Internet 信息资源检索丛书

Internet 信息资源检索和利用

杨晓宁 黄正祥 编著

江苏科学技术出版社

Internet 信息资源检索丛书

Internet 信息资源检索和利用

杨晓宁　黄正祥　编著

江苏科学技术出版社

图书在版编目(CIP)数据

Internet 信息资源检索和利用 / 杨晓宁等编著. —南京:江苏科学技术出版社,2001.2
(Internet 信息资源检索丛书)
ISBN 7-5345-3315-5

Ⅰ.I... Ⅱ.杨... Ⅲ.因特网-情报检索
Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(2001)第 08636 号

Internet 信息资源检索丛书
Internet 信息资源检索和利用

编　　著	杨晓宁　黄正祥
责任编辑	宋　平　刘屹立
出版发行	江苏科学技术出版社 (南京市湖南路 47 号,邮编:210009)
经　　销	江苏省新华书店
照　　排	南京展望照排印刷有限公司
印　　刷	南京通达彩印厂
开　　本	787mm×1092mm　1/16
印　　张	11.25
字　　数	234 000
版　　次	2001 年 2 月第 1 版
印　　次	2001 年 2 月第 1 次印刷
印　　数	1—5 000 册
标准书号	ISBN　7—5345—3315—5/TN·56
定　　价	15.30 元

图书如有印装质量问题,可随时向我社出版科调换。

序

中共中央、国务院《关于深化教育改革的全面推进素质教育的决定》指出“要重视培养学生收集处理信息的能力、获取新知识的能力……”。教育部领导同志谈到“培养高层次创造性人才的几点要求”,其中的第一点就是使学生“永远充满获取新知识的渴望,并善于获取知识,具有较宽广的知识面”。怎样培养学生收集处理信息,获取新知识的能力呢?这和教材的建设有关。目前,理工科学生使用的文献检索教材多数出版于20世纪80年代末、90年代初期,内容有些跟不上信息技术的发展。随着网上搜索引擎、网上虚拟图书馆、网上数据库例如OCLC、Dialog、万方数据库、清华大学期刊数据库等如雨后春笋般相继问世,以及网络查询技术的日益普及,Internet已经为此打造了一个信息资源检索的新平台。

与传统的文献检索形式相比,Internet信息资源检索的优越性表现在两个方面:第一,这是一个跨地区、无国界的信息空间、信息系统。数字化的文字、图片、声像文件无论存储在什么位置,都可以通过Internet互相连接,使用户可以超越时空限制,突破“图书馆”的约束,在任何地方、任何时候都获取这些信息。第二,Internet信息检索以智能为基础,以全文检索为特征,大大突破了传统检索的限制。因此,用户只要经过学习,就可以利用搜索引擎和检索技巧,在庞大的信息流中,通过自己的检索方式,在与系统的交互过程中逐步缩小搜索目标,最终获取所需要的信息。本书作者和省内高校同行为此进行了较长时间的探索,并结合教学实践编写了本教材。

我热切地期望《Internet信息资源检索和利用》的出版发行对于大学生们检索并利用好Internet信息资源起到积极的推动作用。

河海大学副校长 **鞠 平**

2001年1月

前　言

随着Internet规模的逐步扩大、网上信息资源的日趋丰富和图书馆由传统化向电子化、网络化、数字化方向快速发展，网上的各种联机数据库、信息搜索引擎等信息资源已成为高效率、高质量获取信息资源的重要途径。Internet信息资源可分为三部分：1. 非正式出版的信息。如电子邮件、新闻组专题讨论、电子会议、电子布告板新闻，它们分布广，流量大，内容难以控制，质量难以保证。2. 半正式出版信息。如学术团体、商业公司、企业和政府机关等单位介绍性、描述性信息。3. 正式出版信息。如各种网络数据库、联机杂志、电子杂志、电子图书、电子报纸等。因此，检索和利用Internet信息资源已成为教师、科研人员、大学生迫切需要掌握的技能之一，在高校开设“Internet信息资源检索和利用”课程已成为当前的迫切需要。

《Internet信息资源检索和利用》共分四章。第一章介绍Internet信息搜索引擎的检索原理、类型、检索功能及特点。第二章介绍网上主要通用搜索引擎和检索实例。第三章着重介绍OCLC、Dialog、UMI、中国万方数据资源系统等网上综合数据库。第四章对网上其他重要信息资源，例如网上参考工具书、政府报告、工程索引、会议文献检索等，进行了详细地描述。

