置逻辑的运算可以说是带有灵活性的词组比较，即允许有一定范围的相似性计算，也可称为邻近算符检索。位置逻辑检索可以提高检索的查准率。

(5) 其他逻辑操作符：

计算机信息检索中，还常常出现表 1 - 1 的种种逻辑操作符供参考：

表 1 - 1　其他逻辑操作符

操作符 Operator	说明 Description
=	等于
! =或者 < >	不等于
<	小于
>	大于
> =	大于等于
< =	小于等于

等于符“ = ”，将一个值与另一值相比较来决定两数是否相等。当测试字符是否相等时，两个字符必须同为大写或小写。例如：= WILSON，等于同时输入 WILSON 和 Wilson。

不等符“ ! = ”或“ < > ”与上面相反，它意味着子句的第一个值不等于第二个值。例：< > WILSON or ! = WILSON，系统只承认 WILSON 不承认 Wilson。

大于“ > ”和小于“ < ”命令从左读向右，A > B 读成 A 大于 B，A < B 读成 A 小于 B。当大于操作符左右两端的数值相等时，则用“ > = ”和“ < = ”。例：> = 1994，读作“大于等于 1994”，系统查找 1994 年(包括 1994 年)以后的文献；< = 1994，读作“小于等于 1994”，系统查找 1994 年(包括 1994 年)以前的文献；> 1994，读作“大于 1994”，系统查找 1994 年(不包括 1994 年)以后的文献；< 1994，读作“小于 1994”，系统查找 1994 年(不包括 1994 年)以前的文献。

1.6　因特网上的查询系统——搜索引擎

因特网上出现的查询系统——搜索引擎是随着 Internet 网的出现，以及信息科技的进步出现的是日益普及的互联网特有的检索途径。

Internet 已使整个网络逐步堆积起一个前所未有的超级大型数据库。如何在浩翰如海的信息空间快速查找并获取所需信息，已成为信息时代里最为根本的问题之一。为了解决这一问题。各种介绍和如何获取信息的信息查询服务系统应运而生，这就是众所周知的搜

索引擎，它是可以提供信息检索服务的计算机系统。检索的对象包括互联网上的站点、新闻组中的文章、软件存放的地址，及作者、某个企业和个人的主页等。今天我们已很难想象，Internet 网上如果没有搜索引擎该如何查找和获取各种信息。从目前来看，大部分的搜索引擎是英文的，也有不少较好的中文搜索引擎，详见第二章内容。

2 INTERNET 信息检索

随着计算机技术的迅速发展及 Internet 的普及，人们无论从事什么工作都将不可避免地面临一个新的挑战——国际计算机互联网络，也称 Internet，它是目前世界上使用人数最多、连接国家最多、信息资源最多的大型计算机网络，在国际上被称为“21 世纪的信息高速公路和信息平台”，也被誉为是继计算机发明之后的又一场信息革命的标志。几乎所有发达国家的政府、大企业、科研机构、大学，以及很多个人用户都加入了因特网，它是世界上最为开放的网络。这个大网络仍在不断地扩大、成长：每秒钟都有难以计数的信息在网上流动，每分钟都会有新的信息资源出现和更新。Internet 网上的信息资源被人们誉为“世界上最大的图书馆”，人们可以从 Internet 上访问各种信息，如电子报刊、电子新闻、电子报告、会议资料、各种软件资料、图像文件、声音文件等。网上信息更新速度快、费用低廉且大部分是免费的。为使人们能够充分利用丰富的网上信息，Internet 上还提供了各种检索工具，如 YAHOO、搜狐等搜索引擎。总之，Internet 是目前世界上资料最多、规模最大的信息库，是人们获取信息的一个重要来源，而且人们只需在 Internet 网上轻轻点击几下鼠标，所需的信息就会到你的计算机上。

