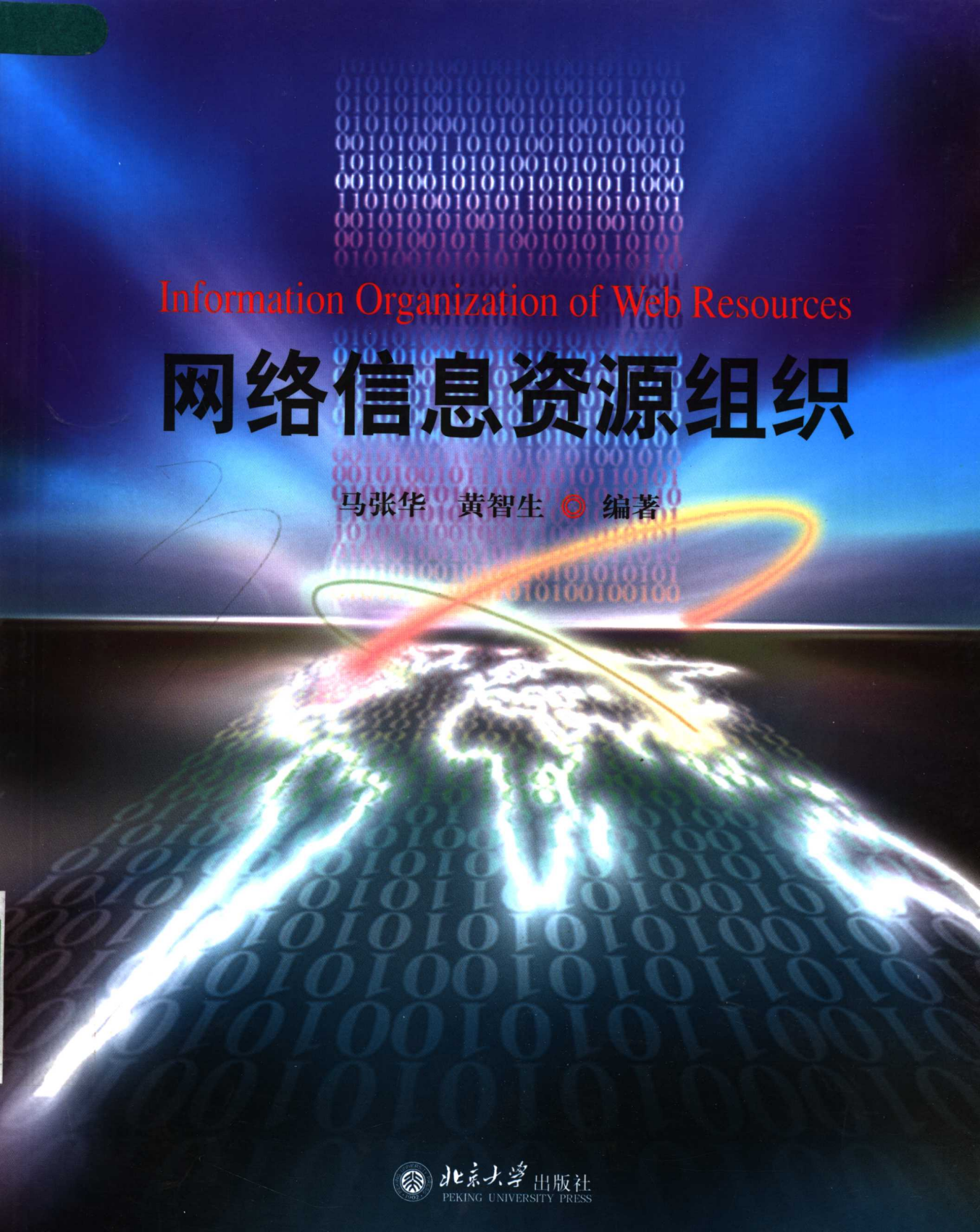
Information Organization of Web Resources
网络信息资源组织
马张华 黄智生 ◎ 编著
北京大学出版社
PEKING UNIVERSITY PRESS

