普通高等学校教材

网络信息检索与利用

Internet Information Retrieval and Utilization

许春芳　主编

吉林科学技术出版社
WWW.JLSTP.COM

网络信息检索与利用

主　编　许春芳
编　委　（按姓氏笔画为序）
王重阳　王　彬　吕　娜　许　鹏
张智刚　李传华　沈洪杰　郭淑艳

吉林科学技术出版社

内容提要

网络信息资源无限、无序，在这浩瀚的信息海洋中，如何快速、准确地获取所要找的信息，是我们每个人都要面临的问题，本书系统、全面地介绍了网络信息资源的基础知识，网络信息检索的基本原理、基本方法，各种网络检索工具的功能、特点和使用方法，分门别类地介绍了学术研究性信息及各类参考信息的检索途径及检索方法。

本书系统全面，自成体系，内容丰富、新颖，实用性强，可作为高等学校各专业本科生、研究生网络信息资源检索与利用的教材与教学参考书，是广大网络信息用户的工具指南。

网络信息检索与利用
主编：　许春芳
责任编辑：赵沫　　封面设计：张智刚
*
吉林科学技术出版社出版、发行
长春市东文印刷厂印刷
*
787×1092 毫米　16 开本　12 印张　280 000 字
2006 年 11 月第 1 版　2006 年 11 月第 1 次印刷
定价：36.00 元
ISBN 7-5384-1838-5

社址　长春市人民大街 4646 号　邮编　130021
电话　0431-85674016
电子信箱　JLKJCBS@public.cc.jl.cn
传真　0431-85635185
网址　www.jlstp.com　　实名　吉林科技出版社

前　言

网络世界瞬息万变，网络信息浩如烟海，如何在这纷繁的网络信息中快速准确地获取所需的信息资料，从而驾驭网络，是我们每个人生活、学习、工作中都要面临的课题。本书通过系统全面地介绍网络信息检索的基本原理、基础知识和基本方法，旨在为读者提供快速检索网络信息的工具指南，使用户掌握网络信息组织、管理的基础知识，网络信息检索的方法、手段和途径，从而获得网络信息检索的基本技能，提高在网络环境下生存与发展的能力。

本书作者结合高等学校学生网络信息检索与利用教学的理论与实践，运用近年来网络信息检索的理论研究成果，根据网络信息资源和网络搜索引擎发展的实际，编写了本书。该书是全军重点网络课程的前期成果。全书共分 8 部分，1. 网络信息资源；2. 网络信息资源检索；3. 网络信息检索工具；4. 中文网络数据库的检索与利用；5. 外文网络数据库的检索与利用；6. 各类学术信息的网上检索与利用；7. 参考信息的网上检索与利用；8. 网络信息检索技术的未来发展；附录：各专业网站介绍。其中第一部分、第八部分由许春芳编写，第二部分由王重阳编写，第三部分由李传华编写，第四部分、第五部分由沈洪杰编写，第六部分由郭淑艳编写，第七部分由吕娜编写，附录由张智刚编写，许春芳编写本书大纲、统稿与主审。

全书在编写过程中参阅借鉴了很多论著与文章，在这里谨向作者致以诚挚的谢意，未在参考书目中一一列出的，亦请作者见谅。

由于网络信息的发展变化之神速及数据库资源变化，加之作者的认知能力有限，书中内容难免有疏漏与谬误之处，敬请读者与业界同行批评指正，以期在今后的学习与教学工作中不断改进。

许春芳

2006 年 9 月

目 录

第1章 网络信息资源

1.1 互联网概述

1.1.1 互联网概况

随着计算机技术与远程技术的飞速发展，在20世纪60年代，出现了因特网，开始主要应用于美国的军事目的，而后发展为民用、教育与学术、商用。特别是到20世纪末，互联网已经相当普及，影响到人们生活的方方面面，使人类生活、学习、工作方式发生改变，从而对于整个世界政治、文化、经济产生了巨大影响。因特网是一个动态发展的概念，发展至今，互联网已经成为人类传播、交流信息的主要渠道，成为人类生活不可或缺的一部分。

国际“联合网络委员会”(FNC)曾在1995年10月24日关于“因特网定义”的决议中指出，因特网是指全球性的信息系统通过全球性的唯一的地址逻辑地链接着，这个地址是建立在“网络间协议”(IP)或今后其他协议基础之上的；可以通过“传输控制协议”和“网络协议”(TCP/IP)，或者今后其他接替的协议与网络间协议兼容的协议来进行通信；可以让公共用户或者私人用户使用高水平服务，这种服务是建立在上述通信及相关的基础设施之上的。