本书的第一章由杨晓宁、何丽梅编写。第二章主要是由杨晓宁根据付天香、彭桃英、吴立志、许发见等同志的网上检索实例编写而成。陈军冰参阅清华大学的OCLC相关资料编写了第三章。黄正祥编写了第四章。何丽梅、赵乃瑄和周建屏编写了大部分实例。杨晓宁、赵乃瑄提供了三个附录。全书由杨晓宁和黄正祥负责汇总统稿。特别感谢施泽华教授在本书的编辑工作中给予积极鼓励和支持。

由于作者自身水平有限，错误及疏漏之处在所难免，恳请广大读者和同行不吝指正。

作　者

2001年1月

目　录

1 Internet信息资源检索概论

1.1 网络与网络信息资源

当代信息技术的广泛应用和发展，尤其是Internet的高速发展，极大地推动了全球信息化进程。根据网络上计算机的使用目的和管理方式，可将网络分为四种不同类型：

(1) 学术研究团体型。一般来说，它是在政府支持下为学术和科学研究及教育界服务的。这些由国家出资建设的网络一般是服务于一个地区或者某个国家的计算机系统，例如美国的BITNET和英国的JANNET，中国的CERNET也是这种形式。

中国教育和科研网络(CERNET)建立于1995年，是原国家计委和原国家教委联合建立的一个为教育科研服务的全国性信息网络，网管中心设在清华大学。除了北京的网管中心之外，第一期工程包括上海、南京、广州、武汉、西安、成都和沈阳等高等院校集中的城市。CERNET的潜在服务对象包括全国1 090所大学的39万教师、10万研究生和220万名在校学生，4万所中学的550万名师生和几万所小学的师生。

同时，国内一些图书馆也已利用网络发挥积极的作用。例如，北京大学图书馆在其主页中，除提供联机公共目录查询外，还挂接一些网络资源服务项目，如FTP资源、免费《科学在线》(Science Online)检索等，其国外联机数据库检索就包括CALIS UnCover网关数据库、OCLC First Search网上专线免费检索和医学文摘库Medline免费检索。清华大学图书馆收集和整理了电子期刊刊名表，分别列出了能获得全文、文摘和目次的电子期刊，并建立链接。在具备访问Internet的条件下，访问者可直接点击电子期刊刊名，根据自己的需要检索该刊。

(2) 公司企业组织型。企业内部专用计算机系统，一般用在一个单位组织系统内，实现部门和部门之间计算机通信和人机对话，例如银行系统等。通常说来，这些系统有严格的安全措施来限制进入某些核心机构。

(3) 合作团体型。这些计算机网络是由一些个人和团体为实现某种共同目标而建立的。例如有关图书的联合目录团体，在这个联合团体内，各个成员图书馆为在网上交换目录信息，建立起相关的机构和共同遵守的规则。

(4) 商业服务型。这些网络具有赢利性质，在付费的基础上，最大程度满足公众和有关团体的各种需要。例如，具有全球广泛用户的美国计算机网络COMSERVE，它不仅提供美国的广泛信息资源，还提供其他国家丰富的信息资源，这样可以广泛地吸引其他国家的网络用户。再如出于商业性目的而成立的Internet服务提供商(Internet Service Provider，简称ISP)，他们通过租用高速通信线路，建立必要的服务器和路由器等设备，向用户提供Internet接入服务，从中收取服务费。一般来说，ISP具有局域网或小型广域网的规模。目前国内的四大互联网络就相当于四个较大的ISP，其中CHINANET就是专门向公众提供Internet接入服务的。此外，其他的一些公司，如中网信息公司、东方网景、瀛海威等都可以提供接入服务。

Internet信息资源按类型可分为四大类：

(1) 电子报纸。网络版世界著名报纸估计有百余种,国内的人民日报、光明日报也有了Internet版。

(2) 动态信息。如政府机构发布的消息、政策法规、会议消息、论文集、研究成果、项目进展报告、产品目录、出版目录、广告等。

(3) 全文期刊。网上有数百种全文文本期刊。

(4) 书目数据库。网上有数万种文摘或目录数据。

Internet信息资源按其文件组织形式又可分为两大类:

(1) 自由文本。全文文摘或题目的非结构化组织,无须规范处理。

(2) 规范文本。按统一标准和格式上网,组成联合编目等规范文本。

Internet信息资源按媒体性质还可分为:

(1) 文本信息。如数据、论文、书刊、目录和数据库、广告。

(2) 图形。如图表、图形、影像、影视。

(3) 声音。

(4) 软件。如免费软件、赠送软件、商品软件及软件升级版本。

伴随网络化和数字化而产生的大数量、多类型、多媒体、非规范、跨地域、跨行业、跨语种网络信息资源,是对原来以相对集中和规范为标志的传统数据库资源的突破性发展。对这一新型信息资源,原有的处理方式已多半不适用,需要发展崭新的自动化信息组织和管理方式。目前使用得较多的网络信息资源组织方式主要有四种:文件方式、数据库方式、主题方式和超媒体方式,其中以文件方式组织网络信息资源较为简单方便。但随着网络信息资源利用的不断普及和信息量的不断增多,以文件为单位共享和传输信息会使网络负载越来越大,而且当信息结构较为复杂时,文件系统难以实现有效的控制和管理,因此,文件只能是网络信息资源管理的辅助形式。数据库技术是对大量的规范化数据进行管理的技术,它可以大大提高信息资源管理的效率。因为数据库的最小存取单位是字段,所以可根据用户需求灵活地改变查询结果集的大小,从而大大降低网络数据传输的负载。因此,数据库方式是当前普遍使用的网络信息资源的组织方式,特别是在大数据量的环境下,其优点是可通过浏览的方式层层遍历,直到找到所需要的信息线索,再通过信息线索链接到相应的网络信息资源。该方式具有严密的系统性和良好的可扩充性,但它不适合建立大型的综合性的网络信息资源系统,只有在建立专业性或示范性的网络信息资源体系时,才显示出其结构清晰、使用方便的优点。超媒体技术是超文本与多媒体技术的结合,它将文字、表格、声音、图像、视频等多媒体信息以超文本组织起来,使人们可以通过高度链接的网络结构在各种信息库中自由航行,找到所需要的信息。这种方式符合人们思维联想和跳跃性的习惯,通过浏览的方式搜寻所需信息,避免了检索语言的复杂性。

1.2 网络信息资源检索

网络信息资源检索是一种基于超文本方式的信息查询工具,用超文本与字符串来表达,这与以线性形式进行组织的传统文本信息的处理方式有较大的不同。它不是以字符而是以结点为单位组织各种信息,一个结点是一个"信息块"。结点内的信息可以是文本、图像、图形、动画、声音或其组合,在信息的组织上采用网状结构,结点间通过关系链加以链接,构成

表达特定内容的信息网络。它对信息的存取可以按照交叉联想的方式从一处迅速跳到另一处，从而打破了原文本系统只能进行顺序线性存取的限制，可以方便灵活地检索信息。

网络信息检索具有以下特点：

(1) 具备网状的复杂信息链接结构，系统能够按照不同查询条件链接结点信息，以供浏览、查询，具有较强的索引功能。

(2) 信息丰富，结点媒体多样化，每个结点都能集成文本、图像、图形、动画、声音、视频等多种媒体，并能用多窗口、图形界面予以表现。

(3) 良好的导航能力，可引导读者在复杂的网络信息中漫游而不致迷失方向。用户可以利用导航机制，了解其所在网络中的位置。

(4) 良好的编辑功能，包括修改、增加、删除结点的能力，对结点内的信息也具有良好的编辑能力，可同时进行多窗口编辑。

(5) 通过网络共享数据库，可使多个用户同时使用库内信息。网络信息资源检索的这些特点主要基于 WWW 服务方式。

1.2.1 WWW 的工作特点

Internet 上的网络信息资源搜索工具大致分为三类：交互式信息服务软件、名录服务软件和索引服务软件。目前在 Internet 运行的交互式信息服务软件主要基于 Gopher 和 WWW；名录服务软件主要基于 Whois、Netfind 和 X.500；索引服务软件主要基于 Archie、Veronica、Jughead 和 Wais。Archie 是为 FTP 资源服务的；Veronica 和 Jughead 是为 Gopher 资源服务的；Wais 可作为 Internet 整个文本信息资源服务的搜索工具。