G354.4
37
2007

网络信息资源组织

Information Organization of Web Resources

马张华　黄智生　编著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

网络信息资源组织/马张华，黄智生编著. —北京：北京大学出版社，2007.2
ISBN 978-7-301-11580-0

Ⅰ.网… Ⅱ.①马… ②黄… 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 012450 号

书　　名：网络信息资源组织
著作责任者：马张华　黄智生　编著
责 任 编 辑：王树通
标 准 书 号：ISBN 978-7-301-11580-0/TP·0897
出 版 发 行：北京大学出版社
地　　址：北京市海淀区成府路 205 号　100871
网　　址：http://www.pup.cn
电　　话：邮购部 62752015　发行部 62750672　编辑部 62752021　出版部 62754962
电 子 邮 箱：zpup@pup.pku.edu.cn
印　刷　者：北京大学印刷厂
经　销　者：新华书店
787 毫米×980 毫米　16 开本　14.75 印张　310 千字
2007 年 2 月第 1 版　2007 年 2 月第 1 次印刷
定　　价：25.00 元

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究
举报电话：010-62752024　电子邮箱：fd@pup.pku.edu.cn

内 容 提 要

本书是对网络信息资源组织的理论、方法比较全面的介绍和讨论。全书共分十章，在概要分析网络信息资源组织的特点、类型、方法等内容的基础上，按组织体系为主线，系统分析了网上分类法、关键词法这两类主流检索系统的特点；同时根据检索系统的构成特点和涉及的领域，分别对元搜索引擎、专业搜索引擎、多媒体搜索引擎、引文检索系统等典型的组织类型进行了剖析，最后，论述和介绍网络环境下本体(Ontology)理论和相应的研究和编制实践，并对网络信息资源组织的发展趋势进行了概要论述。本书内容丰富，层次清楚，重视结合网络信息检索系统的特点对组织机制进行讨论，并适当结合实例分析，有助于对网络信息资源组织知识和方法的系统学习和了解。

本书可作为高等学校图书馆学、信息管理专业、情报学专业、计算机科学专业学生的教材或教学参考书，也可供各种与网络信息资源组织有关的机构和部门，包括各类文献机构、网络公司、企业信息资源管理部门的工作人员，以及网络技术的爱好者学习参考之用。

作 者 简 介

马张华 男，北京大学信息管理系教授，长期从事信息组织、情报检索语言领域的教学和研究，出版有《信息组织》、《文献分类法主题法导论》、《主题法导论》(后两书与他人合作)等数种。

黄智生(Zhisheng Huang) 男，获荷兰阿姆斯特丹大学计算机科学博士学位。长期从事人工智能逻辑与推理，智能多媒体技术，语义网与本体技术等研究，现为荷兰 Vrije 大学(Vrije University Amsterdam)计算机系高级研究员，主持欧盟第六研究框架中的语义网 SEKT 课题“语义网中非一致性本体的推理”的研究。他参与的 ECulturE 系统获 2006 年国际语义网技术挑战赛冠军。

前　言

网络信息资源组织是近年来信息组织研究中发展最为迅速的一个领域。从最初的简单检索工具到互联网上的大型搜索引擎，人们对网络信息资源组织和检索进行了广泛探索和改进，取得了巨大的进展，包括构建了适合网络资源环境的组织体系，发展了类型多样的检索工具，探索了一系列新的适合网络应用的技术方法，使得网络信息资源组织迅速发展成为了信息组织的前沿领域；与此同时，网络资源检索中存在的问题也依然十分突出，涉及网络信息资源组织、检索和应用的各种理论、技术方法的众多问题受到各国学者的广泛关注和重视。显然客观分析和认识网络信息资源组织的实践和问题，不仅对网络信息资源的开发利用十分重要，而且对于了解信息资源组织在电子环境下的发展动向也极有价值。可以这样说，在今天，要探索和发展电子环境下信息资源组织的理论、技术和方法，不了解网络信息资源组织，是无法想象的。鉴于此，从 2000 年开始，北京大学信息管理系开设了网络信息资源组织课程，目的是试图通过对网络信息资源组织实践的研究，分析和认识网络信息资源组织的进展和问题，从而更好地了解电子环境下信息资源组织的特点和方法，推进对信息资源组织的探索和发展。本书即是根据这一课程的需要编写的。

