普通高等教育“十一五”国家级规划教材

21世纪 信息资源管理系列教材

信息检索教程（第二版）

王立清/主编

中国人民大学出版社

21世纪信息资源管理系列教材

信息检索教程(第二版)
知识管理
信息组织与信息构建
现代咨询学原理与应用
信息管理导论
信息法制建设概论
政务信息加工的原理与方法
政府网站建设
政务信息安全管理
计算机文件技术

策划编辑　潘　宇
责任编辑　黄　丽　乔林碧　李　颜
版式设计　楠竹文化＋赵星华
封面设计　李尘工作室

ISBN 978-7-300-09671-1/D · 1873

ISBN 978-7-300-09671-1
9 787300 096711 >
定价：45.00元

普通高等教育“十一五”国家级规划教材
21世纪信息资源管理系列教材

信息检索教程

（第二版）

王立清　主编

中国人民大学出版社
·北京·

图书在版编目（CIP）数据

信息检索教程（第二版）/王立清主编．—2版．
北京：中国人民大学出版社，2008
（普通高等教育“十一五”国家级规划教材；21世纪信息资源管理系列教材）
ISBN 978-7-300-09671-1

Ⅰ．信…
Ⅱ．王…
Ⅲ．情报检索-高等学校-教材
Ⅳ．G252.7

中国版本图书馆CIP数据核字（2008）第136035号

普通高等教育“十一五”国家级规划教材
21世纪信息资源管理系列教材
信息检索教程（第二版）
王立清　主编

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	邮政编码	100080
电　　话	010-62511242（总编室）	010-62511398（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京民族印务有限责任公司		
规　　格	170 mm×228 mm　16开本	版　　次	2004年9月第1版
			2008年10月第2版
印　　张	25.5	印　　次	2008年10月第1次印刷
字　　数	467 000	定　　价	45.00元

作者简介

王立清，女，1965年生于山西大同，汉族，管理学博士，中国人民大学信息资源管理学院副教授，情报学专业硕士生导师。主要研究方向：信息组织与信息检索、政府信息资源管理、信息资源数字化出版等。主持和参与国家社科基金及其他各类科研项目多项。主要著作有《政府网站的构建与运作》。主要论文有《澳大利亚、新西兰政府网站建设的元数据标准》、《新技术革命与图书馆业务流程重组》等。

内容简介

本书为普通高等教育“十一五”国家级规划教材。

主要内容包括信息检索概念和原理、信息检索系统、信息检索效果、检索语言、信息著录和标引、印刷型工具书检索、计算机检索概述、联机检索、网络信息检索、特种文献检索等。尤其对网络环境下检索语言的发展、网络信息检索的特点和方法、网络信息检索工具、网络数据库等进行了重点解析。修订后的本教材继续为读者提供完整的信息检索理论框架，同时补充丰富了信息检索领域新进展的相关内容，新增了信息检索的实践案例，努力突出基础性强、实用性好、新颖性明显、结构清晰、适用面广等特点。

本书可作为高等院校信息管理与信息系统专业及图书馆学、档案学等相关专业本科生信息检索课程核心教材，亦可作为各类信息管理机构工作者的培训教材和参考书，还可供广大信息检索爱好者阅读和参考。

第2版前言

PREFACE

信息检索是一门以信息及其相关检索系统的特点及使用方法为研究对象的课程，主要内容包括信息检索基本理论和原理、信息检索语言、信息检索策略、搜索引擎、网络数据库、印刷型工具书等，旨在培养和训练学生检索与检考、选择与鉴别、挖掘与处理文献信息资源的能力，使学生了解并掌握各种信息检索方法，学会运用各种信息检索系统从各类信息中获取所需要的知识。本课程作为信息资源管理领域一门重要的核心课程，是一门培养大学生的信息素养、提高自我知识更新能力和掌握信息检索技能的实践课程，是21世纪高校信息素养教育的重要课程之一，对于培养学生的信息素质与信息能力，掌握科学的方法进行文献信息的收集、整理、加工和利用，提高学生的自我学习能力和创新能力具有重要的意义。

《信息检索教程》一书初版于2004年9月，至今已近四年。此间，信息检索领域发生了很多的变化，信息资源数量有了极大的增长，全文检索技术得到了进一步发展和完善，搜索引擎不断推陈出新，网络数据库日益丰富，互联网平台成为用户获取信息的重要途径。另外，本书自出版后一直被用作中国人民大学信息管理与信息系统、档案学专业的信息检索课程教材，也被国内其他一些院校所采用。在教学实践中，我们努力追踪信息检索的发展动态，积累信息检索的相关素材和案例，这些都为本次修订奠定了良好的基础。

本次修订对全书结构进行了调整，增加了第6章“计算机信息检索概述”和第12章“特种文献检索”，去掉了原来的第4章“信息检索策略和方法”，将其内容分散至相关章节；全面更新了

搜索引擎、网络资源目录、元搜索引擎、网络数据库等内容，补充介绍了新出现的网络信息检索系统；增加了一定的篇幅阐述搜索引擎、网络信息检索发展趋势；在每章之后增加了相应的案例。期望修订之后，本书能够更加全面、准确、及时地反映信息检索领域的最新发展成就，更好地满足信息检索教学和相关人员参考学习的需求。

本书修订后为12章。其中，第1、6章由王立清负责修订和编写，第2章由汤坚玉负责修订，第3章由王立清和宋薇负责修订，第4、5章由刘丽负责修订，第7、8、10章由汤坚玉和王立清负责修订，第9章由汪娟负责修订，第11章由杨凡和汤坚玉负责修订，第12章由王立清和汤坚玉负责编写。全书最后由王立清统稿。此外，汤坚玉参与了全书的校对工作，宋薇对本书配套的PPT讲稿制作付出了很多。

在网络环境下，信息检索的发展可谓日新月异，限于作者的学术水平，书中难免有疏漏和不妥之处，敬请业界内同行和广大读者批评指正。

王立清

2008年6月于中国人民大学

第1版前言

PREFACE

我们面临的是一个崭新的信息社会，信息已成为当今社会发展的重要战略资源。信息检索作为专门研究信息存储与信息获取的学科，对于提高大学生的信息素养和信息获取能力具有重要的意义。

信息检索源于传统的文献检索。信息技术的进步，尤其是互联网的发展，带来了信息传播与存储方式的极大变革，对信息检索产生了深刻的影响。检索对象从传统的纸质文献扩展为数量庞大、类型多样的信息，其中，数字信息、在线信息、多媒体信息越来越多地受到关注；检索工具除了传统的字典、词典、索引、书目、百科全书、年鉴等纸本工具书外，Google、百度等搜索引擎的使用则更为普遍；检索语言突破了传统的主题法和分类法，自然语言应用显现出强劲的势头；检索效果的评价也将面临新的挑战。

本书力求探究网络环境下信息检索的新变化，系统介绍信息检索原理和方法以及各类型检索工具的使用，反映信息检索领域的新动态新方法，推进信息检索教育。

书中内容分为4个部分：信息检索基础理论、印刷型工具书检索、光盘检索和联机检索、网络信息检索。主要内容包括信息检索的原理和方法与意义、检索评价、检索语言、信息著录与标引、检索策略、印刷型工具书的概况和使用、光盘检索概况及主要光盘数据库检索、联机检索系统的特点及利用、网络检索的特点、网络检索工具的原理和使用方法、网络数据库等。

与现有的同类教材相比，本书具有如下特色：

1. 全面与重点相结合，突出网络检索。本书试图构筑一个

科学的、全面的现代信息检索内容体系，融传统的中文工具书、英文工具书、社科文献检索和现代网络检索为一体，系统阐述信息存储和信息检索整个过程，展现信息检索的全貌。同时，伴随网络环境所引发的信息资源数字化、信息检索在线化趋势，互联网成为全球最大的信息资源宝库，网络检索已成为大学生获取信息的最重要方式，本书重点介绍了网络信息资源的检索，以体现网络环境下信息检索的新特点，吻合当前信息检索的实际。

2. 理论与实践相辅助，突出实用性和可操作性。信息检索重在培养学生获取信息资源的实际能力，本书在介绍检索工具时注重阐述其检索功能，对于重要的检索工具和数据库辅之以检索实例，而且，每章后附思考题。

3. 反映信息检索领域的新成果、新趋势，突出内容的新颖性。本书参考了近年来大量的国内外相关资料和研究成果，注重介绍信息检索在网络环境中的新变化和新发展，包括网络信息资源的特点、传统分类法和主题法的发展、自然语言的应用、网络分类法、网络信息检索标准、元数据、多媒体检索技术等。

4. 脚注详尽、规范，便于相关知识的扩展学习。本书对引用的主要观点、资料及相关内容进行了脚注，并对涉及的网站或网页资料注明了访问日期，在一定程度上起到了扩展本书内容的作用，增强了本书与相关信息资料的关联性，有利于获取有关资料和对问题的进一步研究。

本书第1章由冯惠玲、王立清、李林编写，第2章由冯惠玲、王立清编写，第3章、第4章由王立清编写；第5章、第6章由刘丽编写，第7章由王立清、唐宇薛编写，第8章、第9章、第10章由王立清编写，第11章由王立清、刘丽编写。全书由冯惠玲、王立清统稿。

由于时间仓促且水平有限，书中难免有疏漏、不妥乃至错误之处，恳请广大读者批评指正。

编著者

2004年8月

CONTENTS

目 录

CHAPTER ONE

第1章 信息检索基础

【本章要点】

◇ 解释信息的含义与特征
◇ 论述信息的分类
◇ 介绍信息检索的概念
◇ 阐述信息检索的原理
◇ 讨论信息检索系统
◇ 梳理信息检索方法
◇ 探讨信息检索效果评价

引子

奈比斯特说："趋势就像一匹马一样，比较容易向着正在奔跑的方向奔去。"无疑，以知识经济为显著特征的信息社会，已经成为社会发展的一种无法抗拒的趋势。我们今天所处的时代被人们称为"信息社会"。什么是信息？信息是消息，人们在学习、工作、日常生活中随时随地都在接受和利用信息；信息是资源，它具有使用价值和价值；信息是财富，且是无价之财富；信息是生产力要素，更是一种不可估量的促进生产力发展的新动力。因此，人类社会的发展，科学技术的进步，都离不开信息资源的开发和利用。而且，信息资源的真实状况及开发利用程度，已经成为衡量一个国家经济、文化、科技以及综合国力的重要指标。我们

怎样快速地查找信息和有序地整理信息？信息检索是最快的途径。①

1.1 信息概述

1.1.1 信息的含义与特征

1.1.1.1 信息的含义

信息（information）作为现今社会使用频率最高的词之一，显现着时尚，蕴涵着古老。人类很早就开始了信息活动，古波斯人塔顶设置的“喊话站”和古罗马人的“悬灯”讲述着古代信息活动的故事。我国远古的“结绳记事”和殷商的“烽火告警”，同样展现着信息存储和传递的方式。唐代诗人李中在《碧云集·暮春怀故人》诗中写道：“池馆寂寥三月尽，落花重叠盖莓苔。惜春眷恋不忍扫，感悟心情无计开。梦断美人沉信息，目穿长路倚楼台。琅轩绣瑕安可得，流水浮云共不还。”这是汉语中较早使用“信息”一词的文字记录。南宋陈亮《梅花》诗曰：“一朵忽先变，百花皆后香，欲传春信息，不怕雪埋藏。”清朝的康熙是我国历史上第一个明确使用“信息”一词的皇帝，康熙三十四年（1695），他命大学士派人外出，“惟以侦探信息奏文为要”。当时，“信息”指音信、消息。②

人类历史推进到今天，信息引起了如此广泛、深入、持久的影响，信息化社会、信息高速公路、信息时代、信息系统、信息产业、信息技术……这一切都与信息紧密联系在一起。那么信息到底是什么呢？作为日常用语，信息就是消息，一切存在都有信息，如上课的铃声、网上发送和接收的 E-mail、报纸杂志中的文章、电视播放的新闻等等。对人类而言，人的五官生来就是为了感受信息的，它们是信息的接收器，它们所感受到的一切都是信息。信息技术的发展，将会极大地帮助人类去感知和发现五官不能直接感受的大量信息。

对信息概念进行科学的探讨开始于 20 世纪 20 年代。1928 年，哈特莱（R. V. L. Hartley）在《信息传输》中将信息理解为“选择通信符号的方式”。1948 年，信息论创始人申农（C. E. Shannon）在《通信的数理理论》中将信息定义为“信息是用以消除随机不确定性的东西”。1950 年，控制论创始人维纳

---

① 参见一凡的博客：http://blog.163.com/yifan_998/blog/，2008-03-20。

② 参见孙凡：《信息的由来与发展》，载《情报科学》，1999（2）。

(N. Viener) 将信息的概念引入了控制论，在《人有人的用处——控制论与社会》中指出："人是通过感觉器官感知外部世界"，"我们支配环境的命令就是给环境一种信息"，"信息这个名称的内容就是我们对外界世界进行调节，并使我们的调节为外界所了解时而与外界交换来的东西"①。

关于信息的定义，目前尚无统一的定论。不同领域的研究者从不同的角度出发，对信息有不同的理解和认识。结合信息检索的特点，我们认为信息的含义有广义和狭义之分。广义的信息指自然界和一切人类活动所传达出来的信号和消息，是事物表现的一种普遍形式。从本质上说，信息是事物自身（显示其存在方式或运动状态）的属性，是客观存在的现象。狭义的信息指经过搜集、记录、处理和存储的可供检索的文献、数据和事实。它是人类对客观事物的认识，是实践经验的总结，是认识的结果，是我们检索的对象。

1.1.1.2 信息的特征

信息的特征是指信息区别于其他事物的属性。信息的主要特征如下：

1. 可存储性

信息可以存储，存储和传递是信息的两种基本状态。大脑就是一个天然信息存储器。利用信息的可存储性，人们可以有意识地将流动的信息以某种方式存储在物质媒介上，使信息与物质媒介构成一种依附性很强的、相对稳定的关系。这种稳态的结构可以有效地避免信息的流失，也使我们的信息检索有源可寻。

2. 可传递性

可传递性是信息的本质特征之一，指信息可以通过一定的传输工具和载体进行传递，从而形成信息联系，被人们感受和接收。信息的传递有空间传递和时间传递等不同类型，需要依赖于一定的物质载体，具有动态性和方向性的特征。正是信息的传递使人类掌握了更多的经验和知识，推动了人类文明的进程。

3. 可转换性

信息的可转换性表现在两个层面：一是信息在一定的条件下可以转化成物质、能量、金钱、效益等其他东西，这种转换主要依靠人类对信息的正确利用；二是信息可以从一种形态转换为另一种形态，自然信息可转换为语言、文字和图像等形态，也可转换为电磁波信号或计算机代码。比如，自然语言信息和机器语言信息的转换、不同语种信息的转换、不同载体信息的转换等等。

---

① 岳剑波：《信息管理基础》，2页，北京，清华大学出版社，2000。

4. 可处理性

信息可以通过分类、整序、分析、综合、压缩、扩充等加工处理，而达到便于识别、效用更高的信息。人脑本身就是最佳的信息处理器，可以在感知信息的基础上，进行决策、研究、发明、创造等多种信息处理活动。计算机也同样具有信息处理功能，计算机可以输入各种数据文字等信息，进行相关的处理，以显示、打印、绘图等方式再生成信息。

5. 可共享性

信息的共享性表现为同一种信息可以同时被许多人共同享用，这是信息不同于物质和能量的一个非常重要的特征。也就是说，数个接收者可以获得同一信源发出的同样信息，而在这一过程中，信息的内容不会减少或发生改变。信息可以广泛地扩散和传播，信息交换的双方不会失去原有的信息，而且还会增加新的信息。

6. 可识别性

信息可采取直观识别、比较识别和间接识别等多种方式来感知和识别。信息作为表现事物特征的一种普遍形式，反映了事物的运动状态和存在方式，人类可以通过自身的器官去直接感觉和知觉信息，通过比较去认识信息，借助于先进的信息技术和手段去识别信息。

7. 依附性

信息无法脱离物质而独立存在，在其存储和传递过程中必须依附于一定的物质载体，信息与物质载体构成一个整体。我们将这些信息赖以存储和传递的物质载体称为信息载体，信息载体泛指一切载有信息的物质媒体。

8. 普遍性

信息是物质的基本属性，普遍存在于自然界和人类社会之中，也存在于人类的思维或精神领域之中。只要有物质存在的地方，就有信息的存在，物质普遍存在的属性导致了信息的普遍性。

1.1.2 信息的功能和类型

1.1.2.1 信息的功能

信息具有重要意义，与物质和能源一起构成现代社会发展的三大支柱。在一般情况下，人们将物质和能源称为有形资源或第一资源，把信息称为无形资源或第二资源。物质和能源分别为生产提供材料和动力，信息则为生产提供智力。在当今信息经济社会里，信息为生产所提供智力的重要程度已远远超过第一资源。信息的功能是指信息的功效和作用，主要表现在以下方面：

1. 传承人类文明，推进社会发展

人类所有的知识、故事都是信息，这些信息记载着人类文明发展的轨迹。自古以来，信息的积累和传播，成为人类承上启下的纽带。信息作为人类了解自然及人类社会的凭据，与物质和能源共同奠定了社会发展的基础。

2. 提供决策依据，提高决策效益

现代管理学认为，决策是对管理的未来行动目标作决定，是在两个或两个以上的可行性方案中选择出一个满意方案的过程。决策过程无论是环境分析、准则制定、方案生成还是选择评价，无一不与信息的采集、分析与利用有关。信息能够帮助人们减少应对决策时的不确定性和风险，降低由于缺乏足够准备而造成的损失。社会活动的日趋复杂化，增大了决策的难度，大部分决策都是在面临着多种可能出现的结果中作出抉择的，这就更需要全面、正确的信息作为决策的支持。

3. 保障有效控制，保证系统秩序

控制是保障各个社会组织有利地和高效地获得和利用其资源的监控及实施调节行为的过程。控制的整个过程离不开信息，信息是实现有效控制的灵魂。尤其在信息化水平不断提高的今天，各种社会成员只有在信息交流通畅及时的情况下，依据所获得的准确信息来行使各自的职责，才会使整个系统处于有序状态。

4. 发挥参考作用，推动知识创新

信息是知识的源泉和生产发展的催化剂，信息中包含有大量人类实践活动的成果和教训，人们对各种客观事实和社会现象的解释、论证和总结，比较集中地反映了人类的研究成果。因而，信息对我们的科学研究和社会实践均有着广泛的参考作用。通过相关的信息，可以对所选项目是否具有创新性做出判断，避免重复选题，并参考他人的研究方法，加快科研的进程。

1.1.2.2 信息的类型

从不同层面、不同角度、不同学科领域出发，根据不同的分类标准，可以对信息的类型进行不同的划分。对信息的分类可以从广义信息和狭义信息两个方面来进行理解。

1. 基于广义信息概念的信息分类

从信息的广义内涵来划分信息的类型，即把信息理解为对客观事物存在方式和运动状态的反映。据此有以下分类：

（1）依据信息的产生，可以分为社会信息和自然信息。

社会信息指人类在社会实践活动中，为生存、生产和社会发展而产生、处理和利用的信息，是人类对外界事物的反映、人的思想和情感、人与人之间的联系

等。在一般情况下，我们说的信息更多的是指社会信息。自然信息是自然界中的事物变化、特征以及事物之间的内在关系的反映，如自然景观等。

(2) 依据信息的运动状态，可以分为自在信息、自为信息和再生信息。

自在信息指没有进入人的认识领域，未被反映和把握的纯自然状态信息。自为信息指人这个认识主体所感知的信息，是已被把握的自在信息。再生信息是主体对自为信息经过加工制作后向外界输出的信息，是主体反映客体而形成的观念性信息和思维信息。

2. 基于狭义信息概念的信息分类

从信息的狭义内涵出发来划分信息的类型，即把信息作为检索对象来认识，认为信息是经过搜集、记录、处理和存储的，可供检索的文献、数据和事实。据此有以下分类：

(1) 按照信息的媒体类型，可以分为印刷型信息、缩微型信息、视听型信息和机读型信息四种。

印刷型信息指以纸张为媒体，以手写、石印、油印、胶印、铅印、影印等为手段来记录的信息。虽然网络的发展大大冲击了印刷型信息，进入“无纸社会”的呼声此起彼伏，然而，就世界范围来看，纸质文献的数量持续上升，印刷型信息的生命力仍旧强大，纸质文献在便携、阅读方便、可长期保存和反复使用等方面都显示出其特有的优势和价值。

缩微型信息指以感光材料为媒体，以缩微照相为记录手段的信息，也称缩微复制品，包括缩微胶卷、缩微平片、缩微卡片等。它的优点是信息存储密度高，文献体积小，便于收藏、保存和传递，能安全储存资料。但缩微型信息必须借助缩微阅读机或其他辅助设备才能阅读，不便携带，保存条件要求严格，难于普及。

视听型信息，也称声像型信息，指以磁性材料或感光材料为存储介质，借助特殊的机械设备，直接记录声音和图像，并通过视听设备存储和播放的信息，如唱片、录音带、录像带等。它具有声情并茂、形象逼真、直观性强、动静交替等优点。但是也需要借助一定的设备才能使用。

机读型信息指通过编码和程序设计，以机器语言存储在磁盘、光盘等介质上，并依赖计算机输出的信息。它具有存储量大、查找快速方便的特点。网络信息作为新型的信息类型，是一种非常重要的机读型信息。

(2) 按照信息的加工处理程度，可以分为零次信息、一次信息、二次信息和三次信息。

零次信息指在人际交流中口头携带和传播的信息。包括交谈、聚会、参观以

及人际通过其他直接接触方式形成的信息。零次信息产生于交流的过程，具有选择性和针对性较强、交流速度快、反馈及时等特点。但由于零次信息的出现和传递都带有很大的偶然性，而且未经记录和加工，不便于积累和检验，因而增加了获取难度。

一次信息指未经过加工或粗加工的原始信息资源，也称原始信息，是人们在社会实践活动中直接产生或得到的各种数据、概念、知识、经验及总结。一次信息数量庞杂而分散，主要包括著作、报纸、期刊、会议资料、研究报告、政府出版物、专利说明书、产品样本、标准文献、学位论文等等。一次信息具有价值高、数量大的特点，是最基本的信息，对科学研究和社会实践具有重要的参考和使用价值。

二次信息是以一次信息为依据进行加工整理而形成的信息，是对一次信息浓缩或有序化的产物。在信息检索中，二次信息的主要表现形式有目录、文摘、索引等，有时也称二次信息为检索工具。目录指对图书、期刊或其他单独出版文献的特征进行揭示和报道，并按照一定的方法加以编排的二次信息。目录一般只记录文献外部特征，如书名（刊名）、作者、出版地、出版者、出版时间等，例如综合目录、专题目录、馆藏目录、联合目录等。我们目前在网上图书馆看到的公共联机目录查询系统（OPAC）就是网络环境下常用的一种目录。文摘以单篇或单本文献为报道单位，不仅记录一次信息的外表特征，还要客观地阐明深入的信息内容，是对原始信息的浓缩，有助于我们对原文的了解。索引是将原始信息中的各种知识单元进行抽取，按照一定的原则和方法进行排列的二次信息。这些知识单元可以是篇名、人名、名词术语、关键词、分子式等等。二次信息具有传递信息、报道信息的功能，更重要的是为查找一次信息提供线索。它具有系统性、工具性等特点。

三次信息是在对零次信息、一次信息、二次信息进行分析研究、加工提炼和概括综合而形成的信息。它具有信息量大、综合性强和系统性好等特点，具体包括综述、述评、进展报告、学科年度总结等。其中，综述和述评是三次信息最基本的两种形式。综述，即综合性叙述，将大量分散的有关特定课题的文献、事实和数据进行归纳、分析、综合、筛选，以简练的文字扼要叙述出来，内容十分概括。“述而不作”是撰写综述的一般要求，综述要客观全面地整理、分析和总结现有信息，而且对此不加评论。述评指针对某一学科或某一问题，全面系统地总结各种情况、观点和数据，并给予精辟的分析评价，“有述有评”是述评最为突出的特点。综述和述评能够帮助人们用较少的精力和较短的时间，对有关课题的内容、意义以及历史、现状等有一个简明的了解。

（3）依据信息内容，可以分为经济信息、科技信息、政务信息、文化信息、教育信息、军事信息等。

经济信息指国民经济各部门、各行各业的生产情况和特点，以及各行各业彼此影响制约关系的信息，包含一切经济活动中产生的信息。科技信息指与科学技术有关的信息。政务信息是指一切产生于政府活动中的信息。文化信息主要来自文化领域，包括文学、艺术、出版等。教育信息从教育活动中形成。军事信息是指与国防军事相关的信息。

（4）按信息的出版发行特点，信息可分为正式出版信息和非正式出版信息。

正式出版信息指公开出版发行的信息，主要包括图书、期刊、报纸等。

第一，图书。

据联合国教科文组织的规定，49 页以上装订成册的印刷品称为图书，5～48 页的称为小册子，4 页及以下的称为零散资料。凡正式出版的图书均有国际标准书号 ISBN（International Standard Book Numbers），由 10 位数字分为四个部分组成，依次是组号（地区、语种）、出版者号、书名号和计算机校验位，每个部分之间用“-”隔开。例如：ISBN 7-300-02685-0。第一部分“7”表示中国，第二部分“300”是中国人民大学出版社的代号，第三部分“02685”表示该出版社出版的《情报学概论》一书，第四部分“0”表示计算机校验位。2007 年 1 月 1 日，国际标准书号的格式由 10 位修订为 13 位。2007 年 1 月 1 日以前，各国 ISBN 机构尚没有分配完的 10 位的 ISBN 可以加前缀“978”，一旦现有的 10 位的 ISBN 号用完了，新申请的 ISBN 号全部以 979 开始。13 位的号码与 10 位的号码可以通过算法相互转换。例如《情报学概论》转换之后的 13 位 ISBN 号为：978-7-300-02685-5。

根据用途，图书可以分为阅读类图书和检索类图书。阅读类图书是供读者浏览和研读的，分为科学专著、教科书、论文集、科普读物和文艺作品等。检索类图书是备读者查检用的，包括书目、索引、字典、辞典、百科全书、手册等。根据篇幅和出版形式，图书还可以分为单卷书、多卷书和丛书。图书的内容相对比较成熟、全面，是一种重要的信息。

第二，期刊。

期刊又名杂志，是一种有固定的名称，统一的版面形式，按期出版，标有刊期等序号的连续出版物。正式出版的期刊均有国际连续出版物标准刊号 ISSN（International Standard Serial Numbers），由 8 位数字分两个部分组成，如《图书馆学、信息科学、资料工作》的刊号是 ISSN1005－4189。期刊具有数量大、出版及时、内容新颖等显著特点，期刊登载的文章可以反映学科发展的最新动态

和研究热点，对于科学研究具有重要的参考价值。

第三，报纸。

报纸也属于连续出版物，具有出版周期短、时效性强的特点。报纸类的信息非常丰富，涉及经济、文化、社会、生活各个方面，能够动态地反映出最新的信息。

非正式出版信息，也有人把它称为特种文献或灰色文献，指不经过公开出版物流通渠道、不大量发行、为一部分用户使用的内部文献信息资料。它具有信息量大、形式多样、载体不固定等特点。非正式出版的信息包括：会议文献、学位论文、政府出版物、研究报告、档案、专利文献、标准文献等。

第一，会议文献。

会议文献指在国内外各种学术会议上产生的文献。包括会前文献、会中文献和会后文献。会前文献有会议论文预印本、会议议程和通报等；会中文献主要有开幕词、闭幕词、讨论记录、大会提案和决议等；会后文献有会议录、会议论文集、汇编、报告、会议专刊等。

第二，学位论文。

学位论文是高等学校和研究机构的学生为获得某种学位而撰写的科学论文，一般分为学士论文、硕士论文、博士论文。学位论文是一种重要的原始研究成果，对科研工作有较大的参考作用和较高的利用价值。

第三，政府出版物。

政府出版物是一个非常庞大的信息集合，指国际组织和各国政府部门及其所属机构出版的文件。政府出版物反映政府机构的活动，反映官方的意志和观点，且大部分产生于政府及组织机构的工作过程中，包含大量原始的资料或数据。由于这类信息具有权威性、准确性和经济性等特点，因而备受人们关注。互联网为我们利用此类信息提供了非常便利的网络平台。

第四，科技报告。

科技报告是报道（记录）研究工作和开发调查工作的成果或进展情况的一种文献类型。科技报告具有内容比较新颖、详尽、专深的特点。其中可以包括各种研究方案的选择与比较、成功与失败两方面的体会，还常常附有大量的数据、图表、原始实验记录等资料。在时间上，科技报告发表比较及时，报道新成果的速度一般快于期刊及其他文献。按报告的流通范围，可划分为保密报告、非密限制发行报告、解密报告、公开报告等。在流通范围上，大部分科技报告都有一定的控制，即属于保密的或控制发行的，仅有一小部分可以公开或半公开发表。

第五，档案。

档案是国家机构、社会组织以及个人从事政治、军事、经济、科学、技术、

文化、宗教等活动直接形成的具有保存价值的各种文字、图表、声像等不同形式的历史记录。按来源可分为国家机关档案、企业档案、名人档案等；按内容可分为行政档案、科学技术档案、人事档案、财务档案、诉讼档案等；按记录方式可分为文字档案、图形档案、音像档案；按时间可分为古代档案、近代档案和现代档案；按所有权可分为国家所有档案、集体所有档案和个人所有档案。目前在我国通常将档案分为文书档案、科学技术档案、专门档案和声像档案。档案具有重要的凭证作用和参考作用。

第六，专利文献。

专利文献主要指专利说明书及相关文献，是非常重要的技术信息，通过检索和利用专利文献可以获得有关先进技术的发明及应用的最新信息，对技术创新、成果开发等有积极的借鉴、参考、启迪作用。同时，专利文献还可起到法律文件的作用，在引进国外技术和设备时，通过查阅专利文献可以比较各国、各公司的技术、设备先进程度，核实有关的专利项目，以保护自身利益等。

第七，标准文献。

标准文献是记录技术标准、管理标准和其他具有标准性质的文件的文献形式。它是依照规定程序，经过权威机构批准的文件。标准文献是了解世界各国工业发展情况的重要信息，同时它还能够为研制新产品、改造老产品、改进工艺和操作水平等提供借鉴和参考。

1.1.3 网络环境下的信息变化

1.1.3.1 互联网对信息的影响

1978 年，美国著名情报学家兰卡斯特（F. Wilfrid Lancaster）曾预言：“我们正在迅速地、不可避免地走向一个无纸的社会，计算机科学和通信技术的不断发展，允许我们设想一个整体性的电子系统，在这个系统中研究报告的编写、出版、交流和利用都完全以电子方式进行，因此纸在这个环境中是不再需要存在了，我们正处在从印在纸上到电子化的自然演变的中间阶段。”互联网在全球的迅速普及加快了这一演变过程。互联网是全球性的、最具影响力的计算机互联网络，也是世界范围的信息资源宝库。通过互联网，用户可以实现全球范围的电子邮件收发、信息查询与浏览、文件传输、语音与图像通信服务等功能。目前，互联网已经成为覆盖全球的信息基础设施之一。互联网造就了我们新的工作与生活方式，对整个社会信息交流方式和信息组成结构产生了巨大的影响。

1. 互联网引发了信息新的出版形式——网络出版，形成了新型的网络信息资源

网络出现以来，人们捕捉和获取信息的方式发生了根本性的变革。在出版领

域，人们开始在计算机网络上直接组稿、编辑、出版、制作以及销售。比如，在网页上发布信息；将信息制成文件，以电子邮件的形式定期发送给订阅用户等。网络出版的直接结果是形成了对人类信息获取具有重大影响的网络信息资源。

2. 互联网使人类传统的信息交流方式发生了根本性的变化

具体表现为两个方面：一方面，互联网的信息交流呈现出明显的开放性和广泛性。开放性指互联网面向所有愿意并有条件上网的用户，网上的信息来自各种类型的提供者，并为各种类型的用户使用，可以进行各种类型的信息服务。广泛性是指互联网在很大程度上超出了传统的信息交流的范围，使整个世界成为一个信息交流的整体。另一方面，信息交流方式显现出较强的交互性和实时性。互联网是一种双向式的信息交流活动，用户不仅是网络信息资源的消费者，同时也是网络信息资源的生产者和提供者。网络靠技术的支持，以极高的速度实现了用户的异地信息交流活动，人们之间通过网络所进行的信息交流活动能够实现音频、视频和计算机文本信息的相互配合，实现了跨越时空的实时性信息交流。

3. 互联网推动了信息处理技术的发展

互联网使计算机信息处理技术得到了长足的发展。计算机信息处理技术的核心是数字技术，用数字的形式实现了信息的物理载体（包括纸面）再现、储存、展示等所有需要。数字信号在时间上和数值上均是离散的，使得信息易于存储、分析和传输，并可以进行无限量复制。对多种信息形式（文字、数据、声音和图像）进行综合处理的多媒体技术，给人类思想表达、记录、交流、传播带来了较为深刻的影响。

1.1.3.2 网络环境下信息的新特点

网络环境下信息发生了许多变化，表现出如下新特点：

1. 信息类型多样化

在纸发明以前，人类主要在兽骨上、树皮上等记载信息。造纸术的发明彻底改变了人们简单利用自然物质记录信息的原始方法，纸成为人类社会记载信息的最主要载体，至今印刷型信息仍然是最重要的信息类型。随着现代信息技术的发展，人们开始了对信息海量存储的探索，磁带、磁盘、光盘等成为承载信息的重要介质。随着网络信息数量的迅速增加，互联网日益普及，在网络环境下，形成了印刷型信息、磁光介质型信息和网络型信息鼎立互补的新格局，而且网络信息越来越受到用户的欢迎和喜爱。

2. 信息的数量和内容都得到了极大的丰富

我们处在一个信息爆炸的时代。据统计，全世界平均每小时出现 20 多项新发明，每年要发表 2 000 多万篇科技论文，每年产生 720 亿条信息。在互联网

上，全球每天传递的电子邮件已达到 14 亿封，平均每分钟有 97 万封电子邮件被发送，全球平均每天每 5 个人中就有一人发送或接收一封电子邮件。[①] 在网络时代的今天，信息呈几何级数增长。传统的出版物数量在膨胀，同时，信息发布的自由性和任意性导致了网络信息的激增，现代信息技术为信息内容的展现提供了坚实的技术支持，信息内容更加深入和丰富。

3. 信息在分布上呈现出明显的分散性

纸本文献信息主要集中在图书馆、情报所、档案馆、书店、出版社等场所，人们必须亲临这些机构查询信息，在信息获取的时间和空间上受到了极大的限制。网络信息资源无论在地理上还是在组织形式上都呈现出分散分布的特点，改变了传统的线性组织方式，它采用了超文本和超媒体技术，使浏览者可以十分自由便捷地从一个文件、网页或者网站转移到另一个文件、网页或者网站，并且可以实现不同媒体之间的链接，无论它们是位于同一计算机中，位于不同的计算机中，还是位于不同的国家或者地区中。信息网络技术使信息的收集、编辑、分析、发布在世界范围进行，目前，除了传统的信息收集、出版、销售机构外，还有数据库生产商和相关的信息经营机构等，而且，互联网中任何一个资源服务器上都存储有提供给用户利用的信息。

4. 信息共享程度提高

在传统的环境下，信息只能是在一定时间、空间范围内得到有限共享，信息网络给人类带来了方便的信息获取渠道和信息资源更大程度的共享，为人类提供了一个全新的信息环境。特别是互联网成功地采用了 TCP/IP（传输控制协议和网际互联协议），TCP/IP 采用的互交换技术，解决了不同硬件平台、不同网络产品和不同操作系统之间的兼容性问题。任何计算机只要采用 TCP/IP 协议与互联网中的任何一台主机通信，就都有可能成为互联网的一部分，进行大规模的网络互联，自由地选择利用各种网络服务，顺利地实现信息资源的共享。

1.2 信息检索的概念和原理

1.2.1 信息检索的概念

信息检索（Information Retrieval）一词出现在 20 世纪中期。1950 年，美国

① 参见 http://www.zslib.com.cn/xuehui/2002lw/%E8%83%A1%E4%BF%8A%E8%8D%A3%E8%AE%BA%E6%96%87.doc，2007-07-25。

数学家莫尔斯（Calvin W. Mooers）在一次国际数学会议上发表了论文《把信息检索看作是时间性的通信》，文中提出了“信息检索”。这一时代正是计算机开始应用的时代，在此以前，信息存储和传播主要以纸质介质为载体，信息检索活动也主要围绕着纸质文献的获取和控制展开，信息检索关注的是如何检索利用文献中记载的信息，文献检索一度成为信息检索的同义词。随着计算机技术在检索领域的应用，“情报检索”（Information Retrieval）一词被更多地使用，计算机检索成为主流。由于汉语中“信息”较“情报”的含义更为宽泛，英文information可以理解为“信息”或“情报”，同时，在网络环境下，通信技术与计算机技术紧密地结伴同行，信息载体类型日趋多样化，传统的情报检索研究和文献检索研究逐渐归入信息检索研究这一更具兼容性的概念。

从广义的角度讲，信息检索包含信息存储和信息获取两个过程。信息存储指通过对大量无序信息的选择和收集、著录和标引等方法，建成各种各样的信息检索工具或信息检索系统，使之成为有序化信息集合的过程。获取是存储的逆过程，其实质是根据特定的需求，运用已组织好的检索系统，将特定的信息查找出来。存储是获取的前提和基础，没有存储就没有获取，而获取是存储的目的，二者密切联系，互为依存，缺一不可。本教材立足于信息检索的广义概念，研究信息存储和信息获取的全过程。

狭义的信息检索是指广义的信息检索的后一个过程，即信息获取的过程，相当于人们所说的信息查检等。具体来说，狭义的信息检索指通过一定的方法，从已存储的信息中检索出与用户提问相关的文献、数据和事实的过程，即根据用户的特定要求查找所需信息的过程。

1.2.2 信息检索的原理

实质上，信息检索原理就是将特定的信息需求与存储在检索系统中的信息标识进行异同的比较与匹配，选取两者相符或部分相符的信息予以输出。无论手工检索还是计算机检索，其基本原理都是一样的。也就是说，检索系统对所要存储的信息，按照其外部特征和内容特征进行描述并赋予特征标识，然后存入系统；检索时，将所需信息的特征标识与所存信息的特征标识进行比较。凡是两边标识一致的，就将具有这些标识的信息从检索系统中输出。具体如图1—1所示。

根据检索对象的不同，信息检索可以区分为不同的类型：

1.2.2.1 文献检索（Document Retrieval）

这是信息检索的主体部分，以特定的文献为检索对象，包括全文、文摘、题录等。文献检索是一种相关性检索，它不直接回答用户所提技术问题的本身，只

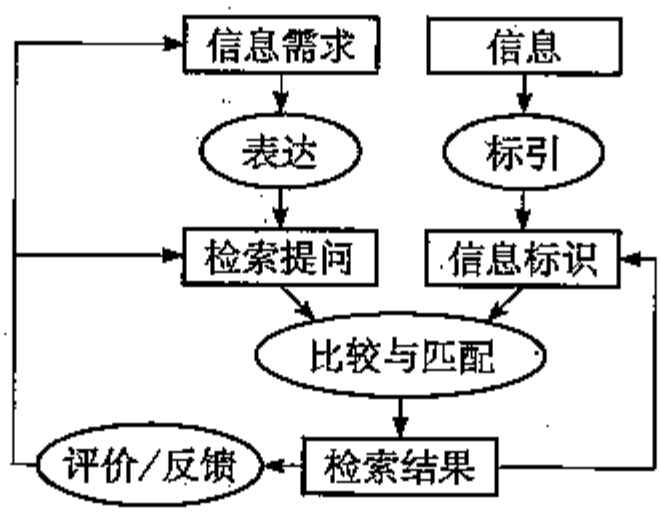

图 1—1 信息检索原理

提供有关的文献供参考。

1.2.2.2 数据检索（Data Retrieval）

以特定的数据为检索对象，包括统计数字、工程数据、图表、计算公式、化学结构式等。数据检索是一种确定性检索，它能够提供确切的数据，直接回答用户所提问题的本身。

1.2.2.3 事实检索（Fact Retrieval）

以特定的事实为检索对象，如有关某一事件发生的时间、地点、人物和过程等。事实检索也是一种确定性检索，一般能够直接提供用户所需的确定的事实。但有时需要对所得到的事实进行必要的分析和推理，才能得到最终的答案。