从物理联系上看，互联网是由成千上万个具有特殊功能的专用计算机通过各种通信线路连接而成。在结构上包括两部分，一部分是连接于网络上的可供用户使用的计算机，即主机(host)，用来运行用户的应用程序，为用户提供资源和服务，网络上的主机又称为结点。另一部分是用来把主机连结在一起并在主机间传送信息的设施，分散在各地的计算机网络通过物理连接形成庞大的计算机网，在这些计算机之间建立了物理上的联系。

从通信角度上看，互联网是通过网络协议(TCP/IP)把各种计算机网络连接起来进行数据通信。这些计算机遵循严密的网络协议进行数据传输。

从信息资源来看，互联网拥有庞大的信息资源，既有教育科研等学术性信息，又有政府、组织等官方信息；既有文化、学习信息，又有休闲娱乐信息；既有企业、事业信息，又有商业产品信息……应该说，互联网是人类巨大的信息宝库，网络用户可以随时通过信息查询工具访问所有资源。

从功能上来看，互联网既可以进行数据传输、信息查询、远程教育、学术交流，又可以进行商业营销、电子支付、网上交易、中介服务；既可以进行医疗会诊、家庭保健、社区帮助，又可以进行网上聊天、兴趣交友、游戏活动等各种休闲娱乐活动服务。

1.1.2 TCP/IP协议

互联网采用统一的通信协议(TCP/IP)，连接着世界范围内的各种计算机系统。TCP/IP传输控制协议(Transmission Control Protocol，TCP)提供传输层服务，国际协议(Internet Protocol，IP)提供网络层服务。是一组协议，包括上百个各种功能的协议，如远程登录，文件

传输和电子邮件等，而 TCP 协议和 IP 协议是保证数据完整传输的两个基本的重要协议。

按协议分层划分，TCP/IP 包括 4 个层次：网络接口层，网间网层，传输层和应用层。

最低层是网络接口层，负责接收 IP 数据包并通过网络发送，或者从网络上接收物理帧，抽出 IP 数据包，交给 IP 层。常用的协议有 ARP(address resolution protocol，地址解析协议)和 RARP (reverse address resolution protocol，逆向地址解析协议)。ARP 将 Internet 的 IP 地址映射成物理网的 MAK 地址，而 RARP 负责将物理网的 MAK 地址映射成 Internet 的 IP 地址。

网间网层负责相邻计算机之间的通信，处理来自传输层的分组发送请求，将分组装入 IP 数据包，填充报头，选择去往信宿机的路径，然后将数据包发送适当的网络接口，处理输入数据包，首先检查其合法性，然后寻径，数据包已到达信宿机，去掉报头，将剩下部分交给适当的传输协议，如数据包未到达信宿机，则转发该数据包，处理 ICMP 报文，处理路径、流控、拥塞等问题。

常用的协议有 IP，ICMP(internet control message protocol，因特网控制消息协议)和 RIP (routing information protocol 路由信息协议)等。IP 的主要功能一是提供无连接，有效的数据分发，二是提供数据包的分组和重组，以支持不同最大传输单元 MUT 的数据链路。ICMP 用来提供错误报告信息，RIP 提供最佳路由和转发数据分组。

传输层提供应用程序间的通信，包括格式化的信息流和提供可靠传输。主要协议有 TCP 和 UDP(ueser datagram protocol，用户数据包协议)，正如网络层控制着主机之间的数据层传递，传输层控制那些将进入网络层的数据。TCP 提供 IP 环境下面向连接的数据可靠传输，提供的服务包括数据流传送，可靠性，流量控制，全双工操作和多路复用等。UDP 是面向无连接服务的管理方式的协议，提供无连接的低可靠的通信服务。

应用层是最低层，向用户提供一组常用的应用程序，比如电子邮件，文件传输访问，远程登录等，如 Telnet 提供远程登录服务；FTP 提供应用级的文件传输服务；SMTP 提供传输邮件服务；TFTP 提供小而简单的文件传输服务；SNTP 简单网络管理协议；DNS 提供域名解析服务；HTTP 提供超文本传输协议。

WWW(World Wide Web)是万维网，环球网或 Web，是一种基于超文本(Hypertext)方式信息服务工具，它将因特网上的位于全世界不同地点的相关数据信息，通过友好的信息查询接口提供给用户，人们只需要提供查询请求，就可以得到想要得到的文本、图像、声音，动画等信息。WWW 与传统的 Gopher、WAIS 最大的区别是提供给用户一篇篇原文，而不是菜单，它直观、简单、实用。WWW 开发了一套标准的、易被人们掌握的超文本开发语言(HTML)、统一资源定位格式 URL 和超文本传输通信协议 HTTP。