WWW(World Wide Web)是 Internet 上最先进的网络信息资源检索系统，它把超文本技术、网络技术和多媒体技术融为一体，并把 Internet 上的信息按一定的规则组织起来，以提供进一步的查询。WWW 以超文本(Hypertext)方式提供世界范围的多媒体信息服务，用户通过 Internet 可以从全世界任何地方调来所希望得到的文本、图像、影视和声音等信息。由于有了 WWW，一个不懂计算机网络的人可能很快地成为通过 Internet 检索信息的行家。

WWW 具有如下工作特点：

(1) 信息检索空间宽广。

WWW 信息服务器采用了超文本技术和统一资源定位器技术，将全世界联机信息资源联接在一起，形成人类知识的结合体。它根据信息说明的需要，直接利用他人的成果，“指引”用户进一步读取存放在其他 WWW 服务器上的相关信息，那些服务器又指引着更多的服务器(也可能指回原来的服务器)，这样，在环球范围内的有 WWW 信息服务器互相指引而形成的信息网便出现了。传统的联机检索系统，由于技术上的原因，使检索只能局限于某一主机的特定数据库，而 Internet 上的信息检索可以同时使用多个主机甚至所有主机的某种资源，用户却不必知道这些资源的具体位置和地址。

(2) 用户界面友好，操作方便。

WWW 是以客户机/服务器的模式进行工作的。一方面，在 Internet 上的一些被称为 WWW 信息服务器上运行着 WWW 服务器程序，用来发布信息。另一方面，在用户的计算机上运行着各式各样的 WWW 客户程序，用来帮助用户完成信息检索。WWW 客户程序主要提供两种基本功能：向用户提供风格统一、使用方便的 Internet 信息检索界面；将用户的信息

检索请求转换成 Internet 查询命令送给网上相应的 WWW 信息服务器进行处理。针对不同的计算机硬件平台和软件平台,已有许多种类和版本的 WWW 客户程序,如 Lynx、Mosaic、Viola、Cello、Macweb、Winweb、Netscape、IE 等等,其中 Netscape 和 IE 是目前使用最广,功能强大、界面友好、使用方便的 WWW 客户程序(浏览器),它们提供了几种在 WWW 中检索信息的机制。

Netscape 和 IE 所浏览的文件以页的形式存放在 WWW 服务器上,每一页都有惟一的统一资源地址标识(Uniform Resource Locator,简称 URL),以便用户检索,所以如果用户知道自己所需信息页的地址,就可以直接在地址栏中键入该地址来进行定位。

用户也可按关键词、页标题进行检索,点击 Netscape Netsearch 按钮,就会看到 Netscape 提供的目录表,其中包括 InfoSeek、Lycos、Yahoo!、WebCrawler 等项,而它们都是 WWW 上的一个检索网点,也可以说是 WWW 上的信息资源检索工具。这些信息资源检索工具提供了不同的信息检索机制,选择任何一项都会链接到这个网点上,用户就可键入主题词或页标题等进行检索。

检索结果通常包括题目、内容简介、URL 地址等,由此可进行更进一步的检索,获得更详细的内容。由于目前 Internet 上比较拥挤,检索人员要掌握一定的检索技巧,灵活运用各种检索策略,正确地选择主题词和叙词,才能提高查准率和查全率,节约上机时间。

Netscape 和 IE 只是一个 WWW 的客户程序(浏览器),它本身并不是一种 WWW 的信息检索工具,但它提供了到 InfoSeek 之类 WWW 信息检索工具(网点)的链接,因而使其具有了多种信息检索的机制。

(3) 信息检索系统设计合理。

WWW 和 Internet 为用户提供的信息是以图形、文本等多媒体方式展现的。用户在计算机屏幕上看到的是一篇篇版面美观,图文并茂的文章,并且这些文章中的一些内容含有对其进一步描述的链接,这是利用了多媒体和超文本技术来实现的。这与传统的基于命令或基于菜单的 Internet 的信息检索界面有很大的不同,用户在网上搜索时可以明显感到主机的硬件平台、操作系统、客户程序和服务程序版本的差异,主机的地理位置、信息的存储方式对用户的操作都没有影响。