1.2.3 信息检索的意义

信息检索是有效获取人类智力资源的重要手段，是连接信息生产者和信息需求者的通道和接口。其主要作用表现在以下方面：

1.2.3.1 信息检索是有效利用信息资源、实现其最大价值的科学方法

人类社会发展到今天，积累了丰富的信息和知识。信息资源管理与开发水平已成为衡量一个国家信息文明程度的重要标志之一。如何在这信息的汪洋大海中找到自己所需要的信息，并充分利用这些信息？信息检索为我们提供了一套较完整的利用和开发信息的方法，包括检索策略的制定、检索工具的选择、检索手段的选择等。技术的进步取决于信息的挑选和积累，搜集、活用和分析信息，然后再发出新的信息，并以新信息为基础，探索事物所蕴涵的更大价值，因此，信息检索是信息分析和科技创新的基础。现代信息技术的发展，推动了信息检索手段的日益现代化，这将大大加快加深社会信息资源的开发速度

和程度。

1.2.3.2 信息检索是再学习的工具，是获取知识的有效途径

我们生活在一个知识经济社会，知识老化周期变短、产品换代加速是知识经济社会一个非常明显的特征。这就要求我们每一个人都必须不断学习新东西，获得新情报，运用新方法，更新自身的知识结构，以适应社会快速发展的步伐。美国工程教育协会曾估计，人们所需知识的20%～25%是学校教育赋予的，而75%～80%的知识是走出学校后，在研究实践和生产实践中根据需要，通过不断再学习而获得的，而信息检索则成为人们获取知识、提高自我的最重要最普遍的形式。人们通过各种途径获取信息，完成知识更新，适应社会的发展，而信息检索正是人们获取知识的有效途径。

1.2.3.3 信息检索能有效地提高科研工作的效率，节省人力物力及时间

对于科学研究工作者来说，信息检索更为重要。一项科研课题无论是在立项之前，或是在研究过程中，甚至在研究完成后，对已有成果的评价方面，都离不开查阅有关文献信息资料。据统计，科研人员大约花40%的工作时间查检文献，如果没有掌握科学的信息检索方法，那么时间还会加长。更有甚者，因为不了解同时期的前沿信息，使全部工作成了“重复劳动”。高效的信息检索可以起到事半功倍的效果，使科研人员掌握相关的进展，避免重复研究，将时间和精力集中于创新工作，多出成果，出好成果。

1.2.4 信息检索的历程

信息检索的发展与人们信息需求的增长以及现代信息技术的进程紧密相关。我们可以将文献检索看作是信息检索的前身，文献的出现蕴涵了文献检索的萌芽，但在人类文明的早期，文献数量不是很大，信息获取主要依靠直接交流来实现。进入20世纪以后，科学技术飞速发展，文献大量增加，信息需求日趋显著，人们获取信息的复杂程度加大，信息检索活动开始从人们的科学研究、科学交流中分离出来，形成了真正意义上的信息检索。追溯时间发展的脉络，信息检索的发展经历了手工检索阶段和计算机检索阶段：

1.2.4.1 手工检索

手工检索直接发源于图书馆的参考咨询工作和文摘索引工作。一方面，19世纪下半叶，美国的公共图书馆和大专院校图书馆的参考咨询工作有了很大进展。20世纪初，多数图书馆设立了参考咨询部门，主要利用参考工具书来帮助读者查找图书、期刊或现成答案。另一方面，随着文摘索引工作的发展，检索刊

物体系逐渐形成，检索工具书日趋完善。约在七八世纪，出现了西方第一部专门的索引，即为《圣经》编的《圣经语词索引》。1665 年，法兰西科学院在巴黎创办了《学者周刊》，这是世界上最早的科学期刊之一，也是以专栏或附录形式出现的最早的文摘刊物。到 19 世纪初，文摘刊物开始脱离附录的形式，走向独立编辑出版，索引也随着报刊文献的增多而得到了很大的发展，并与文摘刊物紧密结合在一起，成为查找科学文献的最重要的手工检索工具。在这一阶段，信息检索逐渐形成一个独立的领域，走向专门化，纸本工具书是这一时期信息检索的主要工具。

1.2.4.2 计算机检索

计算机的诞生带来了信息检索的革命，20 世纪 50 年代初，人们开始研究计算机在信息检索和信息管理领域的应用。1954 年，美国海军兵器中心首先在 IBM701 型电子计算机上成功地建立了世界上第一个计算机文献检索系统，标志着人类开始步入利用计算机进行信息检索的新的历史时期。计算机检索系统将大量文献信息进行整理、存储，并通过通信网络将信息迅速传递给用户，极大地方便了用户的信息查找。用户可以不必到图书馆或情报中心去查找资料，只需要坐在办公室、实验室或家里，在计算机终端接通国际通信网络与情报中心联机后，将选择的检索词或检索式输入计算机，直接向各种机读数据库获取文献信息资料，计算机在很短的时间内就会将检索到的文献在屏幕上显示出来。计算机检索速度快、效率高、及时、全面，突破了地理上的限制。随着计算机技术和网络技术的发展，计算机检索经历了脱机检索、联机检索、光盘检索和网络检索四个阶段。在当今网络环境下，计算机检索将发挥更大的作用，更好地满足人们日益增长的文献信息需求。

1.3 信息检索系统

1.3.1 信息检索系统的概念

信息检索系统是指根据特定的信息需求而建立起来的一种有关信息搜集、加工、存储和检索的程序化系统，其主要目的是为人们提供信息检索服务。信息检索系统有多种形式，如工具书、数据库或搜索引擎等。

美国著名情报学家兰卡斯特向我们展示了信息检索系统的主要工作原理，具体参见图 1—2：

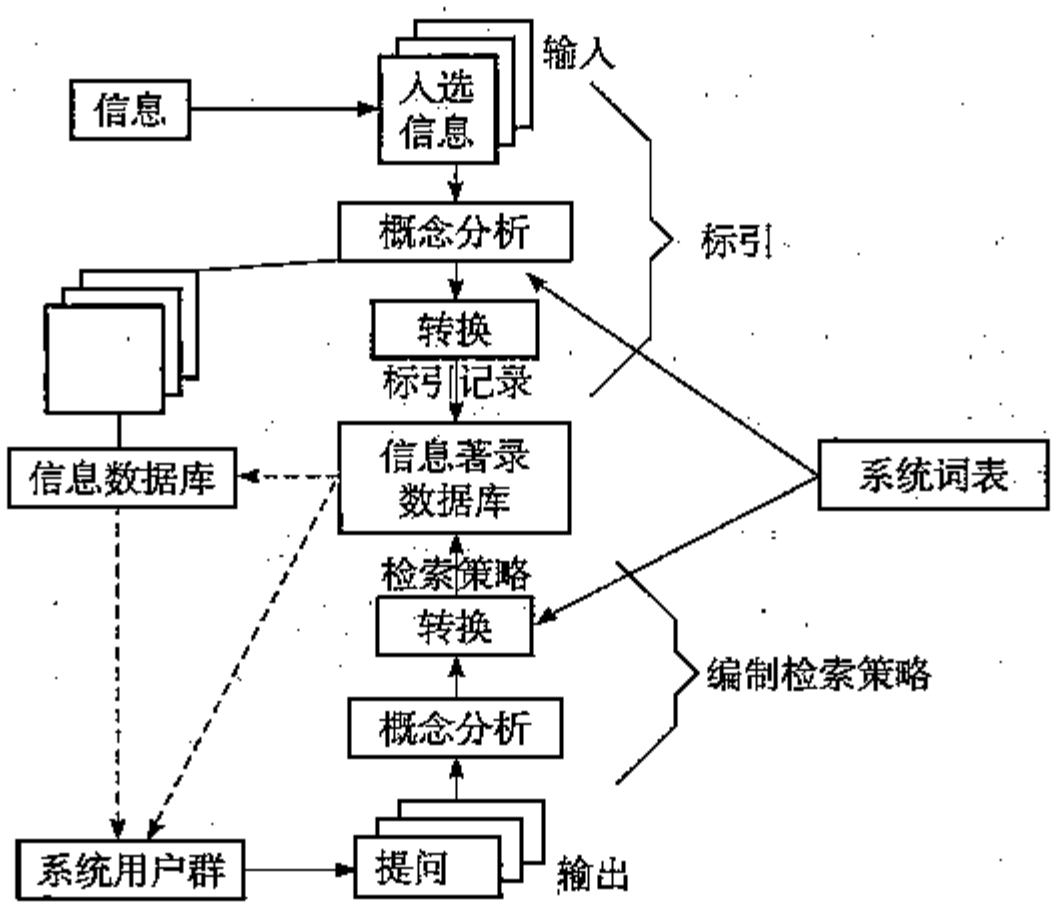

图 1—2 信息检索系统工作原理图

资料来源：F. Wilfrid Lancaster & Amy J. Warner, *Information Retrieval Today*, Information Resources Press, 1993。

信息检索系统包括信息的存储和获取两个部分，分别对应信息的输入和输出过程。具体工作过程如下：

该系统输入的是信息，即采集到的信息，这就意味着要依据特定的目标，按照一定的标准和方法对采集到的信息进行组织和管理，以使信息能够易于识别和理解，便于回答用户的各种提问。其中，标引是信息的组织与管理过程中的重要环节，即通过对信息的分析，选用确切的检索标识（类号、标题词、叙词、关键词、人名、地名等），用以反映该文献内容的过程。标引过程一旦完成，信息就进入某种形式的信息库，而标引记录则进入二次信息数据库，在二次信息数据库中，按便于检索的方式，对标引记录加以组织。

信息检索系统的输入端是针对信息，而信息检索系统的输出端则是针对用户的提问。实际上，系统输出端的操作步骤和输入端非常相似。接受服务的用户群向信息中心送交各种提问，中心的工作人员则为其提问编制检索策略，也可以由用户自己直接对信息检索系统进行提问，编制检索策略。检索策略的编制包括概念分析和转换两个步骤：第一步是对用户提问进行分析，确定用户实际上要找的是什么；第二步是把概念分析转换成词汇，转换成该系统语言的提问，并将检索提问以系统认知的检索式表达出来，这是“检索策略”的核心。检索策略编制出

来后，就以某种方式将其同事先存储好的数据库进行匹配，再将得到的结果返回给提问者。如果提问者对检索结果满意，该过程即告结束。如果提问者对检索结果不满意，则需要修改检索式，调整检索策略，进行再次检索。

1.3.2 信息检索系统的构成

信息检索系统具有对信息的输入功能、存储功能、处理功能、输出功能及控制功能。一般来说，信息检索系统包括 6 个主要的子系统：

1. 信息选择子系统

信息检索系统中的数据主要来自各种公开文献和信息，如一次信息中的期刊、图书、研究报告、会议论文、专利文献、政府出版物、学位论文；二次信息中的文摘、索引和目录等。有些系统还收录了各种机构的内部资料，如实验记录、测试或观测结果、工程设计资料、统计资料等。在网络平台上，收录的可以是万维网资源、Gopher 资源、FTP 资源等等。信息选择子系统主要根据系统的特点和服务的用户群体来搜集相关的信息资源，为系统提供充足而适用的数据来源。

2. 信息索引子系统

在分析和选取信息的内容和形式特征基础上，根据具体的词表和名词规范，来选择准确的信息标识。

3. 词表管理子系统

又称检索语言和名称规范子系统，它的主要功能是管理维护系统中已有的词表，使它与索引等子系统相连接，支持用户的各种词汇查询操作，从提问、对话或其他文本中采集新的词汇信息，以及输出各种形式的词汇数据或词表产品（从个别词目、词间关系、词频数据到整部词表）。

4. 检索子系统

承担接收用户提问、提问校验和进行检索等功能。

5. 用户同系统之间交互子系统

具有与用户进行交流，以便真正明确用户的真实信息需求，明确检索提问，并准确表述等功能。

6. 匹配子系统

将信息标识与检索提问进行相符性比较的子系统。

1.3.3 信息检索系统的分类

按照信息检索的实现手段，可以把信息检索系统分为手工检索系统和计算机

检索系统。

1.3.3.1 手工检索系统

手工检索系统指以印刷型检索工具为基础的检索系统，它可以直接进行利用，不需要依赖任何计算机或其他设备。常用的手工检索系统主要有：

1. 书本式的手工检索系统

指以图书、期刊、附录等形式出版的各种检索工具书和检索刊物，如目录、索引、文摘、百科全书、年鉴和手册等等。

2. 卡片式的手工检索系统

指以卡片的形式出现的检索系统，包括图书馆的卡片式目录等，如一般的图书馆都设有书名目录、著者目录、分类目录和主题目录等。

手工检索系统主要是经过大脑的判断来实施和完成检索，面对的是印刷型载体，符合人们长期以来形成的阅读习惯，而且，可以根据需要及时调整检索策略，达到满意的效果。但手工检索系统收录的范围有限，更新速度慢，检索效率远不及计算机检索系统。

1.3.3.2 计算机检索系统

计算机检索系统指依赖于计算机进行信息检索的系统，主要由三个部分构成，即硬件部分、软件部分和信息数据库。

1. 硬件

硬件是指以计算机为中心的一系列机器设备，一般包括计算机、外围设备以及与数据处理或数据传送有关的其他设备。计算机是硬件的核心，外围设备有外部存储器、输入输出设备等。不同的计算机检索系统在硬件配置上有一定的差异。

2. 软件

又称计算机程序，是指挥和控制计算机各部分协调工作并完成各项功能的程序和各种数据。软件包括操作系统软件、语言编译软件、应用软件和用户软件等。

3. 数据库

数据库是计算机信息检索系统最重要的组成部分。根据 ISO/DIS5127 号标准（文献与信息工作术语），数据库（Database）被定义为：“至少由一种文档组成，并能满足某一特定目的或某一特定数据处理系统需要的一种数据集合。”简单地说，数据库是依照某种数据模型组织起来并存放于计算机存储设备中的数据集合。

对用户而言，计算机检索系统主要是数据库的使用。国际上，一般把数据库分为参考数据库和源数据库两种。

（1）参考数据库（Reference Databases）。

指为用户提供信息线索的数据库，它可以指引用户获取原始信息。参考数据库包括书目数据库（Bibliographic Databases）和指南数据库（Referral Databases）。

第一，书目数据库包含文摘、目录、题录等书目数据，有时又称为二次信息数据库。书目数据库中的数据来源于各种不同的一次信息，是经过加工和提炼的数据。书目数据库的数据结构比较简单，记录格式较为固定。在联机检索和光盘检索中，有许多书目数据库，可以满足用户回溯检索和定题检索的需要。

第二，指南数据库是有关机构、人物等相关信息的简要描述。包括各种机构名录数据库、人物传记数据库、产品信息数据库、软件数据库、技术标准数据库、基金数据库等。

（2）源数据库（Source Databases）。

指能直接提供原始资料或具体数据的数据库。它包括数值数据库、文本—数值数据库、全文数据库、术语数据库、图像数据库和多媒体数据库等。

第一，数值数据库（Numeric Databases）：数值数据库主要用于查询各种有关的数字、参数、公式等。它是一种以自然数值形式表示、计算机可读的数据集合，这些数据是从文献中分析、概括、提取出来的，或从调研、观测及统计工作中直接获得的。它可以直接提供解决问题时所需要的数据，是进行各种统计分析、定量研究、管理决策和预测的重要工具。

第二，文本—数值数据库（Textual-Numeric Databases）：能同时提供文本信息和数值数据。例如产品市场报告数据库等。

第三，全文数据库（Full-Text Databases）：包含原始信息正文或其主要部分的数据库。通过它可以直接检索出原始信息的全文，实现检索的一次到位。全文数据库具有直接性、详尽性、易检索等特点，成为计算机检索领域非常受用户欢迎的源数据库，并得到了迅速发展。

第四，术语数据库（Terminological Bank）：又称电子辞典，专门存储名词术语信息、词语信息以及术语工作和语言规范工作成果的一种源数据库。术语的准确和规范对于学科的发展及专家的交流都有非常重要的意义，术语数据库是一种非常有效的术语控制和规范化的工具。

第五，图像数据库（Graphics Databases）：用来存储各种图形、图像及相关文字说明资料的源数据库，内容主要包括建筑、设计、广告、产品的图片或

照片。

第六，多媒体数据库（Multimedia Databases）：能把文字、数值、声音、图像等不同信息存储在不同媒体上，进行统一处理和管理的数据库。

目前，计算机检索系统是检索系统的主流，主要包括光盘检索系统、联机检索系统和网络检索系统。由于计算机检索系统具有速度快、效率高，数据内容新、范围广、数量大，操作简便，在网络环境中检索不受时空限制等特点，已成为人们获取信息的主要手段之一。

1.4 信息检索方法

信息检索的效率与具体的信息检索方法有很大的关系，运用有效的信息检索方法能够使用户以最少的时间获得最满意的检索结果。总的来说，检索方法有直接浏览法、常用法、追溯法和综合法。

1.4.1 直接浏览法

直接浏览法也称直接查找法，指检索者不依靠任何检索工具或检索系统，从本专业最新核心期刊或其他文献中直接阅读原文或浏览最新目次而获取文献的方法。这是一种最常见的信息资源的获取方式。因为编制检索工具需要时间，半年，甚至长达一年之久，直接浏览则可以及时获得最新文献。但利用这种方法查找的信息不全面、不系统且局限性较大，不能作为查找文献的主要方法。

1.4.2 常用法

指利用检索系统来查找信息的方法，这是目前查找信息的最常用的方法，故而称常用法。常用法包括顺查法、倒查法和抽查法。

顺查，就是由远及近地顺时间查找。利用顺查法，一般需要了解检索课题的背景和发生简况，从而选择比较适宜的检索工具及检索系统，从问题产生的时间开始查起，直到最新的文献信息，这种方法具有查全率较高的优点。例如，已知某项发明或研究的产生年代，现在需要了解它发展的全过程，就可以利用顺查法从最初的年代开始，逐步往近期查找。

倒查，就是由近及远地逆时间查找。倒查法的重点是放在近期信息资源上，以基本满足用户自己的信息需求为原则。使用这种方法可以最快地获得新资料、新信息，这种方法投入劳动比较小，省时省力，检索效率较高，但不如顺查法查

全率高，而且，对检索课题的来龙去脉不易掌握。写作论文做资料准备时常采用这种方法。

抽查，即抽取其中某段时间查找。抽查法关注有关课题的文献信息最可能出现或最多出现的时间段。用这种方法能获得相对集中、具有代表性且能反映该课题发展水平的文献信息，往往能起到事半功倍的效果。它具有检索效率高，检索效果好的优点。但要求用户基本了解该课题的大概情况，能够比较准确地选择出有关信息出现较多的时间段。

1.4.3 追溯法

指从已有的文献信息后所列的参考文献入手，逐一追查原文，从这些新查到的原文后面所附的参考文献再逐一追查，不断扩大检索范围的检索方法。其优点是：在没有检索工具或检索工具不齐全的情况下，借助此种方法，也可以查到一批有关的文献。其缺点是：原文作者引用的参考文献是有限的，不可能列出全部有关文献，而且有的引用文献又与原文关系较小或较远，参考价值不大。单独使用这种方法，还是存在一定的局限性。

美国的《科学引文索引》（Science Citation Index）就是按照这一原理而编制的一种检索工具。比如，它可以从作者途径去检索引用该作者著作的有关文献，它不仅反映出某个作者历来发表了哪些文献，而且也反映出其他作者引用该文献从而发表新的著作的情况。它揭示了科技文献中引用与被引用的客观状况。

1.4.4 综合法

也称分段查找法、循环法或交替法。先利用检索工具查出一定时期内的一批有用的文献，然后依据这些文献后所附的参考文献，利用追溯法查出前一时期的文献，如此分期分段地交替使用常用法和追溯法两种查找方法，直到满足要求为止。这种方法兼有上述两种方法的优点，可以查得全面而准确的信息，适合于查阅那些过去文献量较少的专业资料，并可弥补因检索工具不全而造成的漏检，检索效果较好。

1.5 信息检索效果

1.5.1 信息检索效果评价

信息检索效果是指信息检索系统检索的有效程度，它衡量了检索结果对用户

需求的满足程度，是检索系统性能的直接反映。信息检索效果评价指运用科学的方法，按照设定的指标体系，对信息检索效果进行评价的过程。目前，主要从三个方面进行评价：（1）检索结果有效性评价，主要以查全率和查准率为评价标准；（2）检索系统实用性的评价，包括系统对用户是否需要，是否实用，有多大的实用效果，即检索的社会效果的评价，需要应用社会学方法；（3）检索费用——效率评价，即检索的经济效果的评价，包括检索系统完成检索服务的成本及时间消耗，需要应用经济学方法。

信息检索效果评价对于信息检索系统的建设和发展具有重要意义。它是信息检索系统不断趋于完善的重要依据。获得让客户满意的检索效果是每一个信息检索系统追求的目标，而系统本身又无法完成自我调节，需要外在环境的监督和指引。通过检索效果评价，可以准确地掌握系统的各种性能和水平，找出影响检索效果的各种因素，从而有的放矢地改进系统的性能，提高系统的信息检索能力。

信息检索效果评价的核心问题是建立一套切实可行的评价指标。由于检索效果的评价涉及许多问题，可以选择不同的角度，采用不同的评价方法。美国著名情报学家兰卡斯特提出，用户可以从质量、费用和时间三方面来评价检索系统。质量标准主要通过数据库覆盖范围、查全率、查准率、数据的完整性和准确性来反映。费用标准即检索费用，是指用户为检索课题所投入的费用。时间标准是指花费时间，包括检索准备时间、检索过程时间、获取文献时间等。其中，查全率和查准率是判定检索效果的主要标准。据美国学者克莱弗登的研究，评价信息检索效果的指标主要有六个：收录范围、查全率、查准率、响应时间、用户负担和输出形式。

1.5.1.1 查全率（Recall Ratio）和查准率（Precision Ratio）

查全率和查准率是由美国佩里（J. W. Pery）和肯特（Allen Kent）于20世纪50年代中期提出来的，后经不断改进和完善，现已成为评价检索效果最常用的两项关键指标。确定查全率和查准率最常用的方法是有名的2×2表（见表1—1）。

表1—1 检索结果2×2表

用户相关性判断 / 系统相关性测报	相关	不相关	总计
已检出	a	b	a+b
未检出	c	d	c+d
总　计	a+c	b+d	a+b+c+d

2×2表反映了检索系统在实施某一次检索时所得的结果状况。其中a表示被

检出的相关文献，即查准的信息；b 表示被检出的不相关文献，即误检的信息；c 表示未检出的相关文献，即漏检的信息；d 表示未检出的不相关文献，即正确拒绝的无关信息。信息检索系统中参加检索的全部文献量为（a＋b＋c＋d）。从检索系统角度来看，它们可以分为被检出文献（a＋b）和未检出文献（c＋d）两部分，因为两部分文献反映了检索系统处理是否与检索提问相关，故称之为系统相关性测报。从用户的角度来看，检索系统文档中参加检索的全部文献也可以分为两个部分：一部分与用户需要相符，称为相关文献（a＋c）；另一部分与用户需求不符，称为不相关文献（b＋d）。因为这两部分文献反映了用户判断是否与检索需要相关，故又称之为用户相关性判断。

查全率指检出的相关文献信息量与检索系统中相关文献信息总量的比率，它反映出信息检索系统检出相关文献信息的能力。具体公式如下：

$$\text{查全率} = [\text{检出相关文献信息量}/\text{检索系统中相关文献信息总量}] \times 100\% = [a/(a+c)] \times 100\%$$

查准率指检出的相关文献信息量与检出文献信息总量的比率，它反映出信息检索系统的精确度，说明系统排除干扰，减少噪声的能力。具体公式如下：

$$\text{查准率} = [\text{检出相关文献信息量}/\text{检出文献信息总量}] \times 100\% = [a/(a+b)] \times 100\%$$

虽然查全率与查准率能较好地反映出一个检索系统的检索效果，但是在实际操作过程中，这两个指标也存在一定的局限性。首先，在计算查全率时，一个检索系统中总共有多少相关文献（a＋c）难以确切计算，而只能是大概估算；其次，在计算查准率时，用户对文献的相关性估计与系统的相关性判断不一定是完全吻合的，而且，不同的用户对相关文献的认识也可能不一致，存在着太多的主观成分和一些模糊概念。因此，用上述方法求得的查全率与查准率并不是绝对的，而只能是相对近似地描述检索效果。

传统的情报检索理论认为：查全率与查准率具有互逆相关关系，也就是说，如果提高检索的查准率，就会降低检索的查全率。该论点首先来源于英国学者克里维顿（C. Cleverdon）的著名 Cranfield 实验。美国的兰卡斯特在他的《情报检索系统——特性、试验与评价》一书中也明确提出“查全率与查准率总是相反的关系”，而且根据 50 次检索的调查结果绘制出了有名的经验曲线，反映出查全率和查准率之间的互逆关系。

目前，一些学者对查全率和查准率的关系进行了深入研究，提出这两个指标之间不仅存在互逆关系，而且还可以存在互顺关系和其他关系，并通过检索实

例、理论描述和数学推理等论证了此观点。[①] 证明查全率与查准率之间的关系与检索提问式的结构有关，不同的检索条件下，查全率与查准率之间将呈现以下三种不同的关系：当由于检索策略的变化，使得检索到的相关记录的变化量与全部命中记录的变化量之比小于相关记录与命中记录数之比时，查全率—查准率呈现逆变关系；当由于检索策略的变化，使得检索到的相关记录的变化量与全部命中记录的变化量之比大于相关记录数与命中记录数之比时，查全率—查准率呈现顺变关系；当由于检索策略的变化，使得检索到的相关记录的变化量与全部命中记录的变化量之比等于相关记录数与命中记录数之比时，查全率可能变化，而查准率不变。[②]

1.5.1.2 漏检率（Omission Factor）和误检率（Noise Factor）

漏检率指漏检相关文献信息量与检索系统中相关文献信息总量的比率，它与查全率相对应。具体公式如下：

$$\text{漏检率} = [\text{漏检相关文献信息量}/\text{检索系统中相关文献信息总量}] \times 100\% = [c/(a+c)] \times 100\%$$

误检率指误检（检出不相关）文献信息总量与检出文献信息总量的比率，是衡量信息检索系统误检程度的尺度，与查准率相对应。具体公式如下：

$$\text{误检率} = [\text{误检文献信息量}/\text{检出文献信息总量}] \times 100\% = [b/(a+b)] \times 100\%$$

1.5.1.3 响应时间（Response Time）

响应时间指在一次检索过程中，用户从开始向信息检索系统提问到系统输出检索结果的全部时间。响应时间的长短也是评价检索系统效果的重要指标，直接反映着信息检索的速度。一般来说，响应时间越短，查全率和查准率越高，那么信息检索的效果就越好。如果检索系统速度太慢，系统实用性就会大打折扣。响应时间在很大程度上依赖于检索手段和检索技术的进步，在手工检索阶段，响应时间受检索者主观因素的影响比较大，主要取决于检索者制定的检索策略的优劣，以及对检索工具的选择和对检索工具使用方法的熟悉程度，响应时间一般比较长。在计算机检索阶段，信息检索的响应时间大大缩短，主要由系统对信息处

---

① 此观点可参见：马景娣：《查全率—查准率间存在顺变关系的数学证明》，载《情报科学》，2003（1）；邓汉成、王敏芳、王瑛：《查全率与查准率之间关系的理论研究》，载《情报学报》，2000（4）；邓汉成、王敏芳、王瑛：《从检索实例看查全率与查准率之间的关系》，载《情报学报》，2000（3）。

② 参见马景娣：《查全率—查准率间存在顺变关系的数学证明》，载《情报科学》，2003（1）。

理速度决定，对于网络信息检索而言，用户所处的网络条件和利用的相关设备也在很大程度上影响着响应时间。随着智能检索在信息检索领域的发展，响应时间将会更大程度地依赖信息检索系统的处理速度和运行效率。

此外，还有一些与检索效果相关的指标，如检索系统的收录范围、结果输出形式、易用性、用户负担，以及在网络环境下发展起来的重复链接率、死链接率等。

收录范围又称数据覆盖率，数据库收录范围指标被作为衡量查全率的一项辅助指标，用以揭示数据库的涵盖范围。一个信息检索系统的收录范围直接影响到用户信息需求的满足程度。

输出形式是系统检索出文献信息的展示形式，可能是文献号、题录、文摘或全文等。输出的信息越多且便于浏览，用户越容易作出相关性判断。输出形式影响用户对检索结果的选择和利用。

系统的易用性也称为可存取性，反映了信息检索系统的易用程度。美国情报学家穆斯指出："一个情报系统，如果对用户来说，他取得情报要比他不取得情报更伤脑筋更麻烦的话，这个系统就不会得到利用。"易用性是用户选择信息检索系统的重要因素之一。

用户负担是用户在检索过程中所消耗的物力、财力乃至精力的总和。

结果的重复链接率为检索结果中内容重复的结果数占全部检索结果数的比例，死链接率为检索结果中死链接的结果数占全部检索结果数的比例。

1.5.2 影响信息检索效果的因素

信息检索效果是影响信息检索系统价值的主要因素，更是人们评价信息检索质量的重要指标。影响信息检索效果的因素有很多，几乎与检索系统性能及检索过程有关的各个因素都有关系，其中主要有标引的质量、检索语言的性能、检索途径的数量、检索策略的优劣、检索人员的素质等等。

1.5.2.1 标引的质量

信息标引的正确性对信息检索的查全率和查准率有着直接影响，信息标引的结果是赋予文献信息相关的检索标识，这对于信息存储的质量至关重要。检索标识是组织检索工具和数据库、进行检索的依据，正确的标引可以使同一主题的信息准确而全面地被检索出来。相反，各种标引误差都会对检索效率产生一定的影响。

标引误差主要来自主题分析误差、标引深度误差等。标引人员在标引信息时，首先要对信息进行主题分析，就是从信息的内容出发，分析出应当传递给读

者的主要信息，并从用户检索的角度出发，分析出应标引的主题概念，给出正确的检索标识，以供用户检索。如果对信息的主题内容分析错了或分析得不准确，就不可能给出正确的检索标识，使同一主题的文献分散，造成检索混乱，引起查准率降低。如果标引时主题分析不全面，有用信息没有被提取出来，遗漏了某些主题的标引，以后查找时就会造成漏检，使查全率降低。

1.5.2.2 检索语言的性能

检索语言是将信息标引和检索提问联系起来的重要桥梁，是沟通信息存储和信息检索的纽带，对于特定信息需求和信息检索系统中信息集合的准确匹配具有直接的影响。检索语言用于标引信息内容及其外表特征，可以对内容相同及相关的信息加以集中或者揭示其相关性；将信息的存储集中化、系统化、组织化，便于检索者按照一定的排列次序进行有序化检索；便于将标引用语和检索用语进行相符性比较，保证不同检索人员表述相同信息内容的一致性，以及检索人员与标引人员对相同信息内容表述的一致性。

词表结构对检索有很大影响，如分类表中的类目、主题词表中主题词的范围及专指程度，分类表中的交替类目、参照类目，主题词表中的语义参照系统都会影响到标引词的选择，影响信息存储和查找的准确性。如果词表不以某种方式把所有关联的标引词集中在一起，那些检索人员就不能将与查找要求有关的全部标引词找出来，查全率就会降低。

词表对标引也有很大影响，较好的词表参照系统和等级结构关系能够提高查全率。如果检索语言中的标引词不规范，则可能引起虚假组配现象，从而降低查准率。

1.5.2.3 检索途径的数量

检索途径也称检索入口，主要依据信息的内容特征和外部特征来确定。检索信息内容特征的有分类、主题和全文途径，检索信息外部特征的有题名、著者、文献编号途径等。一般来说，信息存储进检索系统后，该系统能够提供的检索途径越多，越便于检索人员对信息的查找和获取。如果一篇文献在检索系统只提供了一条检索途径，那么就要求检索人员必须找到这唯一的途径，才可能获得这篇文献。如果一篇文献提供了多条检索途径，检索人员只要找到其中一条途径，就可以很方便地找到该文献。例如，《中国期刊全文数据库》提供有篇名、作者、关键词、机构、中文摘要、引文、基金、全文、中文刊名、ISSN、年、期、主题词等13种检索途径，检索人员可以根据信息需求，选择不同的检索途径获取相关的文献。

1.5.2.4 检索策略的优劣

检索策略是进行检索的规划和方案，是影响检索效果的重要因素。检索策略涉及检索人员对检索目的、检索范围、检索系统、检索途径、检索式表达等一系列问题的思考和定位。检索人员应该根据具体的信息需求制定相应的检索策略，比如关于科研立项、科技查新等方面的检索课题，强调的是查全率，因而，应选择大型联机检索系统以及全面收录相关信息的数据库进行检索，尽可能避免漏检而造成重复劳动。对于一般的学术性信息检索，可以选择国内外商业数据库，也可以辅助利用搜索引擎查找相关资料。

在实际检索过程中，会出现检索结果偏离检索目标的情况，这时就需要检索人员能够及时分析失误原因，调整检索策略，通过对检索途径、检索词、检索式表述等方面做进一步的调整，以达到较为理想的查全率和查准率。

1.5.2.5 检索人员的素质

人是信息检索过程中的主体，无论是手工检索系统还是计算机检索系统，都要检索人员来具体实施检索，因此，检索人员的素质对于检索效果有着直接的影响。检索人员应该具备一定的信息检索知识，能够正确地分析检索课题，准确地表达信息需求，掌握信息检索的基本方法，了解计算机操作的基础知识，熟悉有关的信息检索工具和检索系统。

【案例】

安徽安特集团利用网络进行市场信息检索的实例①

安徽安特集团是我国特级酒精行业的龙头企业，全套设备及技术全部从法国引进。其主要产品是伏特加（Vodka）酒及分析级无水乙醇。其中无水乙醇的销量占全国的50%以上。伏特加酒通过边境贸易，向俄罗斯等国家出口达到1万吨，总销售额超过1亿元。

伏特加酒作为高附加值的主打产品，是安特集团利润的主要来源。但是，随着俄罗斯等国家的经济形势日趋恶化，出口量逐年减少，形势不容乐观。安特集团审时度势，决定通过互联网进行网络营销，开辟广阔的欧美市场。集团确定了信息收集的三个方向：

（1）价格信息：生产商报价、批发商报价、零售商报价、进口商报价。

（2）关税、贸易政策及国际贸易数据：关税、进口配额、许可证等相关政

① http://fireking0828.blog.hexun.com/3516751_d.html，2008-06-01。

策，进出口贸易数据，市场容量数据。

(3) 贸易对象：潜在客户的详细信息，包括贸易对象的历史、规模、实力、经营范围和品种、联系方法等。

根据信息需求，安徽安特集团利用互联网平台的相关数据库、搜索引擎、专业的管理机构及行业协会站点、相关生产商和销售商的站点、电子商务交易中心站点，进行了关键词检索和分类浏览，收集了以上三个方面的情报，通过对这些情报的分析和整理，该集团对于世界上 Vodka 酒的贸易状况有了基本的了解，掌握了世界 Vodka 交易的价格走势，认清了安特牌 Vodka 所处的档次水平，也联系了上百家进口商、经销商，可以说基本上把握了国际 Vodka 市场的脉搏，圆满地完成了情报收集和分析的工作。这些工作为以后的网上谈判、选择代理商等网络营销工作打下了良好的基础。

关键术语

信息	零次信息	一次信息	二次信息	三次信息
信息检索	信息检索原理	信息检索类型	信息检索系统	信息检索方法
常用法	追溯法	综合法	信息检索效果	查全率
查准率	误检率	漏检率	响应时间	

思考题

1. 简述信息的特征和功能。
2. 如何理解信息的分类?
3. 简述信息检索的含义。
4. 简述信息检索的原理。
5. 简述信息检索的意义。
6. 简述信息检索系统及其分类。
7. 信息检索的方法有哪些? 各有什么特点?
8. 简述信息检索效果评价的意义。
9. 信息检索效果评价的指标有哪些?
10. 影响信息检索效果的因素有哪些?