1.1.3 IP 地址

连接互联网的每一台计算机都有唯一的表明自己身份的地址，即 IP 地址，因特网网络信息中心(Internet NIC)统一负责全球地址的规划与管理，每个国家又有一个部门统一向国际组织申请 IP 地址，然后分配给用户。每台计算机的 IP 地址按照网络规模大小，分为 A、B、C 三大类地址和 D、E 两类特殊地址。IP 地址的格式为：各类＋网络地址＋主机地址。

A 类地址主要用于大型网络，主机容量为 16 777 216 台，A 类地址首位是 0，网络地址占 7 位，主机地址占 24 位，IP 地址范围为 0. 0. 0. 0～127. 255. 255. 255。

B 类地址用于中型的网络，主机容量为 65 536 台，前两位是 10，网络地址占 14 位，主机地址 16 位，IP 地址范围为 128. 0. 0. 0～191. 255. 255. 255。

C类地址用于小型网络，每个网络最多不超过256台主机，前3位为110，网络地址占21位，主机地址占8位。IP地址范围为192.0.0.0～223.255.255.255。

D类地址用于网络上多台主机同时进行通信的地址，IP地址范围为224.0.0.0～239.255.255.255。

E类地址是一类以备将来使用的特殊地址，IP地址范围是240.0.0.0～247.255.255.255。

1.1.4　DNS域名系统

DNS域名系统(Domain Name Systerm)是Internet采用的另一套字符型的域名系统，它用有一定含义的字符串来标识主机地址，域名系统是为方便解释机器的IP地址而设立的，更加便于人们记忆。它采取分层次结构，按地址域或机构域分层，用圆点“·”将各个层次分开依次为第一级(顶级)域名，第二级域名，……域名可以用一个字母，数字开头和结尾，并且中间的字符只能是字母，数字和连字符，标号必须小于255。

通常DNS与IP是对等的，在使用过程中域名服务计算机将自动完成从域名到IP地址的转换。Internet中每台计算机的域名结构如下：

计算机主机名·说属机构名·计算机网络名·国家顶级域名

如：清华大学的域名为：www.tsinghua.edu.cn，即WWW类服务器·大学名·教育机构·中国。北京图书馆的域名为：www.nlc.gov.cn，即WWW类服务器·单位名·政府部门·中国。

顶级域名(TLD)表示主机所属的国家、地区或网络地址代码。目前分为3类：

1. 国家顶级域名(nTLD)，如.cn表示中国，.us表示美国，.ca表示加拿大。详见国别域名表。

2. 国际顶级域名(iTLD)，供国际组织和联盟注册使用，如.int代表国际组织。

3. 通用顶级域名(gTLD)，目前有13个，见通用顶级域名表。

国家(地区)顶级域名表

域名	国家	域名	国家	域名	国家
al	阿尔巴尼亚	fi	斐济	no	挪威
am	亚美尼亚	fr	法国	nz	新西兰
aq	南极洲	gb	英国(官方)	pa	巴拿马
ar	阿根廷	gr	希腊	pe	秘鲁
at	奥地利	gu	关岛	pg	巴布亚新几内亚
au	澳大利亚	hk	香港	ph	菲律宾
az	阿塞拜疆	hr	克罗地亚	Pl	波兰
bb	巴巴多斯	hu	匈牙利	pr	波多黎各
be	比利时	id	印度尼西亚	pt	葡萄牙
bg	保加利亚	ie	爱尔兰	py	巴拉圭

