网络环境下——数字信息资源的检索与利用

主编◎李晓燕
编著◎李晓燕 梁景汪 杨小月 杨 湖 段文莉 杨硕妍 韩宇峥

吉林文史出版社

网络环境下——数字信息资源检索与利用

主编◎李晓燕
编著◎李晓燕
汪景梁
杨小月
杨　湖
段文莉
杨硕妍
韩宇峥

（吉）新登字 07 号

WANGLUO HUANJING XIA SHUZI XINXI ZIYUAN DE JIANSUO YU LIYONG

网络环境下——数字信息资源的检索与利用

主　　编：李晓燕

编著：李晓燕 汪景梁 杨小月 杨 湖 段文莉 杨硕妍 韩宇峥

责任编辑：于　涉

吉林文史出版社出版发行　787 毫米× 1092 毫米　16 开本　15 印张　400 千字
（长春市人民大街 4646 号）　2003 年 12 月第 1 版　2003 年 12 月第 1 次印刷
长春市华艺印刷厂印刷　ISBN7-80626-621-6/G · 274　定价：40.00 元

目 录

第一章

数字信息资源及其检索概述

引 言

计算机技术和远程通信技术的发展，使 Internet(互联网)在 20 世纪的最后十多年内飞速地蔓延到了全球 240 个国家。它通过统一的通信协议(TCP/IP)，连接了世界范围内不同的信息系统，突破了地域、时空、文化、语言的限制。实现了跨国界、跨领域、实时的信息传递和交换。如今，网络和信息技术已经渗透到了文化、经济、政治、教育、科技、医疗、出版、新闻、体育、娱乐、商业以及社会生活各个领域，对社会、文化的发展产生了巨大的影响，人类认识世界和思考世界的观点、角度、方法都在不断地发生变化。

互联网的开放性及其信息资源共享和交换的能力，吸引了大量用户，越来越多的机构和个人在网上发布、查询和使用信息。据互联网协会(Internet Society，简称 ISOC)统计，1981 年全球提供服务的主机(hosu)不到 300 个，而 2001 年 1 月，已达到 10957.4429 万个，用户达到 3 亿多人，已注册的域名 3600 多万个(其中注册为 .Com 的商业机构约占三分之二)，发展之快，远远超出了最初人们的设想和预测，其影响早已超出国界、文化、语言的界限。

在中国，互联网更是按几何级数的速度增长，据中国互联网络信息中心 2003 年 1 月公布的最新统计报告，截止到 2002 年 12 月 31 日，我国国际出口带宽数已达到总容量 9380 兆(MB)。是 1997 年的 369.2 倍。上网计算机已经有 2083 万台，网上用户人数为 5910 万人：WWW 站点已有 37.16 万个，分别是 1997 年的 69.7 倍、95.3 倍和 248 倍。这些用户和网站的分布情况如下：中国公用计算机互联网(ChinaNET)、中国教育和科研计算机网(CER－NET)、中国金桥信息网(ChinaGBN)、中国科技网(CSTNET)、中国联通互联网(UNINET)、中国网通公用互联网(CNCNET)、中国移动互联网(CMNET)、中国长城互联网(CGWNET)、中国卫星集团互联网(CSNET)。

互联网的发展，使信息环境发生了巨大的变化，信息从产生、传播到使用，其种类和数量都与传统的资源大为不同。主要表现在：

1.信息的产生

(1)信息产生的速度加快，数量动态性持续增长。人们常说的“互联网年”为 3～4 个月，也就是说，互联网上每 3～4 个月增长的信息相当于传统方式的一年的信息量，这种发展将会逐年超过传统传媒工具报刊、广播、电视。

(2)信息的生产和发布不再仅限于官方或正式的出版机构，任何人都可以成为网上信息的发布者，具有很大的任意性和自由性。

2.信息的传播

(1)信息的传播速度大大加快，尽管信息没有集中存放，而是分布在全球各地但由于通过光纤网络传递，过去需要几天才能获取的信息现在几分钟内就可以得到。

(2)信息传播方式发生变比，传统的信息传播方式是通过书刊、广播、电视向受众单向传播，而现在是通过光纤网进行，因而信息与受众的关系也发生了变化。信息与受众之间的关系从单向灌输转为平等交流，受众可以自由选择所需信息。

3.信息的结构

(1)信息的内容发生很大变化，由于学术团体、政府机关、商业部门、个人、民间组织等任何组织或个人都可以在网上发布信息，因而对信息缺乏控制和管理，没有认证和审核，使得各种

信息，包括学术信息、商业信息、个人信息甚至有害信息混在一起。

(2)信息的形式从传统的印刷型图书期刊逐渐变为数字化信息，不再仅仅是视觉和静态形式，而是多媒体和动态的，需要功能强大的计算机软件系统来进行管理和使用。

(3)信息资源的种类也不再仅仅是传统的正式出版物，而是电子期刊/图书、非正式出版物、灰色文献(半公开出版物)、数据库、软件、新闻组、电子公告板(BBS)、FTP 等各类资源共同构成网络环境下的信息资源。

4.信息的使用

(1)通过网络获取信息可以不受任何时间、地点的限制，可以在任何地方、任何时间随时随地从网上获取信息。

(2)信息可以重复使用，可以较容易地进行二次、三次深度加工。

由于信息环境发生的上述变化，使得人们获取信息的途径发生了巨大变化，从网上直接获取信息的用户已逐渐增多。上述这些新的特点与变化都为人们获取和使用信息带来了新的问题：

①信息的选择问题。海量信息从最初给人们带来的惊喜变成了一种令人无所适从、被淹没在信息海洋中的感觉，信息的无序和混杂，使用户不知道如何从网上快速准确地选择对自己有用的、高质量、正确的信息，不知道什么是查询信息的有效途径，有时花了很大精力和时间，却没有任何收获。

②信息的检索问题。信息技术(IT)界和图书馆通过对信息的组织和管理，开发功能强大而友好的检索系统，为用户提供了结构化的信息以及有效的检索工具和检索途径。但是否具备检索能力，能否很快通过检索来获取和利用知识，仍然是目前多数用户面临的困难，因此对用户网络信息资源的检索和利用能力的培养就成为迫切需要决解的问题之一。