CHAPTER TWO

第2章
检索语言

【本章要点】

◇介绍检索语言的概念、功能及类型
◇总结检索语言的主要理论基础
◇阐述分类检索语言的结构与性能
◇分析主题检索语言的原理及性能
◇论述分类主题一体化检索语言的原理、性能及类型
◇探讨网络环境下检索语言的发展

引子

目录、分类、主题词、Index、Thesaurus、Search……这些曾经属于图书情报领域的专用术语，如今已是网络信息世界里任何学科专业人士皆耳熟能详的通用语言。情报检索语言是科学交流中人类自然语言交流与人机交互均能达到共同理解的基础。它凝集了不可胜数的图书馆学家、情报学家对浩瀚文献整序、描述、揭示和传播服务研究的智慧结晶。近年来，信息化浪潮席卷各行各业，情报检索语言顺应时势，如舟，如桨，自然地成为数字化环境里搏风击浪的有效工具。它们在越来越多的领域展现出非常广阔的应用前景。①

① 参见方平、柳晓春：《从图书情报领域迈向信息世界——情报检索语言应用的广阔前景》，载《图书馆》，2004（6）。

2.1 检索语言概述

2.1.1 检索语言的概念

传统意义上的检索语言仅指根据信息检索需要而创制的人工语言，其实质是用于表达一系列概括文献信息内容的概念及其相互关系的概念标识系统。它可以是从自然语言中精选出来并加以规范化的一套词汇，可以是代表某种分类体系的一套分类号码，也可以是代表某一类事物的某一方面特征的一套代码（如化合物的各种代码），用于对文献内容进行主题标引、特征描述或逻辑分类。检索语言又称情报语言、情报存储与检索语言、文献语言、文献工作语言、索引语言、标引语言、标引符号、标识系统等。世界上有多种检索语言，如《杜威十进分类法》、《国际十进分类法》、《冒号分类法》、《中国图书馆分类法》、《中国人民大学图书馆图书分类法》、《汉语主题词表》等。

检索语言由词汇和语法组成。词汇是登录在类表、词表中的全部标识，一个标识（分类号、检索词、代码）就是它的语词，而分类表、词表则是它的词典；语法是指如何创造和运用那些标识（单个标识或几个标识的组合）来正确表达信息内容和信息需要，以有效地实现信息检索的一整套规则。①

随着计算机技术的发展和人们认识的深入，检索语言的概念也发生了变化。目前，检索语言有广义和狭义之分，广义的检索语言泛指信息检索过程中涉及的人工语言和自然语言。人工语言是根据一定的规则人为编制而成的检索语言，它有着严格的使用规则，可用于表述文献主要内容，建立信息检索系统。自然语言是人类交流时使用的语言，不受任何限制，未经加工和规范。将自然语言用于检索，更符合用户日常表达的习惯，也显现出信息检索系统的易用性和亲和力。狭义的检索语言仅指根据信息检索的需要，按照一定的规则对自然语言进行规范，并专门用于信息标引和用户检索的人工语言。

2.1.2 检索语言的功能

广义的信息检索包含信息的存储与检索两方面。在这两个相对应的过程中，信息存储人员和信息检索人员需要遵循一种能共同理解的语言，以保证信息存之

---

① 参见张琪玉：《张琪玉情报语言学文集》，23页，北京，北京图书馆出版社，1999。

有规则，取之有途径。具有这种功能的语言就是检索语言，它在信息检索过程中发挥着重要的作用。

2.1.2.1 标引信息内容特征及某些外表特征，保证不同标引人员表达信息的一致性

检索语言是标引人员对信息内容特征以及部分外表特征进行描述的重要依据，信息标引人员在分析信息的基础上，用检索语言将文献的内容特征和外表特征表述出来，形成信息标识，比如分类号、主题词等，然后将标引记录存放在系统中，以供用户检索使用。信息标引是一个群体行为，只有共同依据检索语言，才能保证标引信息的一致性。

2.1.2.2 对内容相同及相关的文献信息加以集中或揭示其相关性

检索语言采用等级结构、参照系统、轮排聚类法、范畴聚类法等显示概念之间关系的方法，来实现对内容相同及相关的信息加以集中或揭示其相关性的功能。等级结构是显示概念之间关系的一种最重要的方法，它将各种概念按相关性排列成一个具有隶属关系或并列关系的秩序井然的概念等级体系，包括体系分类表、分面类表、词族索引等；参照系统是主题法系统各种语言显示概念之间关系的主要方法，其功能是将具有相关性但因为按照字顺排列而被分散在各处的概念联系起来，参照系统可以显示事物概念之间的全部等同关系、一部分等级关系（主要属种关系）和全部相关关系；轮排聚类法是将表达复杂概念或多因素主题的标识，按它们所表达出来的每个有检索意义的概念因素或主题因素进行轮排，当某一概念因素或主题因素轮排到检索入口位置时，就能使具有同一概念因素或主题因素的概念或主题的标识排到一起，从而起到聚类作用，显露出概念之间的相关性；范畴聚类法可以表明同一范畴的检索词都属于某一学科或专业范围。

2.1.2.3 使信息的存储集中化、系统化、组织化，便于检索人员按照一定的排列次序进行有序化检索

检索语言将表达成千上万个信息主题概念的全部信息标识排列成一个有序的系统。排列信息标识的方法主要有三种：分类排列法，用于号码标识系统；字顺排列法，用于语词标识系统和代码标识系统；分类和字顺结合的排列法，即先按照分类排，再按字顺排，用于语词标识系统（如分类主题目录等）。

2.1.2.4 便于将标引用语和检索用语进行相符性比较

一般来说，任何一种检索语言都有便于将标引用语和检索用语从整体上进行相符性比较（即判断标引用语是否与检索用语完全相符）的功能。大部分检索语

言还可以将标引用语和检索用语从局部上进行相符性比较（即判断标引用语是否与检索用语部分相符）。

2.1.3 检索语言的分类

依据不同的标准，检索语言可以有不同的分类结果，具体如下：

2.1.3.1 按描述文献的特征，可以分为描述文献外表特征的检索语言和描述文献内容特征的检索语言

1. 描述文献外表特征的检索语言

文献外表特征主要指文献的篇名（题目）、作者姓名、出版者、合同号、报告号、引文等，据此作为文献标识和检索依据而形成的检索语言称为描述文献外表特征的检索语言，如题名索引、著者索引、合同号索引、报告号索引、引文索引等，如图 2—1。

- 题名——题名索引
- 著者——著者索引、团体著者索引
- 文献编号
  - 报告号索引
  - 合同号索引
  - 存取号索引
- 其他——引文索引

图 2—1 描述文献外表特征的检索语言

2. 描述文献内容特征的检索语言

描述文献内容特征的语言指主要依据文献内容特征而形成的检索语言，这是检索语言研究的核心部分，具体有分类语言、主题语言和代码语言，如图 2—2。

- 分类语言——体系分类语言、组配分类语言
- 主题语言——标题词语言、关键词语言、单元词语言、叙词语言
- 代码语言——分子式索引、结构式索引、专利号索引等

图 2—2 描述文献内容特征的检索语言

2.1.3.2 按结构或原理，可分为分类语言、主题语言、代码语言和引文语言

分类语言用分类号来表达各种概念，将各种概念按学科性质进行分类和系统编排。分类语言包括等级体系分类语言（体系分类法）和分析—综合分类语言（组配分类法）。它们统称为分类法系统。

主题语言是采用表达某一事物或概念的名词术语，用于标引、存储、检索

的一种检索语言。它包括标题词语言（标题法）、单元词语言（单元词法）、叙词语言（叙词法）和关键词语言（关键词法）等。这些可统称为主题法系统。

代码语言一般只是就事物的某一方面特征，用某种代码系统来加以标引和排列。例如，化合物的分子式索引系统、环状化合物的环系索引系统、有机化合物的威斯韦塞尔线型标注法代码系统等。

引文语言是利用文献之间的相互引证关系而建立的一种自然语言，其标引词来自文献的主要著录项目。它具有选词方便、词汇丰富等特点。这种方法提供了从被引论文来检索引用它的全部论文的途径，从而顺着一种科学思想的发展过程线索找到有关信息。可以将引文语言看作是检索语言的一种特殊类型。

2.1.3.3 按信息标识的组合使用方法，可分为先组式语言、后组式语言和散组式语言

先组式语言指表达信息主题概念的标识在编制词表时就已固定组配好，信息存储和检索人员在标引和查找信息时，依据词表选用组配好的主题标识进行操作，典型的先组式语言有体系分类法和标题词法。先组式语言标识明确，系统性较好，适用于传统的文献单元方式的目录索引，不足之处在于表达专指概念和新概念较困难，灵活性较差，容易造成词表体积庞大、检索途径少。

后组式语言指在词表编制和标引信息时不规定表达主题标识的组配关系，在检索时再根据需要将各个标识进行组配，来表达较为复杂的主题概念。后组式语言能够以较少的语词来表达较多的概念和较专指概念，可以比较及时地表达新概念，概念容纳能力强。同时，后组式语言可以实现概念的多向成族，多途径检索，能够灵活地扩大和缩小检索范围，因此其检索的灵活性好。采用后组式语言所编制的词表体积相对较小。后组式语言的缺点是编制和使用的难度较大，组配语义的明确性较差。比较典型的后组式语言有叙词语言和单元词语言。

散组式语言是指对于复杂的主题标识，在词表中不组配，而是在标引阶段将表达主题概念的若干个标识，根据一定的规则组配在一起的检索语言。散组式语言的性能与其检索标识是否分段轮排有很大关系。如果检索标识能够分段轮排，散组式语言就兼有先组式语言和后组式语言的主要优点，如果检索标识不轮排，性能和先组式语言差不多，只是其专指概念和新概念表达能力有所增强。《冒号分类法》就属于散组式语言，有些标引语言本来是属于后组式语言，但可以作为散组式语言使用，例如叙词语言本来是后组式语言，当采用叙词表在标引阶段将复杂概念的各个标识组配在一起时，叙词语言就可以当作散组式

语言使用。

2.1.3.4 按语言的规范程度，可分为人工语言和自然语言

人工语言有主题法（标题词、单元词、叙词、关键词）、分类法（体系分类法、组配分类法）和某些代码（语义代码、化学代码等）等种类；自然语言是直接取自文献信息本身，不经加工或规范的词语或句子。

此外，检索语言还有一些其他的分类方法。按包括的学科或专业范围，可分为综合性检索语言、专业性检索语言和多学科检索语言；按所用语言文字，可分为单语种检索语言和多语种检索语言。

2.2 检索语言的理论基础

检索语言在创制、发展完善和实践应用过程中，吸收和借鉴了多门相关学科的方法和成果，其主要理论基础有概念逻辑、知识分类和术语学。

2.2.1 概念逻辑

检索语言不论是语词的还是符号的，都是表达一系列概括信息内容的概念及其相互关系的概念标识系统。因此，它们都建立在概念逻辑的基础上。

概念逻辑，是一种科学思维方法，通过明确各种概念及其相互关系而揭示事物的本质属性及各种事物之间的联系与区别。检索语言在建立词汇、显示概念关系以及在文献标引和检索中，都离不开概念逻辑。

概念是事物本质属性的概括。任何概念都有其内涵与外延。某一概念的内涵，即指其所反映事物的本质属性；外延则是指其反映事物的范围。它们之间存在着反变关系，即概念的内涵包含的属性增加，概念的外延便缩小；反之，概念的内涵减少，其外延就扩大。人们给概念下定义，最常用的方法是种差加属的方法，即“被定义的概念（种概念）＝种差＋邻近属概念”。“种差”是指同一属概念下的种概念所独有的属性（即和其他属概念的本质的差别），“邻近属概念”是指包含被定义者的最小的属概念。

2.2.1.1 概念间的关系

概念之间的关系，按其外延是否有相同部分，可归结为相容关系与不相容关系。如果两个概念的外延至少有一部分相重合，则两者之间是相容关系；如果两个概念的外延没有相重合的部分，则两者之间是不相容关系。

1. 相容关系

在相容关系中可以分为同一关系、属种关系、交叉关系、整体与部分关系、全面与某一方面关系、不相排斥的并列关系。

（1）同一关系是指外延相同而内涵不同的概念之间的关系。具有同一关系的概念之间绝大多数是同义词、学名与俗名、同一产品的正式命名与简写等等，例如“计算机”与“电脑”、“乙醇”与“酒精”等即为同一关系。具有同一关系的概念，所指的是同一事物，在检索语言中需要将具有同一关系的概念进行合并，合并后只用一个标识，从而可以避免同一主题的文献被分散标引在多个标识下而造成漏检。

（2）属种关系是指概念之间的外延呈包含与被包含的关系。概念的属种关系具有相对性，包含了另一概念的概念对被它包含的概念来说是属概念；被包含的概念对包含的概念来说是种概念。例如，“自然科学”是“化学”的属概念，“化学”是“自然科学”的种概念。概念的属种关系在检索语言中形成概念的等级关系。

（3）交叉关系是指有部分外延相重合时概念间的关系。具有交叉关系的概念称为交叉概念，例如，“团员”和“党员”。两个交叉概念外延重合的部分一般会形成一个新概念，其内涵等于两个交叉概念内涵之和，这个新概念对原来任何两个概念中的一个来说，都是下位概念。

（4）整体与部分关系是指一个概念表示某一事物，而另一个概念表示该事物的某一部分。例如，“汽车”与“汽车发动机”这两个概念的关系就是整体与部分关系。整体与部分关系是概念隶属关系的另一种形式，整体概念是上位概念，部分概念是下位概念。

（5）全面与某一方面关系是指一个概念表示某一事物的全部问题，而另一概念表示该事物的某一方面的问题。全面与某一方面关系也是概念隶属关系的一种形式，全面概念是上位概念，某一方面概念是下位概念。

（6）不相排斥的并列关系是指同一属概念下，两个以上同级种概念之间的交叉关系。例如“文学家”、“军事家”、“书法家”等这些并列概念的关系，这些概念既具有独立的外延而与其他概念并列，也可以具有其他概念所具有的外延而与其他概念交叉。

2. 不相容关系

在不相容关系中包括互相排斥的并列关系、矛盾关系和对立关系。

（1）互相排斥的并列关系是指同一属概念下，两个以上外延完全不同的种概念之间的关系。例如，“公共图书馆”这个属概念下，“省图书馆”、“市图书馆”、

“区图书馆”之间具有不相容的并列关系。

(2) 矛盾关系是指外延完全不同，其外延总和等于其上位概念全部外延的概念之间的关系。例如，“金属材料”和“非金属材料”这两个概念之间的关系就是矛盾关系，它们互称为矛盾概念，它们的外延之和等于“材料”这个上位概念的全部外延。

(3) 对立关系是指外延完全不同，其外延总和小于其上位概念全部外延的概念之间的关系。例如，“17世纪哲学”和“18世纪哲学”这两个概念之间就是对立关系，它们的外延之和小于“近代哲学”这个上位概念的外延。

2.2.1.2 概念逻辑方法

检索语言在表达各种概念及其相互关系时，普遍地应用了概念逻辑的原理，有效地利用了“概念的划分与概括”和“概念的分析与综合”这两种逻辑方法来建立自己的结构体系。

1. 概念的划分与概括（分类）

即利用概念内涵由反映事物本质属性的概念因素构成，概念因素的增加或减少可以形成新的概念，概念内涵与外延成反变关系等性质，对概念进行划分（缩小）或概括（扩大），形成更为专指或更为泛指的新概念，用以区别客观世界千差万别的事物，并利用划分或概括过程中所产生的概念隶属关系和并列关系，建立某种形式的检索语言结构体系，即概念等级体系，用以显示客观世界千差万别的事物之间的内在联系。这种结构具有很好的系统性。例如，体系分类法就是应用此种逻辑方法的典型。

2. 概念的分析与综合（组配）

即利用在概念的交叉关系中两个概念外延的相同部分可以形成一个新概念，其内涵等于原来两个概念内涵之和，并且它与原来的两个概念具有隶属关系的这种性质，进一步发展为将一个内涵较深的概念分解为两个或两个以上内涵较浅的概念，以及将两个或两个以上内涵较浅的概念合成为一个内涵较深的概念的一种概念逻辑方法，用以建立另一些形式的检索语言结构体系，即概念组配体系。这种结构体系可以提供从多种途径来进行信息检索的功能，而且可以任意选择检索标识的专指度，根据实际需要扩大、缩小或改变检索的范围。例如，叙词语言与组配分类法便是应用概念分析与综合的典型。

2.2.2 知识分类

知识分类是对千差万别的事物做系统研究的重要方法，是对各种事物之间的区别和联系从本质上、原理上进行揭示的重要手段，对信息的系统化具有重要的

价值，其实质是划分知识单元、组织知识体系，包括学科分类和事物分类。学科分类是以信息的学科属性为分类标准，根据各门学科的研究对象的区别和联系，对学科进行区分和组织，确定每门学科在科学整体中的位置，揭示科学的内部结构，建立符合科学发展规律的分类体系。事物分类是根据事物属性的异同将事物划分成类，构成事物分类体系。学科分类是知识分类的主体，事物分类是知识分类的基础。

知识分类应当遵循的两条基本原则是客观性和发展性。客观性原则即对学科或事物进行划分和组织时，应依据知识对象固有的、客观存在的区别和联系。发展性原则是指知识分类应按照学科或事物的发展规律，将知识对象由低级到高级、由简单到复杂进行划分和组织。

检索语言要能适合实践应用和科学研究的需要，除了依据概念逻辑的理论基础外，还必须利用知识分类的成果，把各种概念之间的关系建立在知识分类的基础上。如果说概念逻辑是检索语言的基础，则知识分类便是概念逻辑的基础。检索语言中的体系分类法、组配分类法、叙词法等，都在不同程度上反映了知识分类，例如，分类语言以知识分类为基础建立类目体系，主题语言根据知识分类建立参照系统，编制范畴索引、词族索引等。知识分类体系具有多维性并处在不断变化之中，检索语言反映当代知识分类的程度是其质量的重要标志之一。只有较充分反映知识分类的检索语言，才能保证检索语言在信息检索工作中得到有效运用，并达到较好的检索效果。

2.2.3 术语学

术语是在特定学科领域用来表示概念的称谓的集合，或者说，是通过语音或文字来表达或限定科学概念的约定性语言符号。术语是传播知识、技能，进行社会文化、经济交流等不可缺少的重要工具。研究概念、概念定义和概念命名基本规律的学科即术语学。

检索语言是由概念标识系统组成的，而概念是由术语来表达的，因此，术语是分类表、词表的基本组成要素，检索语言其实就是一个经过精细组织的术语集。检索语言的创制是以术语学的研究成果为基础的。在编制分类表、词表而采用术语时，应当遵循术语学对术语的订名，对术语内涵、外延及相互关系的规定和解释，即以术语学对科学术语的研究成果为依据。在检索语言创制、发展及完善过程中，更多地吸收术语学的研究成果，是提高检索语言质量的要求和重要途径。

2.3 分类检索语言

2.3.1 分类检索语言概述

分类检索语言也称分类法，是将许多类目根据一定的原则组织起来，通过标记符号（分类号）来代表各级类目和固定其先后次序的分类体系。它是直接体现知识分类的概念标识系统，是对概括文献信息内容及某些外表特征的概念运用逻辑方法进行系统排列而构成的。分类法按学科、专业集中文献信息，并从知识分类角度揭示各类文献信息在内容上的区别和联系，提供从知识分类检索文献信息的途径。

分类检索语言主要包括体系分类法和组配分类法两种。

2.3.1.1 体系分类法

信息检索中常用的分类语言是体系分类法，也称等级列举式分类法。体系分类法是基于概念的划分与概括，以学科分类为基础，把概括文献内容与事物的各种类目组成一个层层隶属、详细列举的等级结构体系。具有代表性的体系分类法有《杜威十进分类法》、《美国国会图书馆分类法》、《中国图书馆分类法》、《中国人民大学图书馆图书分类法》等。

2.3.1.2 组配分类法

组配分类法又称分面分类法或分面组配分类法，是基于概念的可分析性和可综合性，即复杂的主题概念可以分析为若干简单的主题概念，若干简单的主题概念可以综合表达复杂的主题概念，将一个复杂的文献主题概念用若干个表达简单概念的标识组配来表达的一种文献分类法。在组配分类法中，一般只将简单主题概念设为类目，标引时，从类目中找出与主题概念相对应的各个类目，用相应的类号组配表达所标引的主题概念。

组配分类法可分为全分面分类法和半分面分类法两种。全分面分类法是纯粹的组配分类法，一般仅在较小的学科或专业范围内使用。半分面分类法是全分面分类法与体系分类法相结合的混合式分类法，如《冒号分类法》、《布利斯书目分类法》等，这种分类法一般是综合性或多学科的分类法。

2.3.2 体系分类法

2.3.2.1 体系分类法的结构

体系分类法的结构分为微观结构和宏观结构。

1. 微观结构

微观结构指分类法中类目的构成结构。类目是表达文献信息内容或主题范围的概念，是构成分类法的细胞。一部分类法由成千上万个类目组成。表示类目概念的名称称为类名。类名规定了类目的含义和内容范围，它可以由单词或词组构成，用来表达学科、专业、事物对象及其组成部分。如，反映学科及其分支概念的“生物学”、“植物学”、“植物形态学”等；反映事物及其种类、构成部分概念的各种物质、产品、民族、语言、机构、人物、事件等。

按照类目之间的关系建立起来的类目集合称为类目体系。类目体系是分类法的核心，它的建立涉及类目的划分、引用次序、类目的排列、类名的确定、类目之间相互关系的处理等。

(1) 类目的划分。

类目划分是把一个类目分为若干个小类，从而揭示这个类目外延的逻辑方法。被分的类目称为母类或上位类，分出来的类目称为子类或下位类。如“教育”可分为“高等教育”、“中等教育”、“初等教育”等。通过类目划分可以明确母类的全部外延，建立若干子类，给设立类目打下基础。

类目的划分是按照一定的分类标准进行的，分类标准指用来作为划分依据的事物属性。类目划分的标准大体上可分为内容标准和形式标准两种。内容标准包括论述的对象、范围、所属学科、涉及的地区或国别、时代、民族、人物事物性质、工艺流程、组织结构、思想倾向、使用的工具、材料、目的及效果、物质运动的形态、社会实践的职能分工等等；形式标准包括编写体裁、语言文字、出版形式、装帧、文献类型等等。在现代分类法中，类目划分一般以内容标准为主要标准，以形式标准作为辅助标准。

类目划分的过程通常是由大到小、由属到种、由整体到部分、由总论到各论、由全面到各方面。这样，通过层层划分，就可以建立起一个逐层展开的分类体系。例如，《中国图书馆分类法》的化学类的划分即是由大到小来划分的：“化学”分为“无机化学”、“有机化学”、“高分子化学（高聚物）”、“物理化学（理论化学）”、“化学物理学”、“分析化学”、“应用化学”等。再如“禾谷类作物”下分为“稻”、“麦”、“玉米（玉蜀黍）”、“高粱”、“粟（谷子、稷）”、“黍（糜子）”、“荞麦”等，是根据由属到种过程来划分的。

(2) 引用次序。

引用次序在体系分类法中表现为分类标准的使用次序，当某一类事物连续划分需要采用几种分类标准时，分类标准的使用次序是否合理，直接影响分类体系，决定着类目体系展开方式。因为首先被采用的分类标准，将能使在该属性上

相同的信息或文献得到最大程度的集中，依据该属性检索信息或文献也最容易，而在后面采用的分类标准都将会不同程度地分散在该属性上相同的信息或文献，并且越在后面采用的分类标准，在该属性上相同的信息或文献就会越容易被分散，其被检索的难度也就会越大。

合理的引用次序应该满足逻辑性原则、符合检索需要的原则和表达性原则。我国著名的检索语言专家张琪玉也曾指出，合理的分类标准使用次序是指：分类体系的层次结构基本符合文献主题概念的层次结构，或者说，类目的层层划分、展开是符合逻辑的；对文献的集中与分散的处理符合读者的检索要求。

在体系分类法中，至今没有一部分类法提出或确立过一种统一的一般引用次序，即使是在同一部体系分类法中，有时也存在着分类标准引用次序的不同。例如，在《中国图书馆分类法》中，法律类就出现了两种引用次序：法律—部门—国家、法律—国家—部门。但是在编制分类表时，应力求做到：优先使用该学科、该事物的主要分类标准；优先使用具有科学认识意义的分类标准；优先使用具有较高检索意义的属性作为分类标准。

（3）类目的排列。

分类法是由许许多多的类目按照一定的顺序排列起来的体系，类目的排列直接影响到分类法的质量。类目的排列应体现出系统性、整体性、等级性、逻辑性、连续性和一致性。尤其是在体系分类法中，类目的排列应坚持相关排列准则，其中，同位类的排列次序反映了客观事物本身发展和联系的排列次序，具有重要的意义。

所谓同位类是指由一个上位类直接划分出来的各个下位类，它们之间不相从属，相互排斥，处于同等地位。同位类的排列是否得当，是否符合规律，会影响到整部分类法的质量。在目前的国内外分类法中，同位类的排列主要采用的序列方法，有按照逻辑顺序排列、按照客观事物发展的顺序排列、按照时间顺序排列、按照空间顺序排列等。此外，还可以依据依存次序、惯用次序、实用次序和字顺次序等其他顺序进行同位类的排列。

（4）类名的确定。

类名是体系分类法的“语词”，起着表达类目含义的作用。类名的选择和确定应坚持科学性、确切性、简洁性。

科学性指采用比较通行的科学名词术语作为类名，一般不采用不通行的同义词、俗称、旧称、不能准确表达全称原义的简称、不通行的译名、近义词等作为类名。如有必要，可将同义词、俗称、旧称等用括号加注于类目名称之后。例如，《中国图书馆分类法》中的“B81 逻辑学（论理学）”、“B82 伦理学（道德

学)”等。

确切性指类名要能准确地、恰当地反映类目的实际内容范围。不能使用概念外延大于或小于类目实际范围的词或词组作类目名称。

简洁性指所用的类名要尽量概括、精练、简短，避免冗长拖沓。同时类名还要做到规范化。

(5) 类目之间相互关系的处理。

分类法是依赖于类目之间的相互关系建立起来的。在体系分类法中，类目之间的基本关系主要有从属关系、并列关系、交替关系和相关关系。

第一，从属关系。

从属关系又称隶属关系，指类目体系中母类与其子类的关系，体现为上位类和下位类的关系，这种关系构成了分类法类目的纵向排列。在分类体系中，上、下位类是相对的。

从属关系包括属种关系、整部关系（整体与部分关系）、方面关系（全面与某一方面关系）。属种关系是类目隶属关系的基本形式，是指同族事物的属与种、类称与特称之间的关系。如“文学作品”是“小说”的上位类，即属概念；“小说”是“文学作品”的下位类，是种概念。整部关系和方面关系的上下位类是一种限定关系。整部关系是指两个不同族的事物一个成为另一个的构成部分的关系，如“生物系统”与“器官”等；方面关系指该事物及其有关的各个内容方面的关系，如“马”和它的下位类“生理、解剖”就属于方面关系。

第二，并列关系。

并列关系又称同位关系，在分类体系中体现为同位类的关系。例如：

F 经济
- F0 政治经济学
- F1 世界各国经济概况、经济史和经济地理
- F2 经济计划与管理
- F3 农业经济
- F4 工业经济
- F5 交通运输经济
- F6 邮电经济
- F7 贸易经济
- F8 财政、金融

第三，交替关系。

交替关系指使用类目和交替类目之间的关系。有的学科或事物分属两个门类，编制分类法时确定归入一个门类，同时在另一个门类设交替，以适应学科的交叉关系，集中同一主题的相关文献。在交替关系的处理中，交替类目不用来类分文献，只起到指向使用类目的作用。例如：

[B035] 国家理论

宜入 D03（“政治理论”下的“国家理论”类）

第四，相关关系。

有些类目之间存在着密切的关系，而这些类目又不属于一个类系，这种类目之间的关系称为相关关系。如“中国共产党”和“中国近代史”。有着相关关系的类目，称为相关类目。相关类目也称参见类目或参照类目。例如：

O212 数理统计

参见 C8（统计学）

2. 宏观结构

按功能分，体系分类法的宏观结构一般由以下四部分组成：类目体系、标记系统、说明与注释、类目索引。

(1) 类目体系。

类目体系是按照类目之间关系建立起来的类目集合。大多数体系分类法的类目体系由主表和复分表组成。

主表是指由基本部类、基本大类、简表、详表逐级展开而形成的类目表。

基本部类是为了便于各种类目的展开而对人类全部知识与事物所做的最基本、最概括的划分，但它不是分类表的类目。基本部类的排列次序称为基本序列。第四版的《中国图书馆分类法》基本部类的划分采用“五分法”，其基本序列为：

马克思主义、列宁主义、毛泽东思想、邓小平理论
哲学
社会科学
自然科学
综合性图书

基本大类是在基本部类的基础上，根据学科发展和文献出版情况所列出的第一级类目，代表着较大的学科或领域，能够使人们对分类表的分类体系有个基本的了解。

简表是由基本大类直接展开的一、二级类目所形成的一种类目表，又称为基本类目表。简表的主要作用有：对基本大类与详表起承上启下的作用，便于用户查找所需要的详细类目；供中小型图书馆或资料室，或者只需要对信息进行粗略标引时分类标引使用。

详表是由简表展开的各种不同等级的类目所组成的类目表，是分类表的正文，也是分类标引的实际依据。详表又称为分类表的主表。

复分表又称为辅助表或者附表，是将详表中按相同标准划分某些类所产生的一系列相同子目抽出来，配以特定号码，单独编列，供主表有关类目进一步细分用的类目表。复分表有通用复分表和专用复分表两种。通用复分表是附在主表之后，供整个分类表有关类目作进一步区分用的表。例如，第四版的《中国图书馆分类法》有八个通用复分表：总论表、世界地区表、中国地区表、国际时代表、中国时代表、世界种族和民族表、中国民族表、通用时间地点表。专用复分表插在主表中的相关位置，供分类表中的某些类目做进一步区分。例如，第四版的《中国图书馆分类法》有 67 个专用复分表，同时在总论表和中国地区表中还各有一个复分表。复分表的主要用途一是缩小类表篇幅、简化分类表；二是增强主表中有关类目的细分程度，并且规范同性质类目的细分。

（2）标记系统。

标记系统是分类语言所有标记符号的集合。分类语言的标记符号即分类号，是用于标记某一分类体系各类目的序数系统。现代图书分类法都使用分类号作标记，一方面是作为类目的代号，固定类目的先后次序，便于标引和组织文献；另一方面可以显示类目之间的相互关系，便于作相符性比较。理想的分类标记应符合简短性、容纳性、灵活性、表达性、助记性和适应性等要求。

分类号有两种，一种是单纯式号码，主要使用的是纯数字标记，《中国人民大学图书馆图书分类法》就采用了纯数字的标记符号作为分类号，例如，“11.”表示 11 大类“历史”；另一种是混合式号码，通常是字母和数字结合使用，《中国图书馆分类法》使用的是混合式号码，如“G35”表示三级类目“情报学”。

标记制度是指由标记符号构成分类标记的基本方法，可分为顺序制、层累制、混合制和分面标记制 4 种。

第一，顺序制。

顺序制指在分类体系确定后，对全部类目不分等级给予顺序号码的编号方法。顺序制具有号码简短明了、便于排检、容纳性强的优点，但它不表达类目之间的关系，所以表达性和助记性差。《美国国会图书馆分类法》就采用了这种标记制度。

第二，层累制。

层累制是一种分类号位数与类目等级相对应的标记制度。一般是用一位数字或一个字母表示一个大类，再加一位数字或一个字母表示下一级类目，如此层层累加。层累制具有较强的表达性和助记性，便于扩检和缩检，有利于文献排架和目录组织。但号码的简短性比顺序制差，分类愈细，类号愈长。在实际使用过程中，绝对遵守层累制原则的分类标记系统是很少的。我国的《中国图书馆分类法》基本上采用了层累制。

第三，混合制。

混合制是一种将顺序制和层累制结合起来的标记制度，一部分用顺序制，一部分用层累制。例如我国的《中国科学院图书馆图书分类法》。

第四，分面标记制。

分面标记制是一种显示类目组配结构的标记制度。根据科学发展的规律，印度图书馆学家阮冈纳赞创立了分面标记法。在他所设计的基本分面公式中，不同的主题方面都有相应的标识符和固定的位置，它们共同组配成一个完整的主题类号。常见的分面标记采用分段组合方式，如果各个节段规定相应的辅助标记，则各个节段可以轮排，以提供更多的检索途径，便于在每段扩充。分面标记制具有较强的表达性和灵活性，但其号码成分复杂，冗长难记，标记的排序能力差。

此外，为了更好地增强分类法标记符号的容纳性、表达性和简明性，分类法在编制过程中，还采用了一些特殊的标记方法：

第一，八分法。

指用 1～8 来标记同位类，当同位类超过 8 个时，不用 9，而是用 91、92、93……98 来标记，即 91 表示第 9 个同位类，92 表示第 10 个同位类，93 表示第 11 个同位类……98 表示第 16 个同位类。以此类推，同位数超过 16 个时，不用 99，而是用 991、992、993……998 来标记。八分法是解决同位类超过 10 个限度时的一种标记方法。

第二，双位法。

体系分类法中，如果同位类超过 18 个，在类目展开时，可以直接采用两位数表示一次划分，这种方法称为双位法，其主要目的是解决号码的扩充问题。

第三，借号法。

这是一种灵活借用上位类或下位类号码的配号方法。当同位类个数超过 9 个，而且只是多出 1 个或 2 个时，可以借用其中某个同位类 1 个或 2 个多余的下位类号，这些类号常是 9 或 8。

第四，组配法。

将两个表示简单概念的类号用组配符号组合成一个复合类号，用来表达分类表中没有列出的复杂概念。

第五，空号法。

空号法是为了适应新类目的增加而采用的一种预留一定数量空号码的编号方法。不过需要指出的是，分类表中出现的空号有时并不一定是预先留着，以备新类目的增加，还有可能是原有类目被删除或者调整而出现空号。

（3）说明与注释。

说明与注释是对分类表结构及使用方法的揭示，用它来进一步阐述分类法的编制原理、特点和使用方法，明确类目之间的关系，确定类目的性质和范围，确定类分图书时的方法等。分类法的说明与注释主要包括编制说明、大类说明和类目注释 3 种形式。

编制说明主要介绍分类法的编制原则、编制过程、类目设置及相关技术处理方法。

大类说明主要介绍基本大类的结构特点和标引规则。分类法编制说明和大类说明对利用分类法具有很大的帮助。

类目注释是对类目的性质或类名的补充说明文字。它是分类法增设新学科、新事物、新理论、新技术等新主题概念的一种重要方法，是分类法增强主题法因素的一种重要手段，也是衔接分类法新旧版之间变化的一种有效方法。类目注释不仅是一部分类法的重要组成部分，也是分类标引人员判断类目涵义、明确类目之间关系的重要依据，正确理解和使用类目注释有助于提高标引质量。类目注释的主要类型包括：定义注释、同义词注释、列类依据注释、内容范围注释、类目关系注释、标引方法注释、增词注释、修订注释等。例如：

D631.42 户籍管理

流动人口管理入此。

参见 C921.3

〈3 版为交替类目，宜入 C921.3〉

（4）类目索引。

分类表一般体积较为庞大，内容复杂，对其不太熟悉的使用者要想准确而快速地查找到相应的类目一般比较困难，这就需要借助于类目索引。类目索引是从类目名称字顺查找相应分类号的类表辅助工具，是分类表的重要组成部分。类目索引的主要作用是帮助不熟悉分类表的使用者从主题名称迅速找到相应类目。另外，类目索引还能集中分类表中被分散的有关同一事物不同方面的类目，弥补分

类表依学科集中，却将同一事物的不同方面分散的不足。值得注意的是，类目索引不能直接用来分类标引，而只能充当辅助工具，从类目索引查得的分类号必须与分类表的类目进行对比，以确定是否准确。

类目索引可分为直接索引、相关索引和主题分类对照索引。直接索引是将分类表中类目及其注释中的有关主题的概念，按其名称字顺排列，表明相应分类号的索引。直接索引的编制方法比较简单，但是难以集中相关事项，也难以反映复杂的专指主题。相关索引是指除了将分类表中的类目和注释中具有检索意义的主题概念按字顺排列外，还集中反映在分类表中被分散了的相关事项。相关索引将一个主题的各个方面以及被这个主题所规定的词，都集中在一个主题标目下，因此，其检索性能比直接索引大。主题分类对照索引是指在主题词表（主要是指叙词表）的主题词后列出对应的分类法的类号，以便从主题词字顺查找相应分类号的索引。主题分类对照索引是沟通分类法和主题词表的桥梁，既是体系分类法的辅助工具，也是分类法和主题词相互转换的工具。

2.3.2.2 体系分类法的特点

体系分类法在实际工作中，主要被用来组织分类排架、统计藏书和建立分类检索系统。

体系分类法的主要特点是：

（1）按学科、专业属性构建类目体系，形成按学科、专业集中文献、信息的知识概念系统，从而能够直接地满足用户从学科、专业出发检索课题的需求，可以达到较高的查全率；

（2）采用等级列举式的概念标识系统来揭示概念之间的相互关系，便于用户“鸟瞰全貌”、“触类旁通”、“层层深入”地查找某一专业的信息，用户也无须事先知道事物或概念的确切名称，就可以在一定的类目下通过浏览查到该领域的相关信息；

（3）采用分类号作为主题的标识，不受语种的限制。

体系分类法的不足之处在于：

（1）修订不便，无法及时增加反映新知识主题的类目。同时，分类表也不可能永无止境地细分下去，如遇到主题十分狭窄的文献，则可能很难找到相应的类目及分类号；

（2）体系分类法采用的是先组定组式标识，难以进行组配检索，使得其检索途径单一，检索效率不高；

（3）采用分类号作为主题的标识，缺乏直观性；

（4）体系分类法是按学科、专业集中文献信息以及线性的分类体系，使得其

处理学科之间相互交叉渗透和综合而形成的新知识领域很困难，难以反映客观实际中多维的知识空间结构。

2.3.2.3 主要体系分类法介绍

目前，国内常见的体系分类法有《中国人民大学图书馆图书分类法》，简称《人大法》，初版于1953年；《中国图书馆分类法》，简称《中图法》，初版于1975年，名为《中国图书馆图书分类法》，至今已出到第4版（1999年），并更名为《中国图书馆分类法》；《中国科学院图书馆图书分类法》，简称《科图法》，1958年由中国科学院图书馆编写，1974年、1979年、1994年分别进行了修订；《中国档案分类法》，初版于1987年，1997年推出第二版。

国外常见的体系分类法有《杜威十进分类法》（Dewey Decimal Classification），简称DC或DDC，初版于1876年；《美国国会图书馆分类法》（Library of Congress Classification），简称LC，初版于1961年；《国际十进制分类法》（Universal Decimal Classification），简称UDC，初版于1905年，1960年出版中文版，所列类目超过21万个。

这里主要介绍《中国图书馆分类法》和《杜威十进分类法》。

1.《中国图书馆分类法》

我国目前广泛使用的分类法是《中国图书馆分类法》。它是由国家图书馆等单位组织全国力量，以学科分类为基础，并结合图书的特性所编制的分类法。它将学科分成五大部类，基本序列是：马克思主义、列宁主义、毛泽东思想、邓小平理论，哲学，社会科学，自然科学，综合性图书，由5大部类、22个大类、6个总论复分表、30多个专类复分表、4万余条类目组成了一个完善的分类体系（见表2—1）。

标记制度采用拉丁字母与阿拉伯数字相结合的混合号码制，用一个字母代表一个大类，以字母的顺序反映大类的序列，在字母后用数字表示大类下类目的划分，数字的设置尽可能代表类的级位，并基本上遵从层累制的原则。

表2—1　　《中国图书馆分类法》基本部类和大类

基本部类	大　类
马克思主义、列宁主义、 毛泽东思想、邓小平理论	A 马克思主义、列宁主义、毛泽东思想、邓小平理论
哲学	B 哲学、宗教
社会科学	C 社会科学总论 D 政治、法律

续前表

基本部类	大 类
社会科学	E 军事 F 经济 G 文化、科学、教育、体育 H 语言、文字 I 文学 J 艺术 K 历史、地理
自然科学	N 自然科学总论 O 数理科学和化学 P 天文学、地球科学 Q 生物科学 R 医药、卫生 S 农业科学 T 工业技术 U 交通运输 V 航空、航天 X 环境科学、安全科学
综合性图书	Z 综合性图书

2.《杜威十进分类法》

《杜威十进分类法》由美国的威尔·杜威编制，是一部在国际上出现最早、流行最广、影响最大的图书分类法。1876 年出版，至 1996 年出版第 21 版，四卷本。卷一为编制说明和通用复分表，卷二、卷三为类表，卷四为索引和使用手册。它依据哲学家培根的知识分类思想，将图书分为十大类：

000 总论	500 自然科学
100 哲学	600 技术科学
200 宗教	700 美术
300 社会科学	800 文学
400 语言学	900 史地

17 世纪英国哲学家培根依据人的心理活动建立了知识分类思想，认为人类的心理活动从低级到高级有三种功能，即记忆、想象和理性，依次产生出历史、文艺、哲学三类知识，有人将《杜威十进分类法》的分类称为倒转培根法。《杜威十进分类法》采用纯阿拉伯数字作为基本标记符号，基本上按照层累制展开。

《杜威十进分类法》的修订和管理工作一直非常出色，这也是它经久不衰的重要条件。而且，《杜威十进分类法》早已推出了电子版（Web Dewey），2003年出版了最新的第22版。

2.3.3 组配分类法

2.3.3.1 组配分类表

组配分类表是由编制说明、基本类表、分面类表和分面公式以及通用辅表组成，它的建立，主要采用了分面分析法。分面分析法是将整个知识领域或某一知识领域按其不同属性分解为若干个不同的分面，每个分面再分解为若干个亚面，每个亚面还可分解为若干个更小的子面，面内列出所属各子目的一种编制分类表的方法。组配分类表由两个层次的分面结构所组成。第一层次的分面结构是对整个知识领域进行分面所形成的基本分面结构。如《冒号分类法》将整个知识领域划分为五种基本范畴：本体、物质、能量、空间、时间，这五个基本范畴就是五个基本分面。第二层次的分面结构是以第一层次的分面结构为依据，对某一知识领域进一步进行分解所形成的分面结构。如《冒号分类法》教育类的分面结构由受教育者、课程、教学方法、教师、教育环境、共同操作及施动者、理论观点、地点与时间、通用复分等组成。