续表

域名	国家	域名	国家	域名	国家
bo	玻利维亚	il	以色列	re	留尼汪
br	巴西	in	印度	ro	罗马尼亚
bs	巴哈马	ir	伊朗	ru	俄罗斯联邦
bz	伯里兹	is	冰岛	sa	沙特阿拉伯
ca	加拿大	it	意大利	se	瑞典
ch	瑞士	za	南非	sg	新加坡
cl	智利	jp	日本	si	斯洛文尼亚
cm	喀麦隆	km	科摩罗	sk	斯洛伐克
cn	中国	kr	韩国	sr	苏里南
co	哥伦比亚	kw	科威特	th	泰国
cr	哥斯达黎加	lk	斯里兰卡	tn	突尼斯
cs	捷克和斯洛伐克	ib	黎巴嫩	tr	土耳其
cy	塞浦路斯	li	列支敦士登	tw	台湾
cz	捷克共和国	lt	立陶宛	ua	乌克兰
de	德国	lu	卢森堡	uk	英国(通用)
dk	丹麦	lv	拉托维亚	us	美国
dm	多米尼加	md	摩尔多瓦	uy	乌拉圭
dz	阿尔及利亚	mo	澳门	ve	委内瑞拉
ec	厄瓜多尔	mx	墨西哥	vn	越南
ee	爱沙尼亚	my	马来西亚	yu	南斯拉夫
eg	埃及	na	纳米比亚	ym	牙买加
es	西班牙	ni	尼加拉瓜	zm	赞比亚
fi	芬兰	nl	荷兰		

通用顶级域名表

域名	含义	域名	含义
. com	商业组织	. firm	商业公司
. edu	教育机构	. store	商品销售企业
. gov	政府部门	. nom	个体和个人
. mil	军事部门	. info	提供信息服务的实体
. net	网络服务机构	. web	与 WWW 相关的实体
. org	非营利性组织	. rec	突出消遣娱乐活动单位
. arts	文化与娱乐实体		

第二级域名是说明主机所属组织的性质或地区，如教育(.edu)，北京(bj)等。

我国的域名一般是四级域名，有的三级或五级，如 www.jlu.edu.cn 吉林大学主机。

高级的域名与低级的子域名不允许重复，大小写字母在域名中没有区别，一台计算机可以有多个域名，但是有一个 IP 地址。

1.1.5　统一资源定位地址(URL)

网络信息资源数量庞大，分散在不同的计算机上，为了使网络查询存放不同计算机上的信息，有一个标准的资源地址访问方法，即统一资源定位地址(URL uniform resource locator)。

URL 是一种统一格式的 Internet 信息资源地址表达方法，它将 Internet 提供的各类服务统一编码，以便用户通过 Web 客户程序进行查询。格式为：信息服务类型，信息资源地址，服务器(host)地址，端口(port)，文件路径(path)。

信息服务类型有以下几种：http://表示 WWW 服务器，主要用来提供超文本信息服务的 Web 服务器；Telnet://表示供用户远程登录使用的计算机；ftp://用于提供各种普通文件和二进制代码文件的服务器；gopher://表示 Gopher 服务器；wais://表示 WAIS(Wide Area Information Server)广域信息服务；new://表示网络新闻组 USENET 服务器。

信息资源地址是提供信息服务的计算机在 Internet 上的域名，如 www.las.ac.cn 是中国科学院文献信息中心的域名；www.moe.edu.cn 是中华人民共和国教育部的域名；www.mii.gov.cn 是中国信息产业部的域名。

1.1.6　互联网提供的服务

互联网以其丰富的信息资源，广泛的地域分布，提供各种各样的信息服务。主要包括电子邮件服务，信息查询服务，网络新闻服务等。

1. 电子邮件服务

通过电子邮件 E-mail 方式，与其他网络用户进行电子信件交流，是一种快捷、简便、高效、廉价的现代交流方式。在使用时，首先需要申请电子邮箱；电子邮箱有免费和收费两种，国内外提供免费电子邮箱的网站很多，如 yahoo，sina，hotmail，sohu，163，263，21cn，126 等等。具体操作是用户先进入该网站；进入“注册新账户”，按照要求填写表格，发送申请，得到确认后就有了自己的电子邮箱地址，利用它就可以发送和接收电子邮件了。

2. 信息查询服务

信息查询工具很多，用户可以利用万维网服务(world wide web)的客户端的 Netscape 和 Internet Explorer 等工具软件便捷地查询超文本信息；还可以利用远程登录服务(telnet)，对远程的各类计算机的各类信息进行查询；又可以利用文件传送服务(FTP)，查询对方计算机的各类信息，下载并上传文件，实现信息查询功能，还可以利用 Archie 信息查询，实现各类文件的查询。

3. 网络新闻组服务

用户可以通过订阅新闻组(newsgroup)的方式，参与不同新闻组的活动。这是志趣相同的用户借助网络展开专题讨论的一种非正式，直接地交流方式。在使用时，用户可以通过安装各类新闻组阅读软件工具，如 Outlook Express 的新闻组软件，Agent 等，先连接新闻组服务器，第一次使用时会自动下载讨论组列表，我们可以选中自己感兴趣的新闻组，选择“定阅”，就可以阅读并发表消息了。