在此针对用户的具体需求再做进一步分析。中国互联网络信息中心 2001 年的调查表明，在上网用户中，按文化程度计算，本科以上文化程度者占 35.8%；按职业分析，专业技术人员占比例最大，为 20.6%。表 1-1 则说明用户的主要需求(仅列主要需求项目)。

表 1-1 网上用户需求分析

	新闻	计算机软硬件信息	休闲娱乐信息	电子书籍	科技教育信息
用户最希望获得哪方面信息	63.5	44.2	44.1	32.8	31.4
网上信息中哪些不能满足用户需求	19.3	19.9	18.3	29.1	21.1

从这些数字可以看出，上网用户中，学历较高者占三分之一强的比例；而用户对学术科研信息的需求量大很大，同样也三分之一左右，但在这方面不满足率并不高。这就说明，对于科研人员、专业技术人员、政府人员、教师、学生来说，更主要的问题是要解决学术和科研信息的和使用。

本书将针对上述需求，通过数字信息资源，主要是学术资源的获取和使用的介绍，解决用户在网络环境下查找学术科研信息所面临的一系列问题，提高用户对信息资源的选择和检索能力，尤其在以下方面取得突破：

(1)阐述数字信息资源的结构与体系；

(2)各类学术资源的定义、特点及其应用；

(3)主要数据库、电子期刊、电子图书的学科范围、发展概况、特点及其具体检索方法；

(4)网络学术信息资源的综合应用。

第一节 数字信息资源的概念与类型

数字信息资源(digital information resources),狭义讲,亦可称为电子资源(electronic resources),指一切以数字形式生产和发行的信息资源。所谓数字形式,是以能被计算机识别,不同序列的"0"和"1"构成的形式。数字资源中信息,包括文字、图片、声音、动态图像等,都是以数字代码方式存储在磁带、磁盘、光盘等介质上,通过计算机输出设备和网络传送出去,最终显示在用户的计算机终端上。

随着互联网的发展,利用网络的传递的数字信息资源的数量每年都以几何倍速增长,我们把这一类数字资源均称网络信息资源(networked information resources),网络信息资源目前在数字信息中已经占有绝对比例。除此之外,到目前为止,仍然存在着大量仅在本地计算机上使用,没有通过网络的信息,如只用于单机的光盘或机读磁带数据库等,我们把这一类资源也归为数字信息资源。

数字信息资源不同于以往的印刷型文献资源和各类视听资料,与之相比,其特点主要有:

(1)存储介质和传播形式发生变化。数字资源可以将传统的图书、期刊中的文字、图片以及各类音像资料中的声音、动态图像融合在一起,利用数字技术进行制作,存储在光盘、磁带或硬盘等载体上。同时以网络作为主要的传播媒介,即转变为光信号,利用网络实现同步传输。不仅传播的速度大大提高,传递的信息量也超过了传统的出版物。例如一张光盘的最大存储量是600兆(MB),一套标准版的《大不列颠百科全书》即可存储在一张光盘上,即使是多媒体版也只需要两张光盘;而现在对数字信息的计算单位已经从"兆"变为"千兆"(GB)甚至"兆兆"(TB),一个数据库的容量通常是以GB或TB为单位计算的。正常速度下,从网上下载一篇几千字的文献最多只需要1分钟左右的时间。

(2)以多媒体作为内容特征。集文本、图片、动态图像、声音、超链接等多种形式为一体,具体、生动、全方位地向用户展示主题,用户可以因此更加深入细致地了解所需信息的内容及特征。

(3)信息资源类型多种多样。既包括数据库、电子期刊、电子图书、电子报纸、专利等正式出版物,以及学位论文、教学软件等灰色文献,也涵盖了新闻组、电子公告板(BBS)等非正式出版的数字信息。信息交流的途径因此不再是单一化的,而是多层次、全方位的。

(4)多层次的信息服务功能。数字信息资源最初产生时主要的服务功能是信息检索,发展到今天,已经产生了一系列的新功能:主动报道,如期刊目次报道服务;文件传递,如FTP服务;信息发现,如网络资源学科导航、分类主题指南等;网上讨论,如BBS、新闻组等。这些服务功能扩展了传统出版物的职能,使数字信息资源得到更大程度、更深入的利用。

(5)更新速度快、时效性强。传统的印刷型出版物一旦出版后,信息的内容就无法更改,必须要修订后出版新版本。而数字信息资源的更新和发布就容易得多,只要有人负责不断跟踪各个领域的最新发展变化,就可以随时修改内容,每月、每周、每日甚至每时更新,及时发布给用户。

(6)具备检索系统,不再像传统文献那样需要逐页翻查,因而使用方便、快捷。特别是经过进一步加工的正式出版物,如电子期刊、数据库等。检索功能均很强大,可以很快找到自己所需的信息。

(7)不受时间、地域限制,即没有收藏地点(如图书馆)、收藏时间(开放时间)的局限,可以随时随地存取。

1.1 数字信息资源的产生与发展

1.1.1 数字信息资源的产生与发展

数字资源伴随1961年美国化学文摘社(CAS)开始发行"化学题录"(Chemical Title)机读磁带而诞生。40年以来,伴随着计算机和网络技术的发展。数字资源从无到有,从少到多,从书目型数据库发展到了全文数据库和电子图书、期刊、报纸乃至多媒体,从本地使用到网上发布,到今天最终成为人们生活和学习不可缺少的重要信息来源。

数字资源产生的最早的形式是数据库。1950年代初，随着电子计算机的产生，人们开始研究计算机情报检索系统，到1958年具有批处理能力的晶体管计算机产生后，计算机文献处理的研究开始有了突破性进展。1960年代初，最早的数据库“化学题录”和“医学索引”（美国国家医学图书馆）相继产生。至1965年，据《Computer-Readable Databases：a directory and data sourcebook》一书统计，已有大约20个数据库可供使用，但这时的数据库存储介质仅限于机读磁带，内容以科技文献书目、索引、文摘为主，检索也是以脱机批处理的方式进行，因此应用并不广泛。