在组配分类表的编制过程中，需要考虑到分面的引用次序与排列次序、标记符号与标记制度等方面的问题。

1. 分面的引用次序与排列次序

分面，又称组面或面，是指某一主题依据某一分类标准划分所得到的一组类目。分面的引用次序是指组配表达主题概念时，各分面被引用的先后次序，即各分面中有关类目的组配次序。同一组配分类表的引用次序应力求明确、规范。世界上不同的组配分类表在规定分面引用先后次序上并不相同。例如，《冒号分类法》是基于具体性递减原则，确定了各领域通用的五个基本范畴的引用次序为：本体、物质、能量、空间、时间，即“PMEST”分面引用公式，并且各大类目规定了具体的引用次序。《布利斯书目分类法》第二版采用的是基于目的性原则，其基本引用次序为：终端产品—种类—部分—材料—性质—过程—操作—施动者—空间—时间。依据这一基本引用次序，各类还制定了具体的分面引用次序。

分面排列次序是指组配分类表中各分面以及分面内各类目的排列先后顺序，其性质与分面的引用次序不同。分面排列次序可以采取与分面引用次序相同或相反的两种排列方法。当分面排列次序与分面引用次序相同时，称为顺排法，《冒号分类法》即采用顺排法。当分面排列次序与分面引用次序相反时，称为倒排

法，《布利斯书目分类法》第二版采用的是倒排法。

2. 标记符号与标记制度

组配分类表中的标记制度主要采用的是分面标记制和回归标记制。

分面标记制，又称分段标记制，是用分面符号把类号分成若干段，使每一段的号码代表主题的一个方面，以实现类目组配结构的一种标记制度。采用分面标记制所标引出的主题能够直观显示类目的组配结构，揭示各个主题因素及其联系。分面标记制所采用的分面符号有两种基本形式：(1) 采用数字或字母作为分面符号；(2) 采用标点符号作为分面符号。例如，《冒号分类法》中五个基本范畴分别用“,”“;”“:”“.”和“‘”表示。

回归标记制，也称为回溯标记制，是通过将分类表中位于前面的分面类目号码直接加在位于后面的分面类目号码之后，组配表达复杂主题概念的标记制度。回归标记制的号码较简短，并且具有较强的表达性和容纳性，不过其配号比较复杂，后面分面类号不能与前面类号相同，并且不能进行分段轮排，不能提供多途径检索。

2.3.3.2 组配分类法的特点

组配分类法的主要特点是：

(1) 通过简单主题概念的组配，一方面可以简化分类表，缩小类表体积，另一方面能够表达各种复杂主题概念和专深主题概念，并且能够揭示主题因素之间的相互关系；

(2) 可以对信息所表达的主题概念进行多方面标引，从而可以实现多途径检索；

(3) 可以较为及时地增补新的主题概念，类表修订灵活、方便。

组配分类法是体系分类法思想的改进，弥补了体系分类法存在的一些不足之处，如难以揭示细小的主题概念、检索途径单一、造成“集中与分散”的矛盾、难以及时修订类表等，但其自身仍然存在不足，主要表现在：类目体系不如体系分类法直观，标引和检索有一定的难度，使用不太方便，非专业人员使用难度较大等，因此，不适合图书馆、资料室组织文献分类排架。

2.3.3.3 主要组配分类法介绍

最早提出分面组配思想的是比利时的奥特莱。他在 1896 年撰写的《论数字分类法的结构》一文中提出了按“观点”分类和把简单概念组合成复杂概念的“组配原则”，并在 1905 年出版的《国际十进制分类法》中大量采用了冒号“:”、间隔号“·”、连接号“—”、圆括号“()”、方括号“〔〕”等分面组配符号，用

以组合简单概念。

系统的分面组配分类理论是由印度图书馆学家阮冈纳赞提出的。他在20世纪30年代编制出版了第一部极具影响的分面组配分类法，50年代初期提出了五个基本范畴和分面标记的思想和方法，随后出版了著名的专著《图书分类导论》，系统地总结了分面分析和分面标记的原则与方法。分面组配分类法突破了传统的等级列举式分类法的理论束缚，其“分面分析”和“分面组配”思想，对世界各国情报检索语言的理论与实践发展产生了重大的影响，各国图书分类法的编制和修订都不同程度地采用了这些原则。1960年，维克里出版了《分面分类法——专业分类表编制和使用指南》，总结了伦敦分类法研究小组编制分面分类法的经验，进一步推进了分面组配分类法的发展。

《冒号分类法》是阮冈纳赞编制的一部分面分类法，初版于1933年，提出了分面标记符号，只用“:”作为分段符号。1939年出版了第2版，采用“八分标记法”。1950年出版了第3版，广泛使用了“焦点”、“面”、“相”等概念，并依然只采用“:”作为分段符号。1952年出版的第4版，提出了五个基本范畴的概念，采用5种不同的分段符号，在很大程度上变革了原来冒号分类法的面貌。1957年出版了第5版，将分类表分为2卷。1960年，又将第5版中的第1卷修订出版，作为第6版。1972年，阮冈纳赞去世。1987年，《冒号分类法》的第7版出版了。

阮冈纳赞认为，图书分类的主要作用，在于给予每一个特定的主题以一个特定的类号，从而使每个不同的主题都能区别开来，并在类号中把主题的组成要素反映出来。而现行各分类法所采用的标记制度，无论是小数制、序数制还是其他方式（如字母），都把类目排成一条直线的方向，从而具有很大的局限性。新类目不能随时插入到它应有的位置；要把已有的类目加以进一步细分时，也往往不能给以最恰当的号码。为了解决这一问题，阮冈纳赞提出了以分析兼综合原则、分面分析和分面标记为核心的分面分类理论。

《冒号分类法》提出了五个基本范畴的理论。它们依次为：本体（Personality）、物质（Material）、动力（Energy）、空间（Space）、时间（Time）。通过这五个基本范畴来分析、归纳和组织文献。每个基本范畴都采用特定的指示符表示，即［P］;［M］:［E］.［S］‘［T］。在第7版中，又将物质面进一步分解成3个方面：物质［M］、物质性质［MP］、物质方法［MM］。

《冒号分类法》在标记制度方面也很有特色，它创立了分面标记制度，使每一特定的主题有一个特定的类号，并在号码中把主题的组成要素反映出来，针对列举式的类表和单线式的标记还提出了一些其他的标记方法，广泛采用了八分

法、百分法（双位法）等。阮冈纳赞还成功地创造了相的标记法，把它运用于不同学科之间的相互联系。《冒号分类法》具有标记表达性强、类表简练、容纳性强、适应性好、易于揭示复杂主题等优点，对今天的知识组织产生了一定的影响。其不足之处有，类目体系不够直观，标记符号种类繁杂、规则繁多，使用起来比较复杂等。因此，《冒号分类法》虽然在理论上对分类语言的发展做出了重大贡献，但在实践中并没有得到广泛使用。

2.4 主题检索语言

2.4.1 主题检索语言概述

主题检索语言又称主题法。它采用语词直接作为文献主题标识，按字顺排列主题标识，提供各种检索词语的途径。主题检索语言从描述事物的特性角度出发，按文献所论述的事物（即主题）集中文献，用规范化的名词术语标引和表达文献的主题概念，用参照系统显示事物概念主题词之间的关系。

主题检索语言与分类检索语言同样都是表现文献内容特征的检索语言，描述和揭示的对象都是各种各样的文献，它们都是建立在概念逻辑、知识分类和术语学的基础上，即利用区分概念的各种逻辑规则来显示词与词之间的关系，利用概念分析与综合的逻辑方法来构造标引语词。在应用知识分类方面主要是应用事物分类原理。

2.4.1.1 主题检索语言的类型

主题检索语言包括很多的类型，根据选词原则、词的规范化处理规则的不同，主题检索语言可分为标题词语言、单元词语言和叙词语言、关键词语言。

1. 标题词语言

标题词语言是一种先组定组式语言，它选择标题词作为文献内容的标识和检索依据，具体表现为标题词表的利用。

2. 单元词语言

单元词语言是以单元词作为语词标识对文献进行标引与检索的主题检索语言，是一种后组式语言。

3. 叙词语言

叙词语言是应用最广的主题语言，叙词语言是以叙词作为文献检索标识和查找依据的一种检索语言，概念组配是叙词语言的基本原理。

4. 关键词语言

关键词语言是直接选用自然语言，基本上不作规范化处理的一种检索语言。关键词指从文献题名或文摘以及正文中抽取的，能够表达文献主题并具有实质意义的未经规范化处理或略加规范处理的自然语言词汇。关键词标引迅速、容易，方便简单。同时，检索点比较多，可以从多个入口进行查找，非常有利于计算机检索系统的使用。

2.4.1.2 主题检索语言的特点

主题法是使用语词标识的检索语言。语词标识几乎都是名词和名词性词组。它具有较好的按事物集中文献和便于从事物出发检索文献的功能。目前，在机检数据库的检索中，主题法是最常用的检索语言，占有十分重要的地位。

主题检索语言与分类检索语言相比，具有明显的优点：

1. 专指性高

主题语言主要以规范化的名词术语为基础，着眼于事物及事物的各个方面。任何一个语词标识都能表达一个或大或小的、不受某一学科统辖、不被各个学科分割、基本上是独立完整的事物概念。凡是论述某一事物的文献，几乎都被标引在表达该事物概念的语词标识之下。从一个语词标识下即能检索到它所表达的事物的比较完全的有关文献。

2. 直观性好

主题语言直接采用主题词作标识，可以直呼其名，依名查检。它不同于分类号，对用户来说，易读、易记、易理解。而且按照字顺排列主题标识，更突显了它的直接优点。

3. 灵活性强

主题语言根据需要对主题词进行灵活组配，故特别适合计算机的逻辑组配功能。主题语言对于从事物出发的比较狭小的检索提问，以及关于新事物、新学科、新概念的检索提问，检索效果特别好。主题语言的主要缺点表现为：由于按字顺排列，所以同一门类学科的文献易被分散在各处，在族性检索方面不及分类语言。

2.4.2 标题词语言和单元词语言

2.4.2.1 标题词语言

标题词语言是主题检索语言中使用最早的一种类型。标题词是从自然语言中选取的、经过规范化处理的、表示事物概念的词、词组或短语。标题词按字顺排

列，词间语义关系用参照系统显示，并以标题词表的形式体现。

标题词语言的主要优点体现在：(1) 词表直接用事物名称列举出表达事物的主题，直观性强；(2) 采用先组定组式方法，因而词表中标题结构固定，含义明确；(3) 用参照系统显示主题之间的相互关系。标题词语言的不足之处是先组定组式的方法使得标题词表检索途径较为单一，无法实现多因素、多途径检索。另外，词表一般收词量大并且专指度相对不足，修订量大。

标题词一般分为主标题和副标题两级，通过主标题词和副标题词的固定组配来构成检索标识，因而只能选用“定型”标题词进行标引和检索，所反映的主题概念必然受到限制，并且无法从多因素、多途径进行检索。尤其是现代科技主题的内涵与外延越来越复杂，几乎不可能用一对主、副标题词完全、确切地表达出来。因此，标题词语言已不适应时代发展的需要，目前已较少使用。比较典型的标题词表有《工程标题词表》。

《工程标题词表》(Subject Headings for Engineering，简称 SHE) 由美国工程信息公司编辑出版，它是和《工程索引》(the Engineering Index，简称 EI) 检索工具配套使用的规范词表，在 1987 年修改补充的基础上，1990 年又作了新的修订，之后定名为《EI Vocabulary》。它的标题词由两级构成：主标题词及副标题词。主标题词表达概念、产品、过程、特征、材料等主题内容，使用名词、动名词，以单元词或复合词的形式出现。副标题词对主标题词起限定和修饰作用，表达主题的某一方面的特征，比如应用、现象、环境、制作、性能、地理位置等等。除了专用副标题词外，SHE 有通用副标题词表。这些通用副标题词不再出现在 SHE 主词表中，它们可以和主词表中的任一词配合使用，体现其通用特性，比如：控制 (control)、模型 (models)、研究 (research)、实验 (testing) 等词有明显的通用性，另外，国家名称也可作副标题词。SHE 词表中全部标题词按字顺编排，标题词下的副标题词再按它们的字顺排序。从 1993 年起，工程信息公司放弃了标题词语言，改用叙词语言编制，由《工程索引叙词表》(EI Thesaurus) 取代了它。

2.4.2.2 单元词语言

单元词又称元词，是从自然语言中选取，经过规范化处理，表达主题概念最小的、最基本的、字面上不能再分的名词术语。例如，“物理”是一个单元词，它表示了一个完整而独立的概念。

单元词表比较简单，它按照字顺，记录了一个检索系统所使用的全部单元词。单元词法采用后组配的方式，在标引时不组配单元词，在检索时才对单元词下所列的文献号进行对比，号码相同的就表示有组配关系。例如：

不锈钢
861081 862522 863519 866330 866332 867573 868582 868996
焊接
862111 866332 863519 863981 864530 869091

如果想查找关于“不锈钢焊接”方面的文献，“863519”号文献和“866332”号文献可以满足我们的查寻需求。在这两篇文献中，同时包含“不锈钢”和“焊接”两个概念。

单元词具有相对的独立性，词与词之间没有隶属关系和固定组合关系，标引时可根据需要加以组配。在单元词法中，组配功能得到了充分的应用。克服了标题词法的不足，比较适合机械检索系统。单元词语言的主要优点是：（1）通过单元词的组配可以表达大量专指概念和新概念，适应性强。（2）不存在词序问题，表达信息或文献标识中每一个单元词都可以作为检索入口，并且通过对单元词的增减，可以自由地扩大、缩小或改变检索范围。（3）单元词词表体积一般比较小，编制、更新和修订所需工作量小。单元词语言的不足是：（1）单元词法的字面分拆和字面组配，容易造成语义失真。（2）单元词法缺乏完善的参照系统，难以满足族性检索的要求。（3）单元词法的直观性和系统性较差。

单元词语言只适用于标识单元方式检索系统，它目前已经发展为叙词语言。

2.4.3 关键词语言

关键词语言作为信息存储和检索依据的一种检索语言，是直接从原文的标题、摘要或全文中抽选出来，具有实质意义的，未经规范化处理的自然语言词汇。但在实践中一般也要对关键词进行极少量的规范化处理。

2.4.3.1 关键词语言原理和优缺点

关键词语言是适应目录索引编制过程自动化的需要而产生的，出现比较早，广泛使用却是近二三十年的事。随着文献量急剧增长，计算机技术和信息处理技术的应用范围不断扩大，传统的手工标引方法越来越不适应情报工作的需要。人们借用计算机来编制索引，以缩短索引的编辑出版时间，加快文献的报道速度，关键词法就成为一种极有效的方法。

关键词语言的原理是：运用关键词语言编制的关键词索引，其关键词按字顺排列构成索引款目，所抽选的关键词都可以作为标引词在索引中进行轮排，作为检索“入口词”进行检索。例如，文献标题可以在相当程度上反映文献内容，因此，可以把文献标题中著者所用的具有实质意义的原词即关键词作为标引—检索

用词，在编制索引时，对关键词进行轮排，这样，文献标题中每一个具有实质意义的词都可以作为检索的入口，可以从多条途径入手对该文献进行检索。

关键词语言的主要优点是：(1) 直观性强。关键词法直接采用自然语言进行标引和检索，表达主题直观，并且符合普通用户的检索习惯。(2) 检索途径多。关键词法采用的轮排方式，可以多途径检索文献。(3) 标引简单。关键词接近自然语言，是由计算机自动抽取的，不用人工标引，不但节省人力，而且可以降低对人员水平的要求。(4) 关键词表达事物、概念直接、准确，不受词表控制，能及时反映新事物新概念。关键词语言的主要缺点是：(1) 不揭示关键词之间的等级关系和相关关系，使得相同主题的信息或文献因作者用词不同而导致漏检，影响查全率。(2) 难以进行族性检索。(3) 由于关键词法采用机械抽词和轮排，有可能会导致不少关键词款目失去检索作用而徒增篇幅。

关键词语言只有与计算机结合起来才能发挥它的独有优势，随着计算机的深入发展与应用，它的作用越来越突出，已被称为使用最为广泛的检索语言。目前，关键词语言已被广泛地应用于手工检索和计算机检索系统的索引编制中，并采取了编制禁用词表和关键词表等方法，以提高关键词抽取的准确性和对词间关系进行控制，提高检索效率。

2.4.3.2 关键词索引的类型

关键词索引的主要类型有题内关键词索引、题外关键词索引、词对式关键词索引等。

1. 题内关键词索引

题内关键词索引（Keyword-in-Content Index，简称 KWIC Index），又称上下文关键词索引，它以文献篇名为基本素材，以篇名中的关键词做索引款目的标目，以关键词的上下文做说明语。如美国《化学题录》(CT) 中的“题内关键词索引”。

题内关键词索引存在着一定的不足，标引词仅来自篇名，数量有限，来源不充分，个别篇名也不一定能真正反映文献内容。标引词不规范，有时会直接影响检索效果。

2. 题外关键词索引

题外关键词索引（Keyword-out-of-Context Index，简称 KWOC Index），改进和精简了题内关键词索引。这时的关键词不局限于从篇名当中抽取，可以根据需要从其他地方抽取。同时，改变了题内关键词索引检索人口在中间的做法，将关键词作为独立标目排在题目的前头。

3. 词对式关键词索引

词对式关键词索引（Paired Keyword Indexing），即将篇名关键词相互组配，

从某一篇名所含的全部关键词中每次取两个来做一个款目的标目。

关键词法的轮排方式在计算机检索中得到了广泛的应用，也正是在关键词法的基础上，逐步产生了自动标引和全文检索，促进了自然语言在信息检索领域的应用。

2.4.4 叙词语言

2.4.4.1 叙词语言的原理

叙词语言是以表示单元概念的规范化语词为基础，以概念组配为基本原理，对文献主题进行描述的后组式检索语言。

叙词是指一些以概念为基础的、经过规范化的、具有组配功能并可以显示词间关系和动态性的词或词组。叙词有这样一些特点：（1）直观性。叙词标识比较直观，按字顺排列，序列明确。（2）规范性。叙词都经过了规范化处理，包括对词义、词类、词形等的规范。（3）组配性。叙词可以灵活、自由地组配在一起，表达各种复杂的概念，比较适于计算机检索，在检索中可以充分采用布尔逻辑检索法、加权检索法等。

叙词语言吸收了其他多种检索语言的原理与方法，吸纳了体系分类法的基本原理，编制了叙词范畴索引和词族索引，从多方面来反映主题词之间的等同关系、等级关系和相关关系等；保留了单元词语言的组配原理，采用了组配分类语言的概念组配来代替单元词语言的字面组配，并取代了单元词语言；吸收了关键词语言的轮排方法，编制了各种叙词索引；采用了标题词语言对语词进行严格规范化的方法，保证了词与概念的一一对应，采用并进一步完善了标题词语言的参照系统。

2.4.4.2 叙词表的编制

叙词表是叙词语言的核心体现。目前，国内的叙词表已有七八十种之多。常用的有《汉语主题词表》、《化工汉语主题词表》、《机械工程主题词表》、《电子技术汉语主题词表》、《国防科学技术叙词表》等。常见的国外叙词表有《INSPEC叙词表》、《工程索引叙词表》、《工程与科学叙词表》等。

叙词表一般由一个主表和若干个附表构成。主表是叙词字顺表，该表将叙词完全按字顺排列，并有标注事项和参照系统。附表主要包括：叙词分类索引、词族索引、轮排索引、双语种对照索引、专有叙词索引等。叙词分类索引也称分类表或范畴索引，便于从学科或专业分类的角度来选用叙词。词族索引也称等级索引，具有属分关系的一组称为一族，构成一个从泛指叙词到专指叙词的等级系

统。轮排索引，也称轮排表，将有相同单词的词组叙词集中在一起，排列在这个单词之下，可以方便人们从该单词出发，查出某一个或全部含有该单词的词组叙词。双语种对照索引如英汉对照索引。专有叙词索引如地区索引、人物索引、机构索引等。

在叙词表的编制过程中尤其要注意以下问题：

1. 主题词的选择与规范

主题词也称叙词，在叙词表中它是表达一定意义的最小词汇单元。主题词不仅反映了一定事物的概念，而且作为事物概念的表达形式而存在。因此，主题词是表达概念的一种形式，而概念则是主题词所表达的内容。

主题词包括普通主题词和专有主题词两种。普通主题词是表示各种事物及其属性的名词，它所表达的常是普通概念，如反映各学科、各种职能活动的基本术语等。专有主题词是表示某一特定事物的专有名词，它所表达的是单独概念，如地名、民族名、语言名、时代和年代、人名、机构会议名称、产品名称、历史事件名称、法规名称、主义、学说、学派、定理等专有名称。

在编制叙词表的过程中，主题词的选择要以所编叙词表规定的专业或职能范围为依据。综合性叙词表和多学科叙词表选词时，各专业、各类职能的名词术语的选用要大致平衡。专业性叙词表的选词，要突出专业特色，兼顾相关专业和相关职能。同时，选词要考虑文献检索的具体要求，以及被标引文献的数量和增长速度。还要考虑被选词的使用频率和检索意义，一般不选用使用频率过高或过低的词作为主题词。对于一个使用频率过高的词，应增选它的下位词；对于使用频率过低的词，可以不选该词，选用它的上位词即可。但对于那些反映新事物、新学科的词，即使开始时可能在文献中出现频率不高，也应给予收录，而对于一些反映旧学科、旧事物的词，即使过去某一时期在文献中出现频率较高，也不一定要选取（对于标引历史资料的主题词表除外）。此外，基本词汇要完备而精练，要注意选用词义明确、符合科学性和通用性的词作主题词。

叙词法规范化处理的内容包括四个方面：

（1）词形规范。是指对自然语言中存在的同义不同形的词语的规范。叙词法中词形的规范需要考虑到同义词与准同义词的规范、词序的规范、词长的规范、汉字形体的规范和外来语词的规范等。

（2）词义规范。是指对自然语言中的多义词、同形异义词进行规范处理。词义规范的内容包括两种类型：一是范围注释，是指对同一主题词在不同学科领域或在不同语言环境下所具有的不同概念进行注释，用来阐明其使用范围；二是含义注释，是指对在某些概念上混淆不清的主题词做简明扼要的说明，用来明确其

含义和用法。

（3）词类规范。是指对主题词选定范围进行控制。比如，主题词一般只能从名词或动名词等具有实际意义并能反映事物本质属性的词中选取，其他的词类应尽量避免或控制使用。

（4）先组度规范。是指对主题词先组程度的规范。叙词语言虽是后组式检索语言，一般是采用组配方式来表达复杂的主题概念，但如果叙词表采用适当的先组词，会提高标引人员标引的一致性，并且会加快标引速度，而过多地采用组配，则可能造成标引的不一致和影响标引速度，因此，叙词表应当对主题词的先组程度进行适当的规范。

2. 主题词之间关系的显示

叙词表的主表是按照主题词的字顺排列起来的，不能直接显示各主题词之间的逻辑关系，展示主题词的语义性。为此，叙词语言采用了多种方法，除了采用词族索引、范畴索引、轮排索引外，最主要的是采用参照系统。

参照系统对于主题词的语义关系的揭示，是通过制订各种符号来加以联系和反映的，具体体现在三个方面：

（1）同义关系，又称等同关系或代用关系，是指两个或多个词所表示的概念相同或相近，并且可以互换的关系。同义关系的规范化处理，是从同义词中选出一词作为正式主题词，其他的词则作为引导词。同义关系用“用”、“代”来表示。

（2）属分关系，又称为等级关系，是指专指度深浅不同的两个主题词之间的关系。属分关系采用“属”、“分”两个参照符号来显示，“属”用于下位主题词指向上位主题词；“分”则用于上位主题词指向下位主题词。“属”与“分”互为反参照。

（3）相关关系，是指主题词之间除了同义关系和属分关系之外的某种比较密切的关系，也称类缘关系。相关关系用“参”来表示。

综上所述，参照系统所显示的叙词之间的关系有：同义关系（代、用）、属分关系（属、分）和相关关系（参）等，具体见表 2—2：

表 2—2　　叙词参照符号及含义

参照符号		术语名称	符号含义	词间关系
汉	英			
Y	USE	用	从非使用词指引到使用词	同义关系
D	UF	代	使用词的同义词	同义关系
F	NT	分	使用词的下分词（狭义词）	属分关系
S	BT	属	使用词的上属词（广义词）	属分关系
C	RT	参	使用词的相关词	相关关系

3. 主题词的组配

叙词语言的组配吸收了组配分类语言的概念组配原理，采用了单元词法的后组方式，超越了单元词的字面组配，实现了概念组配。主题词的组配可分为交叉组配、限定组配和联结组配。交叉组配是指用两个或两个以上具有交叉关系的同性质的主题词组合表示一个复合概念的组配。限定组配又称为方面组配，是指将表示某一事物的主题词与表示事物某一属性、某一方面问题的主题词所进行的组配。限定组配是以概念的限定方式为基础，由泛指的属概念过渡到专指种概念的组配方式，在主题词的概念组配中占绝大部分。联结组配是表示两个概念之间联系的组配方式，其组配并不形成新概念，而只用于揭示概念之间的某种联系。

概念组配是叙词语言的基本原理。概念组配依据概念的分析与综合，与字面组配有时相同，有时不同。通过组配可以增强叙词语言的表达能力，控制词表的词汇量，提升叙词法的匹配能力，提供多途径检索，提高查全率，还可以及时反映新事物、新学科。

2.4.4.3 叙词语言的性能

叙词语言继承和发展了体系分类语言、组配分类语言、标题词语言、单元词语言、关键词语言等多种检索语言的思想、原理和优点，具有多方面的优势，并且已经成为在当今互联网时代应用最为广泛的人工检索语言之一。首先，叙词语言吸收了单元词语言用组配来表达主题概念的方法，但摒弃了单元词语言采用字面分解和字面组配而容易造成语义失真等消极因素，将字面分解和组配完善为概念组配。其次，叙词语言适当借鉴了标题词语言的先组方式，将一些通用的专称、俗语、专指作用很强的词组、专业文献中出现频率较高的经常用以检索的词组等以先组词的方式直接收入叙词表中。叙词法继承了组配分类语言分面组配的思想，形成了自身概念组配思想。体系分类法通过类目的层层划分所形成等级结构、标记制度、编制说明和注释来标识类目之间的相互关系，而叙词语言参考了体系分类法的思想，建立了范畴索引、词族索引和参照系统来揭示主题之间的相互关系。此外，叙词语言还借鉴了关键词语言的轮排技术，通过编制轮排索引，从而加强了叙词语言的族性检索功能，同时也增加了检索途径。

总的来说，叙词语言是一种非常优秀的检索语言。但是，它也有不足之处，主要表现在：叙词表的编制工作难度较大、标引复杂、标引速度慢、族性检索功能不够强、使用人员需要有较多的专业知识等。

2.4.5 主要主题词表介绍

目前，国内外的主题词表有许多，如国外使用最广的综合标题词表《美国国

会图书馆主题词表》(Library of Congress Subject Headings，简称 LCSH)、专业叙词表《医学主题词表》(Medical Subject Headings)，国内比较有影响的主题词表有《汉语主题词表》、《中国分类主题词表》、《社会科学检索词表》、《中国档案主题词表》等。其中，《汉语主题词表》在国内的影响最大。

《汉语主题词表》是我国第一部大型的综合性的叙词表。由中国科技信息研究所和北京图书馆负责主持，1975 年开始编制，1980 年正式出版。分为社会科学、自然科学和附表 3 卷，共 10 个分册，全表收录主题词 108 568 个。其中正式主题词 91 158 个，非正式主题词 17 410 个，词族数 3 707 个，范畴数一级 58 个，二级 674 个，三级 1 080 个。

《汉语主题词表》结构体系比较全面，由主表（字顺表）、附表、词族索引、范畴索引和英汉对照索引组成。图 2—3 为主表款目示例。

Xianxiangguan ——汉语拼音
显像管 [56E] ——叙词和范畴号
Kenescope ——英译名
Picture tube
D 电视显像管 ——非叙词（代项）
监视管
F 彩色显像管 ——下位词（分项）
固体显像管
黑白显像管
S 电子束管 ——上位词（属项）
Z 电子管 ——族首词（族项）
C 显示管 ——相关词（参项）
指示管

图 2—3 《汉语主题词表》主表款目示例

主表（字顺表）包括社会科学和自然科学两部分，是词表的主体部分，由全部正式叙词款目和非正式叙词款目组成，所有款目严格按汉语拼音音序排列。叙词款目是主表的基本单元，每一个叙词款目的结构包含叙词、汉语拼音、英文译名、范畴号、注释项及其语义关系项等。

附表包括 4 种专有词汇表：世界各国政区名称、自然地理区划名称、组织机构名称和人物。世界各国政区名称表收录了世界各国、地区及重要城市名称，中国各省、自治区、直辖市以及部分重要城市和地区名称。自然地理区划名称表收录了世界重要地理区划名称，如山、川、河、湖、海、洋、岛屿、平原、盆地等的名称。组织机构名称表和人物表分别收录重要的机构和人物。

词族索引又称族系索引、等级索引，是将主表中具有属分关系的正式主题词集中在一起，显示词间从属关系的一种索引系统。词族索引用来揭示主题词之间族系关系，满足族性检索的需要。词组索引中，通常以族首词为标目，按照词族中的关系展开各级叙词，以小圆点作为等级符号。示例如图 2—4：

Guangbo Xitong	——汉语拼音
广播系统	——族首词（一级主题词）
·电视广播系统	——二级主题词
··多伴音系	——三级主题词
·无线电广播系统	——二级主题词

图 2—4 《汉语主题词表》词族索引款目示例

范畴索引又称分类索引，是主表中全部叙词的分类索引。它将全部叙词和非叙词按社会科学和自然科学两大范畴划分为 58 个大类，方便人们从分类角度查找与某一范畴内容有关的主题词。

英汉对照索引是将主表和附表中的正式和非正式主题词的英文按字母顺序排列的一种索引，是通过英译名来选择主题词的辅助工具。

2.5 分类主题一体化检索语言

2.5.1 分类主题一体化检索语言概述

分类语言与主题语言是两种不同类型的检索语言，两者各有各的优势和不足，并且不可互相替代。我国著名的图书馆学家刘国钧先生也曾指出：“分类法与标题法各有所长，各有所短，合其双美，离则两伤。”为了把分类语言与主题语言有机地结合在一起，实现两者的兼容，以充分发挥两者各自的优势，满足信息检索的不同要求，国内外的图书情报学者对此进行了不断的研究与探索。1969 年，英国学者艾奇逊（J. Aitchison）和戈默索尔（A. Gomersall）等成功地将《英国电气分面分类法》改编为《分面叙词表：工程及相关学科的叙词表及分面分类法》，标志着分类主题一体化检索语言的诞生。1983 年，熊兴成编制的《常规武器工业分面叙词表》，是我国第一部一体化词表。分类主题一体化检索语言，又称为分类主题一体化词表，是指在一个检索语言系统中，对它们的分类表部分和叙词表部分的术语、参照、标识及索引实施统一的控制，使两者有机地融合为

一体，从而能够同时满足分类和主题标引、检索的需要，发挥其最佳的整体效应。

2.5.1.1 分类主题一体化检索语言的原理

分类检索语言与主题检索语言两者之间的区别主要是表现在形式、结构和应用不同等方面。分类检索语言是按学科、专业集中信息或文献，以分类号作为主题概念的标识，经过类目的层层划分与排列，形成了等级式的学科及逻辑体系。而主题检索语言是按事物来集中信息或文献，以受控的主题词直接作为主题概念的标识，通过参照系统和范畴索引、词族索引来揭示主题词相互关系的字顺系统。

分类主题一体化检索语言的原理建立在分类检索语言与主题检索语言相通的原理基础之上。首先，分类检索语言与主题检索语言都是建立在概念逻辑、知识分类和术语学的理论基础之上，都应用了概念划分与概括、概念分析与综合的方法。其次，所采用的表达信息或文献主题概念的标识在本质上相同的，只是表现形式不同而已，即分类检索语言是用分类号作标识，而主题检索语言是用主题词作标识的。最后，分类检索语言与主题检索语言的处理对象都是语义单元，所类集的内容是相同的，表达的都是主题概念。

2.5.1.2 分类主题一体化检索语言的功能

分类主题一体化检索语言除了单独具有分类检索语言与主题检索语言的功能外，还具有如下功能：

（1）标引人员可以同时完成分类标引和主题标引，通过标引数据之间的对应转换，可以节省人力物力，并且可以减少标引错误和标引不一致性。

（2）用户既可以从学科、专业出发来进行分类检索，也可以从事物主题出发进行字顺主题检索，并且可以加以比较，从而较大程度地提高查全率和查准率。

（3）可以为进行过分类标引而未进行过主题标引的书目数据库通过主题词与分类号的转换而提供主题标引，同样，也可以为进行过主题标引，而未进行过分类标引的书目数据库通过主题词而进行分类标引。

2.5.2 分类主题一体化检索语言的类型

分类主题一体化检索语言按照兼容互换的方式，可分为三种类型：分面叙词表、分类表—叙词表对照索引和集成词表。

分面叙词表是以艾奇逊等主编的世界上第一部一体化词表的名称命名的，它也是最典型、影响最大的分类主题一体化检索语言。它一般由分类表和叙词表两

大部分组成，有的还附有轮排索引及英汉对照索引。分类表和叙词表通过分类号相联系，分类表主要起字顺索引的作用，而叙词表则不仅起着传统叙词表范畴索引和词族索引的作用，还可以直接用于主题标引。我国编制出版的分面叙词表有《教育主题词表》、《农业科学叙词表》、《音像资料叙词表》等。

分类表—叙词表对照索引通常由分类号与主题词对应表、主题词与分类号对应表两部分组成，前者为每个类目列出其对应的一个或多个主题词，后者为每个主题词列出对应的一个或多个分类号。这种对照索引是分类检索语言和主题检索语言兼容互换的工具，通常用作情报检索查词选类的辅助工具，便于分类标引数据与主题标引数据的相互转换。如果在这种对照索引的分类号和叙词下加上原有的注释项和参照项，就可以同时用于分类标引和主题标引。从这个意义上说，分类表—叙词表对照索引也可作为一种特定类型的分类主题一体化词表。我国编制出版的《中国分类主题词表》、《中图法教育专业分类表》等就属于此类型的词表。

集成词表是将某些特定主题领域的若干叙词表和分类表汇编而成的一种词表，可以用于联合分类标引和主题标引，用于实现分类表和叙词表之间的兼容互换。集成词表是一种较为理想的实现兼容与互换的工具。它的主要功能是：(1) 提供一种指示现有检索语言之间的兼容性，以及在数据库检索中便于由一种词表向另一种词表转移的手段。(2) 提供从一种自然语言叙词表转译为另一种自然语言叙词表的辅助工具。(3) 提供一种在采用不兼容检索语言的系统之间传递标引数据的交换工具或中介词典。(4) 向不准备另建叙词表的单位提供一个可供采用的现成的词表用于标引和检索。(5) 为一个包含其他学科的专业叙词表提供边缘学科的词汇及其词间关系的来源。集成词表在我国的典型代表是《中国法 R 类与医学主题词表、中医药学主题词表对照表》。

2.5.3 《中国分类主题词表》简介

《中国分类主题词表》是在《中图法》编委会主持下，经全国约 40 个单位 160 位专家学者的共同努力，历时 8 年编制而成的，于 1994 年 6 月正式出版。《中国分类主题词表》是在《中图法》第三版（含《资料法》第三版）和《汉语主题词表》的基础上编制的我国第一部分类检索语言和主题检索语言相互兼容对照索引式的一体化词表。全表共分两卷 6 册，收录分类法类目 5 万余个，主题词及主题词串 21 万余条，包括《分类号—主题词对应表》和《主题词—分类号对应表》两部分。为提高文献主题标引和文献分类标引质量，由《中图法》编委会组织编写了《〈中国分类主题词表〉标引手册》，1998 年由北京图

书馆出版社出版。

第1卷《分类号—主题词对应表》以《中图法》的类目体系为基础，将《汉语主题词表》的全部主题词以及增加的主题词兼容对应于各级类目之下，起着类目注释的作用，并且在编制过程中，对《中图法》第三版500多处类目和注释进行了增补，可视作一部以主题词作注释的新版《中图法》。该卷分左右两栏编排，左栏是《中图法》的类表，右栏是相对应的主题词和主题词串。其主要功能是文献分类标引和通过分类的途径查找主题词，进而进行主题标引。

第2卷《主题词—分类号对应表》是从主题词到分类号的对照索引体系。它按主题词的字顺排列，其后列出对应的分类号。主题词款目结构与《汉语主题词表》大体相同。在编制过程中，以《汉语主题词表》原有的主题词为依据，进行了大量的增补、删除和修改，增加了14 000个反映新学科、新事物的主题词，删除了一批陈旧过时的词，调整了一些主题词的参照关系，把词族索引直接纳入到主表之中，扩大了检索入口范围和族性、相关性检索的可能性。其主要功能是进行文献主题标引和通过主题查找相关的分类号，作分类标引的辅助手段。

《中国分类主题词表》是分类与主题、先组式检索语言与后组式检索语言相结合的一体化检索语言体系。使用该表不仅可以使分类标引、主题标引在经过同一主题分析、采用同一标引工具的过程中一次完成，而且能够降低主题标引的难度，提高标引的一致性。同时，由于分类号与主题词之间建立了对应联系，有利于在检索系统中实现分类号与主题词之间的相互转换，从而提高检索效率。

《中国分类主题词表》也存在一些不足，如没有编制主题词轮排索引和英汉对照索引，标引组配不够灵活，而且，在分类号与主题词的对应过程中，受到主观因素的影响，也很难做到完全的科学和准确。

2.6 网络信息检索语言

2.6.1 检索语言面临的网络环境

随着互联网技术的快速发展与广泛应用，受控的检索语言面临着巨大的挑战，主要表现在以下四个方面：

2.6.1.1 信息类型的变化

传统环境下，检索语言所面对的信息类型主要表现为文献资料，如图书、期刊、报纸、档案、政府出版物、会议文献等，并且其类型基本上都是纸质文本型

信息。而网络环境下的信息类型多样，除了有电子化的传统文献资料外，还出现了许多新的类型，如联机数据库、软件、博客、论坛、电子邮件等。既有大量的电子文本型信息，也有图形、图像、音频、视频、动画等形式的信息。

2.6.1.2 信息数量与质量的变化

传统环境下的信息数量尽管是庞大的，但仍然无法与网络环境下的信息数量相比。据统计，2006年10月份全球网站突破1亿个。[①] 每个网站一般至少由几十上百个网页组成，而网络信息以网页的形式存在，因此，网络环境下的信息数量大大超出了传统信息数量。另外，网络信息内容范围极其广泛，涵盖了人类所有社会生活领域，有人文科学、社会科学、自然科学等学术性信息，也有大量生活服务、娱乐休闲等非学术性信息。互联网无疑已成为世界上数量最大、内容最为全面的信息资源库。传统环境下的信息在传播之前一般都经过较为严格的规范和控制，因此，信息质量较高。而在网络环境下，任何组织和个人都可以自由地发布信息，不需要经过规范和控制，使得网络信息质量参差不齐，既有质量非常高的学术信息，也存在着大量质量低下的垃圾信息。