1.2.2 目前 Internet 信息资源检索的局限

(1) Internet 是一个全球分布式的网络结构,大量的信息分别存储在各国的主机和服务器上。WWW 上存储着大量有价值的信息,吸引着全世界用户去使用和开发。Internet 上的科技信息分布在世界各地的主机上,一般可以查到 40 万个主页,近百种报刊全文、信息用户新闻和原始技术报告等。但是,网络用户遇到的最大困难在于如何快速、准确地从浩如烟海的信息资源中找到自己最需要的信息。美国 Lycos 公司 1996 年的一项调查显示,80% 的被调查者认为 Internet 非常有用,但他们为查找所需信息花费了大量时间。信息资源的分散给信息检索带来了困难。

(2) 网上信息数量非常庞大。目前,Internet 上每 24 小时信息流量达万亿比特。WWW 网址每 6 个月就增长一倍。用户面对成千上万的链接点,难以找出合适需要的信息。

Internet 的发展十分迅速,信息广泛而混杂,在目前谁都可以提供信息服务的情况下,信息的准确性和完整性难以得到有效保证,很多信息在定期或不定期地更新。Internet 改变了

信息发布和评价的程序，从使用出版者、读者、评论人员共同完成的信息评价工作，变为由后两者承担，缺少了编辑出版这一重要的质量控制。因此，出版自由一方面能使大家看到大量传统媒体上看不到的“灰色文献”，另外一方面又导致网上信息资源膨胀速度加大、信息污染程度加深、信息内容良莠不分、真伪难辨。因此，检索人员必须能够判断、鉴别信息的真伪和时效。在检索某个数据库以前，最好研究一下它的来源，或利用 Internet 信息检索工具系统地浏览所要检索的目标，经过多次选择以确定自己所需要的信息源。

(3) 检索软件智能程度较低。目前主要的浏览器和搜索引擎智能搜索功能不强，只能检索到含有用户指定的关键词的文件，检索不出与用户主题在意义上密切相关但并没有包含这些主题词的文件。

(4) 目前搜索引擎对网上信息资源的全文搜索主要是自然语言的搜索，各种搜索引擎建立的数据库也主要是根据词频的大小来确定，并不根据叙词、标题词来加以控制，类目的设置也不规范和标准，因此，查全率和标准率都得不到保证。

1.3 Internet 搜索引擎的主要检索方式

网上搜索引擎实际上是从 Internet 获得信息的程序。使用搜索引擎时，就是在该网页中按其语法要求输入所想要查找的文本的关键信息，引擎就能根据输入的检索要求输出与之匹配的 Web 地址表。每个搜索引擎都设有一个数据库，里面含有相应的 URL 地址以及其他的网络资源。搜索引擎的数据库由搜索程序通过页与页的链接顺序查找新的地址，不断更新，形成新的数据库。

网上虚拟图书馆(WWW Virtual Library)与网上搜索引擎原理基本一致，虚拟图书馆是 WWW 中的一个课题树，这一项目源于有分支或子目录的许多论题分类。在课题树的最底层是超级链接，在此用户能查寻并引用 Web 文档。这些服务器或具有类似功能的 WWW 服务器大多提供某一学科或领域的多种资源，用户从访问这些虚拟图书馆为起点，通过它所提供的与分布在 Internet 上的各种资源的链接，可以较方便地了解和获取自己感兴趣的信息。此外，用户还能查找国内外图书馆的书目信息，进行图书馆业务工作，如图书的订购、编目，查找各种文献摘要、数据库和免费的电子杂志、报纸等。

不同的搜索引擎和虚拟图书馆操作大同小异，但其内部组织的差异却非常大，目前通常有三种并行的方式来检索 WWW 中的信息。

1.3.1 按专题检索信息

第一种方式是按资源的专题性质进行检索，也就是分类体系检索方式，其目的是简化复杂的信息资源的组织，减少搜寻网上信息的时间。它提供一种可供检索和查询的等级式主题目录，以超文本链接的方式将不同学科、专业、行业和区域的信息按照分类或主题目录的方式组织起来。这种方式比较简单，大多数搜索引擎在其首页都提供分类范畴表，有的还分好几级类目，只需用鼠标在选中的主题上点一下，即可进入下级类目选择，或直接显示相关的站点或文献名称。这种方式基本上只需要点按鼠标操作，只是在最后一级可能需要输入一个关键词来限定一下检索范围，然后逐级浏览，直到找到与自己的需求有关的信息。一些用户在访问 WWW 时，事先并无特定的信息或检索目标，仅仅是希望对某一专业或专题进

行全面的了解,在这种情况下,可以使用按专题检索信息的方法。目前在 WWW 中,有许多按专题介绍和指引 WWW 信息资源的超文本,下面介绍几个按专题检索信息的 WWW 网点。

1) Yahoo!