1965年以后，由于集成电路计算机及硬盘的产生，以及数字通信技术和分组交换网的发展，开始有了数据库联机检索，著名的DIALOG系统以及MEDLINE、ORBIT、BRS、JOIS等相继开始服务，数据库的数量开始成倍增长，到1975年，已达到近300个数据库。数据库生产由政府行为逐步转向商业行为，用户也由原先的以政府机构为主，扩展到更多的图书馆和科研机构，在内容上也开始增加人文社会科学和应用科学等内容。

20世纪70年代以后，卫星通信技术、光纤通信技术、个人计算机的产生和发展给数据库联机检索制造了空前的发展机会，联机检索已不受地域限制，向国际化发展，个人用户开始加入到数据库检索行列中来。数据库的生产由美国向西欧扩展，在短短几年内即增长了10倍，到80年代末，数量已达到3600多个。数据库的容量增加，存储介质增加了光盘，因而也就产生了光盘数据库检索系统；数据库类型也有了变化，除以往的书目、文摘、索引数据库外，全文数据库开始迅速增加，而数值数据库、指南数据库等也已出现。

进入20世纪90年代，网络和信息处理技术的发展，使得基于互联网开发的数字资源及其检索系统有了突飞猛进的增长。仅从数据库即可看出这种发展之体现，以下为美国伊利诺依大学香槟分校图书馆情报学院教授Martha E. Williams对数据库发展的统计和归纳（数据截止到1999年）：

（1）数量：参见表1－2。

表1－2 数据库增长情况

	1975年	1999年	增长倍数
数据库	301	11681	39
数据库生产者	200	3674	18
数据库代理商	105	2454	23
数据记录条数	5200万	128.6万	242

（2）类型：打破了传统的文字、图片的单一类型，增加了集图像、声音、文字一体的多媒体数据库。其中文字型数据库中，在传统的书目、文摘索引、全文、事实数据库基础上又增加了电子图书、电子期刊和报纸以及其他动态信息。全文数据库（含全文电子期刊、图书、报纸等）从1985年占全部数据库比例的28%增长为1999年的50%，而书目型参考数据库从1985年的57%下降为1999年的23%，主要原因是因为互联网和计算机技术的发展为全文数据库提供了传输的方便和海量存储的可能。

（3）内容：主要分布在八大类：商业、新闻/综合、科技/工程、法律、医学/生命科学、人文科学、社会科学及各种交叉学科。其中，商业/经济类数据库的发展近年有所下降，但仍高居首位；科技类有所上升，增加了部分交叉学科的数据库。

（4）信息检索和存取（access）：除去联机检索、光盘检索外，新增了网络检索。从数据库的检索系统可以看出这种变化。光盘数据库占全部数据库的37%强，网络数据库所占比例最大，为44%，且还在增长的趋势中，这说明以互联网作为新的信息传递渠道，信息的存取是自由和交互式的。人们为数据库检索支付的费用比例也在不断增加。

除数据库外，互联网上的其他类型学术资源也发展很快，越来越多的正式出版物正在被放在网上，地域、时间、学科对科学研究的局限已逐步被打破，多数学术期刊已经把作者引用的网页作为正式的信息源列在参考文献中。这种发展主要体现在下列几个方面：

(1)电子期刊:目前在网上的纯电子版期刊和印刷版期刊的电子版已有两万种左右,这些期刊出版周期短,可以检索和重复下载全文,图像与文本结合,包含有多媒体及其他类型动态信息,具备超链接功能,可以向用户主动提供期刊目次报道服务。世界著名的杂志《科学》(Science)、《自然》(Nature)等均已上网服务。中文电子期刊中,目前较为知名、由电子期刊集成商生产发行的有"中国期刊网"、"维普"电子期刊、"万方"电子期刊。不仅在国内有大量用户,也开始向海外扩展市场和扩大影响。电子期刊使学术研究成果以更快的速度得到传播及使用。

(2)电子图书与工具书:可以逐页阅读,并能够快速检索书中的信息。尽管版权问题是困扰电子图书发展的一个主要因素,但电子图书仍以不可遏制的速度发展壮大。例如,著名的《不列颠百科全书》(Encyclopaedia Britannica)同时发行了光盘版和网络版,内容更新更快,检索方便,同时增加了大量图片和多媒体信息,其网络版还链接了上万个其他网站,同时保持了原有的学术质量,不仅吸引了大量用户。也在《电脑杂志》的多媒体百科全书评比中位于榜首。我国的《四库全书》发行电子版之后。由于改变了传统的印刷版必须先查索引、使用不便、需到图书馆使用等缺点,受到了用户的普遍欢迎。而电子图书集成商出版发行的电子图书,如"超星"电子图书、"网上图书馆"电子图书、"书生之家"电子图书、方正"阿帕比"电子图书等,也日益受到用户的关注,并开始广泛发行。

(3)电子报纸:目前主要以光盘和网络版两种形式发行,光盘版多用于发行过期报纸内容,网络版则由于增加了动态新闻、更新及时(许多报纸是以小时为单位更新的,重要内容则随时随地更新),兼之充分发挥了超文本链接技术,组织了大量专题以及多媒体新闻,每天都吸引了大量用户,著名的《纽约时报》、《华盛顿邮报》,以及我国的《人民日报》、《光明日报》等均已可以在网上阅读。

至于其他非正式出版信息,如政府信息、电子公告、搜索引擎、新闻组、FTP 站点等更是以几何级数的速度飞速发展,使得互联网逐步成为信息的海洋、知识的宝库,信息的结构和层次也逐步走向复杂和多样化。