2.6.1.3 信息检索技术的变化

传统环境下的信息检索是依靠目录、索引、工具书、年鉴、图书馆书目卡片进行的，检索到的信息只有文本信息，并且检索范围比较狭窄，检索效率较低。由于计算机技术、多媒体技术、网络技术、数据库技术、人工智能技术、自然语言处理技术等以及相应硬件技术的发展和应用，新的检索技术不断出现，如全文检索技术、多媒体检索技术、超文本和超媒体检索技术、可视化检索技术、联机信息检索技术，这些信息检索技术的应用，大大方便了用户的检索行为，提高了检索功能和效率。

2.6.1.4 信息用户的变化

分类检索语言、主题检索语言的规范性和复杂性，使得传统的信息检索用户一般为专业性情报检索人员，或者是受过专业培训的人员，并且他们进行信息检索的目的多为科学研究需要。而伴随互联网的广泛应用，信息技术和信息检索系统的易用性使得信息检索用户从专业人员扩大为广大的普通用户，用户不受时间、地点、年龄、身份、地位、教育程度等限制，并且用户信息需求也不再仅仅是为了学习和科学研究，而是呈现出多样化、个性化。

---

① 参见 http://tech.sina.com.cn/i/2006-11-04/12471220523.html，2007-12-05。

2.6.2 网络环境下的分类检索语言

分类检索语言比较全面和客观地反映了知识全貌及其内在的逻辑联系，其体系结构的系统性、标识符号的通用性以及族性检索功能，是其他信息检索语言所不具备的，也是无法取代的，而且，分类方法符合人类认识事物的逻辑思维方式。因此，在网络环境下，分类检索依然有着强大的生命力，只是由于分类语言描述对象（信息资源）和利用对象（标引者和检索者）都发生了变化，分类检索语言为适应这些变化必须作出相应的调整，主要表现在形式和内容两方面。

2.6.2.1 在形式上，实现了电子化和体系结构的多维化

1. 分类法的电子化

印刷版分类法由于体积庞大而厚重，翻阅、携带不方便，再加上分类法本身的复杂性，大大影响了标引和检索的速度和效率。另外印刷版分类法的维护管理也比较困难，修改、删除或更新某一类目可能要对整个分类法做很大的调整，费时、费力，更新周期也较长。分类法的电子化由于其直观方便的浏览、显示功能，超文本链接功能和完善的检索功能，可以改变以往手工式标引与检索所带来的不便，提高了标引与检索的速度与效率，减少了标引成本，其修订也较为方便，更新周期缩短。更为重要的是，分类法的电子化是实现信息或文献组织与检索自动化的基础。世界上三大著名分类法——DDC、UDC、LCC 不仅实现了电子化，还推出了网络版。我国于 2001 年也推出《中图法》第四版的电子版。

2. 分类体系结构的多维化

传统分类法的体系是以一种典型的线型结构来揭示类目之间内在关系的，表现出明显的单维特征。尽管传统的分类法也试图采用组配方式或其他方式，改善类目之间简单的单维联系，但结果并不是特别理想。超文本技术在分类语言中的应用，彻底改变了类目之间的线性关系，为分类语言的发展带来了新的生机。超文本技术允许用户在浏览文本信息的同时，随时可以选中其中的“热字”。热字往往是上下文关联的词汇或句子，通过选择热字可以跳转到其他的文本信息。超文本技术为多角度、多途径浏览与检索提供了技术支持，使分类法实现体系多维化有了可能。这样一来，就可以充分利用超文本技术，更好地揭示类目之间的多维关系。

新型的网络分类目录在类目划分标准、横向关系揭示和类目设置方面表现出明显的多维化趋势。网络分类目录打破了传统分类法划分标准唯一的限制，在同一个类目下集中了依照主题对象、学科属性或资源类型等划分的所有下属类目。同时，随着超文本技术的利用，对于多属性主题、交叉学科、边缘学科、总论与

专论、地区与主题、资源形式与主题等横向关系的揭示会变得十分轻松。而传统的分类法由于技术条件的限制，不利于充分地、客观地揭示和反映多维性的知识空间。技术的进步使得多维揭示、多角度设类成为现实，较好地解决了文献信息的集中和分散问题。

2.6.2.2 在内容上，编制方法作了很大的调整，主要表现为聚类标准的主题化、类目划分的随意性、类目排列的非逻辑性和类名的通俗化等

1. 聚类标准的主题化

传统分类检索语言以信息的学科属性作为聚类的依据，网络分类法则是以网络信息的主题作为聚类的主要依据，这样更符合普通用户的使用习惯。

2. 类目划分的随意性

传统分类语言中类目的划分是按照严格的逻辑划分规则，从上而下，层层划分，形成科学、严密的等级分类体系。网络分类语言中类目划分是根据用户的需要和习惯，一般没有按照严格的划分规则，注重开放性和可变性，所划分的类目之间隶属关系比较模糊，类目体系不太严密，列类较杂。此外，网络分类法的划分层次比传统分类法的划分层次要少得多。

3. 类目排列的非逻辑性

传统分类检索语言同位类的排列注重类目之间的内容关系和逻辑联系，系统性、整体性、科学性、逻辑性、连续性和一致性较强。传统分类检索语言同位类排列主要采用的序列方法有：按照逻辑顺序排列、按照客观事物发展的顺序排列、按照时间顺序排列、按照空间顺序排列等，此外，还可以依据依存次序、惯用次序、实用次序和字顺次序等其他顺序进行同位类的排列。网络分类检索语言中同位类的排列不局限于使用逻辑排列法，往往为了方便、快速、自动地排列类目，而使用字顺排列、按重要性排列等方法。

4. 类名的通俗化

传统分类检索语言中类名的确定有比较严格的规范，以力求类名的科学、统一、准确、规范，并且一旦类名确定下来，在较长时间内不会改变，类名稳定性高。网络分类语言中类名的确定是从用户的角度出发，关注普通用户的一般思维方式、检索习惯和需求特点，力求简单和通俗易懂，并且类名会根据社会发展和用户检索习惯而进行调整，因此，类名的稳定性不强。

在新的网络环境下，分类语言沿着两个方向继续得到发展。

一个方向是积极地调整传统分类法自身，以满足信息资源数量的迅速增长。2003 年 6 月最新推出的《杜威十进分类法》第 22 版的电子版增加了大量印刷版中没有的类目，对 004～006 数据处理计算机科学、301～307 社会学与人类学、

340 法律、510 数学、540 化学、610 医药与健康、900 历史与地理等作了重大修订和补充。同时，一些网站采用了《杜威十进分类法》、《国际十进制分类法》、《美国国会图书馆分类法》以及其他的综合性分类法、专业分类法来对互联网资源进行组织和整理，为用户提供服务。但这些分类法为适应网络资源，对原来的类目进行了必要的调整，对类目级别的深度进行了适当控制。

分类语言发展的另一个方向是抛开传统的分类法，重新建立新的分类体系，即网络分类目录。它继承了传统分类法层层划分、从总到分逐级展开的基本思路，但不遵循以学科分类为基础的分类原则，不再使用分类号作为信息分类标识和依据，直接使用语词来形成网络分类目录。类目的设置是采用主题与学科相结合的方式，类目体系在体现科学性的同时，更注重追求实用、易用、通用和灵活。

2.6.3 网络环境下的主题检索语言

主题语言使用语词对信息进行揭示和组织，直接用语词标识信息内容，可以较好地满足用户的特性检索需求，主题语言在网络环境下仍然是一种重要的检索语言。互联网的普及、网络信息资源的迅猛增长和信息检索用户的改变对主题语言的发展产生了较大的影响，网络环境下的主题检索语言主要表现为以下几种形式：

2.6.3.1 传统主题检索语言在网络信息检索系统中的应用

传统主题检索语言在网络信息检索系统中的应用主要表现在标题语言、叙词语言和关键词语言在网络信息检索系统中的应用。采用标题词表、叙词表的网络检索系统一般是学术性较强的专业网站，如图书馆网站的书目信息检索系统、网络联机数据库检索系统等。采用这些检索系统一般都能得到较高的查准率和查全率，检索效率高。例如中国人民大学图书馆的联机目录的主题检索，该系统依据《汉语主题词表》和《美国国会图书馆主题词表》分别对收录的中西文文献进行了主题标引，对每一种文献都给予了相应的主题词，通过主题检索途径，可以检索到与该主题相关的文献。关键词语言在网络环境中的主要应用形式是搜索引擎的关键词检索。搜索引擎的易用性和良好的检索性能，使得其已成为目前世界上最流行的检索方式。

2.6.3.2 辅助词表的应用

除了主题词表外，一般检索系统还编制了辅助词表。辅助词表的主要类型有后控词表、禁用词表、同义字词典和反义字词典等。

后控词表也称为词间关系表，是利用受控语言的基本原理和方法编制的自然语言检索用词表。后控词表的主要特点是：（1）词汇只用于检索，不能用于标引。（2）词汇控制不如先控词表严格，一般只对同义词、近义词及不同书写形式进行控制，适当处理部分与相关关系。（3）收词丰富，包括同义词、近义词、俗称、缩写、流行俚语等。（4）动态性强，及时更新和增补新主题概念。（5）有较强的灵活性和自由度，标引工作简单，编制简便。（6）具有面向文献和用户的特点，文献保障和用户保障能力强。

后控词表只对系统的输出阶段进行控制，它的控制处理相对受控语言检索系统比较简单，无需标引人员花费很长时间去分析文献的主题概念，选用合适的检索词进行标引、归类，检索者也不必花太大的精力分析检索要求，考虑用符合标准的检索语言来表达自己的信息需求。后控词表展现了比较完整的语义关系，用户通过浏览词表选用检索词，大大减轻了构造检索策略的负担，提高了检索速度，节省了检索时间。后控词表作为自然语言与受控语言相结合的产物，对于提高自然语言检索系统查全率和查准率具有重要的意义，为用户准确选词、精确检索、扩检和缩检、改变检索范围、进行相关检索提供了捷径。

目前，国内外对后控词表展开了一系列的研究，并取得了一定的成果，比较有代表性的有美国国防技术信息中心（DTIC）科技报告全文检索系统（http：//www. dtic. mil）、美国教育资源信息中心（ERIC）数据库全文检索系统（http：//www. eric. ed. gov）、生物科学情报社（BIOSIS）的检索系统（http：//www. biosis. org）、我国中文食品工业文献检索系统、北京国际贝斯有限公司的基于互联网的汉语后控全文检索系统等的后控词表。

禁用词表也称停用词表、禁用词典、禁用单元词表等，英文表达为 Stop List 或 Stop Words，是将一些单独使用时无检索意义，或者出现频率过高的词作为检索系统的非检索用词，如英文检索系统常见的禁用词有 in、at、of、about、up、out、is、are 等，以对检索词的有效性进行控制。禁用词表按收入词的类型，一般可分为两种类型：普通禁用词表和条件禁用词表。普通禁用词表是指由在任何情况下都无实际检索意义的词所构成的词表，如包含介词、连词、代词等的词表。综合性网络信息检索系统中编制的禁用词表通常就是普通禁用词表。条件禁用词表是指由在一定条件下才成为禁用词的词汇构成的词表。这种禁用词表主要出现在专门性的检索系统中，因此，不同网络检索系统在理论上其禁用词表不会兼容。条件禁用词表除了收录条件禁用词外，通常还包括无实际检索意义的普通禁用词。

在自然语言中，存在着大量词汇间的同义、近义和反义关系，用户在网络信息检索时所选择的检索用词往往不统一和不规范，用户也不可能把反映主题概念

的全部同义词、近义词、反义词等都作为检索标识进行检索，如果把用户的检索标识转换成规范化叙词，或者更进一步扩展出所有可能出现的同义词、近义词和反义词，显然会较好地提高查全率。同义字词典和反义字词典就是基于这种目的而编制的，这两种词典都是基于主题语言，显示概念等同关系原理而产生的。同义字词典和反义字词典除了能够提高用户的查全率外，对自然语言与受控词表的一体化、系统与检索用户的交互选择、自动扩展的智能化检索的发展也起到非常重要的作用。

2.6.4 自然语言在信息检索中的应用

随着计算机技术的发展、计算机信息检索系统的广泛使用，自然语言在信息检索领域的应用开始流行起来。传统的采用受控语言（人工语言）的信息检索系统要求检索者必须具备一定的检索理论和实践技能，熟悉系统的检索功能与操作命令、检索语言的特点及有关的检索策略与检索技巧等方面的知识。随着信息资源的海量化、信息需求的不断扩大，越来越多的非专业人士开始涉及信息检索领域，人们开始不满意传统受控语言的严格与规范，渴望信息检索更加简洁和易用。自然语言处理技术在信息检索领域的应用带来了一场大变革，促进了新的自然语言检索方式的产生和发展。

自然语言指直接取自文献本身，不经加工和规范的语言，它包含词、词组或句子，没有繁琐规则的约束，不添加任何人工的色彩。

自然语言具有许多人工语言所没有的优点：

（1）自然语言检索方便，不受人工语言的种种限制，不需要复杂的检索规则，使用者能够较快适应，易用性突出。

（2）自然语言采用从文献中直接抽词的方式，避免了人工标引过程中的失真现象。全文检索技术的发展在很大程度上推进了自然语言检索的发展。目前，一些数据库和搜索引擎采用了自然语言检索，允许用户直接采用自然语言进行检索，用户可以输入类似“What is Jamestown?”、“When did Web searching start?”等问题。

（3）自然语言非常容易吸纳新的词语、新的概念。采用自然语言检索新出现的事物可以获得较好的检索效果，使用者可直接使用这一新词作为检索入口，不必像传统的人工语言那样，必须将该词先转换成另一规范词，再进行检索。

自然语言信息检索系统与受控语言信息检索系统相比有明显的不同。受控语言信息检索系统是在文献信息和用户信息需求输入系统之前进行控制，控制的工具是人工编制的词表或分类表，而且需要对检索的课题进行主观的思考和分析，

检索结果的优劣在很大程度上取决于用户对规范化词表或分类表的掌握程度及经验技巧。文献信息的输入（前控）和检索提问的输入（后控）都采用同一词表，前控和后控的程度相等。但自然语言检索系统主要是在输入系统之后，在系统内部进行控制，文献信息输入时基本上不做处理，而主要依赖后控，即将自然语言转换为系统的提问，并对有同义、近义、相关等关系的词进行组织。同时，自然语言检索系统与计算机自动标引和自动分词等技术紧密地联系在一起。

自然语言在信息检索中的应用主要表现为使用关键词的全文检索。

全文检索是指不经过任何标引，而直接通过计算机以自然语言的形式在文本中进行匹配查找。文本中任何字符和字符串均可作为检索入口。因此，全文检索是一种不依赖叙词表而直接使用自由词的检索方法。全文检索具有直观性、详尽性和广泛的适应性等特点，采用全文检索技术的检索系统一般具有如下优势：(1) 方便易用。全文检索是采用用户所熟悉的自然语言进行检索，用户使用检索系统之前一般不需要进行培训和学习。(2) 查全率高。全文检索系统可以对文本中任意字符都进行匹配检索，不受标引限制，因此，可以实现较高查全率，并且用户可以直接查看文本的任何章节、段落、句子、词或字，而不只是索引或摘要。(3) 检索功能强大。一般的全文检索系统都能进行布尔检索、截词检索、位置检索、相关检索等，能够满足用户不同的检索需求。

正是全文检索系统的强大优势，使得其已成为广大用户喜爱的检索方式。国外著名的全文检索系统有：ProQuest 系列数据库、EBSCO 数据库、SpringerLink 全文数据库、AltaVista 全文搜索引擎等。国内比较著名的全文检索系统主要有：《中国期刊全文数据库》、《中文科技期刊数据库》、万方数据资源系统等。

【案例】

《中国图书馆分类法》部分类目

G27 档案学、档案事业
- G270 档案学
  - G270.7 档案工作自动化
  - G270.9 档案学史
- G271 档案管理
  - G271.2 组织机构
  - G271.3 规章制度
  - G271.4 统计方法
  - G271.6 档案工作者

G272　收集和整理
　G272.2　收集
　G272.3　鉴定
　G272.4　修复与整理
　G272.5　分类法与主题法
　G272.6　编目
G273　保管和利用
　G273.2　保管与典藏
　G273.3　保护
　G273.5　流通、利用
G274　公布、出版
G275　各种类型档案工作
　G275.1　历史档案
　G275.2　文书档案
　G275.3　技术档案
　G275.9　其他档案
G276　特种档案工作
G277　文书工作
G278　建筑和设备
G279　世界各国档案事业
　G279.1　世界
　G279.2　中国
　　G279.20　方针政策及其阐述
　　G279.21　档案事业组织与活动
　　G279.27　地方档案事业
　　G279.29　档案事业史
　G279.3/.7　各国

关键术语

检索语言	概念逻辑	知识分类	术语学
分类检索语言	体系分类法	组配分类法	主题检索语言
标题词法	单元词法	关键词法	叙词法
分类主题一体化检索语言	网络环境	辅助词表	自然语言

思考题

1. 什么是检索语言？
2. 简述检索语言与概念逻辑、知识分类和术语学的关系。
3. 检索语言的功能有哪些？
4. 检索语言可以分为哪些类型？
5. 试述分类检索语言的特点和类型。
6. 简述分类检索语言的结构。说明类目之间的关系及表现形式。
7. 什么是标记符号？标记制度有哪些？
8. 试述引用次序在分类体系建立中有什么作用。
9. 常用的分类法有哪些？重点介绍其中两种。
10. 试述主题检索语言的特点和类型。
11. 分别简述单元词法、标题词法和关键词法。
12. 叙词法有什么特点？
13. 比较《汉语主题词表》和《中国分类主题词表》。
14. 试述网络环境下分类语言的新特点。
15. 简述网络环境下主题语言的新发展。
16. 什么是辅助词表？辅助词表的主要类型有哪些？
17. 与人工语言相比，自然语言在信息检索系统中的应用有什么优点？
18. 什么是全文检索？全文检索的主要优势是什么？

CHAPTER THREE

第3章 信息著录和标引

【本章要点】

◇介绍信息著录和标引的含义与作用
◇叙述信息著录和标引的发展
◇介绍元数据的定义与作用
◇比较机读目录MARC和都柏林核心元数据集DC
◇介绍网络信息描述自动处理方法——自动标引与自动分类

引子

信息著录和标引是使信息序化的过程。它通过揭示信息的内在以及外在特征，将分散无序的信息重组，规范控制信息流向，以便用户有效利用。传统的信息著录和标引着重关注由人工对传统文献信息源进行描述，如图书、期刊、报纸、特种文献等。互联网的发展造成信息资源海量生产和无序分布。为了更有效地利用网络信息，人们开始采用元数据标准以描述网络资源，并开发出海量信息的自动处理技术。

3.1 信息著录的含义和标准

3.1.1 信息著录的含义和作用

3.1.1.1 信息著录的含义

信息著录简称著录，是指在组织检索系统时对文献内容和形式特征进行选择和记录的过程。信息著录是组织检索系统的基础，是信息存储过程中的一个重要环节。

著录的对象是信息，包括图书、期刊、文件、网络资源等等。信息著录的结果是款目或称记录。款目是指依据一定的标准和方法，对一种文献或一种信息源的内容、价值和物质形态进行描述而形成的一条记录。款目由一条条著录项目组成。著录项目是用于揭示文献内容和形式特征的记录事项。比如，我国国家标准《文献著录总则》规定了九大著录项目，依次为：题名和责任者项、版本项、文献特殊细节项、出版发行项、载体形态项、从编项、附注项、文献标准编号及有关记载项、提要项。每个大项又包括若干个小项。都柏林核心元数据集涵盖了15个元素，包括题名（Title）、创作者（Creator）、主题及关键词（Subject and Keywords）、描述（Description）、出版者（Publisher）、其他贡献者（Contributor）、时间（Date）、类型（Type）、格式（Format）、标识（Identifier）、来源（Source）、语言（Language）、关联（Relation）、范围（Coverage）、版权（Rights）。

准确性和规范化是信息著录的基本要求。准确性要求著录结果要全面、客观、准确地揭示文献或其他信息源的内容特征和形式特征。规范化要求信息著录坚持标准化著录原则，按照统一的著录项目、著录格式、标识符号等进行著录。信息著录的质量在很大程度上决定着输入信息的质量，直接影响到检索效果。比如，在著录的过程中，提供的著录项目不完整，内容不准确，会影响信息检索系统的质量，造成对某些信息的误检和漏检。尤其是在计算机检索过程中，著录直接影响到数据库数据的质量，著录过程中一个字符或一个空格的多少，都有可能对检索产生负面的影响。

3.1.1.2 信息著录的作用

信息著录的目的是为了报道和检索信息，通过著录可以浓缩文献信息的特

征，起到揭示文献、报道文献，帮助人们快速地了解文献，进而选择自己所需文献的作用。信息著录的作用具体如下：

1. 揭示功能

信息著录主要反映的是文献本身所具有的特征，在对文献全面系统分析、选出最具有代表性的特征后，通过概括而精练地叙述内容特征，以及简略而准确地描述形式特征等，将每一种文献的主要信息浓缩于只字片言中，使读者无须找到具体文献，就可以方便地了解文献的基本信息，如外表特征、内容主题等，并以此来决定对原始文献的取舍。①

2. 组织功能

信息著录之后所形成的一个个款目或记录，是编制目录的基础，也是组织数据库数据的基本单元。文献编目工作包括信息著录和目录组织两个步骤，信息著录从分析文献的内容特征和外表特征开始，到记录下各种与文献报道和检索有关的信息为止，经过一系列工序、采用多种方式与手段，最终形成记载文献相关信息的款目或记录。目录组织则是将这些款目或记录按照一定的组织规则编排在一起，最终形成相应的检索工具或数据库。信息著录的质量直接影响着目录组织工作的效率，也对目录的质量有着重要的影响。

3. 检索功能

存储是信息检索的第一个阶段，也就是将表达文献特征的具有检索意义的标识加以记录并组织起来形成手工检索工具或计算机数据库，而所有表达文献特征的标识需要通过著录和标引来完成。作为信息著录结果的款目或记录，记载了反映文献特征的可供检索的各个标识。例如，从计算机检索来看，信息的查找是通过对数据库的搜索实现的。建立数据库是信息存储的主要形式，数据库是各种数据的逻辑集合，数据是构成数据库的基本元素，而一个个数据记录则是通过著录形成的，如果没有这些数据，计算机数据库的检索也就成了无源之水、无本之木。

3.1.2 信息著录的标准

信息著录标准是指在描述信息过程中所要依据的规则和条例，是实现信息著录标准化的前提和根本。信息著录标准包括国际标准和国家标准。

3.1.2.1 信息著录标准的发展

信息著录标准一直受到信息工作者的重视，世界各国都为信息著录标准的编

---

① 参见蒋新：《“文献著录”新议》，载《江苏图书馆学报》，1994（6）。

制和推广付出了多年的努力和积极的探索。早在16世纪，西方一些国家的图书馆学者们就曾讨论过有关集中编目和统一著录的必要性。1969年国际图书馆协会与机构联合会在丹麦哥本哈根主持召开了“国际编目专家会议”，着重讨论了制定国际书目著录标准的必要性和可能性问题，并最后达成协议，决定成立一个工作小组负责起草供单行出版物著录用的国际标准规则。1974年3月正式出版了《国际标准书目著录》（*International Standard and Bibliographical Description*，简称ISBD）条例，即ISBD（M）。70年代前半期，开始制定供连续出版物（Serials）、地图资料（Cartographic Materials）和非书资料（Nonbook Materials）著录用的标准，即ISBD（S）、ISBD（CM）、ISBD（NBM）。另外，还出现了其他一些标准。美国国会图书馆以此条例编制了机读目录，在国际上产生了广泛影响，从而使文献著录标准化工作取得了重大进展。

在我国，文献著录标准化工作也取得了一定的进展。1979年“全国文献工作标准化技术委员会”正式成立，负责文献著录标准化的研究工作。经过几年的努力，相继颁布了与《国际标准书目著录》（ISBD）基本一致的《文献著录总则》（GB3972.1—83）、《普通图书著录规则》（GB3792.2—85）、《连续出版物著录规则》（GB3792.3—85）、《非书资料著录规则》（GB3792.4—85）等国家标准。1985年中国图书馆学会编印了《西文文献著录条例》。1984年图书统一编目款目卡开始采用《普通图书著录规则》，上海地区于同年开展“随书配卡”工作，实行了中文图书统一编目。1996年颁布了《中国机读目录格式》，为我国机读目录制定了标准，同年又对我国各类型的文献著录标准做了进一步修订，并编写了《中国文献编目规则》等等。自1983年颁布《文献著录总则》至今，我国已正式推行和发布多种信息著录标准及相关标准和规则，如《国际标准书目著录》、《UNIMARC格式和手册》、《规范数据款目著录规则》、《普通图书著录规则》、《文献主题标引规则》、《中国机读规范格式（试用本）》等。

信息著录标准的制定和颁布，为信息著录的标准化奠定了一定的基础，同时也为数据库资源共享创造了最基本的条件。信息著录标准化进程中，近年来受到广泛关注的热点之一，是网络信息资源的描述问题，亦即元数据的记录问题。与传统信息资源相比，网络资源在资源类型、结构、形式、描述环境、描述主体等方面存在着不同，其描述规范也因此呈现多样化。

3.1.2.2 《文献著录总则》

1983年7月，我国正式颁布了《文献著录总则》，旨在根据各种类型文献的共同特点，确定文献著录原则、内容、标识符号、格式等的统一规定。具有指导作用，为信息著录提供原则性的框架，并不作为文献著录的直接依据。不同类型

文献的著录在依据总则的基础上，制定有相应的具体条文，作为文献著录的直接依据，如《普通图书著录规则》、《连续出版物著录规则》、《非书资料著录规则》、《档案著录规则》、《古籍著录规则》、《地图资料著录规则》、《检索期刊条目著录规则》、《文后参考文献著录规则》等。它们共同构成了我国比较完备的著录标准体系。

1. 著录项目

《文献著录总则》依据《国际标准书目著录》（ISBD），规定了九大著录项目，每个大项下又设置了若干个子项。

（1）题名与责任者项。

包含6个子项：正题名、并列题名、副题名及说明题名、文献类型标识、第一责任者、其他责任者。

正题名是文献的主要题名，著录时按照规定信息源所载题名如实照录。并列题名指同一种文献具有两个或两个以上不同文字的对照题名，未被选作正题名的其他文种题名就作为并列题名著录于正题名之后。副题名一般指对正题名做解释的书名。文献类型标识表示著录文献的所属类别。第一责任者和其他责任者是对文献作者的描述。

（2）版本项。

包含版次及版本形式、与本版有关的责任者两个子项。版次及版本形式指文献排版次数，如第2版、第3版、修订版等。它的不同意味着文献在内容上经过修订后重新出版，说明文献的内容和形式发生了变化。

（3）文献特殊细节项。

该项未设子项。主要是著录连续出版物的卷期起讫、图的比例尺和投影法，或是其他文献的特殊问题。

（4）出版发行项。

包含：出版地或发行地、出版者或发行者、出版日期或发行日期、印刷地、印刷者、印刷日期。

（5）载体形态项。

包含的子项有：数量及其单位、图及其他形态、尺寸或开本、附件。附件指在形态上与文献主体部分分离，但必须与文献主体结合使用，并作为一个整体入藏。如随书的光盘等。

（6）从编项。

包含的子项有：正丛编名、并列丛编名、副丛编名及说明丛编名文字、丛编责任者、国际标准连续出版物编号（ISSN）、丛编编号、附属丛编。

(7) 附注项。

该项未设子项，主要是对描述文献形式的著录正文进行补充和说明。

(8) 文献标准编号及有关记载项。

包含4个子项：国际标准文献编号（ISBN）、中国标准文献编号、装订、价格。

(9) 提要项。

没有子项，主要是对文献的内容进行简要的介绍。

2. 著录级次

著录级次指著录文献的详简程度。《文献著录总则》把著录项目分为主要项目和选择项目两种。主要项目包括：题名和责任者项的正题名、第一责任者，版本项，出版发行项的出版发行地、出版发行者、出版发行日期，载体形态项。选择项目包括主要项目之外的所有其余著录项目。

按著录的详简程度分三级：(1) 简要级次：款目仅著录主要项目，又称第一著录级次。(2) 基本级次：著录主要项目的同时，还著录了部分选择项目，也称第二著录级次。(3) 详细级次：著录主要项目和全部的选择项目，也称第三著录级次。

3. 著录格式

著录格式是指款目中各个著录项目的排列次序和表达方式。分为卡片式款目著录格式和书本式款目著录格式两种，具体的著录格式如图3—1和图3—2：

正题名＝并列题名：副题名及说明题名文字［文献类型标识］/第一责任者；其他责任者．—版次及其他版本形式/与本版有关的责任者．—出版发行地：出版发行者，出版发行日期（印刷地：印刷者，印刷日期）

页数或卷册数：图；尺寸或开本＋附件．—（丛编名/责任者，国际标准连续出版物编号；丛书编号．附属丛编）

附注

国际标准编号；中国文献标准编号（装订）：价格

提要

Ⅰ．书名 Ⅱ．著者 Ⅲ．主题 Ⅳ．分类号

图3—1 卡片式款目著录格式

正题名＝并列题名：副题名及说明题名文字［文献类型标识］/第一责任者；其他责任者．—版次及其他版本形式/与本版有关的责任者．—出版发行地：出版发行者，出版发行日期（印刷地：印刷者，印刷日期）．—数量及其单位：图及其他形态；尺寸或开本＋附件．—（丛编名/责任者，国际标准连续出版物编号；丛书编号．附属丛编）．—附注．—国际标准编号；中国文献标准编号（装订）：价格

提要

图 3—2　书本式款目著录格式

目前，在多数信息机构并存着两种目录，一种是传统的卡片式目录，如书名目录、著者目录、分类目录，另一种是机读目录。在 20 世纪八九十年代，中国信息机构开始从传统的卡片式目录向机读目录过渡。例如，北京大学图书馆的卡片式目录的收录时间跨度为：自北京大学图书馆的前身京师大学堂藏书楼成立至 1995 年 9 月 5 日。对于 1990—1995 年 9 月 5 日期间入馆的图书，既建立了机读目录数据，又制作了传统的卡片目录。自 1995 年 9 月 5 日起，不再制作传统的卡片目录，只建立机读目录数据，供读者通过计算机进行网上查询。

3.2　机读目录与元数据

3.2.1　机读目录

机读目录（MARC）即机器可读目录的简称，来自英文 Machine-Readable Catalogue，是利用计算机识读和处理的目录。机读目录是描述文献著录项目的国际标准格式，是实现计算机处理书目信息及资源共享的基础。

机读目录最早产生于美国，1963 年，美国 G. W. 金等人发表关于美国国会图书馆书目系统自动化的报告。1966 年 1 月，产生了《标准机器可读目录款式的建议》，制定了 MARC—1 格式。1967 年经过调整和改进，推出了 MARC—2 格式。1968 年 7 月开始了正式的 MARC 计划，1969 年 3 月向全国发行 MARC—2 格式的英文图书机读目录磁带，称为美国机读目录格式 USMARC。1977 年，为了进一步协调、促进国际交流，统一各国机读目录格式，国际图书馆联合会在 USMARC 基础上主持制订了“国际机读目录通信格式”，即 UNIMARC。到 20

世纪80年代末，英国、德国、法国、加拿大、丹麦、意大利、挪威、瑞典、澳大利亚、日本以及拉丁美洲和非洲的一些国家，共20多个国家和地区进行了机读目录的研究和开发，建立了机读目录系统，生产和发行机读目录产品。中国机读目录研制于20世纪70年代，1979年成立了北京地区机读目录研制小组，依据UNIMARC格式和《文献目录信息交换磁带格式》(GB2901-82)，结合我国实际情况，编制了《中国机读目录通讯格式》讨论稿，1992年正式出版了《中国机读目录通讯格式》，即CNMARC。

机读目录的出现，有力地推动了信息机构文献工作的自动化和标准化。就目前国内具体使用情况来看，关于中文文献的著录主要采用的是CNMARC，以共享我国图书馆和信息部门的中文书目记录；西文文献的著录则采用USMARC，以共享国外权威的西文书目记录。这里重点介绍CNMARC。

CNMARC（China Machine-Readable Catalogue）按照UNIMARC格式设计原则制定，并结合了汉字的特点，标识系统和数据代码规定比较详细，目前广泛地应用于计算机编目。

一条CNMARC记录由记录头标区、地址目次区、数据字段区、记录分隔符组成。记录结构如表3—1所示：

表3—1　　CNMARC记录结构

记录头标区	地址目次区	数据字段区	记录分隔符

1. 记录头标区

按ISO2709规定，每条记录以24位字符长的记录头标开始。它包括的数据有记录类型、书目级别、记录的完备程度，以及记录是否遵照国际标准书目著录(ISBD)规则等。记录头标区包括的数据元素如表3—2所示：

表3—2　　CNMARC头标区数据元素

数据元素名称	字符数	字符位置	说　明
记录长度	5	0～4	记录长度用整个记录的字符数表示，包括记录头标区、地址目次区和全部数据字段，由计算机自动生成。
记录状态	1	5	反映记录处理状态。CNMARC共定义了5种记录状态： c=更正过的记录； d=删除的记录； n=新记录；

续前表

数据元素名称		字符数	字符位置	说　明
				o＝已有较高层次的记录； p＝曾为不完整的预编记录。
执行代码	记录类型	1	6	反映记录类型。具体定义如下： a＝印刷型文字资料； b＝文字资料（手稿）； c＝乐谱（印刷型）； d＝乐谱（手稿型）； e＝测绘制图资料（印刷型）； f＝测绘制图资料（手稿型）； g＝放映和视听资料（电影、胶卷、幻灯片、录像）； i＝非音乐录音资料； j＝音乐录音资料； k＝二维图像（图片、设计图等）； l＝计算机载体； m＝多媒体资料； o＝多载体配套资料； p＝混合型资料； r＝三维制品和实观教具； u＝拓片； v＝善本书。
	书目级别	1	7	说明书目的级别。具体如下： a＝分析级（组成部分）； c＝合集； m＝专著； s＝连续出版物。
	层次等级代码	1	8	表示控制类型。如下： ＃＝未定义层次等级关系； 0＝无层次等级关系； 1＝最高层次的记录； 2＝低于最高层次的记录。
	未定义	1	9	
指示符长度		1	10	由1位十进制数表示，一般取值为2。由计算机自动生成。
子字段标识符长度		1	11	由1位十进制数表示，一般取值为2。由计算机自动生成。
数据基地址		5	12～16	指示第一个数据字段相对于记录开头的起始字符位。数据基地址由系统自动生成。

续前表

数据元素名称		字符数	字符位置	说 明
记录附加定义	编目等级	1	17	说明书目记录内的书目信息或内容标识的完整程度。具体如下： #＝完全级； 1 ＝次级 1（表示编制机读记录时未查对原文献）； 2 ＝次级 2（表示该记录是在版编目记录）； 3 ＝次级 3（表示记录的项目不全）。
	著录编目格式	1	18	反映记录所采用的著录编目格式，指明著录字段200～225字段是否遵守国际标准书目著录（ISBD）规则。 #＝完全采用 ISBD 格式； n ＝非 ISBD 格式； i ＝部分采用 ISBD 格式。
	未定义	1	19	
记录目次区		4	20～23	数据字段长度（20）一般取值为 4； 起始字符位置（21）一般取值为 5； 执行定义部分（22）一般取值为 0； 未定义字段（23）为“#”。

2. 地址目次区

是关于该记录数据字段区记录情况的有关数据，包括字段号、字段长度、起始字符位置等。

3. 数据字段区

由一些可变长数据组成。数据字段区的功能块有 10 个，每个功能块设立了一些相关字段，具体如表 3—3 所示：

表 3—3 CNMARC数据字段区功能块

功 能 块	含 义	定 义 的 主 要 字 段
0——标识信息块	标识记录或编目文献的标识号	001 记录控制号 005 记录版本标识 010 国际标准书号（ISBN） 011 国际标准连续出版物号（ISSN） 020 国家书目号 021 版权登记号 022 政府出版物号 091 统一书刊号

续前表

功能块	含义	定义的主要字段
1——编码信息块	记录定长编码数据元素	100 一般数据处理 101 作品语种 102 出版国别 105 专著编码数据 106 文字形态特征 110 连续出版物编码数据
2——著录信息块	包括《国际标准书目著录》（ISBD）所规定的主要著录项目（附注项和国际标准书号除外）	200 题名与责任说明项 205 版本项 210 发行项 215 载体形态项 225 丛编项
3——附注项	著录自由行文附注。主要是针对著录项目、检索点以及文献形式与内容的各个方面进行说明	300 一般性附注 304 题名与责任者附注 305 版本沿革附注 306 出版发行项附注 307 载体形态项附注 310 装订与获得方式附注 320 书目附注 327 内容附注 328 学位论文附注 330 提要或文摘
4——款目连接块	揭示相关书目之间的关系	410 丛编 411 附属丛编 421 补编（增刊） 422 正编（正刊） 423 合订（合刊） 430 继承 431 部分继承 432 替代 433 部分替代 434 吸收 435 部分吸收 436 合并 437 分自 453 译为 454 译自

续前表

功能块	含义	定义的主要字段
5——相关题名块	记录文献正题名以外与编目文献有关的题名	500 统一题名 501 作品集统一题名 503 统一惯用标目 510 并列题名 512 封面题名 513 附加题名页题名 515 逐页题名 516 书脊题名 517 其他题名 540 编目员补充的附加题名 541 编目员补充的翻译题名 545 章节题名
6——主题分析块	记录包括主题法和分类法等各种不同体系构成的主题数据	600 个人名称主题 601 团体名称主题 602 家族名称主题 604 作者/题名主题 605 题名主题 606 普通主题 607 地理名称主题 610 非控制主题词 660 地区代码 675《国际十进制分类法》分类号（UDC） 676《杜威十进分类法》分类号（DDC） 680《美国国会图书馆分类法》分类号（LC） 686 其他分类号 690《中国图书馆分类法》分类号 692《中国科学院图书馆图书分类法》分类号
7——知识责任者块	著录对书目记录所描述的文献的产生负有某种形式和知识责任的个人和团体的名称	700 个人名称——主要责任者 701 个人名称——等同责任者 702 个人名称——其他责任者 710 团体名称——主要责任者 711 团体名称——等同责任者 712 团体名称——其他责任者 720 家族名称——主要责任者 721 家族名称——等同责任者 722 家族名称——其他责任者
8——国际使用块	包括书目记录的来源等	801 数据来源
9——国内使用块	包括馆藏信息等	905 信息

3.2.2 元数据

元数据的英文为 Metadata，意为关于数据的数据。在互联网中，元数据是指描述任何互联网数据和资源，促进互联网信息资源的组织和发现的数据，以协助对网络资源的识别、描述、位置指示。

3.2.2.1 元数据的作用

元数据具有描述、定位、搜寻、评估、选择等多种功用，可以连贯有效地描述、管理、编目网络资源，以便用户更方便地找到资源，并找到更多的相关资源。具体作用表现为：

（1）定位和检索。借助于元数据，人们可以准确地检索和确认所需要的资源。可以说，这种作用是推动元数据发展的最重要的力量。

（2）著录和描述。为了实现高的查全率和查准率，需要对网络资源的数据单元进行详细、全面的著录和描述，描述数据单元的元数据名称叫做元数据元素（Metadata Element），包括内容、载体、位置与获取方式、制作与利用方法等多个方面。