本书力图对网络信息资源组织的基本理论方法以及常见检索类型提供一个较为全面的分析和讨论。全书在概要分析网络信息资源组织的特点、类型、工作机制、方法等基本内容的基础上，首先按组织体系为主线，系统分析了网上分类法、关键词法这两种主流检索工具的体系结构和特点；同时，按照检索系统的构成特点和涉及的领域，分别对元搜索引擎、专业搜索引擎、多媒体搜索引擎、引文检索系统等具有一定典型性的检索系统进行了剖析，使得典型组织方式得到概要的介绍；随后，系统论述和介绍网络环境下本体(Ontology)理论与相应的研究和编制实践，并在最后对网络信息资源组织可能的发展趋势进行了概要论述。为了有利于从不同层面、不同角度对网络信息资源组织特点的了解和把握，本书一方面重视从整体角度对网络信息资源组织的基本特点、组织机制等的分析和讨论；另一方面也十分重视按组织体系的特点展开论述，如分类法部分按先组式的特点对网络环境下类目体系的发展进行分析，关键词检索系统则按后组式检索工具的特点，分别按其组成成分、要素展开论述；同时，重视典型实例的剖析，如对网络分类系统、元搜索引擎、专类搜索引擎、多媒体搜索引擎、引文索引等都结合选择的典型网络工具加以讨论，以便可以通过这一方式增加对相关特点的分析和讨论；对于 Ontology 的论述，本书在第八、九章进行，前一章着重本体以及语义网的理论论述，后一章侧重介绍网络本体语言与工具，包括典型本体语言和编制规范，以及本体编制应用与维护等相关内容。对各种组织形式中涉及的相关计算机实现方面的技术问

试读结束：需要全本请在线购买：www.ertongbook.com

题，本书则通过注释的方式，利用相关引用文献加以说明，以便于了解。

本书由北京大学信息管理系马张华和荷兰阿姆斯特丹 Vrije 大学计算机系黄智生合作完成，其中，黄智生撰写第二章、第八章、第九章；马张华撰写第一章、第三至七章；第十章由两人共同撰写；附录中的网络检索工具大事记由赵经纶、窦曦骞同学整理。网络信息资源组织涉及面广、专业性强，且发展极其迅速，即使要跟踪某些分支，也需要付出极大的劳动，要全面掌握各种相关内容几乎是不可能的。此外，由于商业竞争的原因，许多研究或技术是保密的，只能根据有限的公开资料去进行分析，所有这些，无一不增加了研究的困难。我们希望通过这种不同知识领域之间的合作，能够在一定程度上减少专业知识上的局限，拓展对网络信息资源组织的认识。限于水平，此书肯定存在许多不足之处，希望能够得到同行的批评指正。

本教材的编写和出版得到了北京大学出版社教材出版基金的资助。陈文广、王继民等老师在本书的写作过程中提供了不少帮助；北京大学出版社王树通编辑在本书出版过程中给予了热情支持和帮助，谨在此一并表示感谢。

作　者
于北京大学
2006 年 10 月

目　录

第一章　网络信息资源组织概述 …… (1)
第一节　网络资源组织应考虑的基本要素及其影响 …… (1)
一、网络资源组织应考虑的基本要素 …… (1)
二、网络资源组织的要求 …… (3)
第二节　网络检索工具的类型 …… (5)
一、按检索机制划分 …… (5)
二、按检索系统构成特征划分 …… (6)
三、按检索内容划分 …… (6)
四、按工作机制划分 …… (7)
五、按检索资源类型划分 …… (8)
第三节　搜索引擎的工作机制 …… (9)
一、信息采集模块 …… (9)
二、信息存储模块 …… (12)
三、信息检索模块 …… (13)
第四节　网络资源组织特点 …… (15)
思考题 …… (17)
参考文献 …… (18)
第二章　网络资源的检索与表达 …… (20)
第一节　信息检索中的数据表达 …… (20)
一、数据检索的抽象模型 …… (20)
二、网络信息检索的特征 …… (22)
三、数据互操作性 …… (23)
第二节　网络信息检索协议 Z39.50 …… (23)
一、Z39.50 信息检索模型 …… (24)
二、Z39.50 数据操作过程的实现 …… (24)
第三节　万维网与超文本标识语言 HTML …… (26)
一、超文本标识语言的主要组成 …… (26)
二、超文本对信息检索技术的影响 …… (27)
第四节　可扩展标识语言 XML …… (28)