1.1.2 中国的数据库业

我国的数据库研究工作始于 20 世纪 70 年代中期,80 年代初开始自建数据库的工作。90 年代以后,我国的数据库业得到了蓬勃发展,一些大型数据库开始以光盘形式进入市场,如国家专利局的"中国专利数据库"、国家标准局的"中国标准数据库"、重庆维普公司的"中文科技期刊数据库"、万方数据公司的"中国企业、公司及产品数据库"等。1993 年我国第一家数据专业制作公司——北京万方数据公司宣告成立,标志着我国专业化数据库企业工作的开始。1996 年,由中国学术期刊(光盘版)电子杂志社发行的"中国学术期刊"(光盘版)开始正式出版,这是我国的第一个电子版全文期刊产品;1999 年,在其基础上发展起来的"中国期刊网"开始在网上发布和提供服务,说明我国的数据库产业已开始由小规模的光盘生产向大规模的网络数据库形式发展。

到目前为止,我国的数据库内容几乎覆盖了科技、工程、经济、商业、金融、财政、交通、税务、文教、新闻出版、能源和国家事务诸方面(参见表 1 - 3),其中经济与社会方面数据约占 55%左右,科学技术方面为 45%左右,事实型、全文型数据库比重明显增加,商情数据库也日益增多,数据库总数已达到几千个,数据库总容量约占世界数据库的 1%。虽然与世界数据库业仍有较大距离,但总体规模扩大,不断向高水平、大规模发展,能拥有一定数量、对外提供服务、保持自身健康持续发展的数据库也在不断增加。

表 1 - 3 我国数据库学科分布情况

学科	专业	数量		百分比(%)
基础学科	数、理、化	17	119	11.46
	天文、地球	82		
	生物	20		
农林医	农林	50	92	8.86
	畜牧兽医	6		
	医药卫生	36		

学科	专业	数量		百分比(%)
工程技术	工程技术基础学科	46	255	24.7
技术	测绘、矿山、冶金	32		
	能源、动力与电器	22		
	计算机、通信、自动控制	38		
	纺织、食品、化学工程	25		
	水利、交通、土木建筑	64		
	航空与航天技术	20		
	环境科学	8		
经济	经济	297		28.61
社会科学	管理、统计	70	275	26.5
	图书情报	95		
	党政、社会及其他	110		
合计		1038		100

1.2 数字信息资源的类型

1.2.1 数字信息资源的类型

数字信息资源的范围非常广泛,其类型多种多样,划分标准也有很多种。

1.按照数字资源的性质和功能划分

借用印刷本文献的划分标准和名称,可分一次文献、二次文献、三次文献型资源。

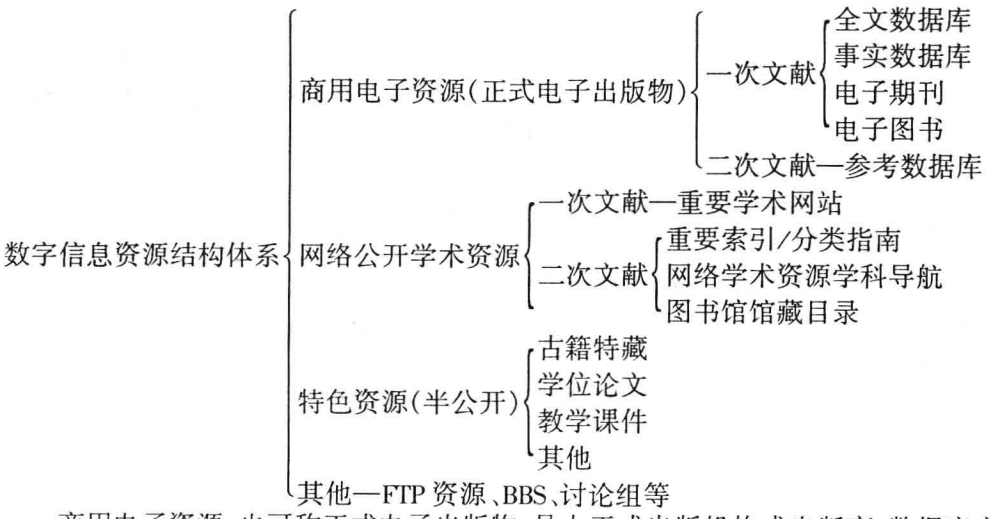
一次文献:即原始文献。指反映最原始思想、成果、过程以及对其进行分析、综合、总结的信息资源,如事实数据库、电子期刊、电子图书、发布一次文献的学术网站等。用户可以从一次文献中直接获取自己所需的原始信息。

二次文献:指对一次文献进行加工、整理,便于利用一次文献的信息资源,如参考数据库、网络资源学科导航、搜索引擎/分类指南等。二次文献可以把大量分散的一次文献按学科或主题集中起来,组织成无数相关信息的集合,向公众报道原始信息产生和存在的信息;同时也是一种有效的检索工具,供用户查找信息线索之用。

三次文献:指对二次文献进行综合分析、加工、整理的信息资源,如专门用于检索搜索引擎的搜索工具,比较典型的是 Web Crawler,被称为"搜索引擎之搜索引擎"(search engine of search engine),即"元搜索引擎",当用户进行检索时,反映出来的结果是各搜索引擎的检索结果。

2.按照数字资源的生产途径和发布范围划分

数据信息资源结构图



商用电子资源:也可称正式电子出版物,是由正式出版机构或出版商/数据库商出版发行的,在数字学术信息资源中所占比例最大,包括各类数据库和电子期刊、电子图书。其特点是:学术信息含量高;具备检索系统,便于检索利用;出版成本高,必须购买使用权才可以使用,因

此并不是面向社会公众免费开放的。

网络公开学术资源:这部分也可以说是半正式出版物,完全面向公众开放使用,包括各种学术团体、行业协会、政府机构、商业部门、教育机构等在网上正式发布的网页及其信息,亦属于一次文献类型;使用这部分信息主要依靠搜索引擎/分类指南、网络学术资源学科导航等二次文献资源;用于提供使用图书馆印刷型馆藏的联机公共目录(OPAC)也属于这部分范畴。