（3）资源管理。利用元数据全面地描述网络资源，不仅有利于检索，同时也有利于实现对资源有效、安全的管理，这些元数据元素包括权利管理（Rights/Privacy Management）、数字签名（Digital Signature）、资源评鉴（Seal of Approval/Rating）、存取管理（Access Management）、支付审计（Payment and Accounting）等方面的信息。

（4）资源保护与长期保存。利用元数据全面地描述网络资源，不仅有利于现实的管理和查询，还有助于网络资源的长期的历史保护，这些元数据元素包括详细的格式信息、制作信息、保护条件、转换方式、保存责任等。

鉴于元数据的上述作用，如果对于网络上所有资源（网站、网页、文档、服务）都用相同的元数据元素进行描述，对每个网络资源形成一条由这些元数据元素组成的元数据记录，将这些元数据记录集中管理起来，那么将在很大程度上较好地解决网络资源的可检索性、可管理性、可交换性等问题。因此，很多国家和地区都在致力于元数据标准的制定与完善。其中影响最为深远、使用最为广泛的是国际标准都柏林核心元数据集（Dublin Core Elements Set）。

3.2.2.2 都柏林核心元数据集（Dublin Core Elements Set）

都柏林核心元数据集是一种跨领域的信息资源描述标准。这里的信息资源被定义为“任何具有标识的东西”。都柏林核心元数据应用的资源类型没有根本性

的限制。

1995年，在美国俄亥俄州的都柏林召开了由OCLC和美国超级计算机应用中心主持的第一届元数据研讨会，与会代表来自信息管理和信息技术领域，他们一致认为有必要产生一个简单的描述网络上文件类对象（DLO）资源的元数据集，并最终产生了一个包括13个元素的都柏林核心元数据集。此后历届都柏林核心元数据研讨会的理论探讨和实际应用，推进了都柏林核心元数据集走向成熟，在原有13个元素的基础上，又增加了范围和版权两个元素。目前，都柏林核心元数据集共包括15个元素。具体如表3—4所示：

表3—4　　都柏林核心元数据集

元数据元素	标识	定义	解释
题名（Title）	Title	赋予资源的名称	使资源为众所周知的有代表性的正规名称
创作者（Creator）	Creator	创建资源内容的主要责任者	创作者包括个人、组织或机构
主题及关键词（Subject and Keywords）	Subject	资源内容的主题描述	用以描述资源主要内容的关键词语或分类号码表示的有代表性的主题词
描述（Description）	Description	有关资源内容的说明	该说明可以包括但并不限于：摘要、内容目次、内容图示或内容的文字说明
出版者（Publisher）	Publisher	使资源以现有形式被获得和利用的责任者	如包括个人、组织或机构的出版者
其他贡献者（Contributor）	Contributor	对资源内容负有责任的实体	贡献者包括个人、组织或机构
时间（Date）	Date	与资源生命周期中的一个事件相关的时间	资源产生或有效使用的日期、时间。推荐使用ISO 8601［W3CDFT］定义的编码形式，使用的是YYYY—MM—DD形式
类型（Type）	Type	资源内容方面的特征和类型	类型包括种类、功能、体裁或作品集成级别等描述性术语。推荐从可控词表（如Dublin Core Types［DCT1］）中选用有关术语。对于资源物理或数字化方面表示，采用“格式”项描述
格式（Format）	Format	资源物理或数字化的表现形式	格式可包括媒体类型或资源容量。也可用于限定资源显示或操作所需的软件、硬件或其他设备，如容量包括数据所占空间和存在期间

续前表

元数据元素	标识	定义	解释
标识（Identifier）	Identifier	依据有关规定分配给资源的标识性信息	推荐对资源的标识采用符合正式标识体系的字符串和数字组合。如正规标识系统包括统一资源标识（URI）、统一资源地址（URL）、数字对象标识（DOI）以及国际标准书号（ISBN）、国际标准刊号（ISSN）等
来源（Source）	Source	可获取现存资源的有关信息	当前资源可能部分或全部源于来源元素所标识的资源。建议使用正规标识系统确定的字符或号码标引资源来源信息
语言（Language）	Language	资源知识内容使用的语种	推荐使用由 RFC1766 定义的语种代码，它由两位字符（源自 ISO639）组成。作为可选项，可选用两字符的国家代码（源自 ISO 3166）。如“en”表示英语，“fr”表示法语
关联（Relation）	Relation	对相关资源的参照	推荐对关联的标识采用符合正式标识体系的字符串和数字组合
范围（Coverage）	Coverage	资源内容的领域或范围	范围包括空间定位（地名或地理坐标）、时代（年代、日期或日期范围）或权限范围
版权（Rights）	Rights	持有或拥有该资源权利的信息	版权项包括资源版权管理的说明，或者是对这一信息的服务的参照。版权信息通常包含知识产权、版权或其他各种各样的产权

都柏林核心元数据集中的 15 个元素都是可选择、可重复和可扩展的。也就是说，不同国家、地区、行业、文件类型在应用时可以根据需要挑选其中的部分和全部元数据元素，也可以增加其他必要的元数据元素。目前世界上有很多国家和部门都将都柏林核心元数据集作为一项基础标准。我国图书馆学界较早认识到元数据标准的重要性，较早开发元数据标准的领域。1997 年，中山市图书馆开始了“数字式中文全文文献通用格式”的研究，该标准全部采纳了国际标准都柏林核心元数据集的 15 个元数据项目，并增加了记录控制号（Record），共 16 个元数据项目。[①] 1997—2000 年开展的国家重点科技项目“中国实验型数字式图书馆”，强调了要采用国际标准，其中包括都柏林核心元数据集。

① 参见莫少强：《数字图书馆元数据和资源共享的研究与实践》，载《图书情报工作》，2002（1）。

3.2.3 都柏林核心元数据和机读目录的比较

在网络环境下，都柏林核心元数据以其简单、灵活、具有语义互操作性和可扩展性等优点，在网络信息资源的描述和著录中表现出强劲的势头。传统的机读目录格式以其揭示内容深入、详尽，并在信息存储和检索领域应用历史已久的优势，依然保持着不可动摇的地位，是一种国际性的书目著录标准。

就某种意义上说，机读目录和都柏林核心元数据都是元数据，在著录文献的相关信息上，都是数据的数据，目的均是将文献的相关信息格式化。它们都是用来描述信息资源的主题、内容特征，并通过所著录的信息来提供检索的依据。都柏林核心元数据和机读目录的区别和联系主要表现在以下几个方面：

1. 著录的对象不同

都柏林核心元数据的著录对象是网络资源或数字资源，它的设计原则中具有可扩展性、可选择性、可重复性和可修饰性的特征，有利于揭示各种类型的数字资源的内容和其他特征。机读目录格式比较适用于传统的出版物、图像、缩微制品、视听资料、数据库等。随着 USMARC 中的 856 字段的引入，也可用于描述电子资源。机读目录的使用环境主要限于图书情报机构和网上的公共查询目录。

2. 数据的形式不同

都柏林核心元数据集包括 15 个元素，它在应用中是可以选择、可重复和可扩充的。限定词与元素之间的关系是不确定的，限定词使用非常灵活，结构较为简单、灵巧。MARC 格式主要由头标区、目次区、数据区 3 部分组成。数据结构比较严谨，比较复杂。

3. 著录的主体不同

都柏林核心元数据著录简单明了、语义明确，它使创建者和信息提供者可以无需经过培训就能自己进行资源描述。机读目录是一种复杂的、详细描述的格式，对内容著录的规定严格，对使用者要求高，只有经过培训的专业编目人员才能使用。

4. 著录的详简程度不同

都柏林核心元数据的著录相对比较简单，只有 15 个元数据元素，在信息描述过程中，可以任意选用它们，也可以重复使用，顺序可以任意编排，还可以根据具体情况进行某些补充。而机读目录的编目规则追求的是详尽、细致的著录，有严格的著录规则。

5. 标识的方法不同

都柏林核心元数据直接采用单词或词组的形式作为标识，表达直观，语义明

确。机读目录的字段采用3位阿拉伯数字作为标识，子字段采用一位英文字母或阿拉伯数字作为标识，标识不具备语义。

3.3 信息标引的含义和步骤

3.3.1 信息标引的含义和质量控制

3.3.1.1 信息标引的含义

信息标引是指在分析文献内容的基础上，用某种检索语言把文献主题以及其他有意义的特征标识出来，它是文献存储与检索依据的一种文献处理过程。简单来说，标引就是依据检索语言，确定文献标识的过程。对纸质文献而言，信息标引工作主要是通过对文献主题的分析，手工赋予文献分类号、主题词、关键词、人名、地名等标识，以便编排检索工具和提供用户检索。随着计算机在信息检索领域的广泛使用，自动标引成为一种趋势。

标引具有十分重要的作用，在检索系统中处于承上启下的地位。它既是文献存入检索系统的依据，又是从检索系统中查出文献的依据，是整个检索系统的关键之一。标引目的是为快速、准确地检索所需文献提供方便。G. 佩林（G. Perrin）曾经指出："从文献中获得一些短语有时显得比一些句子更为重要，因为这些短语不光是代表句子和语言文章意思的单元，而且还是人们进行读、写和思考的单元。如人们在检索过程中，通过对几个单词或短语的迅速浏览来决定该文献的重要程度。"①

按照标引的依据可以将信息标引分为分类标引和主题标引。分类标引的目的是揭示文献的内容及形式，以便将同类的文献集中在一起，把不同的文献区分开来，根据文献之间的关系组成一个系统，科学地组织和管理文献。分类标引的作用是编制分类目录和分类索引，组织分类排架，进行分类统计，便于读者进行族性检索。目前，国内大多数信息机构采用《中国图书馆分类法》作为分类标引依据。主题标引则按照文献的主题内容，在主题词表中选取符合其内容的主题词，对文献进行标识。主题标引的意义在于它是一种以主题词作为文献主题标识和查找依据的检索方法，是当前国内外普遍采用的方法，特别是机检用得更多。检索时只要找到某一个主题词就可以获得有关该主题的全部文献，它能较好地满足特

① 转引自顾敏、史丽萍、李春玲：《自动标引综述》，载《黑龙江水专学报》，2002（3）。

性检索的需要，有利于文献情报工作的自动化、现代化和网络化。

3.3.1.2 信息标引的质量控制

信息标引质量的优劣，直接影响到计算机的查全率和查准率，影响到用户利用检索系统的效率。尤其是随着大型数据库的开发和利用，对于海量的数据，必须组织大规模的集体标引才可能完成任务。因而，关于信息标引的质量控制就显得尤为重要。影响标引质量的因素有许多，包括技术因素和管理因素等。控制标引质量应考虑以下因素：

1. 标引深度

标引深度即标引全面性，指把一篇文献所论述的各个主题内容提炼出来，给出检索词并对其进行标引的完善程度。它通常指一篇文献被赋予主题词的平均数量。例如，CALIS（全国高等教育文献保障系统）联合目录规定：在一般情况下，标引深度不超过10个主题词。标引深度是从对文献主题内容揭示的广度来衡量标引质量的一个因素。因此，标引越充分，检索点越多，也就越能较全面地揭示文献的主题，从而提高查全率。在主题标引过程中，尤其是对于多主题的文献，不能仅从题名分析进行拆分与组合，而应该全面了解文献的内容，以便准确地揭示文献的显性主题和隐性主题，把每个主题及其各个方面都反映出来。只有这样，才能全面客观地进行主题标引。例如，《法国重农派学说的中国渊源》一书，是一本研究中国近代思想史方面的著作，因而在标引时应同时标引出“重农学派—影响—中国”和“经济思想史—近代—中国”。

2. 专指度

标引的专指度是指检索标识表达信息内容的精确程度。它是以对主题概念揭示的精确度来衡量标引质量的一个因素。标引的全面性与准确性是相辅相成的，准确性的直接后果就是查准率的高低，在准确基础之上的全面性才是有意义的。在主题标引过程中应选用恰当的、具有较高专指性的主题词去揭示主题内容。在分类标引中应将文献归入最专指、最切题的类目。

3. 一致性

一致性是指选用表达文献主题内容所需标引词的一致程度。在进行主题标引时，由于标引人员的更换或多人同时标引，产生对同一主题内容的文献标引不一致的情况，会使同一主题内容的文献分别标引成若干的主题，且由于标引人员对文献内容认识深度的不同，从而导致主题标引词过多或过少，进而也就影响查全率和查准率。

3.3.2 信息标引的步骤

无论是分类标引还是主题标引，信息标引的步骤都包括主题分析和概念转换。也就是首先要对文献的内容进行分析，明确文献的主要内容和主题概念，然后用分类法或主题法将其充分、准确地表达出来。

3.3.2.1 主题分析

主题是用以表达文献所论述和研究的具体对象和问题。主题分析即要弄清楚文献讨论的中心思想是什么，以确定被标引文献的主题概念。主题分析的重点是对文章的主题类型、主题要素及其相互关系进行分析。通过分析，明确该文章属单主题类型，还是属多主题类型，再从主题结构上弄清构成主题的要素有哪些，其中关键的主题要素是哪几个。

1. 主题类型和结构

主题类型可以从不同的角度划分：

(1) 依据主题数量的多少可以分为：单主题和多主题。

单主题是指一篇文献只研究一个事物（对象），或一个事物（对象）的一个方面或几个方面，如《信息检索导论》、《情报学概论》均是单主题。在单主题文献中，根据其主题概念的数量和关系，又可分为单元主题和复合主题。只需要一个基本概念就可以概括的主题称为单元主题，如《生物学》；包含两个或两个以上基本概念的主题称为复合主题，如《生物科学发展战略》。

多主题是指同时研究两个或多个独立的事物（对象）。如《广播、电视简明技术手册》、《汽车和拖拉机的维修与使用》这两篇文献就是多主题。

(2) 依据主题的显露程度可以分为：显性主题和隐性主题。

显性主题是指文献明确阐述表达的主题。如《中国经济体制改革》这篇文献明确包括“中国”和“经济体制改革”两个概念。而隐性主题则是指在文章篇名中没有直接用语词加以描述，而是隐含在正文中。例如，《加压素治疗休克引起冠心病》这一文章，除了休克、加压素、冠心病这些直接的主题概念外，还隐含着致病化学因素、药物副作用等主题概念。

主题结构指构成文献主题和各个基本主题的因素以及它们之间的相互关系。主题结构分析的目的是判断复合主题的中心、动态部分和限定部分，用以把握主题的主要成分和次要成分，从而对文献内容所涉及的主题概念进行提炼、精选、取舍，最终确定出检索标识。

按照《文献主题标引规则》提供的主题分析结构模式，我们可以把所有文献主题因素归纳成五个基本方面：主体因素、通用因素、位置因素、时间因素、文

献类型因素。

（1）主体因素。即文献研究和论述的关键性主题概念，包括各种事物、学科、问题、现象等具有独立检索意义的基本概念。

（2）通用因素。是指对主题概念起限定和修饰作用的因素，是一些没有独立检索意义的主题词，如研究、综述、影响、发展、方法、设计等。

（3）位置因素。是指文献研究的事物（对象）所处的地理位置。

（4）时间因素。是指文献研究的事物（对象）所发生的时间。

（5）文献类型因素。如文集、丛刊、年鉴等。一般来说，期刊标引没有这一项。

在上述五种因素中，主体因素是文献论述的关键内容，应作为析取主题概念的重点，其他因素则应视文献论述的具体情形和检索的需要，做出适当的取舍。

在主题结构中，各个基本主题因素之间的相互关系主要表现为这样一些：应用关系、影响关系、从属关系、比较关系、因果关系等。这些关系在分类标引和主题标引过程中应给予正确的处理。

2. 主题分析方法

主题分析是对文献的内容特征和外表特征分析的过程，内容特征是其根本的依据，外表特征是其辅助依据。分类标引和主题标引均是对文献内容所表达的主题进行提炼和概括，这是它们的相同之处。分类标引中的主题分析更多地强调分析文献的内容性质，而主题标引中的主题分析要明确文献所论述的对象事物及其方面问题，并注意其他有检索意义的信息。

主题分析方法一般有两种：

一种是先找出文献论述的对象，再进一步查明是论述了对象哪个方面的具体问题，可以按照事先设计好的主题结构模式提炼相关主题要素，分析主题要素之间的关系。一般来说，主题结构模式可表述为“主体因素—通用因素—空间因素—时间主题—文献类型因素”。例如，《红塔集团跨世纪发展战略思考》一书，其主体因素是“红塔集团”，通用因素是“战略”，因而，分类宜入 F426.89，主题标引应为：烟草工业—发展战略—玉溪地区。

另一种是先找出文献所涉及的各种概念，并查明它们之间的相互关系。文献中包含哪个因素就分析哪个因素，有的因素不止一个就要全部分析出来，以便筛选和进行匹配，然后，再将各个因素按照主题结构模式进行分析。例如《经济全球化与证券经营机构风险管理》一书，涉及的概念有：经济、证券机构、风险管理等，主体因素是“证券机构”，然后通过概念转换，分类号为 F832.51，主题标引为：证券市场—风险管理—研究—中国。

在主题分析过程中一定要客观地全面地反映文献的固有特征，不能进行主观臆断。同时，标引人员还应充分考虑用户的检索需要，分析选定对用户有实际意义的主题概念（包括隐含的主题概念）。

3.3.2.2 概念转换

主题概念转换是以主题分析为基础，将确定的主题概念赋予检索标识的过程。概念转换的结果是形成检索标识。分类标引的概念转换主要依据主题分析的结果，查找分类表，将相应类目的分类号作为检索标识赋予被标引的文献。主题标引的概念转换是依据主题分析的结果，查找主题词表，将相应的主题词作为检索标识赋予被标引的文献。

主题概念转换按其复杂程度可以分为两种：（1）直接转换：这种转换比较简单，标引人员从词表中直接选择与主题概念对应的分类号或主题词即可；（2）分解转换：将复杂的主题概念首先进行分解，然后再选择相应的主题词或分类号。

概念转换结束后，还要进行标引结果的审核，即审核文献的分类或主题检索标识是否正确，包括文献主题分析的正确性、充分性，检索标识的正确性等。只有这样才能保证文献标引的质量。

3.4 分类标引和主题标引

3.4.1 分类标引

文献分类标引是指依据一定的分类检索语言，对文献内容的学科性质及其有检索意义的形式特征进行分析、归纳，赋予文献分类检索标识（分类号）的过程。目前，国内主要依据《中国图书馆分类法》来进行分类标引。

3.4.1.1 基本原则

人们在长期的分类标引实践中，总结出了许多分类的基本原则，具体如下：

1. 学科属性原则

文献分类标引应以文献论述的中心内容的学科属性作为分类的主要标准，以其他形式特征，如：地区、时代、体裁、民族等文献类型作为辅助标准。只有在不适于以学科属性为区分标准时，才考虑以形式特征为辅助标准。这是文献分类标引中最主要的一条基本原则。例如：《集成电路手册》分入 TN4—62，《钢铁是怎样炼成的》（苏联现代长篇小说）分入 I512.45。

2. 专指性原则

文献分类标引必须符合专指的要求。也就是说，要将文献分入最恰当的类，而不能分入范围大于或小于文献实际内容的类目。要区分总论与专论，不要将专论性的文献归入总论类。还要区分是阐述一般原理的，还是阐述具体问题的，不要把研究具体问题的文献归入阐述一般原理的类。例如，《信号处理》是论述信号处理一般原理的著作，入 TN911.7，*Firewalls and Internet Security: Repelling the Wily Hacker*（《防火墙与网络安全：防止黑客侵入》）一书应入 TP393.08，而不入更泛指的类目 TP309。

3. 实用性原则

文献分类标引必须使文献尽其用，即要根据读者的需要将文献分入最大用途的类。对于交叉学科或是内容涉及多个学科的文献，应利用互见分类、分析分类等方法，对重要的分类检索点，予以充分揭示。此时应优选一类号作为主要分类号。例：根据作者的写作目的与其主要服务对象将《最新汉英旅游词典》归入 F59—61；《莫泊桑短篇小说选》是法汉对照读物，依“最大用途”分入法汉对照读物 H329.4：I565.44；《长江流域环境经济发展研究》是交叉学科文献，入 F127.5 互见 X196。

4. 系统性原则

文献分类必须体现分类体系的逻辑性、系统性，凡是归入下位类的书必须具有上位类的属性，体现它们之间的从属关系。也就是说，凡能归入某一类的文献，一定也能归入其上位类。例如，《神经网络》从人体生理角度论述神经网络原理，入 R338；《神经网络原理》从科学计算角度论述人工神经网络原理，入 TP183。

5. 一致性原则

文献分类标引必须遵循一致性的原则，即是说要将内容相同的文献集中归入同一个类目，而不要分散于有关各类。对于个别难以确定类属的主题，可以通过讨论，建立分类规范文档，人为地将其集中到某类，以确保文献分类标引的一致性。《老年经济学》入 F069.9，不入 C913.6—05；《改革学》入 D0，不入 C93—03。

3.4.1.2 各种类型主题文献的分类标引规则

依据文献主题类型的不同，可有不同的分类标引规则：

1. 单主题文献的分类标引规则

（1）简单地对某一事物或问题进行综合研究的文献，应按事物或问题的学科属性归类。例如，《计算机网络基本知识》一书入 TP393。

（2）从某一学科角度论述某主题的文献，应根据研究角度归入有关的学科类

目。例如，《烟草栽培生理》入 S572.01。

（3）从几门学科来综合论述一个主题的文献，依论述该主题的主要学科归类。例如，《中华国宝》多媒体光盘包括“艺术珍宝”、“古建筑”、“风景名胜”、“珍稀动物”和“珍稀植物”五大部分，入 K92。

2. 多主题文献的分类标引规则

对于多主题文献的分类，必须对各主题进行分析，分清主次，然后依其最能体现该文献内容实质的或在内容中起主导作用的主题归类，必要时对另外的主题做附加分类；如果文献所论及的几个主题具有同等的检索意义，则分别标引。此时可选择篇幅较大者或篇幅居前者的类号作为主要分类号。例如，《激光在医学和生物学中的应用》入 R312；《猪鸡鹅鸭快速饲养法》入 S83，互见 S828。

3. 相关关系主题文献的分类标引规则

（1）应用关系的主题是指一个主题应用到另一个或几个主题中，或者几个主题同时应用到一个主题之中。论述一种（或多种）理论、方法、工艺、材料、设备、产品等在某一主题或学科方面的应用的文献，归入应用到的主题或学科所属类目；论述一种理论、方法、技术、材料等在多个主题方面应用的文献，按理论、方法等本身的学科属性归类；某一（些）事物或学科应用到另一个事物或学科，而产生的交叉学科主题的文献，一般归入应用到的领域中的有关类目。如：《计算机辅助机械制造》入机械制造 TH16，《集成电路应用手册》入集成电路 TN4。

（2）影响关系主题是指文献内容的几个主题，一个对另一个或对多个主题产生影响，或者多个主题对另一主题产生影响等。论述一个主题对另一主题影响的文献，归入被影响的主题所属的类目；论述一个主题对两个或两个以上主题影响的文献，一般归入产生影响的主题本身所属的类目。如：《环境污染与生物》入环境生物学 X17。

（3）因果关系的主题是指文献涉及几个主题，其中一个主题是另一个或多个主题的原因，或者一个主题是另一个或多个主题作用的结果。在标引时，一般分入结果方面的主题所属的类目，如果所造成的结果是多方面的，且能区分出重点主题方面，则分入重点主题所属的类目，否则按原因方面主题所属的类目归类。如：《现代生活与过敏病》入 R593.1；《地震对人类和自然界带来的危害》入 P315.9。

（4）从属关系的主题是指文献各主题之间具有包含关系、属种关系或整体与部分的关系。标引文献时，一般依较大主题的学科属性归类，必要时对次要主

题进行分析分类。若较小主题是论述重点，则依较小主题的学科属性归类。例：《植物油脂化学与油脂化学》依大主题油脂化学归类入 TQ641；《农业植物与花卉》一书的论述中心为园艺观赏植物花卉，因此，依小主题花卉归类入 S68。

（5）论述两个主题相互比较的文献，一般按著者重点论述后所赞同的主题归类，必要时为另一个主题作互见；如果是多个主题之间的比较，若在分类法中有包含这些主题的类组成的概括性类目，则归入该类目中。如：《什么是唯物论 什么是唯心论》入唯物论 B02；《美国、日本经济发展比较研究》入上位类 F112.2。

（6）并列关系的主题是指文献同时论述两个或两个以上各自独立的主题。对具有两个并列主题的文献，归入到能概括其内容的上位类；无共同上位类的，依其论述重点、写作目的或篇幅较多的主题归类；若重点不明确，则按前一个主题的学科属性归类，并为另一个主题作互见分类。对于同时涉及三个或三个以上的并列主题的文献，一般可根据其涉及的范围，将其归入到共同的上位类或概括性的类目。如：《诊断学与内科学精要》依前一个主题“诊断学”归类，入 R44，后一个主题“内科学”作互见分类，入 R5。①

3.4.2 主题标引

主题标引指依据一定的主题词表，对文献的内容先进行主题分析，再赋予文献语词标识的过程。目前，国内主要采用《汉语主题词表》及其相关的专业词表进行主题标引。

3.4.2.1 选词规则

（1）文献主题标引应选用词表中的正式主题词标引。词表中的非正式主题词只起指向正式主题词的作用，本身不得直接用于标引。如：《全国高等学校图书馆工作会议文集》标引为：院校图书馆—图书馆工作—中国—文集［在《汉语主题词表》中，“大学图书馆”用（Y）“院校图书馆”］。

（2）文献主题标引应该首先选取与文献内容主题概念相对应的、最专指的主题词。如：《心电图诊断技巧》标引为：心电图—诊断，不能标引为：电诊断。

（3）文献内容的主题概念在词表中没有相应的最专指的主题词时，可选用与其最直接相关的、最邻近的主题词进行组配标引。如：《石英电子钟表修理大全》标引为：石英钟—电子钟—维修—手册。

---

① 参见 http://www.calis.edu.cn/calis/lhml/lhml.asp?fid=FA0308&class=2，2007-03-12。

（4）文献内容的主题概念在词表中没有恰当的主题词组配，可考虑选用一个最直接的上位主题词进行上位标引，或近义的主题词进行靠词标引。当文献内容的主题概念采用上位主题词、近义主题词或组配标引都不合适时，可增补新的专指主题词进行标引。如，《视听新潮流：家庭影院》标引为：家庭影院。

（5）新增词应遵循一定的原则，必须是词形规范、概念明确、具有较重要的检索意义或具有较广泛的组配作用，如“远程教育”、“光盘刻录机”；新增词应是比较成熟、稳定、具有生命力的主题概念，如“因特网”；新增词应是词表中明显漏收的重要主题概念，如“电力电子学”等。

（6）各类名称主题词可直接作为正式主题词来使用，如地理名称、个人名称、机构名称、作品名称、会议名称、节目（栏目）名称、产品、设备、仪器、仪表等名称、大型系统名、数据库名、应用程序名、计算机语言名称等。①

3.4.2.2 组配规则

在主题标引过程中，将两个或两个以上的主题词按照一定的逻辑关系加以组织以表达文献主题的，称为组配标引。组配标引是主题标引中准确揭示文献主题的一种基本的标引方法。组配标引能以较少的主题词完整确切地表达主题概念，提高标引的专指性，并提高检索效率。组配标引的关键是解决好主题词之间的组配问题，主题词的组配应遵循一定的规则。

（1）主题词的组配必须是概念组配，而不是字面组配。组配的主题词之间，存在着概念限定或概念交叉的关系。例如：“熊猫洗衣粉”这一主题，应该用“熊猫牌商品”和“洗衣粉”组配，而不能用“熊猫”和“洗衣粉”两词组配。

（2）当表达一个复杂主题概念有几种组配形式可选择时，应优先采用交叉组配法。只有不能进行交叉组配时，才可使用限定组配法。如：《介质光波导》应标引为：介质波导—光波导，而不能标引为：介质—光波导。

（3）应选用与主题关系最密切、最邻近的主题词进行组配，不能选用泛指的主题词越级组配。如：《中国人民解放军财务史》标引为：中国人民解放军军史—军队财务，而不能标引为：中国人民解放军—军队财务。

（4）主题词组配标引的结果，必须概念清楚、确切，具有单义性。如：《知识经济浪潮》应直接增补“知识经济”一词，标引为：知识经济—概论，而不能标引为：知识—经济—概论。

（5）当一个标题中的主题词涉及不同主题因素时，主题词的组配次序一般按

---

① 参见 http://www.calis.edu.cn/calis/lhml/lhml.asp?fid=FA0309&class=2，2007-03-01。

照“主体因素—通用因素—空间因素—时间因素—文献类型因素”确定。当一个标题中同时出现多个主体因素主题词时，一般按对象、方法、材料、过程、条件等次序排列。如：《压力容器焊后热处理》标引为：压力容器—焊后处理—热处理。

3.4.2.3 主题词组配标引的形式

(1) 概念交叉组配。又称同级组配，指两个或两个以上具有概念交叉关系的主题词进行组配，来表达一个主题内容。该组配一般表现为同级主题词之间或事物与事物之间的组配。如，“青年工人”主题中，“青年”与“工人”是交叉关系，且词表中没有“青年工人”这一主题词，用“青年”与“工人”组配表达这一主题概念，这种组配方式就是交叉组配。在进行概念交叉组配的过程中，首先将要标引的复杂主题概念分解成若干简单的主题概念，这些简单的主题概念都是该复杂主题概念的属概念，并且在词表中均有对应的正式主题词。然后将这些主题词组配成一个更专指的主题概念。如，《教育社会心理学》可分解为“教育心理学”和“社会心理学”两个简单概念，它们都是“教育社会心理学”的属概念，且都是正式主题词，然后组成“教育心理学—社会心理学”。这两个概念外延的重叠部分即为“教育社会心理学”，它是“教育心理学”和“社会心理学”共有的种概念。

(2) 概念限定关系组配。也称复分组配，是由一个表示事物的主题词和另一个或几个表示事物的部分、属性或方面的主题词组合起来表达一个新专指概念的组配方法。限定关系主要表现为事物与其各个方面之间的关系，而不是事物与事物之间的关系。具体特征表现为：几个概念之间，即几个主题词之间，存在主次关系、偏正关系，偏、次概念的主题词对正、主概念的主题词，从时间、空间、属性、特征等各个方面进行限定、细分、修饰和说明，以达到使描述和表达的主题概念更加专指。如，用“工业经济”与“经济政策”组配标引“工业经济政策”，用“汽车”与“发动机”组配标引“汽车发动机”。

(3) 连接关系组配。是一种特殊的概念限定关系，指复合主题中主体因素之间具有相互关系、比较关系、应用关系、影响关系、因果关系等。主题标引对这种关系的描述，一般都需要使用一个具有某种概念连接功能的主题词作为中介。主题词的排列顺序为：反映相互关系、应用关系、影响关系、因果关系中的主体、影响因素、应用学科、原因等方面的主题词置于前面；具有连接功能的主题词置于中间位置；其客体、被影响、应用到的学科、原因等方面的主题词置于最后。如：《水对植物生长的影响》标引为：水—影响—植物生长。

3.5 自动标引

3.5.1 自动标引概况

自动标引指直接通过计算机的操作处理，赋予检索标识的活动。在网络环境下，传统的手工标引已经无法适应信息存储的需要，自动标引由于具有较强的处理能力，能够适应信息数量迅速增长的需要，处理速度快，可以在一定程度上克服手工标引人员由于主观因素而导致的标引误差，增强标引结果的一致性，标引成本相对较低等优点，呈现出明显的优势，逐渐被广泛应用。

自动标引的发展起始于 20 世纪 50 年代末。1957 年，美国 IBM 公司的卢恩（H. P. Luhn）发表了两篇文章，首次将计算机技术引入文献标引领域，开创了自动标引的先例。在 60 年代，卢恩等研制的以计算机为编制手段的关键词索引法，曾广泛应用于《化学题录》等大型专业索引刊物的编制。此后，70 年代美国国防部文献保障中心（Defense Documentation Center）采用的机助标引系统、90 年代美国 NASA 宇航信息中心使用的机助赋词标引系统等，都是结合自动标引研究成功建立的人机结合的实用系统。

我国研究人员 70 年代末开始研究汉语文献自动标引问题，在 TK—70 计算机上建立了一个试验系统，借助词典对文献题名进行切分，然后使用一套组词规则，将切出的小词组成专指的关键词输出。90 年代中期以后，就开始逐步出现供实际使用的自动标引系统。1996 年，中国医科院情报所就采用人机结合的方式建立生物医学文献数据库。随着计算机技术的发展，自动标引技术得到了很大发展，并取得了显著进展。

自动标引有多种形式，从标引深度来分，有全文自动标引和题名自动标引；从选用的标引词来分，有叙词自动标引和关键词自动标引；从标引方式来分，有自动赋词标引和自动抽词标引；从标引形成标识来分，有主题自动标引和分类自动标引。

全文自动标引指对文献的全文进行自动标引，标引方式包括单词标引、短语标引和语义标引等。题名自动标引是指以题名作为标引源，比如，我国一些档案部门就主要采用了题名关键词自动标引。

自动赋词标引指在计算机自动标引过程中，使用的标引词选自预先编制的词表，而不是来自文献本身。自动抽词标引是指用计算机从文献文本中抽出标引用

词（即能表达文献主题概念的词）的一种自动处理过程。

3.5.2 自动标引方法

人们建立自动标引系统的最终目的是利用机器从输入文献中自动生成能够用以代表文献特征的标识，以方便检索。依据自动标引采用的理论，自动标引的方法主要有统计标引法、语言分析标引法、人工智能标引法等。

3.5.2.1 统计标引法

统计标引法是自动标引各类方法中使用历史最长、应用范围最广的一种。该方法的理论基础是著名的齐夫（Zipf）定律，它建立在较成熟的语言学统计研究成果基础之上，简单易行，具有一定的客观性和合理性，在自动标引中占有较重要地位。根据其处理方法的不同，统计标引法有词频统计法、加权标引法、n-Gram 标引法和统计学习标引法等。

1. 词频统计法

词频统计法认为：一个词在一篇文章中的出现频率是这个词对于这篇文章重要性的有效测度。它依据齐夫定律，即将某一篇较长的文章（500 字以上）中每个词出现的频率按照递减顺序排列起来（高频词在前，低频词在后），并用自然数给这些词编上等级序号，频次最高的是 1 级，其次是 2 级，3 级……如果用 f 表示词在文献中出现的频次，用 r 表示词的等级序号，则有 $f \times r = c$（c 为常数）。卢恩在齐夫定律的基础上，提出了自动抽词的基本思想。将词的出现频率按等级排列，以一定的标准排除高频词与低频词，剩下的就是最能代表文献主题内容的词。目前，词频统计法多与其他的标引方法综合使用。

2. 加权标引法

加权标引法包括以下几种方法。

（1）逆文献加权标引法。

这种方法在标引时，不仅考虑词在一篇特定文献中的出现频率，而且考虑在整个文献集合的文献频率。标引词的权重与其出现频率一致，与其文献频率成反比。词的出现频率是针对文献集合中某确定的文献，词的文献频率则是对整个文献集合而言。在一篇特定的文献中，特征词的出现频率越高，说明它与该文献的主题相关的程度越高。在一个文献集合中，非特征词的文献频率一般较高，几乎出现在所有的文献中，而特征词的文献频率一般较低。

（2）词区分值加权标引法。

这种方法的基本思想是从词区分文献的能力出发来设计标引词的权重，标引词的权重与其区分值成正比。词区分值显示了对文献的“分离”能力，如果一个

词能够较好地反映出文献集合中各文献的差异，那么这个词区分文献的能力就较强，否则这个词区分文献的能力就较弱。

逆文献频率加权标引和词区分值加权标引主要依赖于词的频率特征和词的区分能力，它们的主要缺陷是与用户的相关性无关。

（3）词相关性加权标引法。

这种方法根据检索结果给出的相关性反馈来确定标引词权重。

（4）价值测度加权标引法。

这种方法还要考虑相应的效益和费用。

词相关性加权标引法和价值测度加权标引法在考虑词在一特定文献或整个文献集合中的频率特征的同时，还考虑了标引词在相关文献集合和无关文献集合中的频率特征，以及检索结果的效益值。

3. n-Gram 标引法

n-Gram 标引法原理简单，处理容易，以 n 字符串为统计对象，将其统计得分赋予该串中心字符，然后选择包含得分超过特定阈值字符的单词和短语作标引词。n-Gram 法产生于英语，能适用于多种语言，包括基于字母的西文，以及东方语系的日文。

4. 统计学习标引法

统计学习标引法通过一个学习过程建立标引词与其相关词和不相关词的关系，并以此为基础确定标引词的标引值。这种方法由学习和标引两个过程组成，首先汇集肯定和否定训练集合，统计在集合中出现的单词的词频，选择促进词和削弱词，确定两个平均标引值之间的中值，这样就可以得到标引词的关系和阈值，进行标引。

3.5.2.2 语言分析标引法

标引的对象是由自然语言构成的文献，因而，我们可以从语言学的角度去探索自动标引的方法。语言分析标引法包括句法分析标引法和语义分析标引法。

1. 句法分析标引法

句法分析标引法是从语法角度上确定句子中每个词的作用（如主语还是谓语）和词之间的相互关系（如是修饰还是被修饰）。句法分析一般通过与事先准备好的解析规则或语法相比较而实现。句法分析包括浅层句法分析和深层句法分析两种。浅层句法分析只是把句子解析成较小的单元，但不揭示这些单元之间的句法关系；深层句法分析则充分分析和揭示句子的语法特点和反映的主题内容。比较有代表性的句法分析标引法有基于深层结构的标引法、FASIT 标引法等，它们都以深层句法分析为基础。

2. 语义分析标引法

语义分析标引是分析词在特定的上下文中的确切含义，以选择与主题含义相同的标引词描述文献和提问。比较有代表性的语义分析标引法有潜在语义标引法、相信函数模型和语义矢量空间模型等。

3.5.2.3 人工智能标引法

人工智能是计算机科学的一个分支，专门研究怎样用机器理解和模拟人类特有的智能系统的活动，探索人们如何运用已有的知识、经验和技能去解决问题。实现自动标引的目的是让机器从事标引工作中的脑力劳动，即让计算机模拟标引人员完成标引文献的工作。

人工智能标引法在标引中的具体技术是专家系统，专家系统又称知识库系统。专家系统的知识表示方法主要有产生式表示法、语义网络表示法和框架表示法。比较有代表性的人工智能标引法有 JAKS 标引法、WorldViews 标引法和 MedIndEx 标引法等。

3.5.3 自动分类

自动分类是指由计算机系统自动提取信息的特征项，依据一定的算法，将信息按内容或属性归到一个或多个类别的过程。主要包括自动归类和自动聚类两个部分。自动归类是指计算机系统按照一定的分类标准，将待分信息划归到不同类目的过程。自动聚类是指由计算机系统按照待分信息的各种特征，将具有相近或者相同特征的信息聚合在一起的过程。二者主要区别在于自动聚类不需要事先定义好分类体系，而自动归类则需要确定好类别体系。

3.5.3.1 自动归类

1981 年，侯汉清从计算机管理分类表、计算机分类检索、计算机自动分类、机编分类表等四个方面探讨了自动归类的问题，拉开了我国关于自动归类研究的序幕。

自动归类开始于辅助分类标引系统的研制，出现了一些有影响的辅助分类标引系统和自动归类系统。如，1984 年广东省中山图书馆开发的计算机辅助图书分类系统、1995 年南京大学信息管理系推出的档案自动分类系统、1995 年杭州应用工程技术学院推的中文文献自动分类系统、1997 年山西大学计算机系推出的金融档案自动分类系统、1999 年上海交通大学推出的基于神经网络优化算法的中文自动分类系统等。自动归类系统从其实现的技术来分可以分为两种：