一、一个简单的例子 …… (28)
二、可扩展标识语言 XML 文件的结构与功能 …… (29)
三、名字冲突与名字空间 …… (30)
第五节 元数据,资源描述框架与 Dublin Core …… (31)
一、元数据 …… (31)
二、网络资源与网络资源描述框架 …… (32)
三、Dublin Core 网络元数据语言的发展 …… (32)
四、元数据规范的作用以及网络使用现状 …… (33)
思考题 …… (35)
参考文献 …… (36)
第三章 网络分类检索系统 …… (37)
第一节 网络分类的概述 …… (37)
一、网络分类的意义 …… (37)
二、网络分类检索系统的类型 …… (38)
三、网络分类检索工具的组织特点和要素 …… (39)
第二节 传统文献分类法在网络中的使用 …… (40)
一、传统分类法在网络资源组织中的使用概况 …… (40)
二、文献分类法在网络资源组织中使用的特点 …… (42)
第三节 网络分类搜索引擎的结构特点 …… (48)
一、网络分类搜索引擎编制概况 …… (48)
二、网络分类搜索引擎的编制特点 …… (49)
第四节 网络环境下分类法的发展、问题和前景 …… (54)
一、网络环境下分类法的发展 …… (54)
二、分类搜索引擎相关问题讨论 …… (56)
三、网络分类法的发展前景 …… (59)
思考题 …… (60)
参考文献 …… (61)
第四章 典型网络分类检索系统剖析 …… (62)
第一节 BUBL LINK 剖析 …… (62)
一、发展概况 …… (62)
二、大类体系 …… (63)
三、类目体系展开特点 …… (64)
四、结合采用多种检索形式 …… (65)
五、概要评价 …… (67)

第二节　Yahoo! 主题指南剖析 …… (67)
一、发展概况 …… (67)
二、大类体系 …… (68)
三、类目体系展开特点 …… (69)
四、类目收录和处理规范 …… (72)
五、概要评价 …… (72)
第三节　Open Directory 剖析 …… (73)
一、发展概况 …… (73)
二、大类体系 …… (74)
三、类目体系展开特点 …… (75)
四、横向关系揭示 …… (75)
五、编制和管理方式 …… (77)
六、概要评价 …… (78)
第四节　搜狐分类目录剖析 …… (79)
一、发展概况 …… (79)
二、大类体系 …… (80)
三、类目体系展开特点 …… (81)
四、横向关系揭示 …… (82)
五、概要评价 …… (83)
第五节　新浪分类体系剖析 …… (83)
一、发展概况 …… (83)
二、大类体系 …… (84)
三、类目体系展开特点 …… (85)
四、横向关系揭示 …… (87)
五、概要评价 …… (87)
思考题 …… (88)
参考文献 …… (88)
第五章　网络主题检索系统 …… (89)
第一节　网络主题检索系统概述 …… (89)
一、网络主题检索系统的特点 …… (89)
二、网络主题检索系统的类型 …… (90)
三、关键词搜索引擎的组织特点和要素 …… (92)
第二节　索引模块的结构组成 …… (93)
一、关键词搜索引擎索引的构成 …… (93)