特色资源:也属于半正式出版物,主要基于各教育机构、政府机关、图书馆的一些特色收藏制作,在一定范围内分不同层次发行,不完全面向公众发行,有时需要特别申请。例如教师的教学课件(CAI),只在校园网内的教学范畴允许使用。

其他资源如 FTP 资源、新闻组、BBS、电子邮件等属于非正式出版物。

3.按照数字资源的载体划分

可分为光盘数据库、网络数据库等,我们将在本章第二节中做详细介绍。

4.按照数字资源的学科划分

主要包括的学科有:农业、人类学、天文学、艺术、航空航天、生命科学、商业/经济、儿童、计算机、教育、电子工程、环境、地理、地质、地球物理学、政府出版物、医学/健康、历史、新闻、法律、图书馆/信息科学、文学、数学、音乐、网络信息/组织/服务、海洋学、营养学、物理学、大众文化、体育/运动、宗教、科学、社会和文化、旅游、气象/气候/天气预报、饮食等,基本可以归纳为:学术信息,用于科学研究、科研信息发布等;教育信息,如教学课件、远程教学、学校信息、与专业、学科、教学相关的信息等;文化信息;政府信息;商业信息;有害和违法信息等等。

1.2.2 主要信息资源类型介绍

下面我们对本书将要重点讲述的资源做一简要介绍:

(1)参考数据库(reference database),指包含各种数据、信息或知识的原始来源和属性的数据库。数据库中的记录是通过对数据、信息或知识的再加工和过滤,如编目、索引、摘要、分类等,然后形成的。到目前为止,参考数据库主要是针对印刷型出版物开发的,目的是指引用户能够快速、全面地鉴别和找到相关信息。

参考数据库主要包括:书目数据库、文摘数据库、索引数据库。书目数据库主要是针对图书进行内容的报道与揭示的,如各图书馆的馆藏机读目录数据库;文摘和索引数据库则相对期刊论文、会议论文、专利文献、学位论文等进行内容和属性的认识与加工,如"科学引文索引"(Science Citation IndeX)、"化学文摘"(Chemical Abstracts)、"工程索引"(Engineering Index)、"生物学文摘"(Biological Abstracts)、"中国人民大学书报资料中心复印报刊资料索引总汇"等数据库。

(2)全文数据库(full-text database),即收录有原始文献全文的数据库,以期刊论文、会议论文、政府出版物、研究报告、法律条文和案例、商业信息等为主。如美国的 LEXIS - NEXIS 数据库、"学术期刊图书馆"(ProQuest Aaademic Research Library)及"中国人民大学书报资料中心复印报刊资料全文数据库"等。

(3)事实数据库(factual database),指包含大量数据、事实,直接提供原始资料的数据库,又分为数值数据库(numeric database)、指南数据库(directory)、术语数据库(termina logical database)等,相当于印刷型文献中的字典、辞典、手册、年鉴、百科全书、组织机构指南、人名录、公式与数表、图册(集)等。数值数据库,指专门以数值方式表示数据,如统计数据库、化学反应数据库等;指南数据库,如公司名录、产品目录等;术语数据库,即专门存储名词术语信息、词语信息等的数据库,如电子版百科全书、网络词典等。

全文数据库与事实数据库在 20 世纪 70 年代曾被合并称为源数据库(source database)。意指文献信息来源的数据库。后随着这两种类型的数据库的发展,逐渐分离。

(4)电子图书(electronic books),最初的电子图书主要以百科全书、字典词典等工具书为主,但近年来发展迅速,已涉及到了很多学科领域,文学作品、学术专著所占比例越来越大,电子图书正在逐步发展成为比较主要的数字信息资源。

(5)电子期刊(electronic journal,简称 e-journal)。包括:与纸本期刊并行的电子期刊,如著名

的“科学”(Science)、“自然”(Nature)、中国电子期刊杂志社的期刊等;纯电子期刊,如“数字图书馆杂志”(D-Lib Magazine)。

(6)电子报纸(electronic newspaper),目前网上已有数千种报纸供用户使用。同电子期刊一样,电子报纸同样也有印刷型报纸的电子版和纯电子报纸两种类型。

(7)搜索引擎/分类指南(search engine),是目前利用互联网开放信息的常用工具,也可以称得上是互联网开放信息的索引目录。搜索引擎主要是使用一种计算机自动搜索软件,在互联网上检索,将检索到的网页编入数据库中,并进行一定程度的自动标引,用户使用时输入检索词,搜索引擎将其与数据库中的信息匹配,然后产生检索结果。例如常用的 Yahoo、Hotbot、Alta Vista、Excite、Google、天网、悠游等。分类指南是将搜索到的网页按主题内容组织成等级结构(主题树),用户按照这个目录逐层深入,直到找到所需文献。通常搜索引擎与分类指南是结合在一起的,例如 Yahoo、新浪、悠游等。

(8)网络学术资源学科导航:将互联网上的开放信息加以甄别、筛选和科学整理,按学科组织起来,构成完整的学科导航系统,为教学、科研、技术人员提供各类学术信息。与搜索引擎/分类指南不同的是,网络学术资源的学科导航库通常是由图书馆单独或联合建设的。

(9)FTP 资源:FTP 含义是 File Transfer Protocol,意为文件传送协议,是互联网上最早应用的协议之一,它可以使用户远程登录到远端计算机上。把其中的文件传回到自己的计算机上,或把自己计算机上的文件上传到远端计算机系统上。所谓 FTP 资源,是指互联网上的开放 FTP 站点,这些站点允许用户登录上去,从中下载各类数据、资料、软件等。有些搜索引擎,如天网(http://e.pku.edu.cn),可以专门用来检索网上的 FTP 站点。

(10)其他:如网站、BBS、新闻组等,也可以给用户提供一些有用的知识或动态信息。

第二节 数字信息资源的检索

数字信息资源的检索,是指通过检索系统,采用一定的技术手段,根据一定的准则,在数据库或其他形式的网络信息资源中自动找出用户所需相关信息。简单地说,是一个信息存取(information access)的过程,是人、计算机和网络共同作用下自动完成的。