2.1 概述

在高度信息化的社会里，如何有效地快速处理大量的信息已成为人们日益关注的话题，Internet 网作为一个先进的信息传输和处理工具，正以一种前所未有的高速悄悄地进入我们的日常生活。人们用各种名称来称呼 Internet，如互联网络、交互网、网际网、全球信息资源网，等等。目前，Internet 已通达全球百余个国家，用户数达 6000 万个，据 Internet 协会的估计，到 2001 年，将会有 50 亿台以上的计算机与 Internet 联网。

2.1.1 INTERNET 的作用

随着 Internet 的迅速发展，它的应用领域不断扩大，正日益渗透到文化、教育、政治、经济、科研等各领域当中，成为我们日常生活和学习中不可缺少的组成部分。那么，Internet 能为我们做些什么呢？用户最常使用的网络服务有，电子邮箱：71.65% ，搜索引擎：50.40%，软件上传或下载服务：44.16%，各类信息查询：39.31%，网上聊天室：25.47%，新闻组：16.99%，BBS 电子公告栏：16.32%，网上游戏娱乐：13.64%，免费个人主页空间：13.49%，网上寻呼机：13.17%，网上炒股：8.50% 等等。以上的因特网用途可归纳为以下三方面。

一是获取信息。我们生活在信息时代，需要从各种渠道获取信息，通常通过广播、电视、

报刊等媒体来获取。而 Internet 给我们带来了全新的感受，WWW（万维网）上凝聚了 Internet 的精华，展示了 Internet 最绚丽的一面，上面载有各种精美丰富的多媒体信息。通过 WWW 方式，人们不仅可以浏览文字内容，更能够看到丰富多彩的图片，甚至能够听音乐、看电影。另外，WWW 具有独特的链接方式，使您只需点击一下相关单词、图片或图标，就可以迅速从一个网站进入另一个网站，从而达到获取信息的目的。用户在网上最主要获得的信息是新闻：65.52%，计算机软硬件信息：51.70%，电子书籍：38.04%，休闲娱乐信息：38.79%，科教信息：31.43%，金融证券资讯：21.22%，求职招聘信息：19.25%，商贸资讯：17.26%，各类广告信息：12.79%，旅行信息：11.94%，医疗信息：9.39%等。

二是相互交流。生活中我们经常要与人交流，相互沟通，通常采用写信、打电话、发传真等方式，现在 Internet 上也可以实现这些功能。

（1）电子邮件。电子邮件即 E-mail，可以实现远距离通信。今天 E-mail 已成为人们网上互通信息的最普遍的手段之一。E-mail 不仅可以传输文本文件，还可以传输含有声音、图片或其他程序产生的文件。你可以在几秒到几分钟之内，将你的信件送往世界各地的邮件服务器中。而且，无论您走到哪里，只要有一台能连上网的计算机，就能随时收到您邮箱中的信件。

（2）网上打电话。利用网络通讯是 Internet 最强大的功能。目前通过网络电话有两种方式：a. 从网络到网络：通话双方同时登录上网，并且使用同一种软件（如 Iphone、Netmeeting），就可以实现语音的交流，这种形式的网络电话是免费的。b. 从网络到电话机：用市话费拨打国际长途，这是 Internet 上最流行的活动之一。

（3）网络聊天。你可以进入提供聊天室的服务器，与世界各地的人通过键盘、声音、图片等多种方式进行交谈。

（4）新闻讨论组。新闻讨论组是一个世界范围的电子公告板，用于发布公告新闻和各种文章，供大家使用、讨论和发表评论。您可以加入到你感兴趣的专题讨论组中，阅读他人的文章或发表自己的观点，与大家一起进行讨论。