二、网络资源的数据特点 …… (94)
第三节 关键词搜索引擎的查询、检索提供和优化 …… (96)
一、查询界面 …… (96)
二、检索排序和算法 …… (98)
三、检索优化 …… (100)
第四节 词汇控制 …… (102)
一、索引单元的选择与检索句法 …… (102)
二、词间关系控制 …… (106)
第五节 链接控制 …… (113)
一、链接分析的意义 …… (113)
二、PageRank 算法 …… (114)
三、HITS 算法 …… (115)
四、网络社区的识别与应用 …… (116)
第六节 关键词搜索引擎的特点和发展前景 …… (117)
一、关键词检索系统的特点 …… (117)
二、发展前景 …… (118)
思考题 …… (119)
参考文献 …… (120)
第六章 元搜索引擎 …… (122)
第一节 元搜索引擎概述 …… (122)
一、发展概况 …… (122)
二、元搜索引擎工作的机制 …… (123)
三、元搜索引擎的特点 …… (125)
四、元搜索引擎的关键要素和著名元搜索引擎 …… (126)
第二节 Dogpile 评介 …… (132)
一、发展历史 …… (132)
二、资源和检索特点 …… (133)
三、检索提供及其改进 …… (135)
四、黄页搜索与白页搜索 …… (137)
五、概要评价 …… (140)
第三节 Vivisimo 评介 …… (141)
一、发展概况 …… (141)
二、检索特点 …… (142)
三、Vivisimo 的自动聚类技术 …… (143)

四、概要评价 …… (145)
思考题 …… (146)
参考文献 …… (146)
第七章　专业、专门搜索引擎 …… (148)
第一节　专业搜索引擎 …… (148)
一、专业搜索引擎的类型和特点 …… (148)
二、Scirus 剖析 …… (150)
第二节　多媒体搜索引擎 …… (156)
一、多媒体搜索引擎概述 …… (156)
二、Yahoo! 多媒体检索剖析 …… (158)
第三节　引文索引及其在网上的应用 …… (165)
一、引文索引概述 …… (165)
二、CiteSeer. IST 剖析 …… (166)
思考题 …… (174)
参考文献 …… (174)
第八章　本体技术与语义网 …… (176)
第一节　概念与本体 …… (177)
第二节　描述逻辑 …… (178)
第三节　语义网的基本概念与基本思想 …… (180)
一、语义与网络 …… (180)
二、语义网的层次性 …… (181)
思考题 …… (183)
参考文献 …… (184)
第九章　网络本体语言与工具 …… (185)
第一节　网络资源框架与网络资源框架模式 …… (185)
第二节　网络本体语言 …… (189)
第三节　语义网规则语言 …… (193)
第四节　本体的应用和管理维护 …… (196)
一、本体的产生 …… (196)
二、本体的演化与管理 …… (197)
三、本体映射 …… (198)
四、本体的使用 …… (198)
思考题 …… (199)
参考文献 …… (200)

试读结束：需要全本请在线购买：　www.ertongbook.com

第十章 网络资源组织发展趋势概览……（202）
一、网络资源组织的现状和进展……（202）
二、网络资源组织的发展趋势 ……（204）
思考题……（209）
参考文献……（209）

附录Ⅰ 常用术语缩略语……（211）
附录Ⅱ 网络检索工具发展大事记……（212）
主要参考文献……（220）

第一章　网络信息资源组织概述

网络信息资源组织，亦称网络资源组织，是指根据使用的需要，对网络信息资源进行选择、处理、序化，并以适当的方式加以提供的活动。网络资源组织是信息资源组织理论方法在网络资源中的应用，是网络信息资源检索、利用的基础和必要条件。

互联网的出现和使用是20世纪人类发展中一个影响深远的事件。尤其是20世纪90年代初万维网(World Wide Web)的出现，其基于超文本语言的传输方式，以及支持多媒体等格式的特点，使其迅速成为网络信息应用的主流，极大推进了网络的普及和应用。互联网的发展使信息传播形式发生了巨大的变化，同时也是对信息组织理论和技术的一个挑战。互联网不仅是信息交流的空间，同时也是网络信息资源存储的空间。这一空间以众多的服务器为存储平台，依据通讯协议和规范，通过通讯线路，以超文本链接等方式将它与客户端连接，成为一个虚拟而又实际的存在。用户可以利用这一空间自由发布、提供、存储和交流信息。这一状况，给网络资源的组织、管理和开发利用提出了新的课题。如何有效进行网络资源的组织，直接影响到网络资源开发利用的效果。为此，各种网络检索工具应运而生，成为广泛关注的焦点。围绕这一工作，计算机界、文献界、图书馆界进行了许多努力，包括制订各种标准和规范，探索适用的组织方式和技术方法，用以改进网络检索工具的编制，优化其性能，取得了巨大的进展。目前，这一过程仍然在不断发展之中。