数字信息资源的检索源于信息资源的大量产生和飞速增长。它的基本工件原理包括两方面:一方面,为保证用户全面、准确、快速地获得所需信息,要对海量原始数字信息进行收集、加工、处理,对其重新进行规范化的组织和管理,使之从无序变为有序,从分散变为集中,从广泛性变为具有针对性(如针对某一学科或某一特定人群),从不易识别变为特征化(例如标出原始信息的名称、主题、创作者等),以便于识别和查找。这些加工整理过的信息存储成为数据之后,即以数据库或其他形式的资源存在。另一方面,对用户所表达的信息需求进行分析、并与数字资源中的信息进行匹配运算,自动分拣出二者相符合一致的部分,输出给用户,即为检索结果。在网络环境中,数字信息资源的检索是由人通过计算机(包括硬件环境和软件)即检索系统来进行的,因此我们也可以称之为计算机信息检索。与传统的文献检索相比,它提高了检索效果和检索的准确性,节约了人力和大量时间,逐步深入到了社会生活的各个方面。

广义地讲,数字信息资源的检索包括上述两方面的含义,即信息的“存”与“取”两个方面;狭义来看,则重点指后者。

2.1 数字信息资源的检索系统

同数字信息资源的检索可以分为广义和狭义来理解一样,其检索系统同样也包含了这两方面的意义,广义的检索系统也就是现代意义上的数字图书馆系统,它包括信息的采集、加工、组织、存储、管理、发布和检索等诸方面及其相应设备,从数据库的物理构成和功能来分析检索系统时,这些方面是密不可分的;而如果单从存储和用户检索方式来理解,则可以单指检索,即“取”信息这一部分。

2.1.1 检索系统的构成

(1)从物理构成来讲,检索系统由硬件、软件、数据库三部分组成:

①硬件(Hardware):也可以说是硬件环境,是和计算机检索有关的各种硬件设备的总称,如大型计算机主机(服务器)、存储器(硬盘或光盘)、网络(广域网、局域网或存储区域网等)、输入输出设备(键盘、打印机、鼠标等)、计算机终端或个人计算机(PC)等。

②软件(Software):与计算机检索相关的数据库系统软件及相关应用软件。包括:信息采集、存储、信息标引加工、建库、词表管理、用户检索界面、提问处理、网络发布、数据库管理等模块。随着网络和计算机技术的发展,软件的开发平台、程序语言的持续升级,用户功能需求的增加,这一部分的具体结构也在不断发生变化。

③数据库(database):数据库是指按一定方式、以数字形式存储、可通过计算机存取、相互关联的数据集合。数据库的特点是:重复数据少;可以共享数据资源,以最优的方式提供一个或多个应用服务;数据具有独立性,其存放独立于应用程序之外。由于数据库中的信息都经过了详细、精心的选择和加工。主题化、有序并能够提供多种检索途径,因此相对互联网上无组织和大量无用的信息来说,检索结果准确,时间少,价值高。从发展的角度看,以网络为中心的分布式数据库系统是今后的发展趋势。

(2)按照功能划分,广义的检索系统(数字图书馆系统)又可以分为以下几个模块:

①信息采集模块(collection):本模块的任务是连续、快速地采集各类信息,为数据库提供充足的数据来源。在传统的信息采集工作中,这项工作主要是收集印刷型文献中的信息,因此以人工为主,计算机只起辅助作用,如录入、扫描(包括扫描后的光学字符识别)、视音频采集等。现在,随着原始信息的数字化,智能型软件(如搜索引擎应用的机器人软件)正在逐步取代人工的工作,越来越多的信息采集工作是由系统自动进行的,只由人工进行质量控制。

②信息存储模块(repositories):如同传统图书馆要有书库来收藏书刊一样,本模块的功能是对数字资源进行存储和管理,数字资源按照不同类型,如文字、声音、图像、数字等,按不同的格式(如 GIF、JPEG、DOC、HTML 等)被存储在不同的数据仓库中。存储介质包括磁带、磁盘、光盘。从根本上讲,存储方式决定了应用方式,存储方案决定了整个系统的扩展性和灵活性。

③标引著录模块(description):即对信息的内容和特征进行分析,然后给予一定数量的标识,作为信息组织、存储与检索的基础。例如信息的名称、创作者、主题、分类、出版/生产时间、出版/生产者、关键词等,都可以作为信息的描述性标识。

传统的标引著录多针对印刷型文献,因此采用人工标引录入的方式,速度较慢。而数字化信息的产生正在使标引著录逐步走向计算机化、智能化,信息的著录格式也逐步由以往常用的机读目录格式(MARC)转为多元化的元数据(metadata)格式,使得著录更具有专指度和更为快速。

④规范模块(authorities):指对信息特征和用户提问的语言形式做出规定,如主题词表、人名规范、地名规范、时代名称规范等,目的在于:一是使用户的检索更具准确性;二是逐步形成一个知识网络,通过相关信息的提供,使用户的检索更为完整。在传统的计算机检索系统中,规范模块是封闭的、各种规范也是相对独立的,只能由信息的加工者使用、修改和维护。而在网络环境下,各种规范模块已逐步融合成为网络知识组织体系(networkeknowledgeorgan-ization systems),其中的一部分,特别是主题词表部分已逐步开放,自动积累吸收用户的词汇,因此更加完善。规范系统在使用中,可以独立于数据库系统之外,与数据库系统挂接使用。

⑤内容发布模块(publish):将数据库内容传递到网络上,让用户以常规手段(如通过浏览器)查询浏览。这方面的技术涉及到网络协议、媒体特性、易用性、信息导航、语言转换等许多方面。

⑥检索模块(access):也就是狭义理解的检索系统,即将用户的需求进行分析,并和数据库中的信息匹配运算,再反馈给用户所需的检索结果。检索模块一般包含有:a)检索界面:即人机接口;b)检索功能:如简单检索、复杂检索、浏览、图像检索等;c)检索途径:如题名、作者、主题、文摘等检索入口;d)检索技术:如布尔逻辑、组配检索、截词符、词根检索、位置算符等;e)检