4. 网络教育与商业服务

利用虚拟技术，通过网络开展网络远程教学，开设虚拟课程，开办虚拟大学等。

各类商业部门可以通过建立网站，对其服务、产品等的宣传，通过发布广告，网上销售，专业交易市场，网络零售，网络折扣店，专卖店，网络招聘等开展网络商业服务，网络用户可以在网上进行购物、交易等。

5. 网络资源服务

利用远程网络资源提供各种服务，如 ASP(Application Servics Provide)服务，通过因特网向客户提供应用软件及增值服务。

6. 网络休闲娱乐保健服务

通过网络社区，网络旅游，游戏娱乐，家庭服务，网络医疗保健，宽带媒体等对网络用户提供休闲、娱乐、医疗、保健服务。

7. 网络金融服务

通过电子支付，电子银行等开展网络信贷支付，保险支付，商业支付，网上证卷等各种金融服务。

1.2 互联网信息资源

1.2.1 互联网信息概述

互联网信息资源是指互联网上各种信息资源的总和，由于互联网的开放性、易获性，使网上发布信息十分容易，信息发布具有自由性、随意性、匿名性、虚拟性，因此，网络信息资源具有如下特点：

1. 数量庞大，内容丰富

随着计算机存储容量呈指数递增，连接各类计算机的互联网的总容量大得无法估量，数量增长迅速，内容涉及方方面面，十分广泛。

2. 种类繁多，来源分布广泛，分散，无序

既有各个学科专业领域的学术信息，又有各种休闲娱乐消遣等信息；既有来自权威部门的各类统计数据，历史资料，又有每日新闻，动态时事信息；既有来自权威部门的论据严密，论点鲜明的学术信息，又有网上过客的无评无据的随意聊天信息，真可谓各类信息云集，无所不有，无所不在。信息资源呈全球性分布，分布在不同的国家，不同的地域，数据呈无序分散状态。

3. 变化频繁，信息良莠混杂

有的信息内容翔实、可靠，有的信息无凭无据、虚假，需要我们对其加以鉴别。信息瞬息万变，出版周期快，每时每刻都在更新，变化，信息时效性强。

4. 信息表现形式多样，信息检索快捷

有文字、声音、图片、电影、动画等等，是各类超媒体信息的集合；信息检索快捷，交互性强，获取方便，各种搜索引擎功能完备，使网上信息便于利用，信息共享性强。

1.2.2 互联网信息种类

丰富的互联网信息分布在整个网络之中，可以依据不同的分类标准，对其进行划分。按信息发布来源划分，有学校、科研院所信息资源，企业信息资源，政府信息资源，行业机构信息资

源，个人信息资源等等。按媒体类型划分，有文本信息，图像信息，音频信息，视频信息4大类；按信息的访问权限划分，有开放信息与保密信息之分。按人类交流信息的方式划分，可分为非正式出版信息，半正式出版的信息，正式出版的信息，正式出版的信息又可以分为电子图书，电子期刊，电子报纸，网上数据库等。

1. 按照信息的加工程度划分

(1)一次网络信息资源

一次网络信息资源是指网上原始信息，包括电子图书、电子期刊、电子报纸、网络新闻组、电子论坛、各种网络上首次发表的信息等。这类信息资源数量极其庞大，是人们利用的主要资源。

(2)二次网络信息资源

二次网络信息资源是指对一次网络信息资源的搜集、加工、整理，形成一次信息的使用工具，如搜索引擎、虚拟图书馆，它是检索网络信息资源的入口和工具。

(3)三次网络信息资源

三次网络信息资源是指对二次网络信息资源进行加工、整理后所形成的工具指南，如元搜索引擎。

2. 按信息资源所采用的网络传播协议和信息加工层次划分

可分为：Web信息，Telnet信息，FTP信息，用户组信息等等。现详细介绍以下几种类型：

(1)万维网信息资源

万维网英文缩写是WWW，全称为world wide web，起源于1989年3月，由欧洲粒子物理实验室所(the Eurpean Laboratory Particle Physics，ERN)发展出来的主从结构分布式超媒体系统。WWW信息资源是互联网上最主要，最常见的形式，是指建立在超媒体基础上，集文本、图像、声音、动画为一体，以直观的用户界面展现和提供信息的网络资源形式，浏览器与Web服务器之间采用超文本传输协议HTTP进行相互通信，以响应和传输WWW客户机与服务器的用户请求。Web服务器的信息是用专门的一种编程语言HTML来描述的。HTML文档由文本，格式代码和到其他文档的链接所组成。WWW浏览器是一种应用于WWW的网络软件，是一种客户端的程序。不仅可用于与WWW服务器的连接，更主要是帮助用户浏览，阅读和查找WWW信息资源。目前，用户熟悉的有Netscape，Internet Explorer(IE)，Mozilla，Opera和Firefox浏览器等。它们既可以浏览文本信息，又可以显示图形、图像和声音、动画等。WWW服务器还可以使用多种协议工作，访问Gopher，Wais，FTP，News和其他类型的网络资源，可以满足应用需求。