网络资源组织是网络资源开发利用的基础，同时也是信息资源组织发展的前沿。对于网络资源组织的相关内容，国内外进行了许多讨论，但是，以网络资源组织作为对象，对其特点、类型及其基本理论方法进行系统研究的工作，仍然比较薄弱。本书力图通过我们的努力，在这一方面能有所推进。

第一节　网络资源组织应考虑的基本要素及其影响

一、网络资源组织应考虑的基本要素

对于网络资源组织相关因素及其特点，国内外学者均进行过许多讨论[1,2]。信息资源组织通常是根据资源对象特点、用户需求，在一定应用环境下建立起来的。网络资源组织同样涉及这些相关内容。与传统信息资源组织相比，网络资源组织具有如下特点。

1. 分布式

与传统文献检索系统具有集中的文献资源集合不同，网络资源是以分布的方式存在于

许多计算机和网络平台上的，相互之间是通过网络，以超文本方式等链接的。它们是分散的，没有预先明确的布局，同时受到网络联结和设备条件的影响，需要采用相应的采集和处理的方法。

2. 数量大

网上集中了有史以来最大的可供采集处理的资源数量，并且一直在急剧增长。根据Steve Lawrence 和 Lee Giles 1999 年在《自然》杂志发表的文章估计，其时万维网上可索引的网页多达 8 亿，但当时没有一个搜索引擎的覆盖量超过其总量的 16%[3]。时隔 6 年，Antonio Gulli 和 Alessio Signorini 研究发现[4]，至 2005 年 1 月底，全世界可检网页已高达约 115 亿。大型搜索引擎中，Google 收入的网页量超过 80 亿，MSN 测试版约为 50 亿，Yahoo! 约 40 亿，Ask/Teoma 约 20 亿。据中国互联网络信息中心调查，至 2001 年 4 月底，中文网页总数为近 1.6 亿条，其中动态网页和静态网页之比为 10∶6[5]；至 2004 年底，全国网页总数为近 8.7 亿条，其中静态网页数 4.7 亿条，动态网页数约 4 亿条，去掉重复后约 6.5 亿条[6]。从这些数字可以看出，网络搜索引擎需要处理的资源数量远远超出了传统数据库的规模，对数据库容量和处理能力形成了巨大的挑战，同时也影响到存储结构改进等诸多问题。

3. 种类多样

网络资源几乎包括了各种现存资源类型的数字化文本。

(1) 内容范围上，几乎涉及人们感兴趣的所有领域。包括：学术、政府、商业、教育、文化、生活、休闲等各种信息。其中，商业、休闲领域占的比重最大。

(2) 形式多样，涉及多种媒体形式。除传统出版形式外，还包括多种资源类型，如：① 全文型；② 事实型；③ 数值型；④ 书目文献型；⑤ 实时活动型，如投资行情和股票分析、网上购物、娱乐、聊天、讨论组等；⑥ 多媒体信息，如图片信息、音频、视频、音乐、影视、声像等。此外，除了必须处理多种媒体类型，还涉及不同种格式的处理。

(3) 性质上，除正式出版文献，同时也存在大量由单位发布的半正式出版文献和各种非正式出版文献。

(4) 语言文字上，网络资源涉及多种不同的文字，如英文、中文、日文、韩文……以及特定文种中的不同字集，如中文的简体、繁体等。

(5) 存在形式上，既有静态资源，也存在着动态资源。

这就需要根据不同的特点和对象，采用适当的方法组织检索系统，包括根据情况建立综合性、专业性、专门性、多语言检索系统等。

4. 动态性

由于网络资源自由发布的特点，其资源存在着巨大的不稳定性，各种资源和数据可以方便地加入、调整和中止，此外还存在链接调整和因文件范畴变化等引起的重新定位问题。这些都增加了网络资源的不稳定性。据 Alexandros Ntoulas 等[7]对 154 个网站的研究，发现每周约 8%的网页被新网页代替，一年后只有 20%的网页仍然存在。中国互联网络信息中