索结果:打印、存盘、结果格式、二次检索;f)提问处理:也可称匹配运算,即处理和运算用户的检索式。在网络环境下,为了更好地根据用户需求完善数据库系统,检索模块同时具备交互功能,即自动收集、累积用户的检索需求,再由管理系统进行分析。

⑦服务模块(service):这是在传统检索系统基础上发展起来的新功能,即不仅向用户提供检索,也在信息资源基础上,根据用户需求,为用户提供一些可定制的服务,以及由系统主动向用户提供新的服务内容,如在检索系统中提供培训教程,由系统定期、自动发送最新期刊目次(E-mail alert),根据用户反馈回来的请求为用户提供文献传递服务(document delivery)、虚拟咨询(virtual reference)服务等。

⑧管理模块(administration):主要指管理客户端,即对用户和用户行为进行管理和调查分析。主要包括三个部分:一是对用户的管理,如用户类型、用户认证、用户权限、IP(Internet Protocol)地址控制、用户名和密码、并发用户限制等。二是运用数学和统计学方法,对用户行为的各种相关信息进行累积、加工、分析,生成各种状态报告,提供给数据库生产者、系统开发者和用户,以便对数据库及其系统进行修改、完善,使其不断得到提高。如用户使用统计报告,就可以通过对用户使用情况的统计数字,来分析用户是否很好地利用了数据库,其中反映了什么问题。三是监控系统使用情况,如观察用户有无违反版权规定、恶意下载(abuse)现象,并对违法用户进行相应处罚。管理模块同样包括为机构用户(如图书馆)提供的业务管理服务。如下载MARC格式的书目数据,简单地修改页面设置,查看本馆用户的使用统计,下载出版物列表等。

(3)按存储设备和用户检索方式划分检索系统类型:

联机数据库检索系统、光盘数据库检索系统、网络数据库检索系统。本节会在后面的内容中逐步详细介绍这几种类型。

2.1.2 检索系统的评价

使用一个检索系统,要相应地对其进行评价,从而确定一个检索系统是否做到了功能全面、界面友好和服务周到。

(1)检索功能:主要是指系统提供给用户的各种检索途径和检索入口,可供选择的越多,相对用户就越方便。比较关键的问题是如何使各种功能配置合理,并在检索系统首页上选择用户最易接受的缺省(default)功能。这方面不同的检索系统使用的方法不尽相同。本书将在今后各章中逐步介绍。

(2)检索技术:即系统是否允许用户使用各种检索技巧,以便更准确和快速地找到自己所需信息。

(3)检索结果:即用户是否得到了内容全面、下载和使用均比较方便的检索结果,例如显示格式包含的内容是否全面;检索结果数量较多时是否允许在翻页的同时标记记录;是否提供存盘、打印、E-mail发送等多种下载功能;检索结果是否与其他资源之间存在链接,为用户提供查找到其他资源的捷径等。

(4)用户服务:主要是指在检索功能之外,系统还为用户提供了哪些服务。具体包括:检索帮助文件是否完整、详细、易查;是否可以记录读者的检索历史,以便用户随时可以利用和翻看以前的检索结果;有无词表、名录等常用参考工具,可随时查阅;允许用户对检索界面做一些小的调整,更方便使用;电子期刊提供最新目次报道服务;网上提供培训教程,便于用户自我培训等。

2.2 联机数据库检索

联机检索(online retrieval)是指用户利用计算机终端设备,通过通信线路或网络,在联机检索中心的数据库中进行检索并获得信息的过程。

联机系统由联机检索中心、通信设施、检索终端三部分组成。联机检索中心是该系统的中枢部分,主要包括中央计算机(硬件)、数据库、系统和检索软件等部分。中央计算机又称为"主机",其功能是在系统和检索软件支持下完成对信息的存储、处理和检索。通信设施由通信网(电话网、专用数据库网等)、调制解调器及其他通信设备组成,终端则可以使用传统的终端机

或个人计算机。

联机检索的工作原理是:用户用电话或专用线接通检索中心,在终端键入指令,将信息需求按系统规定的检索命令和查询方式经过通信网络发送到系统的主机及数据库中,系统将用户的请求与数据库中的数据进行匹配运算,再把检索结果反向送回到检索终端。

联机检索的特点是:

(1)数据库数量多,信息量大,内容丰富。以 DLALOG 系统为例,目前已有数据库 300 多个,记录 3 亿多条,内容广泛,涉及自然科学、人文及社会科学多个领域。检索时可以一次检索多个数据库,检索范围广泛全面。

(2)数据库更新快,每日可随时进行更新,可以很容易检索到最新文献。

(3)数据库和系统集中式管理,安全性好,可以在存储设备上直接处理大量数据,但主机的负担重,网络扩展性差。

(4)检索模式:主仆式,即所有的工件都在主机上进行,一旦主机瘫痪,所有系统都处于瘫痪状态,因此对主机的性能要求极高。

(5)信息组织模式:普通线性文本,包括:按照文档号组成的顺排文档;按照记录的特征标识(如题名、作者等)组成的倒排文档。文档的基本组成单位是记录,文档之间没有任何关联。这种信息组织模式有利于高效、准确的检索,但很难建立知识的体系。

(6)检索机制:检索功能强,索引多,途径多,所有的数据库使用统一的命令检索,因此可以同时保证查全、查准,检索效率和检索质量高。但系统要求必须使用统一的检索命令。用户必须记住各类检索指令并且能够灵活综合运用,因此必须由专业人员检索,例如在图书馆或专业信息机构中,都有专门人员负责联机检索。这种检索机制不利于在网络环境下扩展为大规模的使用。

(7)检索费用高:每下载一条记录都要支付相关费用,包括记录的显示或打印费、字符费、机时费、通信费(由于系统连接需通过通信线路或网络进行,需支付高额通信费用),检索时必须快速进行,一般用户因此望而却步,不敢使用。