三是网上商机。利用 Internet 进行网上商业活动，是目前发展最迅速的服务焦点之一。从个人角度来说，我们每个人都有展示自我的愿望。今天，Internet 使愿望变成了现实，你可以把自己的个人资料制成主页，链接到一些知名网站的免费主页存放空间，推销自己，以寻求适合自我发展的职位。从政府、企业的角度来看，因特网是一个对外宣传的阵地，各国政府纷纷利用因特网来进行宣传。企业可在网上进行广告宣传，在因特网上做广告有着巨大的优势：不受地域、时间的限制，享受全球用户 24 小时的主动查询，而费用却极为低廉，特别适合中小型企业进行产品广告宣传。开设网上虚拟商店，进行虚拟经营等。

伴随着网络技术的发展，今后还会出现许多新的服务和功能，Internet 将给人类的生活、工作带来根本性的变化。

2.1.2 中国的互联网络

Internet 在中国的发展起步较晚，但是由于政府的高度重视及人们的积极参与，使中国互联网络的发展达到了惊人的高速。

1994 年，中国作为世界上第 71 个成员国加入 Internet；同年 4 月，中科院高能物理所率先以 64Kbps 的速率与 Internet 网建立了连接；邮电部开通了北京、上海两个 Internet 网的国

际出口；经过 3 年的建设，先后建成了中国科学技术网（CSTNET）、中国公用计算机互联网（CHINANET）、中国教育和科研计算机网（CERNET）、中国金桥信息网（CHINAGBN）等四大互联网络，到 1997 年底，四大网络之间实现了互通。中国用户可以方便地通过四大网络接入国际 Internet。1999 年，中国联通公用计算机互联网（UNINET）经国务院批准，成为第五家公用互联网单位。

目前，我国国际线路的总容量为：351M，连接的国家有美国、加拿大、澳大利亚、英国、德国、法国、日本、韩国等。分布情况如下：

中国科技网（CSTNET）：10M

中国公用计算机互联网（CHINANET）：291M

中国教育和科研计算机网（CERNET）：8M

中国金桥信息网（CHINAGBN）：22M

中国联通互联网（UNINET）：20M

下面分别对这 5 家公用互联网单位予以介绍。

中国科技网

中国科技网的网址是：http://www.cstnet.net.cn

中国科技网是在中关村地区教育与科研示范网（NCFC）和中国科学院网（CASnet）的基础上建设和发展起来的覆盖全国范围的大型计算机网络，是我国最早建设并获国家正式承认具有国际信道出口的中国四大互联网络之一。

中国科技网始建于 1990 年，并于 1994 年 4 月最早实现了我国与国际互联网络的全功能连接，同时在国内开始管理和运行中国顶级域名 CN。中国科技网现有多条高速国际信道连到美国、日本及法国，通过这些信道进入 Internet 国际互联网络。CSTNet 有 4Mbps 的信道连到美国，128Kbps 的信道连到日本。到 1999 年，国际信道升级至 10M。

中国科技网作为最早进入 Internet 国际互联网络，并拥有丰富信息资源的国家级科技信息网，对于我国网络事业的发展起到了积极的推动作用。图 2－1 为中国科技网主页。

中国公用计算机互联网

中国公用计算机互联网（上海）网址：http://www.sta.net.cn

中国公用计算机互联网（ChinaNet）由信息产业部（原邮电部）主管，1994 年 8 月邮电部建立了北京、上海两条 64Kbps 专线，通过中国公用数据网 ChinaPAC、ChinaDDN 向全社会提供中国公用 Internet 服务。1996 年 1 月，中国公用计算机互联网（ChinaNet）全国骨干网建成并正式开通，全国范围的公用计算机互联网络开始提供服务。由于邮电部门在我国通信领域具有一言九鼎的地位，它的介入使我国的 Internet 进入高速发展的时期。据《通讯产品世界》1995 年 12 月的统计：1995 年 3 月，中国在 Internet 网上的装机数量仅为 400 台，用户数仅为 3,000 人。到 7 月份，上网计算机猛增到 6,000 台，用户达 40,000 人。到 1999 年 6 月，上网人数已达 400 万。图 2－2 为中国电信上海站点主页。