心 2004 年底对中文网络的调查发现[8]，中文网页一周更新量达 10.36%，而一年的更新量接近 90%。另据 Alexandros Ntoulas 等的测定，链接的变化幅度更大，每周大约会产生 25%的新链接，因此如果网络工具更新周期超过一周，就不能充分反映网页的实际链接情况。这与传统文献一旦完成，相对稳定的特点形成鲜明的对照。这就需要以适合的方式对资源进行定期的查核和更新。

5. 非结构化和冗余数据

首先，网络资源的存储，目前主要是按照分布式的超文本结构，通过对服务器的联结进行的，其发布和存储方式是不可能预先规划和缺乏控制的，不仅仅存在资源质量参差不齐问题，而且数据重复严重（如镜像和复制），语义（Semantic）冗余严重；其次，到目前为止，网络文献作为网络资源组织处理对象，仍然不具备充分结构化的特点。以网络文本文献为例，虽然其使用的 HTML 等标识语言，具有描述文献结构的功能，但由于网页资源类型多样、内容表述方法、体例格式各异，缺乏统一的揭示模式；同时许多网页制作者往往并不详细揭示网页成分，标识不够充分。因此目前多数网页充其量只是半结构化的，只能揭示网络最基本的一些信息，如包括部分标题、作者信息，某些网页甚至连主要成分也未作明确区分，因此无法通过软件加以识别。

6. 缺乏质量控制

由于网络资源来源多样，包括大量非正式出版文献，质量缺乏保证。如，有的资源内容陈旧过时，有的文字粗糙、质量低下，许多资源未进行应有的编辑处理，因而可能存在虚假数据，此外还存在许多输入、语法错误等。这就使得如何根据网络资源的特点，确定适合的质量控制标准，并解决好质量控制标准的可操作性和自动处理问题，成为网络资源组织的重要内容。

7. 检索需求多样

网络使用的对象为终端用户，与传统检索系统的用户相比，服务范围更加广泛，检索需求也更加多样。从用户对象的角度看，网络终端用户涉及社会各个领域、各种文化水平、各种年龄的对象，文化层次不齐，通用性用户多；就检索需求而言，包括娱乐休闲、实用需求、学术目的等各种方面，要求网络系统的资源和组织方式具有更强的多样性和适用性；从用户行为特点上讲，由于网络资源有较大的可替代性和选择余地，因此网络资源组织要求能够提供迅速、直接、通用的组织形式。

二、网络资源组织的要求

上述特点对网络资源组织提出了新的要求，要有效进行网络资源组织，必须根据相关特点，探索相应的解决方法，包括：

（1）针对网络资源分布式特点，要求建立适用的搜索和采集机制，以便能够解决分布式资源收集的问题，有效发现和处理有价值的资源对象；

(2) 针对网络资源数量大的特点,需要发展适合海量资源的存储形式、高效处理能力和适合的检索方法,包括发展分布式存储、压缩存储等存储技术,探索通过联合编目方法和自动处理技术等,解决对海量数据的处理,同时针对在海量资源组织和检索时检准率更加重要的特点,改进检准措施;

(3) 针对资源种类多样性,发展适合的组织形式以及识别和处理技术,包括发展种类多样的系统,探索各种资源类型的识别和处理技术,如 PDF 处理技术、多媒体分析、处理技术等,不断发展对资源的处理能力;

(4) 针对非结构化和冗余数据,要求根据其结构揭示特点加以应用,并发展相似文献识别技术,对重复网页进行识别;

(5) 针对网络资源动态性的特点,要求发展相应访问方法,以使网络变动能够及时得到反映,使得新的内容能迅速纳入;

(6) 针对质量缺乏控制的资源状况,建立适合的质量评价机制,并根据网络的特点,结合多种因素予以解决;

(7) 针对网络终端用户的特点和使用环境,充分发展通用检索技术,通过增加系统的易用性、直接性改进使用效果,同时发展面向对象的、适合不同用户的检索技术和个性化检索技术,并通过查询优化形式的应用,改进检索效果。