(8)检索界面单一,过于呆板。

自 20 世纪 70 年代以来,联机数据库检索系统发展异常迅速,盛极时曾有 DIALOG、STN、LEXIS - NEXIS、ORBIT 等多个大型检索系统,为用户提供了高质量、远胜于传统手工查询的信息服务,几乎每个图书馆或信息服务机构、中大型公司里都有专门进行联机检索、为用户或本机构决策提供信息服务的检索专家。

但进入 80 年代末、90 年代初以后,互联网的迅速发展,导致越来越多的用户在网上自行寻找自己所需的信息,而联机检索由于对检索人员的要求高、费用贵等原因,开始进入衰退时代,几家著名的联机检索公司逐渐被并购或倒闭,仅存的 DIALOG 公司、LEXIS - NEXIS 公司等,被并购后仍保留了原有的系统名称和品牌,但也相继推出了基于互联网的网络检索机制,以提供普通检索用户使用。目前,联机检索的方式虽然仍然存在,但与后来居上的光盘检索、网络检索相比,用户量较少,大部分使用者仍然是检索专家。

2.3 光盘数据库检索

光盘数据库通常是指 CD - ROM 数据库。CD - ROM(computer Disc read - only memory),意为只读光盘,轻便、灵活、体积小、容量大,一张只读光盘的最大存储量为 600M,可存储文字、图片、图像、声音等。

光盘数据库检索产生于 20 世纪 80 年代末,最初是在微机上,利用微机的光盘驱动器进行单机检索。以后随着数据库容量和光盘数量的增加,逐渐发展出了联机光盘检索。

单机光盘检索系统由微机、光盘驱动器(光驱)、光盘数据库、系统软件等组成,自成系统,在微机上即可检索数据库,可供单个用户进行本地检索。由于单机检索可支持的同时检索的光盘数量有限,使用的数据量较小,通常使用者为个人用户。当一个数据库有若干张光盘时使用单机光盘检索就很不方便,必须不停地在光驱上退盘、插盘。因此,在数据库飞速发展的今

天.一般图书馆、信息服务机构都使用联机光盘检索。

联机光盘检索是指把单用户系统发展成多用户的局域网系统,通过网络(如校园网)连接各个用户终端,用服务器管理多组光盘数据库及其检索系统。联机光盘检索系统由若干台微机、光盘驱动器、光盘服务器、光盘数据库、检索系统软件、管理系统软件构成,主要性能如下:

(1)光盘服务器:在整个光盘检索系统中起着主控作用,当终端用户访问光盘塔上的数据时,服务器传输映射命令,控制光盘塔上的光驱工作,再把光盘塔上查询到的数据反传给客户端、光盘服务器可选用性能好、高配置的专用 PC 服务器。

(2)光盘驱动器:主要指塔式光盘驱动器(光盘塔)、光盘库。光盘塔由若干光驱(标准配置为 7 个、14 个、28 个等)组成,可同时支持几十张甚至上百张光盘的检索,实现数据共享,统一管理光盘数据。光盘库(jukebox)可同时存储大容量、多盘片光盘(几百张),并同时读取若干张光盘(4 光驱、6 光驱等)。二者的不同之处在于:光盘塔可存储的光盘容量有限,但数据均为在线数据(online data),不需再次调用光盘即可检索;光盘库的光盘存储量则比较大,但数据为半在线数据(nearly online data),必须通过索引盘调用数据光盘才能使用,检索时间长,检索效率比较低。考虑到这两方面的缺陷,综合二者的优点,现在又发展出了磁盘阵列,即把若干硬盘挂接在光盘服务器上。将光盘数据拷入硬盘中,做虚拟光盘检索,这样既可以实现大容量数据存储,也可以缩短读取数据的时间;当然,展现在用户客户端上的,仍是光盘检索的形式。

(3)软件系统:包括服务器端和客户端软件,以及数据库检索系统。服务器端软件最常见的是基于 windowsNT 开发的光盘服务器操作系统,主要用于管理光盘数据库、调度光盘数据、记录和统计用户使用情况。客户端软件主要用于接受用户请求,提供各种检索途径,将用户请求发送到服务器端,并将检索结果显示给终端用户。数据库检索系统主要用于管理不同数据库的数据,接收用户请求,进行匹配运算,再将数据返回到客户端。数据库检索系统和客户端软件通常因数据库的不同而不同。

联机光盘检索系统的特点是:

(1)由于存储介质和空间的限制,数据库数量没有联机检索多,信息量不够大,多以二次文献(文摘、索引)为主。

(2)数据库系统建立在用户方,出版商必须寄送光盘给用户,因此更新速度慢,一般为月更新或季更新。这方面不如联机数据库和网络数据库,后两者的数据库更新可以随时进行,频率通常为日更新和周更新。

(3)与网络数据库检索相比,数据库和系统集中式管理,负担重,数据库和用户越多,响应时间越长。

(4)检索模式:以客户端/服务器方式为主,客户方在微机上运作,这种检索模式与联机数据库相比,检索效率提高了很多。

(5)信息组织模式:普通线性文本。

(6)检索机制:检索功能强,索引多,不同的检索系统使用不同的检索命令。具备命令检索和菜单检索两种方式,后者对非专业人员来说,易学易用。

(7)系统访问通过局域网就可以进行,不受大的网络环境影响,不需支付网络通信费用。

(8)检索环境宽松,不存在联机检索的通信费、机时费、数据费,检索费用低。

(9)用户界面比较友好,较为直观。

光盘数据库检索从 20 世纪的 80 年代后期开始,经历了大约 10 年左右相对兴盛发展的时期。从 90 年代后期开始,随着互联网的发展,特别是一次文献数据库业(电子期刊、电子图书、事实数据库等)的壮大,光盘数据库逐步暴露出其局限性,无法提供大数据量的存储和处理大用户量的访问。因此,在网络比较发达的地区,已逐步为网络数据库检索取代。目前,光盘数据库仍在局域网条件较好、广域网发展尚不成熟的地区广泛使用。

2.4 网络数据库检索

网络数据库(web - database)检索是指用户在自己的客户端上,通过互联网和浏览器界面