(2)远程登录信息资源

远程登录信息资源(Telnet)是指借助远程登录(Remote login)在网络通信协议(Tele Communication Network Protocol)的支持下，在远程计算机上登录，实时访问，使用远程系统中的资源。这些资源既包括硬件资源，如超级计算机，精密绘图仪，高速打印机，高档多媒体输入/输出设备等，又包括软件资源等。远程登录信息资源有些是开放式的，有些是不开放的。后者需要用户名和口令，当输入远程计算机的域名和IP地址后，需要输入用户名和口令，登录后，需要按照访问权限使用其资源。如美国的商用联机情报检索系统DIALOG，OCLC，DataStar等等。

(3)文件传输服务信息资源

通过文件传输服务协议FTP(file transfer protocal)进行系统间文件传输的信息资源。既

可以从远程计算机上下载文件，又可以上传文件。这些文件包括各类游戏软件，杀毒软件，通信软件，各类图片，电子书刊等等。

(4)用户组信息资源

网上各类用户组，包括新闻组(Usenet Newsgroup)，邮件列表(Mailing List)，专题讨论组(Disscussion group)，兴趣组(interest group)，辩论会(Conference)，等等，是网络用户交流的电子论坛。信息内容直接，具有开放性，涉及各种各样的观点、讨论、动向、新闻、成果，是获取信息直接、便捷、非正式的一种方式，用户可以通过加入该组，对读到的文章随笔、回信、转信等功能参与交流。

1.3 网络信息资源的组织与管理

1.3.1 网络信息资源的分类与主题体系

网络信息资源内容丰富、广泛，需要对其加以分类组织，以便于用户查询。分类是依据事物的属性或特征加以区别或类聚，并将区分的结果按照一定的次序进行组织。网络信息资源分类也是为了更好地揭示内容，便于利用。

网络信息资源分类与图书分类存在很大差异，首先，分类对象不同，网络信息资源往往涉及多学科，交叉学科多，学科分界不清，娱乐性内容多，传统知识性学科少，商业性内容多，新兴学科内容多。而图书分类内容清晰，大多已经形成较为规范的主题内容，学科门类相对单一，类目设置上突出了教育、娱乐、旅行、医疗健康等与日常生活密切相关、用户普遍感兴趣的类目，适应了网络信息用户的广泛性和大众化，方便用户查询和检索这方面的网络信息，弱化了科学技术、学术性类目的设置。另一方面，分类目的不同，网络信息的分类目的是满足用户从不同的角度查找需要，更通用，更直接，而图书分类揭示图书主题内容，具有鲜明的逻辑性，层层展开线性排列，用一系列的线形符号来表达。图书分类大多以学科为中心设立类目体系，网络信息资源分类大多采用以主题为中心或主题与学科结合的两种设类方式。其中，学科与主题结合的方式，可以使类目在具有直接性的同时增加全包容性，使用更普遍。因此，对传统分类体系中详尽展开的类目，如自然科学、应用技术门类等，分类搜索引擎一般只设置了概括性类目。

因此，网络信息资源采用的分类体系与图书分类不尽相同，图书分类大多采用标准的分类法，大多以学科为中心设立类目体系，体系分类法和分面组配法，如《中国图书馆图书分类法》，《国际十进图书分类法》等等，类目设置标准、严谨，具有相对稳定性，一贯性。而网络信息资源的分类体系，更注重用户便捷检索信息的效率，一般具有直接揭示主题特征的特点，大多采用主题与学科相结合的方式，类目设置相对直观、明了，类目更新及时，有些类目会随着网络用户的爱好、使用频率、社会热点问题的变化等等随时调整，通常采用自然语言来表达，类目间有的按字顺排列或使用频率排序，而不是按逻辑关系顺序排列，类目之间逻辑性不强，类目归属不尽合理。