网络资源组织因素及部分解决方案可以简单列举如表 1-1。可以看出,网络资源组织因素的特点决定了,这些问题的解决需要涉及多个领域的不同层次的内容,包括组织形式、处理机制、技术方法等方面。在某种程度上,正是网络资源组织面对的一系列新因素及其问题,推动了相应技术方法的探索和改进,使得网络资源组织成为一个全新的、发展迅速的领域。

表 1-1 网络资源组织因素及部分解决方案

网络资源组织因素	解 决 方 案
分布式	建立数据采集机制;联合编目等的使用
数量大	分布式存储、压缩存储、缓存等存储技术的发展;联合编目平台和自动处理技术的应用、数据挖掘方法的发展;结合链接因素、用户因素、地址因素等提高检准率
资源多样	发展多种类型的检索工具,包括专业、专门搜索引擎;发展多种资源识别和处理技术,如 PDF 处理技术、多媒体分析、处理技术等
动态性	定期复核机制,动态增补技术
非结构化和冗余数据	相似文献识别技术;已有结构化数据的识别和应用;引入链接因素、用户因素等;发展新的检准措施
质量缺乏控制	建立质量评价机制,引入链接因素、用户因素、地址因素等评价资源的重要性
用户多样性	建立通用性检索工具;发展面向对象的、适合不同用户的检索技术;个性化检索技术和检索优化形式的发展

第二节　网络检索工具的类型

网络信息资源是根据其存储的特点，通过超文本链接等方式建立联系的，这一联系方式使得网络资源在组织上存在着比较大的自由和多种可能性，可以形成多种层次的组织结构，如网页-基层网站-整体网站形成的组织系统；也可以通过相关性联系的揭示，建立关联系统，如通过友情链接形成的，具有揭示相似资源功能的横向系统等。但其中，以广大网络资源为组织和处理对象的网络检索工具，即搜索引擎，在网络检索中应用最为广泛，是网络资源开发利用中使用的基本组织形式，对了解网络资源组织的特点最具有代表性，最值得引起关注。

网络搜索引擎，是根据用户需要和资源特点建立的。用户需求和资源类型的多样性，造成了检索工具的类型多样。按照不同的划分标准，通常可以将常见的网络检索工具分别划分为如下类型。

一、按检索机制划分

按检索机制，网络信息检索工具总体上可以划分为分类检索系统、主题搜索引擎，此外还存在按照引用关系建立的引文索引等检索工具类型。

分类检索工具亦称为网络分类目录、主题指南(Subject Directory)、分类搜索引擎等，是一种依据资源的主题对象，按照一定的等级和次序建立的浏览工具。这类检索工具一般由人工编制，将经过选择的资源区分为相应的主题类，并按照其关系和等级逐级展开。用户可以根据展开的等级对其进行浏览检索。这类检索工具可以依据传统的文献分类法编制，也可以根据网络资源的情况重新进行构建，后者的规模更大，用户也更为普遍。由于这类系统的资源通常是精选的，文献质量高，可以通过类目关系对资源进行浏览查找，是网上进行系统检索的重要工具。依据传统文献分类法编制的系统如各国图书馆基于《杜威十进分类法》(Dewey Decimal Classification，简称 DDC)、《美国国会图书馆图书分类法》(Library of Congress Classification，简称 LCC)、《国际十进分类法》(Universal Decimal Classification，简称 UDC)等编制的众多网络分类目录[9]；主题指南则有 Yahoo! 主题指南，Open Directory 以及国内的搜狐、新浪分类目录等。

主题检索系统亦称为主题搜索引擎(Search Engine)，是一种以表达资源内容、特征的词语为检索标识，直接进行资源检索的工具。其中，以关键词搜索引擎系统使用最普遍、影响最大。这类系统通过自动搜索软件收集资料，资源数量大，可以直接通过自然语言词汇的匹配进行检索查找，速度快，直接性好，在网上使用最为广泛。这类系统的问题主要是如何提高检准率，提供高质量文献。国外的 Google、Yahoo! 搜索、Ask/Teoma，国内的百度等均属于这一类型。此外，也存在着依据控制词表编制的主题检索系统。这类检索工具一般由文献单位编制，其资源集合有较强的针对性和学术性的特点。常见的有依据《美国国会标

试读结束：需要全本请在线购买：　www.ertongbook.com