该站点提供了分类的 Web 目录,共分 14 个大类主题:艺术、商业与经济、计算机和 Internet、教育、娱乐、政府、健康、新闻、消遣、参考工具书、地区信息、科学、社会科学、社会和文化,以超文本指南的方式将主题词链接起来。按 Yahoo! 的目录表来查找信息是一种最直观的且不加任何限制的查询方法,因为 Yahoo! 把 Internet 的主要信息资源都按主题分类排列于目录中,用户可任选一项进行查询。如果用户对体育感兴趣,就可移动鼠标至"Recreation and Sports(消遣与体育)"项,并按鼠标左键,这时 Yahoo! 就按其超链接式目录系统进入下一层目录中,显示全部有关的体育信息,接下来继续点击"Sports"项,进入下一层目录,用户可看到有关"Sports"的详细资料。如用户想看看"NBA"的情况,就点击"Basketball",进入下一层目录后,再点击"National Basketball Association[NBA]"即可。Yahoo! 就是这样通过逐层接近的方式,帮助用户达到最终目的。由于 Yahoo! 在这些方面做得很出色,所以一般公认 Yahoo! 代表了这种检索工具的最高水平,成为这一类检索工具的代表。

2) WWW 虚拟图书馆

在向用户提供按专题检索 WWW 信息方面,WWW 的发源地——位于瑞士的欧洲核子物理研究中心(GERN)一马当先,提供了 WWW 虚拟图书馆专题信息服务,这项服务后来由 WWW 的世界性组织 W3 提供。这是全球最著名的按专题链接 WWW 信息的一个"树根"。它的 URL 地址是:

http://www.W3.org/hypertext/datasources/bysubject/overview.html

WWW 虚拟图书馆是一个基于 WWW 超文本的按专题检索信息的服务,是分布式 WWW 信息目录服务。它提供信息总目,总目中的每个链接分别指向另外的超文本提供的专题信息目录,而这些提供专题信息目录的超文本并不都是由 W3 制作和由 W3 的提供者提供。这些目录分布在世界各地的许多不同结构的 WWW 服务器上。通过一级一级目录的链接,用户最终可以调出不再含任何链接的文献类文本,当然也可访问传统的 Internet 信息服务。

3) Galaxy

商业网络通信服务公司(EINet)开发的 Galaxy 是另一个按专题检索 WWW 信息的超文本。WWW 虚拟图书馆是学术机构间相互合作产生的信息服务,而 EINet 公司的 Galaxy 则是商业公司为了提供自身知名度向公众提供的免费信息咨询服务。EINet 公司开发的 Galaxy 的超文本的 URL 地址是:

http://www.einet.net/galaxy.html

它的 WWW 页面也是一个总目式的 WWW 超文本。Galaxy 的另外一个特点是通过表格操作可向该服务器提交增补主题内容的建议。

1.3.2 按关键词检索信息

第二种方式是通过查询关键词在 WWW 上检索信息,即关键词检索方式。关键词检索方式是索引式搜索引擎,提供对关键词、主题词或自然语言的查询,用户在搜索框中输入检索词或检索表达式,搜索引擎会返回一组指向相关站点的超级链接。AltaVista 是这类搜索

引擎的典型代表。它支持对关键词、主题词和自然语言的查询，用户可以选择对万维网(WWW)或新闻组(Usenet)查询，包括对篇名、URL、链接、主机、图像、文本等的检索。在其首页不提供类表，但是它的一个子站点 www.hot100.com 列出十多个类别的各 100 个热门站点名录供用户直接选择。要进入这个子站点，可以直接输入 www.hot100.com 或者点击 AltaVista 首页下部的"Our Network"，然后点击下一页的"Web21"。在提供按关键词检索信息的网点中，大多数采用各种自动化程序(即所谓的蜘蛛)来不断地搜索 WWW 网，并更新它们的数据库。这一点与 Archie 和 Veronica 有些类似，只是它们针对于不同的 Internet 资源。Internet 上还有一些现成的 WWW 资源索引数据库供用户使用，下面介绍几种可按关键词检索信息的网点。