1. 基于词的归类技术

从理论上讲，文本自动处理是以概念为基本单元，而词是概念的基本组成部

分，是信息的载体。因此，这种方法是根据那些可以代表文章主题内容的词汇对文章进行类别判定的一种方法。它一般来说包括三个过程：首先，选择一种分类体系，如《中国图书馆分类法》，利用现有分类法、词表、同义词典等工具，形成归类底表。其次，抽取表达主题内容的关键词，并将其与预先设计的“分类号关键词”所形成的矢量空间模式进行匹配，找出每个关键词涉及的分类号。最后，把所有分类号进行逻辑运算、归并、整理，结合各种复分表索引库，根据级别，得出每个分类号的权值级别。级别最高的类，即为该文献应归的类。词汇归类是三个过程综合的结果。目前主要的词汇归类方法有：逻辑分析法、词频分析法、字面相似度分析法等。

2. 基于知识的归类技术

基于知识的文本自动归类方法主要依赖于一个明确的知识库。基于知识的分类技术的显著特点是需要手工建造知识库，主要依赖语言学知识，需要编制大量的推理规则，实现相当复杂，而且其开发费用相当昂贵。最近的研究工作表明，在一定的领域内，基于知识的系统能够进行快速、准确的分类。

3.5.3.2 自动聚类

自动聚类指的是由计算机系统按照被考察对象的内部或者外部特征，按照一定的要求将相近、相似或者相同特征的对象聚合在一起的过程。自动聚类是一种重新组织信息、发现知识和挖掘数据的重要工具，能集中内容相同或相关的信息并揭示其相关性，使之有序化并加以控制，提高信息检索的精度和效率，改善检索结果的输出。

目前，聚类技术被大量应用于数据库、Web 文本文档和搜索引擎中，已引起当今国际人工智能与数据库界的广泛重视。在聚类检索中，聚类技术是智能地解决用户个性查询问题的基础，能够用于知识发现、分类聚集和优化查询、解决检索结果过多过杂的问题等。

自动聚类的实现方法一般包括四个步骤：

（1）网页表示。包括特征抽取和特征选择。一般是对网页特征进行特征加权，将网页特征表示成计算机能够处理的数学向量。其中有两个因素影响了特征的权值。一是该词的词频，另一个是该词在网页中出现的位置。

（2）相似度计算。主要根据网页表示的距离函数来定义。

（3）聚类。根据网页表示和相似度计算的结果，按照一定的规则将聚类网页分成不同的类。在这样的类目体系中，主题相近、内容相关的文献信息聚在一起，而相异的则被区分开。

（4）给出聚类的标识。在最后形成的每一类中抽取一定具有代表性的特征，

作为该类的标识。一旦检索到标识中的某一条信息，则可通过这条信息把聚类中的其他文献信息全部检出，从而实现高效率的知识挖掘式的检索。

自动归类和自动聚类都是在信息标引技术的基础上，用计算机系统进行文本自动分类的过程，且广泛运用于搜索引擎领域。两者的区别是：自动归类需要确定一个后台的分类表，根据既定的分类规则，或者由计算机通过训练学习到分类知识，为待分类文献确定一个或者多个类别。自动聚类则不需要事先定义好分类体系，完全依靠数学分析方法提取出类目，并根据类目集聚相似的对象。比如Vivisimo自动聚类系统（http：//vivisimo.com）等。相比较而言，自动聚类比自动归类在技术上更容易实现，可以运用到单个搜索引擎或者元搜索引擎中，聚类效果也更加显著。

【案例】

文献主题标引的实例

以下实例均选自Calis的《中文文献主题标引原则》[①]，依据的词表是《汉语主题词表》，并以《中国分类主题词表》（1994）作为补充。

1.《高等数学导论》

标引主题词：高等数学—概论

说明：这是一个非常典型的单主题文献标引例子，对于这类文献的主题标引可直接用一个主题词（包括单一概念词或复合概念词）表达的单元主题，不必进行组配标引。

2.《计算机数据通信基础》

标引主题词：计算机通信—数据通信

说明：该文献表现的是一个具有交叉关系的复合主题，应采用交叉组配方式进行标引。

3.《高等教育理论研究》

标引主题词：高等教育—教育理论

说明：这是一个具有限定关系的复合主题，应采用限定组配方式进行标引。

4.《超级市场连锁经营计算机管理》

标引主题词：连锁商店—企业管理—计算机应用

说明：该文献为应用关系的复合主题，词表中有专指的应用关系主题词“计算机应用”，可直接选用。

---

① http：//www.calis.edu.cn/calis/lhml/lhml.asp?fid=FA0309&class=2，2007-03-15。

5.《水对植物生长的影响》

标引主题词：水—影响—植物生长

说明：该文献为影响关系的复合主题，需选用“影响”一词进行组配标引。

6.《比较文学：东方与西方》

标引主题词：比较文学—东方国家—西方国家

说明：比较关系的复合主题，如果词表中有专指的比较关系主题词如“比较文学”，可直接选用。

7.《形式逻辑与数理逻辑比较研究》

标引主题词：形式逻辑—对比研究—数理逻辑

说明：比较关系的复合主题，如果词表中没有专指的比较关系主题词如“比较研究”，则选用“对比研究”一词进行组配标引。

8.《血型与性格》

标引主题词：血型—关系—性格

说明：相互关系、相互作用的复合主题，应进行组配标引。

9.《汽车发动机》

标引主题词：汽车—发动机

说明：该文献为整体与部分关系的复合主题，应进行组配标引。

10.《国内流行中外彩色电视机电路大全》

标引主题词：彩色电视—电视电路—电路图

说明：论述事物某一方面的图书，标引事物与其方面的概念。

11.《彩色电视机电路分析与维修》

标引主题词：彩色电视—电视电路—电路分析

彩色电视—电视电路—维修

说明：论述事物的两个方面的图书，应分组标引事物与其两个方面的概念。

12.《彩色电视机选择、使用和维修》

标引主题词：彩色电视

说明：论述事物三个或三个以上方面的图书，一般只做整体标引。

13.《物理化学中的胶体化学》

标引主题词：物理化学

胶体化学

说明：从属关系的复合主题，如果同时论述上位主题概念和下位主题概念，应视为多主题，进行分组标引。上位主题词和下位主题词之间，不得进行组配标引。

14.《高等数学 II 一元微积分与微分方程》

标引主题词：微积分—高等学校—教材

微分方程—高等学校—教材

说明：从属关系的复合主题，如果只是重点论述下位主题概念，则只应标引下位主题概念。

15.《激光在生物学和医学上的应用》

标引主题词：生物学—激光应用

医学—激光应用

说明：多主题文献要分解为单主题，进行分组标引或分组组配标引。

16.《猪鸡鹅鸭快速饲养法与疾病防治》

标引主题词：养猪学

猪病—防治

鸡—饲养管理

鹅—饲养管理

鸭—饲养管理

鸡病—防治

鸭病—防治

鹅—禽病—防治

说明：多主题文献一般不选用上位词标引，但如果标引深度过大，可重点选择其中几个并列的单主题进行标引，同时再标引一个它们的上位主题。

17.《战斗在“一二·九”运动的前列》

标引主题词：一二·九运动（1935）

说明：关于历史事件的标引，如果词表中已有该历史事件的专指词，可按词表形式直接标引。

18.《湘江战火：长沙会战纪实》

标引主题词：抗日战争时期战役战斗—国民党军—长沙市

长沙会战（1939）

说明：关于历史事件的标引，如果词表中没有该专指词，可用历史事件内容与其发生的国家组配标引，并视其重要性做自由词标引[①]。

19.《中国古代建筑史》

标引主题词：建筑史—中国—古代

---

① 自由词标引是指标引时不使用词表，由标引人员根据文献内容自拟检索词。

说明：关于专门学科史的标引，可以相应的专指主题词如“建筑史”、“医学史”等标引。

20.《中国古代四大发明》

标引主题词：技术史—中国—古代

说明：关于专门学科史的标引，如果词表中没有相应的专指主题词，则以学科名称的主题词与“学术史”或“历史”进行组配标引。综述自然科学史的文献或兼论自然科学史和技术史的文献，用主题词“自然科学史”标引。工业技术史方面的文献，以学科主题词和“工业史”、“技术史”等专指词进行组配标引。

21.《鲁迅年谱》

标引主题词：鲁迅—年谱

作家—年谱—中国—现代

说明：个人传记、评传、年谱等以人物名称和写作形式的主题词进行组配标引，并以人物所属学科、写作形式以及所在国别等进行组配标引。

22.《东海污染调查报告：1978—1979》

标引主题词：海洋污染—调查报告—东海—1978—1979

东海—海洋污染—调查报告—1978—1979

说明：以地区及其自然和社会现象为研究对象的文献，除从学科主题概念角度标引外，还要以地区名称作为地名主题进行标引。

23.《伦敦与巴黎日记》

标引主题词：游记—欧洲—现代

日记—中国—清代

说明：地理类游记，以游记为主体因素，以内容所涉及的地区为位置因素进行组配标引。

24.《论生活、艺术和真实》

标引主题词：文学—关系—现实生活

说明：文学理论，以文献所论述的主题内容进行标引。

25.《壮志凌云》

标引主题词：军事题材—长篇小说—中国—当代

说明：文学作品（包括小说、诗歌等）一般从作品的题材、体裁，以及作者所属的国别、时代角度进行标引。

26.《哥德巴赫猜想》

标引主题词：报告文学—中国—现代

陈景润—1933—1996—生平事迹

科学家—生平事迹—中国—现代

说明：描写真人真事的报告文学、通讯报道、特写等，除从作品体裁、作者所属的国别、时代进行组配标引外，还要从人物传记角度加以标引。

27.《英汉对照赠言手册》

标引主题词：英语—汉语—对照读物

说明：关于综合性语言对照读物、注释读物，两种语言的，以非汉语语种（包括外语和少数民族语）为主体因素，“对照读物”或“语言读物”为文献类型因素进行组配标引；三种或三种以上语言的对照读物，以“对照读物”或“语言读物”为主体因素与各语种进行组配标引。

28.《梨》（果树栽培丛书）

标引主题词（单书著录）：梨

梨—栽培

标引主题词（整套著录）：果树栽培丛书

果树栽培—丛书

说明：丛书按著录方式分为分散著录与综合著录两种。分散著录（单书著录）的丛书，按学科内容进行分散标引；综合著录（整套著录）的丛书，应以该丛书的学科内容概念为主体因素，“丛书”为文献类型因素进行综合标引。无法确定丛书的学科内容时，可不予标引。

29.《中国植物志 第四十九卷第二分册·被子植物门·双子叶植物纲》

标引主题词：植物志—中国

被子植物门—中国

双子叶植物纲—中国

说明：多卷书一般以整套著录进行综合标引，当其正题名由共同题名和附属题名组成时，应同时标引整套图书和单卷书的主题。

30.《新华字典》

标引主题词：汉语—字典

说明：综合性单语种词典、字典，以语种为主体因素，“词典”、“字典”为文献类型因素进行组配标引。

关键术语

信息著录	信息著录标准	信息标引	元数据
机读目录	都柏林核心元数据集	自动标引	自动分类

思考题

1. 信息著录的含义和作用是什么？
2. 信息著录的标准有哪些？
3. 什么是机读目录？试说明 CNMARC 的结构。
4. 什么是元数据？元数据有哪些作用？
5. 试比较 CNMARC 和元数据。
6. 简述信息标引的含义和步骤。
7. 分类标引应坚持哪些原则？
8. 主题词组配标引的形式有哪几种？
9. 自动标引的主要方法有哪些？
10. 如何评价自动分类？

CHAPTER FOUR

第4章
参考工具书概述

【本章要点】

◇介绍与分析参考工具书的概念
◇叙述我国工具书的产生与发展
◇分析参考工具书的功能与特点
◇阐述参考工具书的种类
◇介绍参考工具书的结构及排检方法
◇讨论参考工具书的评价与选择
◇探讨参考工具书的数字化趋势

引子

网上有篇文章，名为《漫谈文学翻译必备的工具书》，曰："章克标先生的《翻译难》一文，其中谈到'在"文化大革命"中，我所有的藏书，全都被毁灭，片甲不留。我翻译所用的工具书，有很多字典、辞典、参考书、百科全书之类，工作遇到问题，就可查看，得到解决。人的知识有限，翻译离不开好辞典和各种参考书，这些都没有了，只好歇手。所以我是很明白翻译的难处的'。对此我深有同感，搞文学翻译这活儿确实需要许多工具书辅助才能较好地完成……拉拉杂杂写来，归根结蒂，还是章克标先生那句话：'人的知识有限，翻译离不开好辞

典和各种参考书'。"① 伴随着社会的发展，文献资源与日俱增。面对海量的知识信息，人们在欣喜的同时也多了几分迷茫，因为个人的需求不一，能力有限，人们不可能阅读所有的文献，掌握全部的知识。参考工具书正好充当了助手的作用。它是书山探路的向导，学海泛舟的指南。学习中有了它，犹如治学有良师；工作中利用它，如同做事有帮手。作为一种特殊的图书，工具书一直得到人们的青睐。

4.1 参考工具书的概念与特点

4.1.1 参考工具书的概念

图书从使用的角度可分为两类：一是为了获取知识或者欣赏，而从头到尾细细阅读的书；二是为了查阅特定的材料或解决疑难问题，才去翻检或查阅的书。后者相对于前者，是作为工具专供翻检的书，就像人们进行生产活动或其他活动时离不开必要的工具一样，读者在读书、治学和解疑过程中会经常使用这类书，因此，人们称之为参考工具书或工具书。

参考工具书是根据特定的需要，广泛汇集有关的知识资料或文献信息，按便于检索的方法编排，以备查考的工具性图书。例如，词典、手册、年鉴等，它仅供查找有关知识及知识线索。

"参考工具书"这一概念的产生与发展是与其对应的实体的产生与发展同步的。伴随着这类特殊文献的产生与发展，用以表述其概念的"参考工具书"得以出现，并不断发展变化。人们从不同的角度去理解参考工具书的概念。

有的从工具书的用途去阐述，如，何多源的《中文参考书指南》称参考工具书为"参考图书"，"专备人检阅以解答其问题者，如字典、辞典、百科全书、年鉴、书目……"。《现代汉语词典》将工具书解释为："专为读者查考字义、词义、字句出处和各种事实而编纂的书籍，如字典、词典、索引、历史年表、年鉴、百科全书等。"

有的从工具书的性能去概括，如来新夏认为："工具书，顾名思义是指作为我们读书工具的一类图书。"②

---

① http://www.ksw123.com/2006-06-24/115108416731655.shtml，2008－04－09。

② 来新夏：《略论工具书》，载《四川图书馆学报》，1981（4）。

更多的人从工具书的特点去论述，如武汉大学图书馆学系编写的《中文工具书使用法》认为："工具书是根据一定的社会需要，以特定的编排形式和检索方法，为人们迅速提供某方面的基本知识或资料线索，专供查阅的特定类型的图书。"

还有的从工具书的特殊矛盾出发去探讨，如袁逸认为："工具书是为了解决社会知识信息量的增长与特性检索之间的矛盾，使人们便捷地找到全面而精当的特需资料或线索，而对已有一定范围的知识信息进行搜集、整理，并按规范的符号系统或知识体系特殊编排和检索的专供查阅的图书。"①

有的从工具书的文献属性去归纳，如王世伟的《中文工具书教程》认为，工具书是"将若干具有完整独立概念的信息按特定的符号系统或知识体系编排，专供查阅的二次或三次文献"。

4.1.2 我国工具书的产生与发展

我国工具书的起源可以追溯到先秦时代。传说中夏禹铸"九鼎"的图纹、周代《山海图》及战国中王墓出土的"兆域图"，可看成原始的图录；周宣王时太史编的《史籀篇》，可说是字书的萌芽；周代的"牒记"及《周谱》的"旁行斜上法"和"古六历"，可视为表谱的渊源；孔子编纂的《诗经》目录、《尚书》目录及序录，则可视为目录的滥觞。

汉代是我国封建经济和封建文化巩固与发展的时期，也是工具书正式产生并奠定基础的时期。经过魏晋南北朝时期的初步发展，唐宋时期印刷术的发明，科举制度的推行，文学、史学的发达以及宗教的盛行，直接促进了工具书的发展繁荣。辽金元时期，由于战乱多，统治者崇尚武功，对于文事无暇修治，因此工具书编纂较少。但韵书、类书、政书、地图等的编纂也取得了一定的成绩。明清时期的工具书种类多，部头大，体例较新，出现了许多集大成的著作，是古代工具书集大成的时期。

近代以来，在社会政治、经济以及传统文化和外来文化影响下，工具书的内容、类型和编排方式都发生了显著的变化。内容方面从集中于古代典籍的整理和诠释，转而注意记述一些新事物。如字典、词典方面，在《康熙字典》的基础上编成的《中华大字典》，其增收的1 000多字为近代方言和科学用字。《辞源》和《辞海》是具有现代意义的两部大型百科词典，它们收录了大量的新名词术语（包括科技词语），在一定程度上满足了当时知识界的需要。工具书的类型也增加

---

① 袁逸：《给工具书下个定义——兼评几种有关的工具书定义》，载《四川图书馆学报》，1986（4）。

了，如反映新情况和新材料的年鉴、手册一类工具书相继问世。各种综合性、专科性和统计性年鉴大量出现，外语词典、文摘也开始出现，索引有较大的发展，定期出版了索引刊物。工具书的排检方法也有革新，如出现了四角号码法、注音字母法、中国字庋撷法、笔画笔形法等。

新中国成立后，工具书的编纂不仅注意继承传统辞书编纂的成功经验，且已注意借鉴国外工具书编纂的技术方法。新出版的工具书，不仅在数量上远远超过历代，而且在质量上也发生了根本性变化，更新了内容，改革了排检方法。1979年以来，工具书的编纂出现了一个崭新的局面，出现了百科全书热、年鉴热、文摘热、名录热、词典热及影印古旧工具书热。

目前，我国工具书的编纂呈大型化、系列化、多样化和现代化趋势发展。各类型工具书都有大部头作品问世。系列化表现在工具书的配套成龙、相辅相成。多样化是工具书编纂的又一趋势。首先，表现在工具书类型多样。其次，工具书品种比较齐全，涉及各个学科门类和多个语种。现代化主要包括两个方面，一是编纂的标准化，二是编纂出版的电子化。

4.1.3 参考工具书的功能

“工欲善其事，必先利其器。”参考工具书是我们看书学习、进行科学研究不可缺少的工具。利用参考工具书，有助于解决查考字词、文句、成语、诗词、图书、报刊、论文、人名、地名、报告、法规、条约、纪年、典章制度等方面的问题。如果使用得当，能迅速而准确地查到所需要了解的知识、资料以及文献的线索，为读者节约大量的时间和精力。它在一定程度上解决了人类知识的无限丰富与人们寻求特定知识信息的要求之间的矛盾。参考工具书的功能是多方面的，主要表现为以下几点：

1. 查检资料，答疑解难

这是工具书最基本的功能。人们在学习或工作中，总会遇到各种各样的疑难问题，工具书为我们解决这些问题、寻求答案提供了便利。比如，碰到一个字不会念或不知道它的意思，我们会想到查字典；想了解某一事实时，百科全书会给我们一个答案。

2. 指引门径，辅助治学

当代社会，知识爆炸，信息泛滥，学科分类日趋细化，学科交叉日渐凸现，任何人都没有可能掌握所有的知识，即使是自己的专业，也必须不断进行知识更新。这时候，就需要我们翻检查阅各种工具书，了解新知识或新知识的线索，为我们的学习和研究指引门径。林语堂先生曾说过：“无论古今中外，治学工具之

书，皆指示修学门径，节省时间，且可触类旁通，引人入胜。”①

3. 提示线索，提供参考

从信息类型来说，工具书属于二次、三次信息，它是对一次信息的加工、整理和综合。工具书是人们获取原始文献的桥梁和通道，利用工具书，能够较快地查出有关资料，达到尽可能多地占有信息资料的目的，收到事半功倍的效果。工具书还能够反映国内外学术信息，展示各学科研究之进展，在为人们提供信息线索的同时，也提供参考依据。

4. 汇集知识，传承文化

工具书作为一种特殊类型的图书，汇聚了人类知识的结晶，浓缩了各门各类的知识。比如，由于年代久远及收藏不善、校刻不精等缘故，历史上的文献资料往往有遗佚或脱误，而其中亡佚的古籍片断，散见于一些工具书尤其是类书中，通过利用相关的工具书，可以辅助校勘辑佚，重新整理传承。

上面谈的几点功能，只是就工具书的整体而言，具体到每一部书，其作用可能各不相同。因此，我们除了应从总体上把握工具书的性质、特点和作用外，还需要详细了解每一部工具书的特殊作用。

4.1.4 参考工具书的特点

参考工具书是一种查检资料、解决问题的工具，它不同于一般的图书，具有如下特点：

1. 内容丰富，概括性强

工具书既广采博收，又高度浓缩，它是在大量普通图书的基础上，对人类已有的知识进行提炼浓缩而成的信息密集型文献，能提供完整详尽、系统概括的基本知识和高密度的信息。现在的工具书体系既蕴涵了历史文化遗产的精华，又反映了现代科学技术的成就，可谓是一座集成度很高的信息宝库。

2. 编排特殊，便于查检

工具书十分讲究科学的编排形式和高效率的检索方法。它一般把收选到的大量知识、信息组成大小不等的单元，并要求自身有严谨的结构体系，以覆盖有关的知识领域和文献范围，做到以简驭繁，具有高度的逻辑性和组织性。工具书的排检，或按部首，或依笔画，或用号码，或以音韵，或分类分主题序列，或以年、月、日为次，或依地域分编，一目了然，一索即得。而且注意配备辅助索引，为读者提供多种检索途径。

---

① http://www.ncl.edu.tw/cl_ebook/doc/list_8_1.htm，2003-07-03。

3. 准确可靠，权威性强

工具书一般收录的是比较确定的、可靠的、公认的、权威的观点和概念，而且论述精练、出处详明。此外，它还经常采用修订、再版等形式及时更新和补充新的内容，以适应社会发展和科学技术进步的客观实际。

4.2 参考工具书的种类与排检方法

4.2.1 参考工具书的种类

参考工具书的范围很广，名目繁多，内容丰富，人们根据不同的标准将其划分为不同的类型。按语种分，可分为中文参考工具书和外文参考工具书；按规模分，可分为大型参考工具书、中型参考工具书和小型参考工具书；按时代分，可分为古代参考工具书和现代参考工具书；按收录范围分，可分为综合性参考工具书和专科性参考工具书；按学科内容分，可分为社会科学参考工具书和自然科学参考工具书；按载体形态分，可分为印刷型的书本式、报刊式、幅页式、附录式、卡片式参考工具书与非印刷型的缩微式、音像式、机读式参考工具书；按提供实质信息的方式，可分为直接提供信息的参考工具书和间接提供信息的参考工具书两大类；按性质和功用分，可分为线索性参考工具书、词语性参考工具书、资料性参考工具书、表谱性参考工具书、图录性参考工具书、边缘性参考工具书；按编制目的、收录内容和功能用途分，可分为书目、索引、文摘、字典、词典、年鉴、百科全书、手册、名录等。了解工具书的各种类型，熟悉各种工具书的特点和功能，对于我们全面认识和有效使用工具书具有重要意义。

值得注意的是，工具书类型的划分不是绝对的。由于工具书具有多种功能，所以各种工具书之间不同程度地存在着交叉现象。例如：百科词典既具有百科全书的基本性质，同时具有词典的形式。人名录、地名录接近于人名词典、地名词典，笔名别名录可作为索引使用。近年来编纂的大量著作词典，每一条目相当于一本专著的内容摘要，兼有书目和文摘的功用。至于年鉴，则显示出更明显的多功能性，特别是专业性年鉴具有多种工具书的职能，兼有书目、索引、文摘、词典、综述、名录、指南等功用。

4.2.2 参考工具书的结构

参考工具书的结构是指参考工具书的整体构成形式和各部分的基本体制，大

体上由序和跋（前言和后记）、凡例或说明、正文、辅助索引和附录补遗等几部分组成（如表4—1）。

表4—1 参考工具书的结构

组成部分	注 解
序和跋（前言和后记）	是对工具书的简介和评论文字，一般说明工具书的编纂目的、编纂宗旨、编纂过程、编纂过程中的某些具体问题，以及收录范围、内容特点、使用价值等。
凡例或说明	揭示工具书编排与使用要旨，是利用工具书的先导，主要介绍其选择内容的原则、解字释词的体例、词条的编排方法及细则、查检法、特定符号等。
正文	工具书的核心部分，是查阅的主要内容，其特点和性质都体现在这一部分。
辅助索引	工具书一般选择一种编排方式对正文加以组织，同时，通过索引提供辅助检索途径。辅助索引是供查询正文部分的各种索引，能提供多种有效的检索途径，它是衡量工具书质量好坏的重要指标之一，辅助索引越多，检索途径就越广，越能从多角度挖掘正文的内容，使用就越方便，检索效率也就越高。
附录补遗	是为来不及列入正文内容或为区别于正文内容而设的部分，附于工具书正文之后，主要包括与正文有关的一些文章、图表、地图、对照表、汇编、索引等资料，或补充正文遗漏处及须订正的事项，或将分散在正文条目下的相关部分列成表目，有助于查考或理解原文，有其特定的重要意义。

4.2.3 参考工具书的排检方法

参考工具书是专供查检的，所以它的编排方法应该是周密而简便的。任何工具书都是对所收资料进行科学化、系统化的组织，并以一定的顺序排列而成。编排的目的是为了检索，检索则需按编排的规则去进行。排与检是互相联结、密不可分的。对工具书的编纂者来说是编排方法，而对工具书的使用者来说则是检索方法。所以统称为排检方法。

工具书的排检方法比较多，大体上可分为三种：按字顺编排、按内容编排和按自然顺序编排。按字顺编排，是按照汉字字形或读音将工具书中的条目加以编排，具体包括形序法和音序法。这是工具书的重要编排方法，大多数工具书都采用这种方法。按内容编排，指根据知识单元、文献线索的内容性质将工具书中的条目加以编排，主要有分类编排法和主题编排法。按自然顺序编排，是根据事物发生发展的时间顺序或事物产生所处的地理位置将工具书中的条目加以编排，包

括时序法和地序法。

4.2.3.1 形序法

形序法是根据汉字形体结构的特点加以排列的方法。主要包括部首法、笔画法、笔顺法和号码法。前三种方法多半结合运用，互为补充。在使用形序法的过程中，必须注意现行简化字与繁体字的对应。

1. 部首法

部首法是根据汉字的形体结构的特点，利用其偏旁（汉字的各个组成部分）的同一性来编排条目的方法。汉字的形体结构，除少数属独体字外，大多是合体字，合体字多为由形旁（也称义符）和声旁（也称声符）组成的形声字，彼此之间具有一部分相同的形体，把这些形体相同的部分归为一类，称为偏旁或部首。在独体字中，有的本身就是部首。

部首法是我国工具书最普遍最常用的一种编排方法。它利用汉字形体结构的特点把数量庞大且又极不规则的汉字分组归类在不同的部首里，以便人们从字形来求音求义。特别是对于不知读音的汉字，只要确定其部首，即可查得。但由于目前通用的几种部首法定部原则不同，具体规定不一，对于多部首和部首不明显的字，往往难以确定部首，查阅比较困难，因此，使用部首法编排的工具书应注意以下几点：

（1）熟悉新旧部首法的定部规则和特殊规定。旧部首法以《康熙字典》为代表，其定部原则一般是依字义定部，部首与字的本义相符或相关，不考虑部首的不同位置。由于以字义归部，一些部首与该部所收字的偏旁写法不完全一致，不少部首带有附形，增加了取部的困难。为准确取部，应熟悉部首的附形，了解取部的一些特殊规定。新部首法以 1989 年版《辞海》为代表，其定部原则是依字形定部，不追求字的本义及原始字形的归类排列。在依字形定部的总原则下，对于如何取部还有些具体的规定。使用以部首法编排的工具书时，应先仔细阅读其排检说明或凡例，了解其部首的归类原则及具体规定。

（2）同一部首内的各字一般都按除去部首后的笔画数排列，同画数的按起笔笔形排列。

（3）注意利用“难检字表”或其他检索途径。某些字结构复杂，难以分解，或一个字有多个部首，不易确定。查检这类形体特殊和难以确定部首的字，可以利用“难检字表”（多按笔画多少排列）。如果有其他检索途径，也可加以使用。

（4）学一点文字学知识，有助于判断一个字的部首。部首法的定部或多或少地反映出一定的文字学原理，所以，应尽可能地多了解些文字学知识，这对于一个字部首的判断很有帮助。

(5) 根据繁简不同的字体来确定部首。汉字有繁简之分，所以，按繁体字编排的工具书应用字的繁体部首，按简体字编排的工具书应用字的简体部首。“繁简字对照表”或“简化字总表”，可以用于以繁查简或以简查繁。

2. 笔画法

汉字由笔画构成。笔画法就是以笔画数目的多少为排列次序，笔画少的在前，笔画多的在后，同笔画的字归在一起的排检方法。这种方法，识字就会使用。缺点是很多字的笔画不容易数准确，手写体又和印刷体不同，检索时很麻烦；同笔画的字太多，检索速度很难提高。

采用笔画法编排的工具书较多，有《中国人名大辞典》、《中国古今地名大辞典》、《十三经索引》等。笔画法还作为辅助检索方法使用。

掌握笔画法要注意辨清一个字的笔画。汉字的繁体字和简化字，笔画不同；印刷体和手写体，新字形和旧字形，笔画也往往不同。遇到一个字笔画难以确定时，要多查几个近似的笔画。

笔画相同的字，或按部首归类，或按笔形顺序排列。使用时需先分辨清楚，以加快查找速度。

3. 笔顺法

汉字的基本笔形是点（丶）、横（一）、竖（丨）、撇（丿）、捺（乀）5种。两种或两种以上的基本笔形连用又组成复杂的折笔（如𠃌等）。汉字书写时起笔只用丶、一、丨、丿、𠃌（包括竖折、撇折）5种笔形。利用笔形顺序作为排检方法的，有的只用第一笔的笔形（起笔笔形法），有的利用各笔的顺序排列，而且笔顺也不统一，有4种笔形（丶一丨丿；丶丨丿一；一丶丿丨）、5种笔形（一丨丿丶𠃌；一丨丶丿𠃌），还有7种笔形的。

笔顺法虽然简单，但因书写习惯不同，笔顺和起笔有时很难确定，就是在现行的一些工具书中，某些字的笔顺也有分歧。因此，仅有少数几种工具书利用笔顺法编排，如陈德芸编的《古今人物别名索引》。现在主要用起笔笔形作为笔画法的补充。

使用笔顺法检字，应掌握汉字笔顺的一般规则。根据汉字的结构特点，其笔形顺序通常是：先横后竖（如“丰”），先撇后捺（如“人”），自上而下（如“号”），由左到右（如“以”），先内后外（如“这”），由外及里（如“同”），先中后旁（如“幽”），先离后交（如“里”），左上点先（如“为”），右上点后（如“戈”），中间点后（如“丹”），竖折底后（如“区”）。文化部、文字改革委员会1964年12月联合公布的《印刷通用汉字字形表》（文字改革出版社，1986），规定了6 196个字的字形和笔顺，应该作为确定笔顺的规范。

4. 号码法

号码法实际上是形序法的一种变形。它把汉字分解为若干种笔形，用阿拉伯数字作为代码，然后把每个字的笔形代码连结为号码，再依号码大小排列。这种方法的优点是能把没有严密次序的汉字按号码井然有序地排列起来，取号的位置固定，不管部首、不数笔画、不论读音，只要记住笔形代码、位置次序，检索迅速，使用便利。其缺点是学习和掌握比较困难，只有经过反复练习，才能运用自如。号码法包括四角号码法、中国字庋撷法等。其中，四角号码法影响最大，流行最广，为 20 世纪 20 年代各种工具书所广泛采用。

四角号码法是根据汉字是方块字这一特点提出的。一个字有四角，用数码标示四个角的笔形，联成四个数字的号码，依号码顺序排列和查检汉字。它是一种单一式的排检方法，不必与其他方法结合使用。

四角号码法的优点是可以见字知码，查检迅速，较为简捷。但是取号规则过于繁复，取笔方法也有新旧之分、繁简之别，变形的笔形较多，也不易掌握，对笔形稍有误解即难查到。但因四角号码法有较大的使用价值，所以应熟悉其取号规则和方法，多查多练，努力掌握。有些工具书用四角号码法编排正文或编制辅助索引，如《中国丛书综录》、《二十四史纪传人名索引》。

“四角号码检字法”最早用于商务印书馆 1928 年出版的《四角号码学生字典》，以后得到普遍应用。后来为适应简化字及字形规范，对四角号码作了一些修改，改称“四角号码查字法”，一般称为新四角号码法。目前出版的工具书，采用新法、旧法的都有。旧法笔形依通行的手写体，而新法笔形则以《印刷通用汉字字形表》的规定为准。二者的取角方法有所不同。《现代汉语词典》和《四角号码新词典》附有《新旧四角号码对照表》，列出了主要的不同点，可以参看。

四角号码法的构成分为三个步骤：首先，把汉字的笔形分为十类，用 0 到 9 十个数字表示，每类下包括多个笔形。可以利用口诀帮助记忆十种笔形的代码：“横 1 垂 2 3 点捺，叉 4 插 5 方框 6，7 角 8 八 9 是小，点下有横变 0 头。”其次，将字的四个角的笔形转换为四个数字。最后，按字的左上、右上、左下、右下四角的顺序取号，把笔形数字连接起来，就组成了这个字的四角号码。为了避免重码字过多，再取一个附号。四角号码法有许多具体规定，熟悉掌握这些具体规定，对准确取角很有帮助。四角号码法按以下顺序排列汉字：按号码大小由小号到大号排列；同号码的字按附号顺序排列；四角和附号均相同的字，再按各字所含横笔的数目顺序排列；条目首字相同的，再以条目第二字的头两角的号码大小为序。

4.2.3.2 音序法

音序法是根据汉字的读音及表示读音的语音符号的顺序排列汉字的方法。这种方法比较简捷，检索速度较快，但不知读音就无法检索，需要辅以其他检索途径。音序法主要包括汉语拼音字母法、注音字母法和声韵法三种，使用最普遍的是汉语拼音字母法。

汉语拼音字母法是按汉语拼音字母顺序排列汉字的方法。我国1958年公布的《汉语拼音方案》，采用26个拉丁字母作为汉语拼音，排列次序依照国际习惯，从A到Z。在26个字母中，除I、U、V三个字母不做音节头，其他23个都可做音节头，分成23部，排列时，先以首字母的ABC音序排，若首字母相同再按第二个字母音序，以此类推，如果两个字或词相当，再按阴平、阳平、上声、去声、轻声顺序排列。

汉语拼音字母法使用拉丁字母作为拼音字符，简化了编排方法，检索方便、准确，促进了语言标准化，利于推广普通话，并且符合国际上工具书的编排规则。但要求读者必须学会普通话，语音要求标准，否则发音不准，不易查找。另外，如果读者遇到的字词，不知其音，也不明其义，就无法使用音序法检索。此外，汉字同音字很多，但每一种工具书对同音字的编排方法不尽相同，有的按笔画多少排序，有的按部首排序，这也是利用汉语拼音字母法时需要注意的问题。

4.2.3.3 分类编排法

分类编排法指按学科或按事物性质的系统性分类排列的方法。使用工具书时，必须熟习分类体系，了解排列顺序。可分为以学科分类为基础的文献分类法和依照事物性质排列的分类法两种。

1. 以学科分类为基础的文献分类法

我国的文献分类法经历了古代的七分、四分到现代图书分类法的发展过程。汉代时，刘向、刘歆父子著《别录》和《七略》，将图书分为六艺略、诸子略、诗赋略、兵书略、术数略、方技略，再加上辑略作为总序置于首位，故称为“七分”法。后来，班固的《汉书·艺文志》、王俭的《七志》和阮孝绪的《七录》继承和发展了七分法。西晋初年，荀勖撰《中经新簿》，改七分法为甲、乙、丙、丁四部，创立了四分法。唐代魏征的《隋书·经籍志》，又以经、史、子、集定四部之名，奠定了四分法的基础。清代纪昀等编纂的《四库全书总目》继续丰富和完善了四部分类法。经部主要收录儒家经典著作及其注释和阐述著作，史部主要收录正史及各种体裁的历史著作和相关政治、法律、外交、地理、教育等类著作，子部主要收录诸子学派的著作，集部收录文学作品、文学评论著作。该分类

法至今仍是类分古籍的主要分类法。《中国古籍善本书目》采用修订的四部分类法，分为经、史、子、集、丛书五部48类。

到了近代，我国受《杜威十进分类法》等西方图书分类法的影响，开始突破传统的四部分类法，涌现出了一些有影响的分类法，如刘国钧的《中国图书分类法》、皮高品的《中国十进分类法》和杜定友的《世界图书分类法》等。新中国成立后，图书分类法有了很大发展，形成了三部影响较大的分类法，即《中图法》、《人大法》和《科图法》。其中，《中图法》使用范围最广，不仅被广泛用以类分图书，而且被用来编制目录、索引、文摘、百科全书、专科词典等工具书。如，《中国国家书目》、《全国报刊索引》等。

此外，有些分类编排的工具书，虽然也是按照学科体系归类的，但并未采用通行的文献分类体系，而是根据条目的内容及实际编排的需要，自行拟定分类体系。使用这类工具书要先翻看一下分类表（即该书的目录），然后根据需要决定查找哪一类。

2. 依照事物性质排列的分类编排法

即按同一类事物性质归类集中的编排方法。古代的类书、政书和现代的年鉴、手册及某些辞书多采用此法编排。

这种方法是我国第一部词典《尔雅》所开创的。《尔雅》分为19篇，后面16篇就是分类词汇。如“释亲”就是把有关家族关系的词汇集在一起加以解释。后来一些解释词语的书也沿用这种排列方法。如《骈字类编》专收双音节词，按第一个字的性质分类编排，分为天地、时令、山水等13门，每门再分若干细目，把1 604个字头分别归入各类，字头下再列词语。

《尔雅》开创的体例也为后代编纂类书、政书所借鉴。类书按所采事、文的内容分门别类编排材料。其缺点是，由于对事物性质认识的局限性，某些事物的归类在今天看来并不恰当，因此查找起来会发生困难。例如，有关桥梁的材料，《艺文类聚》列入水部，《古今图书集成》列入考工典；有关薪、炭的资料，《艺文类聚》列入火部，《古今图书集成》列入草木典。现代的年鉴、手册和一些辞书的类目也是按事物性质、系统归属来立类的。如《中国年鉴》、《中国古代名句辞典》等工具书。

由于古今分类标准极不一致，资料的归类相差较大，使用按分类法编排的工具书时，应先了解所用工具书的分类体系，以确定所查资料的具体类目，并注意相关类的查阅。

4.2.3.4 主题编排法

主题编排法是按既定的主题汇集和编排资料的方法。它将不同学科领域论述

和研究同一问题或同一事物的文献资料集中于同一主题下，按字顺加以排列。

主题编排法在国外是比较常用的方法，几乎每一种检索工具都有以主题词顺序排列的检索途径。学术著作也大都附有主题索引。在我国，仅有部分书本式主题索引，一些工具书的辅助索引也采用了主题编排方法。