网络信息资源分类主要采用以下方法：

1. 等级式主题分类体系

网络信息资源的等级式主题分类体系基本上采用等级结构，不依学科分类，按事物划分类

目，一个主题为一个类目，类目与类目间按字顺或其他人为的顺序排列，一个类目又分为若干个细目。这种主题分类体系被许多网站与搜索引擎所采用，如新浪（www. sina. com）、21CN搜索引擎（search. 21cn. com）等均采用这种方法。

2. 分面组配分类体系

这种网络信息资源的分类体系是由多个分面组成，每一个分面的类目可以与其他分面的类目组配，表达专指的概念。分面组配分类体系利用分面控制词表，进行网络信息资源的组织、检索与存取。雅虎中国（cn. yahoo. com）、北极星（www. beijixing. com. cn）、搜狐（www. sohu. com）、网易搜索（dir. so. 163. com）、Google网页目录（www. google. com）等均采用这种分类方法。它所提供的是两个分面，一是地域分面，分省级行政区、地市两级类目；另一个是主题分面。查询时可将两个分面的类目进行组配，检索与检索要求相符的网络资源，也可以通过地域分面检索各个主题的网络资源。

3. 学科分类体系

这种网络信息资源的分类体系是将科学、技术的各个学科、领域及其分支设为类目，类目的设置既参照图书馆分类法的学科体系结构，也考虑网络的实际需求。如中文搜索引擎“网络指南针”就提供一个学科分类系统，设有表示学科的一级类目，按类名的拼音字顺排列，下设若干个二、三级类目。

下表揭示了国内外大型网站与搜索引擎的曾经设置的主要大类体系，一定程度地反映了网络信息资源的分类体系。

国内外大型网站与搜索引擎的大类体系表

网站与搜索引擎	大类体系
Yahoo	娱乐、艺术与人文、休闲与生活、计算机与互联网、商业与经济、教育、政府与政治、健康、新闻与媒体、参考资料、地区、科学、社会科学、社会与文化
Hotbot	艺术与娱乐、商业与货币、计算机与互联网、游戏、健康、家庭、新闻与媒体、参考、地区、科学与技术、购物、社会、运动、旅游与娱乐
Magellan	汽车、商业、就业、计算机、教育、娱乐、游戏、健康、家庭、星相、生活方式、新闻聊天、购物、运动、旅游
Open Directory	艺术、商业、计算机、游戏、健康、家庭、儿童、新闻、娱乐、参考、地区、科学、购物、社会、运动、世界
LookSmart	娱乐、工作与收入、购物、计算、人群与聊天、运动、生活方式、旅行、图书馆、个人生活
Ask jeeves	汽车、艺术与娱乐、商业、计算机、游戏、健康、住家与家庭、新闻、娱乐、房地产、参考、地区、科学、社会、运动、旅行、世界
搜狐	娱乐休闲、工商经济、公司企业、文学、体育健身、医疗健康、生活服务、社会文化、社会科学、国家地区、电脑网络、教育培训、艺术、新闻媒体、科学技术、旅游交通、政法军事、个人主页

续表

网站与搜索引擎	大类体系
新浪	娱乐休闲、计算机与互联网、商业经济、教育就业、文学、艺术、体育健身、医疗健康、生活服务、社会文化、科学技术、社会科学、政法军事、新闻媒体、参考资料、个人主页、国家与地区、少儿搜索
网易	娱乐休闲、电脑网络、经济金融、医疗健康、文学作品、艺术分类、生活资讯、体育竞技、教育学习、情感绿洲、政法军事、少儿乐园、社会文化、新闻出版、旅游自然、科学技术、公司企业、个人主页
焦点	新闻、国家地区、艺术、娱乐休闲、旅游、女性天地、文学、教育就业、公司、为您服务、电脑、医疗保健、资料、工商经济、网络、社会文化、科技、个人主页、体育、政法军事
悠游	新闻媒体、电脑、互联网、金融贸易、休闲天地、生活资讯、体育运动、视听娱乐、文学娱乐、旅游交通、医疗保健、人物、图书出版、购物、工商企业、科学技术、政治军事、社会人文、地区
蓝帆	计算机与网络、娱乐休闲、医疗健康、旅游交通、体育健身、文学、艺术、新闻媒体、综合参考、生活服务、教育就业、商业经济、人物明星、社会文化、科学技术、政治军事、社会科学、国家与地区
奇摩	艺术文化、视听娱乐、运动体育、教育学习、科学技术、地区地域、图书出版、网络指南、生活信息、商业金融、计算机通信、医疗保健、大众媒体、社会人文、政治行政、休闲天地
找到啦	生活信息、休闲旅游、社会文化、工商经济、教育学习、自然科技、政府机关、影视娱乐、投资理财、计算机网络、新闻媒体、社会文化、艺术人文、医疗保健
TOM. com	教育就业、娱乐休闲、生活服务、旅游与交通、国家与地区、计算机与互联网、工商经济、体育健身、医疗健康、公司企业