1) Lycos

Lycos 是由美国卡内基·梅隆大学开发的按关键词检索 WWW 信息的工具。其 URL 地址为：

http://www.lycos.com/

2) WebCrawler

它是由华盛顿大学的 Brian Prinkerton 开发的，其 URL 地址为：

http://www.webcrawler.com/

3) InfoSeek

其拥有者为 Steve Kirsch for InfoSeek Corporation，它的 URL 地址为：

http://www.infoseek.com/

InfoSeek 也是 Internet 用户经常光顾的检索网点。

4) Excite

Excite 是由 Architext Software 公司开发的一个检索网点，它的 URL 地址为：

http://www.excite.com/

Excite 收集了 5 000 万页网页数据，它的检索方式由 Excite search、Excite city-net、Excite live 和 Excite reference 组成。Excite search 采用主题词检索；Excite city-net 帮助用户查看城市(美国)以决定旅行计划；Excite live 提供各种消息，包括运动、新闻、股市行情、电视节目、天气、电影评论等；Excite reference 提供找人、电子邮件、地图、共享软件和字典服务。Excite 的最大特点是采用一个称为"智能概念抽取"的专用查询软件，允许用户使用自然语言提问，例如"How to stay healthy by eating well"或"learn to speak chinese"等。不过目前本服务只能处理简单的 and 和 or 布尔逻辑检索，还不能处理高级查询功能。

5) OpenText

OpenText 可以对 100 万个 WWW 节点、FTP、Gopher 服务器进行全文索引，索引总量达 7.68 亿个词语。OpenText 提供六种产品及服务，它们是：在线 Internet、在线搜索、OpenText 索引、在线网络和服务、训练与支持。OpenText 提供简单查询、高级查询和加权查询。简单查询是基本的关键词与 and 和 or 匹配；高级查询可指定位置进行关键词查询，例如对全网、摘要、网页、第一标题、URL、超文本链接等进行布尔、邻接和字段查询；加权检索可对单词和词语加权，而且可以指定检索位置。输出结果按相关性排序，高级查询给出命中数。其 URL 地址为：

http://www.opentext.com/

6) Inktomi

Inktomi 于 1996 年 2 月由 Inktomi 公司推出,它支持 280 万个 WWW 主页的全文索引,其服务包括网络搜索、产品、合伙人、公司和技术五类。Inktomi 的高性能网络服务系统是美国加州大学发明的,该技术把一般的商业工作站组成多机并行处理系统,这使得 Inktomi 成为一个大型、快速、规模任意伸缩的查询引擎,具有性能成本比高、容错能力强的特点。其 URL 地址为:

http://www.inktomi.com/

1.3.3 按地区检索信息

第三种方式是按 WWW 服务器的结构和它们位于世界上的物理位置进行按地区的链接和查询,链接的常常是按国家、地区继而按机构排序的超文本。

1) 通过"虚拟旅游者"检索信息

"虚拟旅游者"(The Virtual Tourist)是一个按地区访问 WWW 的超文本页面,它不但提供文字说明,还配备含链接的世界地图。通过使用鼠标器点击世界地图的不同部分,用户可以非常直观地进行按地区检索信息。它的 URL 地址为:

http://www.vtourist.com/webmap/

"虚拟旅游者"实现了下述目标:

(1) 它为 WWW 世界提供了基于世界地图的信息检索界面。

(2) 它按国家和地区,将地理信息、旅游信息和文化信息汇集在一起。

(3) 它在建立按地区检索信息的服务时采取的方法是超级文本链接,并且依靠各个国家和地区的有关结构相互合作实现。

(4) 它有效地利用了 WWW 的图像链接功能,只要用户操纵鼠标点击图中含链接的部位,就可以进一步获取有关该地区的详细信息。

2) 通过 Gopher 检索信息

许多 Gopher 服务器都提供按地区检索信息的功能。我们可以用以下 URL 地址的 Gopher 服务器来按地区检索信息:

gopher://gopher.tc.umn.edu/11/other%20gopher%20and%20information%20servers/

1.4 搜索引擎的检索功能

同一般的数据库检索系统一样,搜索引擎通常也能提供以下 10 种常见的数据库检索功能,包括布尔逻辑检索、字符串检索、截词检索、字段检索、限制检索和位置检索等。但要注意的是,并非每一种 WWW 检索工具均能提供全部检索功能,也并非每一种检索功能在各个不同的 WWW 检索工具中的表现均完全相同。按照这 10 种检索功能在各种 WWW 检索工具中受支持的程度划分,排在最前的是布尔逻辑检索和字符串检索,几乎所有的 WWW 检索工具都支持这两项功能;排在最后的是位置检索,除 Alta Vista 外,几乎所有 WWW 检索工具目前都不支持此项功能;居中排列的是截词检索、字段检索和限制检索,它们受支持的程度因不同的检索工具而异。

1) 布尔逻辑检索

几乎所有的搜索引擎都具有布尔逻辑功能,尽管各个搜索引擎所使用的表示符各有差异,但所执行的操作相同。常用的布尔运算符包括 and、not、or。and 称为逻辑“与”,因为 and 联接的搜索词将同时出现在搜索结果中,使用 and 可以查找更多的信息。or 称为逻辑“或”,使用 or 可以查找范围更宽的信息,因为 or 联接的搜索词不必同时出现在搜索结果中。not 称为逻辑“非”,使用 not 可以剔除不需要的信息,因为 not 后面的搜索词不会出现在搜索的结果中。

2) 限定词(+,-)检索

限定词检索用在检索结果中必须包含或不能包含某词语的场合,大多数系统都具有该项功能。

+(required):把加号(+)放在一个词前表示在所有检索结果中都必须包含该词。如检索式“+ billiards + rules”,找到的是 billiards rules(台球规则)方面的资料,而“billiards + rules”则在检索结果中一定含有 rules,但不一定是有关 billiards 的。

-(prohibited):把减号(-)放在一个词前表示在任何检索结果中都不能包含该词。如检索式“+ billiards-equipment-supplies”,将排除有关 billiards 设备(equipment,supplies)方面的资料。

3) 字段检索和限制检索

字段检索和限制检索常常结合使用,这是因为限制检索往往是对字段的限制,而字段检索本身是限制检索的一种。在一般的数据库检索中,题名、叙词、标识词、文摘这四大主题字段采用后缀符限制,如“/ti”、“/de”、“/id”、“/ab”;而其他非主题字段则采用前缀符限制,如“au =”、“py =”等。但在 WWW 检索工具中,字段检索一律表现为前缀符限制形式,如属于主题字段限制的有:“Title:”、“Subject:”、“Keywords:”、“Summary:”等;属于非主题字段限制的有:“image:”、“text:”、“applet:”等。此外,作为一种网络检索工具,WWW 检索工具还额外提供了许多新的、带有典型网络检索特征的字段限制类型,如:主机名限制(host:)、超链限制(anchor:)、域名限制(domain:)、URL 限制(url:)、Link 限制(link:)、网址限制(site:)、新闻组限制(newsgroups:)、Email 限制(from:)等。如在中文雅虎中,“t:”指定仅查询网站名称,“u:”指定仅查网址(URLs)。ChinaByte 提供了 < in > title 检索,它可以将搜索限制在浏览窗口顶端地址标签部分。这些字段限制功能限定了检索词在数据库记录中出现的位置。由于检索词出现的位置对检索结果的相关性有一定的影响,因此,字段限制检索可以用来控制检索结果的相关性,以提高检索效果。在 WWW 检索工具中,目前能提供较丰富的限制检索功能的有 AltaVista、HotBot。例如:想查找题目中提到三峡的文献,可以输入title:three + gorges sanxia;想了解国外有多少网点做了河海大学网址的链接,可以输入link:www.hhu.edu.cn。

4) 词语检索

在一串词的前后加双引号(“ ”)或用连字符(-)连接,限定检索结果中的词语必须以同样的顺序出现,并且相邻。如输入“stupid pet tricks”或stupid-pet-tricks。

5) 截词检索

截词方法是利用某个词的一部分来检索,以便查找到一系列相关词的信息。这项技术可以简化搜索过程,减少构造检索策略的工作量,扩大检索范围,也启发在某种程度上检索出非确定信息。在带有该功能的网址中中英文截词方法往往是不同的,如北极星的中文截