上述分类体系的特点是直观、通用，突出了与日常生活密切相关的类目，如有些网站设置了娱乐、保健、聊天、彩铃、天气、汽车、房地产等类目，学术性、科技性类目数量少，其相关类目深度浅，这些大类体系下的二级类目也具有同样特点。国外大型网站与搜索引擎也有的直接采用国际上通用的大型图书分类法，如杜威十进分类法(DDC)，国际十进制分类(VDC)及各国的分类法。如 OCLC 的 NetFirst 网站采用《杜威十进分类法》和《美国国会图书馆主题词表》。

关于网络信息主题体系，传统的主题法包括标题词法、单元词法，叙词法和关键词法，而网络信息大多采用关键词法，这些关键词一般是机器自动抽取。各种搜索引擎均提供关键词检索功能，有的还提供高级查询功能，如 google 等。而大型网络专业数据库大多采用叙词法，既选用标准主题词法，通过规范的主题词表选词进行查询。如 EBSCO 出版公司的学术期刊全文数据库、DIALOG 的各种专业数据库等。

1.3.2 元数据(Metadata)

元数据是用来组织与检索信息资源的数据(data that deseribes data)，是描述各类电子数据的内容、形式、特征和属性，存储相应的检索路径，从而达到对网络资源的组织、分类、索引等

目的，它为用户提供了一种标准化的语言或交换方式，为各种形态的信息资源提供规范的描述基准和方法，便于用户检索与查询，是一种提高查全率与查准率的有效工具。

按照组织信息资源的功能划分，可以将元数据分为以下类型：描述型元数据，结构型元数据，存取控制型元数据和评价型元数据。描述型元数据用来描述和识别网络信息资源的数据，如 MARC、Dublin(DC)。结构型元数据是用来描述网络信息资源的内部结构的数据，如位置信息，章节等。存取控制型元数据是用来描述网络信息资源被利用的条件，期限及知识产权状况的数据。评价型元数据是用来描述和管理数据在信息评价体系中位置的数据。

元数据的格式目前有 20 余种，国际上有影响的有以下几种：USMARC(US Machine Readable Catalog)美国机读编目格式；DC 元数据格式(Dublin Core Element Set)都柏林核心元素集；TEI 元数据格式(Text Encoding Initiative)电子文本编码标准；FGDC 元数据格式(Federal Geographic Data Committee)地理空间元数据标准；CDWA 元数据格式(Categories for the Description of Works of Are)艺术类作品描述格式；EDA 元数据格式(Enscoded Archival)档案编码格式。

任何元数据格式的内部结构都是多层次的，由内容结构，句法结构和语义结构几个部分共同构成一个完整的元数据格式。

内容结构(Content Structure)：对元数据格式所采用的内容描述性元素，结构性元素，技术性元素等构成元素进行准确定义和描述。

句法结构(Syntax Structure)：是描述元数据结构的规则和文法，如元素选取使用规则，元素描述方法等。

语义结构(Semanitic Structure)：规定元数据元素的具体描述方法。

1.3.3 都柏林核心元素集(Dublin Core Eleaceats Set)

都柏林核心元素集是旨在推动电子资源发现的最小的元数据元素集。它原是为作者生成对万维网资源的描述而设计的，都引起了国际间的广泛注意和承认，是目前世界上使用最广泛的元数据格式，具有极强的适应性和灵活性。

目前，都柏林核心元素集共包括 3 类 15 个元素，其不含子元素、命名域或其他限定词：

1. 资源内容描述类元素

题名(titce)，主题(snbpect)，描述(deseription)，来源(Source)，语言(Loaguage)，长联(rdlation)，覆盖范围(Courage)。

2. 知识产权描述类元素

创作者(Creater)，出版者(Publisher)，其他参与者(Contributor)，权限管理(rights)。

3. 外部性描述类元素

日期(data)，类型(tgpe)，格式(format)，标识(identifier)。

都柏林核心元素集的 15 个元素可选择、可重复、可扩展。关于它的应用及发展相应的配套方案在不断研究。如以自动资源发展工具收集的形式提供元数据，简化元数据记录的编制工具等等。

都柏林核心元素集的作用就是通过对网络信息资源属性信息的描述，提高网络资源检索的查全率和查准率。