使用主题索引主要应注意：

第一，各种索引的主题标引方法不同，标识也不同，检索之前要了解清楚。主题词表是使用主题索引的工具，利用主题词表可以准确地选定主题词，根据参照项还可扩大或缩小检索范围。主题词的概念不能太宽或太窄。同时，要注意语词的规范化。一般选用名词作为主题词。

第二，使用关键词索引时应选择文献中的重要词语，并注意多查同义词、近义词，因为关键词并不规范，容易漏检、误检。

4.2.3.5 时序法

时序法是按事件、事物发生、发展的时间次第性的顺序加以组排的方法。这种排检法常用于年表、历表、大事记、年鉴、年谱等工具书。如，《中国历史纪年表》是严格按照年、月、日的自然顺序来组织和记载历史事件的，《中外历史年表》、《中华人民共和国大事记》等则是按照事件发生、发展的时间顺序编年组织的，《历代人物年里碑传综表》是按照人物的生卒年依次组织的。时序法的特点是线索清晰、检索方便。利用时序法查找资料时，应首先确定事件所发生的时间或人物所处的时代，也可根据有关工具书查出具体时间，然后再利用有关的工具书进行查检。

4.2.3.6 地序法

地序法是按地理区划或行政区划的顺序编排的一种方法。此法主要用于编制和检索地理和地方资料的工具书，如中外地图集、地方志、地名录等参考工具书。《中华人民共和国地图集》、《中国地方志联合目录》、《中国名胜词典》、《最新中国期刊全览》等，皆用此法排检。利用地序法，可以根据某一地名或信息所在地区，方便地查检有关资料。

4.2.4 参考工具书的评价与选择

使用参考工具书，首先要能识别它们的优劣，对其进行客观评价。在选择参考工具书时应从以下几个方面考察。

1. 编纂目的

可根据有关的简介、书评等材料和工具书的前言或序言等了解其编写意图。

同时还可以试查几个自己熟悉的问题，看其实际内容和表现方法是否同预定的目的相符。

2. 权威性

主要考察编著者和出版者的水平和资历、工具书的版次等。这时要了解编著者在本领域的学术地位和学术水平、资料的来源、出版者的出书经验和声望。对于版本而言，古代编纂的工具书一般选用较早的或近年影印的或经过精校的版本。因为古代工具书流传的时间较长，翻刻的次数较多，翻印一次，就会增加不少新的错误。至于近年影印的本子，不会增加排印的错误，并且大都经过精选精校，有的还增加了新编的索引，使用更加方便。而现代编纂的工具书却往往新版比旧版好。因为这些工具书再版时大都经过原编者或有关专家的认真修订、改错和增补。

3. 知识性

要考察其收录的知识、信息和文献资料等内容是否正确，收录是否完备，该收的不漏，不该收的不滥收，内容是否繁简得当，材料来源是否真实、可靠，解释引证是否客观、准确，语言表述是否严谨、规范，是否有拼写和语法错误。

4. 思想性

工具书内容是否含有政治问题或者封建迷信观点，对各种事物的褒贬处理是否恰当，是否带有编辑者个人的主观色彩。

5. 服务对象

工具书总是针对不同层次需求而编纂的。一部优秀的儿童词典或少年百科全书，对少年儿童是好书，却不能满足科研的需要。

6. 收录范围

要了解工具书涉及的学科范围、时间和地域界限、资料类型和来源等。同时要考察其内容的时效性，并同类似的工具书进行比较，比较其全面性、正确性、新颖性。

7. 编排方式

工具书的编排方式直接关系到其使用效果。要考察工具书编排方式是否恰当，是否有完备的索引和参照系统。

4.3 参考工具书的数字化

4.3.1 数字化参考工具书的优势

随着计算机技术、存储技术和网络通信技术的飞速发展，涌现出越来越多的

各种类型的数字化参考工具书（也有人称“电子工具书”），包括书目、索引、百科全书、字典、词典、年鉴、地图、名录等，主要以网络版工具书和光盘版（CD-ROM 或 DVD-ROM）工具书为代表。数字化工具书受到了人们的普遍欢迎。据报道，在国外，《不列颠百科全书》（网络版）在其提供免费查询的头一周内，查阅人数超过 1 000 万人次。而在国内，据中央电视台报道，《金山词霸》在 100 天内销售了 110 万套。目前，已呈现出数字化工具书与传统的印刷型工具书并驾齐驱的趋势。具体说来，数字化工具书的优势主要体现在以下几点：

（1）存储容量巨大。如《中国大百科全书》共有 77 859 条，12 568 万字，出版了 74 卷大开本的印刷版，而光盘版（1.1 版）只有 4 张光盘；《人民日报》50 年全部图文资料只需 16 张光碟。

（2）内容更丰富。虽然许多数字化工具书是以相应的印刷版为基础，但它们不拘泥于印刷版的内容，在电子版中增加了许多新的内容，如增加新条目、综合其他类型工具书的内容、建立同互联网上相关站点的链接等。随着多媒体技术的发展，集文字、图像、声音、影像于一体的检索工具也在不断出现。

（3）检索功能更强、检索迅速。除了保留印刷型工具书原有的检索途径外，数字化工具书往往会利用先进的检索技术，增加许多新的检索功能，如支持模糊检索、组配检索、超文本检索等，还可以实现全文检索。

（4）使用更方便。通过网络数字化工具书可以实现信息的异地传输和检索，且可供多个用户同时使用。读者还可以轻而易举地实现多种工具书的并发查找，并对检索结果进行精确的统计分析。检索结果可以直接打印或复制在磁盘上。

正因为具有众多的检索入口、多层次的检索方式、强大的检索功能，而大大拓展了检索的深度与广度，提高了检索的速度与准确度，数字化参考工具书得到迅速发展。目前，国内外综合的、大型的、优秀的、著名的印刷版检索工具基本上都已经电子化、网络化，并以加速度在向更广泛、更深细的领域扩展。可见，数字化是参考工具书未来发展的一大趋势。

4.3.2 数字化参考工具书举例

4.3.2.1 中文工具书参考咨询系统（http://dlib.zslib.com.cn:8080/was40/tool/tool.htm）

中文工具书参考咨询系统（如图 4—1）是中国目前最大的中文工具书知识库，由广东省中山图书馆研制，是经文化部立项的科研项目，是我国文献资源建设的基础工程之一。该项目在研究中文工具书知识描述规则和建立中文工具书知识数据库的基础上，利用中文全文检索系统建立了一个中文工具书知识库查询系

统。这个系统按照特定的知识描述规则，汇集了字典、词典、百科全书、政书、年鉴、手册、书目、索引、表谱、图录等有关资料，现已存储5万多种，共300多万字的中文工具书的各种信息，涵盖社会科学和自然科学的各个学科领域，可以模拟甚至代替图书馆工作人员解答有关中文工具书的各种咨询问题。该系统对每种工具书除著录书名、作者、出版者、出版时间、开本页码等字段外，还有知识分类号、知识描述词、收录内容、功用等内容；系统软件采用TRS全文检索技术，可从书名、著者、知识分类号、知识描述词、条目名称、收录内容、功用等途径进行全文检索。用户只要输入拟查找的问题，就可以得到有关内容的图书全文，或提供相关工具书的线索。

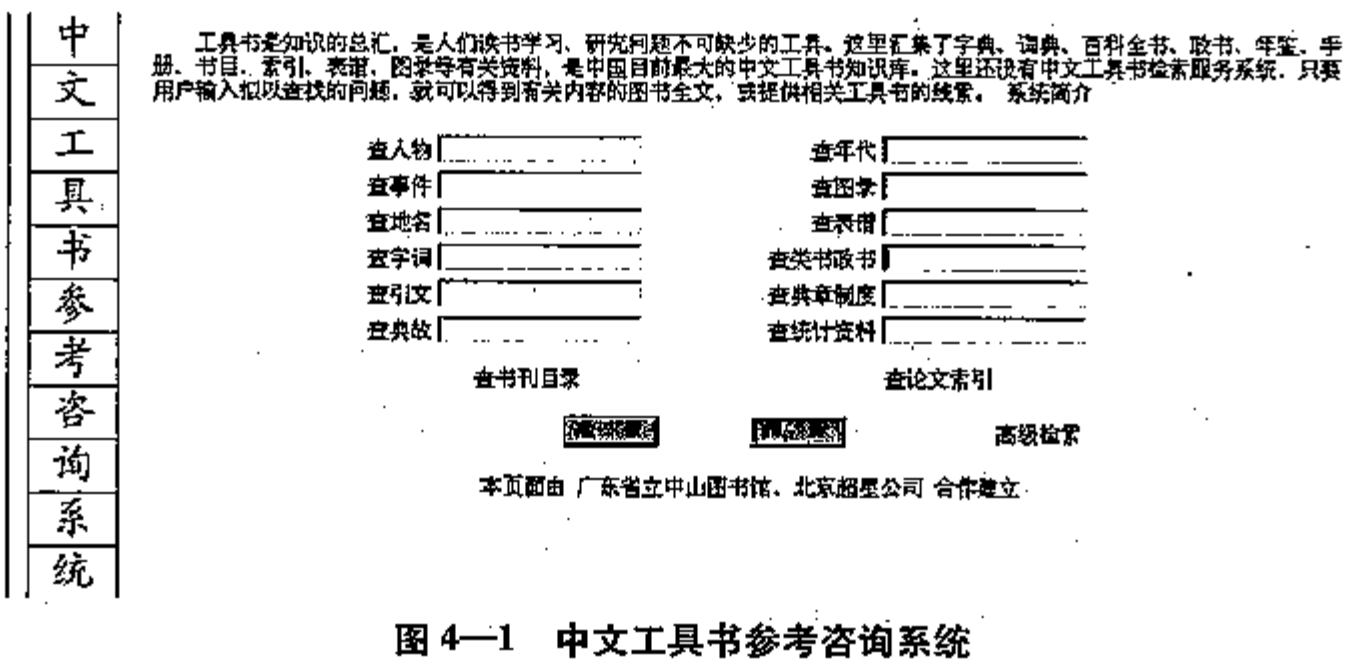

图4—1　中文工具书参考咨询系统

中文工具书参考咨询系统将收录的相关工具书按照用户的查询习惯分为查人物、查事件、查地名、查字词、查引文、查典故、查年代、查图录、查表谱、查类书政书、查典章制度、查统计资料等12大类，如，“查人物”中的相关文件资料共有286篇，包括《中国出版人名词典》、《科学名人传》、《科学家列传》等。“查事件”中的相关文件资料共有21篇，包括《当代百科大辞典》、《简明历史事件辞典》、《中国百科年鉴》等。“查地名”中的相关文件资料共有52篇，包括《中国历史地名词典》、《中国地名大辞典》、《外国地名手册》等。其中，有些工具书可以直接在网上阅读全文，有的只是提供了工具书的题录，指引用户选择自己需要的工具书。

该系统可以方便地查找各种类型的工具书，整个体系的编排也很符合用户查询工具书的习惯，而且，题录标引也比较规范，对于部分工具书还直接提供了全文阅读。

4.3.2.2 《不列颠百科全书》网络版（http://www.britannica.com）

《不列颠百科全书》（*Encyclopedia Britannica*，简称 EB）全套共 32 册，被认为是最权威的大型综合性百科全书，是 ABC 三大百科全书之 B（A 指《美国百科全书》，C 指《科里尔百科全书》），我国知识界过去习惯称之为《大英百科全书》。1768 年首次出版，历经多次修订和再版，已有 200 多年的历史，1974 年推出了第 15 版。《不列颠百科全书》的条目均由世界各国著名学者、各个领域的专家撰写，对主要学科、重要人物事件都有详尽介绍和叙述，200 多年来，《不列颠百科全书》一直是英语世界的知识宝库，其学术性及权威性已为世人所公认。Encyclopedia Britannica Online 是《不列颠百科全书》的网络版，在保留原百科全书的质量和特点的基础上，又增加了许多新的功能，如：功能强大的检索功能和按照主题字顺排列的浏览功能、互联网指南（由百科全书的编辑们搜集、选择和推荐一些最好的互联网站点）以及一些选自最顶尖杂志和报纸上的文章、相关产品等。该工具书两周更新一次，需付费使用，但任何读者都可以申请 3 天的免费试用。界面如图 4—2。

图 4—2 《不列颠百科全书》网络版

该系统提供关键词检索，支持通配符（“?”代表一个字符，“*”代表多个字符）检索、布尔逻辑（AND、OR、NOT）检索、词组检索（用引号括起）、自然语言检索，还可以使用圆括号确定检索符运算的优先次序。该系统也可以提供多种浏览功能，包括字顺浏览、世界地图浏览、时间浏览和主题浏览，用户可以根据自己的需要来进行选择。

4.3.2.3 百科全书网（http：//www. encyclopedia. com）

它是一个可免费浏览百科全书的网站，向读者提供57 000多条来自《哥伦比亚百科全书（第6版）》的最新文章，每篇文章还可链接到报纸和杂志上的相关文献。在该网站的主页上，有两种检索途径可供读者使用：（1）通过检索词的英文字顺进行检索；（2）通过在检索框内输入具体关键词进行检索。通过检索，读者便可浏览到具体的内容。同时该网站还为读者提供与该词相关的其他的关键词，以便对该词的内容有更加全面的了解。界面如图4—3。

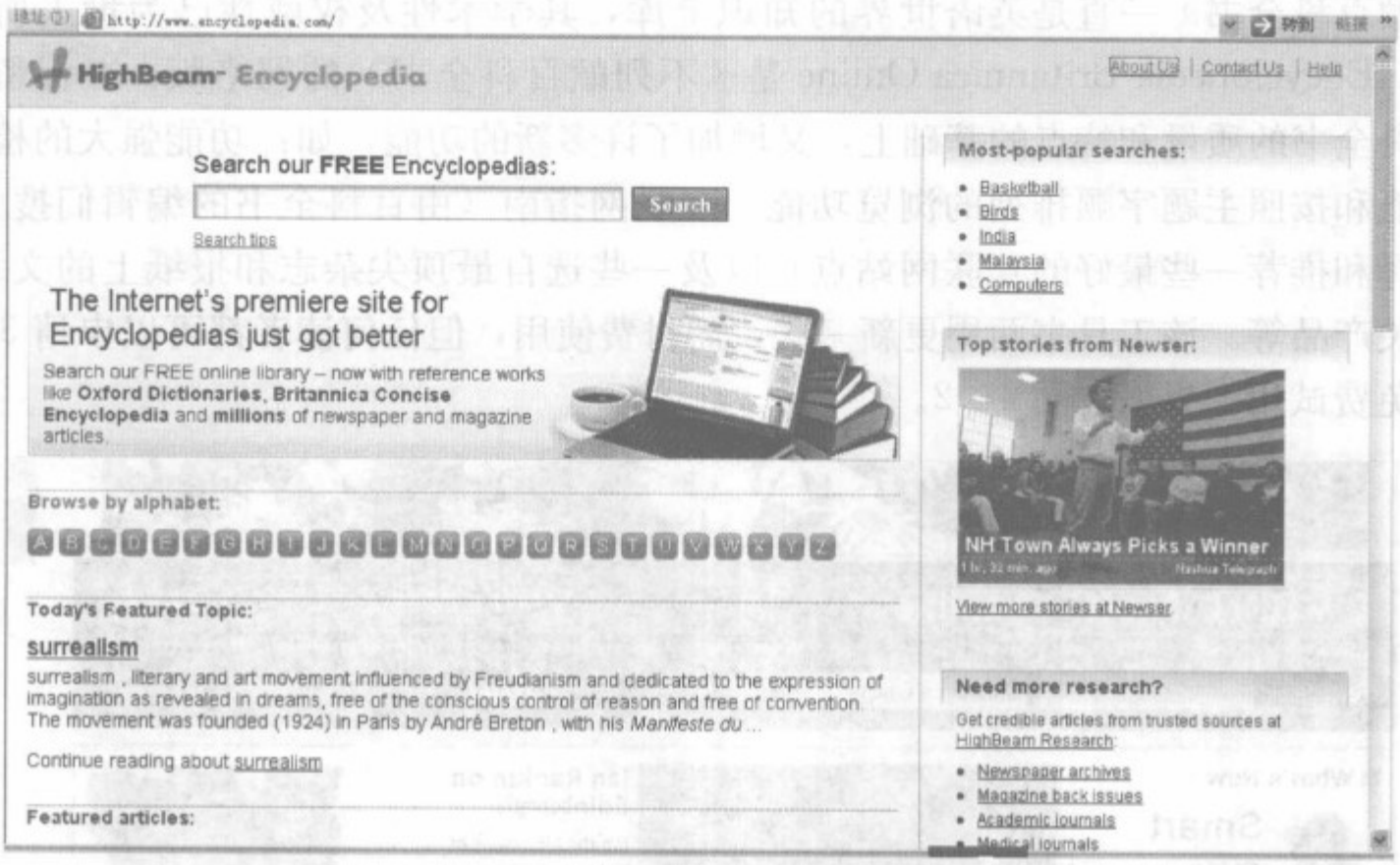

图4—3 百科全书网

4.3.2.4 《世界百科全书》网络版（http：//www. countryreports. org）

收集世界各国的商业、旅游、求学等信息。界面如图4—4。

4.3.2.5 中国网上114（http：//www. china-114. net）

该网站由上海禾升科技有限公司创建。在该网站的主页上，读者可通过三种途径免费检索到全国许多企事业单位的具体信息：（1）按地区分类检索有关单位的信息，即以各省市自治区名称为检索入口，点击你需要查找的具体省市自治区名称，就可首先查到该省市自治区的具体单位名称及产品。再点击具体单位的名称，就可以检索到该单位更加具体的信息，即法人代表、联系电话、单位邮编、地址、人员情况、电子邮件、产品服务等。（2）按行业分类检索有关单位的信息，即以行业名称为检索入口，点击你需要查找的具体行业名称，就可检索到该

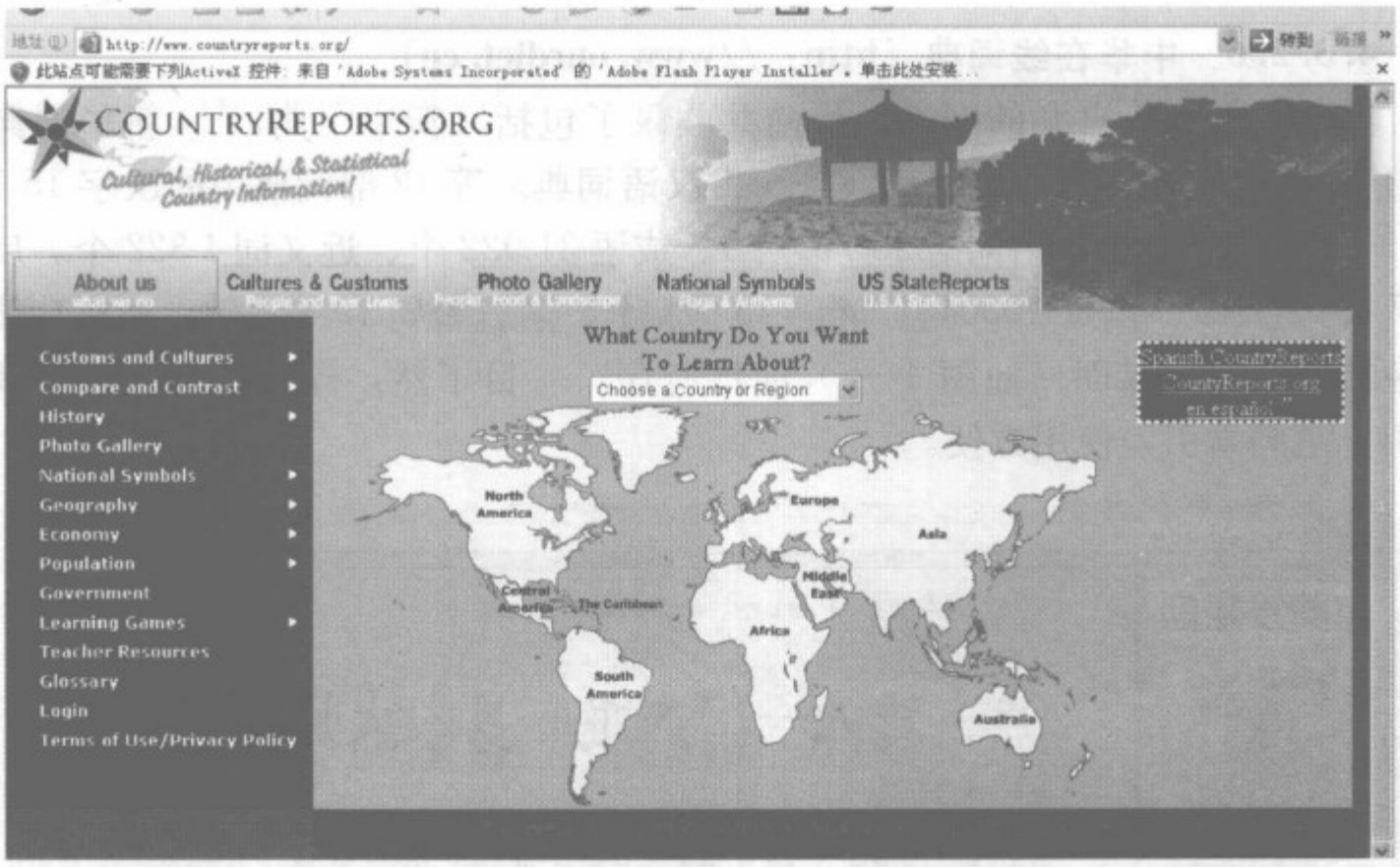

图 4—4 《世界百科全书》网络版

行业单位的具体信息（内容同上）。(3) 按单位名称、单位地址等字段进行检索。用户在主页的输入框中输入单位名称或地址并确定后，则可浏览有关信息。同时，还可以免费检索到一些产品服务信息。界面如图 4—5。

图 4—5 中国网上 114

4.3.2.6 中华在线词典（http：//www. ourdict. cn）

这是一个个人自办的网站，目前共收录了包括《新华字典》、《新华词典》、《现代汉语词典》、《现代成语词典》、《古汉语词典》等12部词典中的汉字15 702个、词语36万个（常用词语28 770个）、成语31 922个、近义词4 322个、反义词7 691个、歇后语14 000句、谜语28 071个、名言警句19 424句，所有功能都是免费使用。其界面（如图4—6）简洁有序，一目了然，提供拼音索引、部首索引、笔画索引，使用方便。

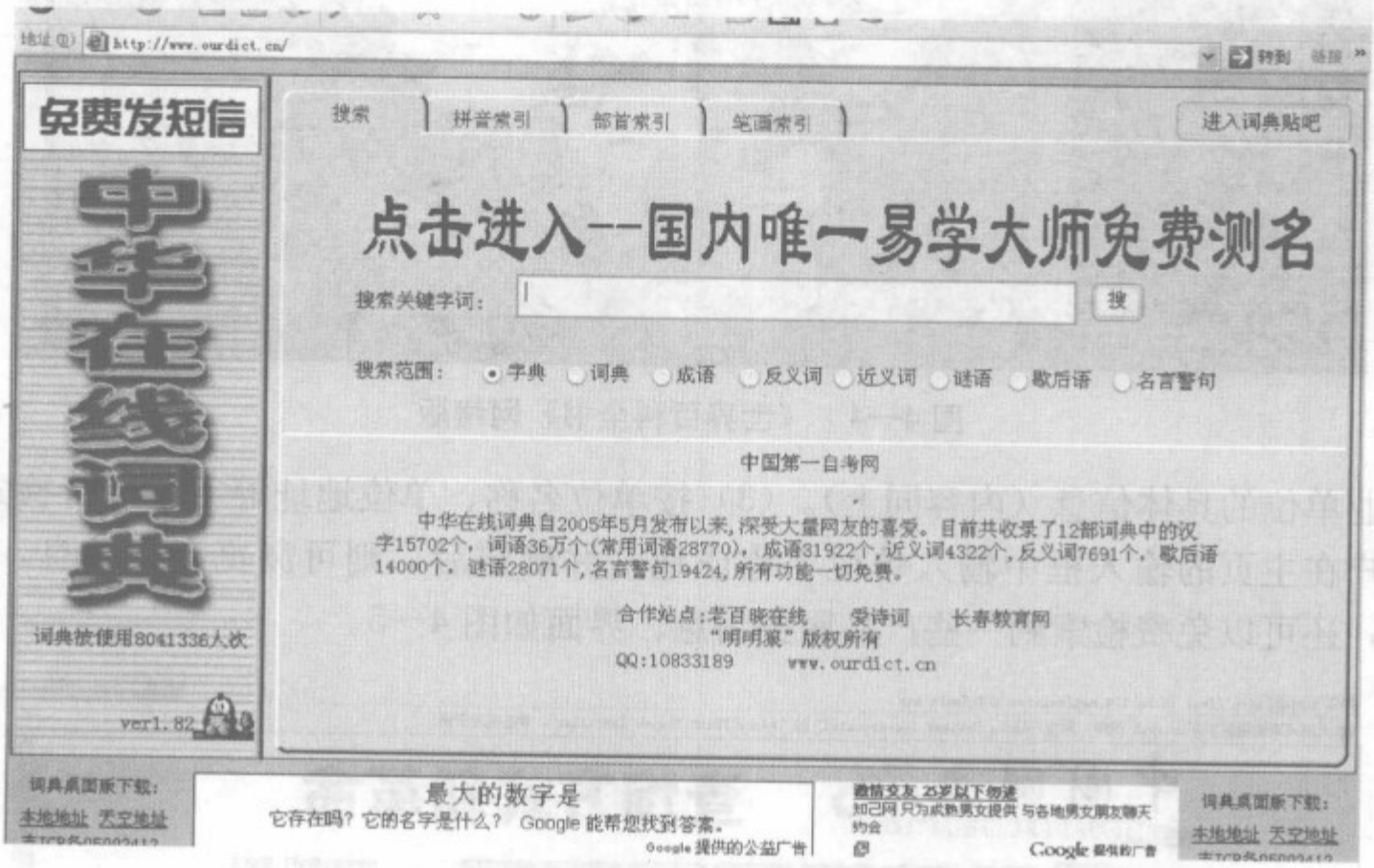

图4—6 中华在线词典

4.3.2.7 牛津参考工具书在线（http：//www. oxfordreference. com）

提供的多为人文学科工具书。收录了牛津大学出版社所有学科100种以上的字典、辞书与百科全书，范围包括语言、一般参考、人文及社会科学（艺术及建筑、文学、政治及社会科学、宗教与哲学等）、经济与商业、法律、科学与医学（生物科学、计算机科学、地球与环境科学、一般科学、物理学与数学等）。含多种查询功能，每个词汇都有链接以供互为参照，另含有超过1 000个网站的链接以作为线上补充参考资源。内容随时更新。其界面如图4—7。

4.3.2.8 Xreferplus电子参考工具书在线（http：//www. xreferplus. com）

Xreferplus含全世界最主要的25个大型出版社（包括麦克米兰、剑桥大学出版社、Elsevier、Blackwell、企鹅等）的130本各种类型参考书，包括百科全

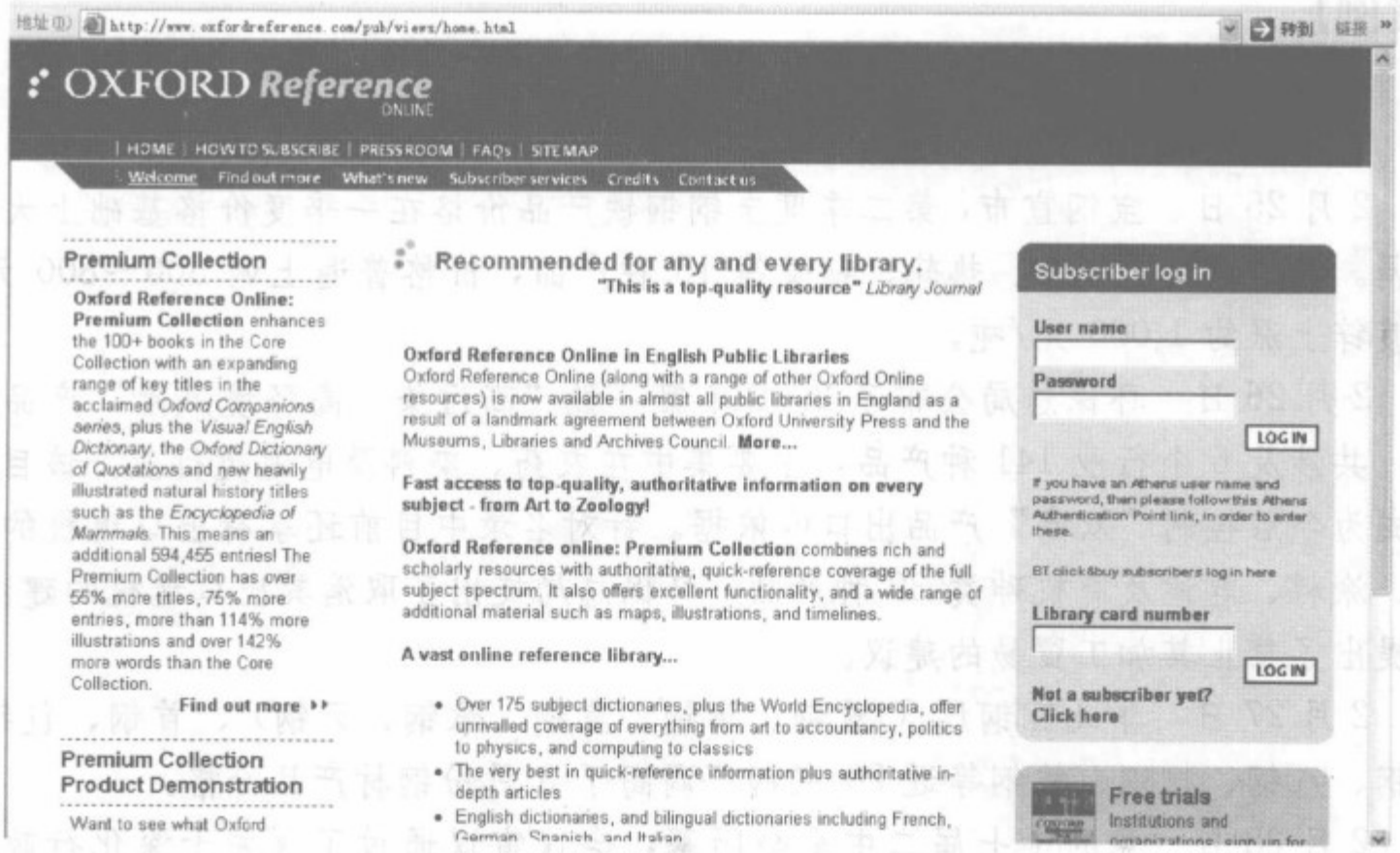

图 4—7 牛津参考工具书在线

书、字词典、索引典、引用语辞典，内容超过 150 万个词条，主题涵盖所有学科领域。其界面如图 4—8。

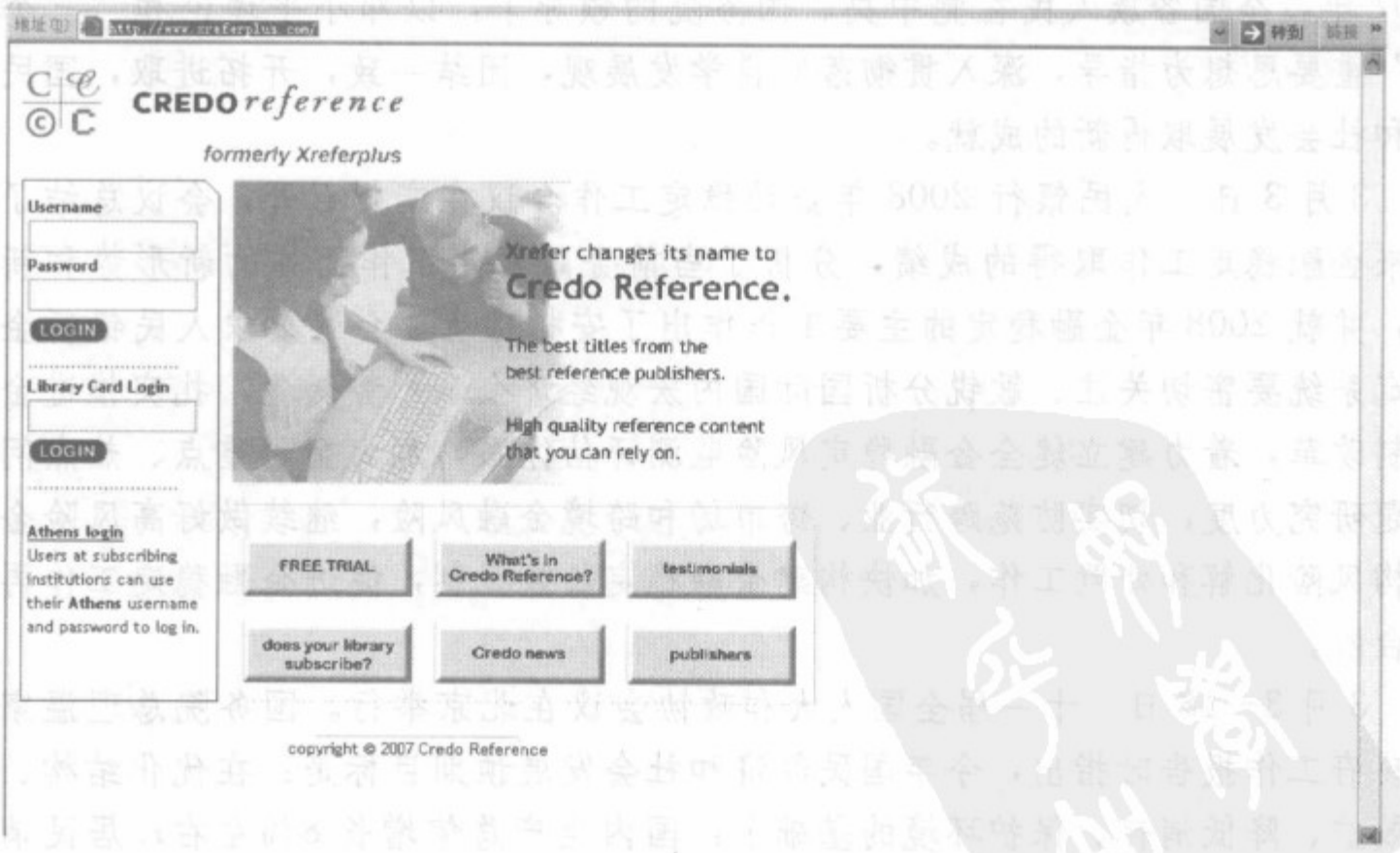

图 4—8 Xreferplus 电子参考工具书在线

【案例】

时序法应用之一：经济大事记（2008年2月20日—3月19日）[①]

2月25日　宝钢宣布，第二季度宝钢钢铁产品价格在一季度价格基础上大幅上调。其中钢坯、线材、热轧、钢管等15种产品，价格普遍上调500～800元，热镀锌上涨约1 000元/吨。

2月26日　环保总局公布了2008年第一批“高污染、高环境风险”产品名录，共涉及6个行业141种产品，主要集中在农药、染料及电池类行业。该目录将成为今后控制“双高”产品出口的依据。针对名录中目前还享有出口退税的农药、涂料、电池及有机砷类39种产品，环保总局提出了取消其出口退税的建议，并提出了禁止其加工贸易的建议。

2月27日　华北五钢厂（首钢、唐钢、宣钢、承钢、天钢）、首钢、包钢、太钢、八钢、柳钢、杭钢等近20家钢厂调高了3月份钢材产品价格。

2月28日　党的十七届二中全会闭幕，会议审议通过了《关于深化行政管理体制改革的意见》和《国务院机构改革方案》，同意把《国务院机构改革方案》提请十一届全国人大一次会议审议。

中华人民共和国国家统计局发布《2007年国民经济和社会发展统计公报》，2007年，全国各族人民在党中央、国务院的领导下，以邓小平理论和“三个代表”重要思想为指导，深入贯彻落实科学发展观，团结一致，开拓进取，国民经济和社会发展取得新的成就。

3月3日　人民银行2008年金融稳定工作会议在广州召开，会议总结了几年来金融稳定工作取得的成绩，分析了当前金融稳定工作面临的新形势和新任务，并就2008年金融稳定的主要工作作出了安排部署。会议要求人民银行金融稳定系统要密切关注、敏锐分析国际国内宏观经济金融形势变化，扎实推进金融体制改革，着力建立健全金融稳定风险监测评估体系，加大金融重点、热点问题专题研究力度，切实防范跨行业、跨市场和跨境金融风险，继续做好高风险金融机构风险化解和处置工作，加快构建金融稳定长效机制，促进金融稳定工作再上新台阶。

3月3—18日　十一届全国人大和政协会议在北京举行。国务院总理温家宝作政府工作报告时指出，今年国民经济和社会发展预期目标是：在优化结构、提高效益、降低消耗、保护环境的基础上，国内生产总值增长8%左右；居民消费

---

① 本刊资料室：《经济大事记（2月20日—3月19日）》，载《宏观经济管理》，2008（4）。

价格总水平涨幅控制在4.8%左右；城镇新增就业1 000万人，城镇登记失业率控制在4.5%左右；国际收支状况有所改善。防止价格总水平过快上涨，是今年宏观调控的重大任务。

3月17日 住房和城乡建设部表示，将采取措施防止房价大起大落。一是明确政府的责任，把价格管理、监管纳入到政府的目标责任制，实行问责制。二是采取综合措施，力求总量基本平衡、结构基本合理、价格基本稳定。三是加强市场的监管。

3月19日 中国人民银行决定从2008年3月25日起，上调存款类金融机构人民币存款准备金率0.5个百分点，至15.5%。

关键术语

工具书 参考工具书 参考工具书概念 参考工具书功能
参考工具书特点 参考工具书种类 参考工具书结构 参考工具书排检方法
参考工具书评价 数字化参考工具书

思考题

1. 简述参考工具书的功能。
2. 和普通图书相比，参考工具书有何特点？
3. 参考工具书有哪些类型？
4. 参考工具书主要有哪几种排检方法？简述各自的特点。
5. 与传统的印刷型参考工具书相比，数字化参考工具书有哪些优势？

CHAPTER FIVE

第5章 参考工具书使用

【本章要点】

◇ 图书与知识型信息检索（包括书目、索引、文摘、字典、词典）

◇ 数据与事实型信息检索（包括年鉴、百科全书、手册、名录）

引子

参考工具书内容丰富，种类繁多，体例各异。而用户的信息需求一般是具体的、特定的。如果说工具书是开启知识宝库的钥匙，就要对号开锁，要学会根据面临的问题正确地选择工具书。工具书包罗万象，从古到今，从文到理，从一首绝唱到民间俗语，从一个公式、定律到百科全书。要对工具书了如指掌，了解什么问题、什么要点、什么知识、什么内容可以查找哪一类或哪一本工具书，并非一朝一夕的事，这既需要有理论知识，又要有实践技能。[①] 不同类型的参考工具书具有不同的特点，适合于检索特定的信息资源。在检索实践中，相应于不同的信息需求，可以有针对性地采用特定类型的参考工具书，以便于提高检索效率和检索质量。参考工具书成千上万，除了要知道哪类问题该用哪类工具书解决外，还要熟悉一些具体参考工具书的内容范围、特点、编排结构，以及它们之间的相互关系，包括内容和时间上的联系。这样，就能在使用参考工具书的过程中，驾

① 参见徐军英：《自觉学习与继续学习——从大学生利用工具书谈起》，载《江西图书馆学刊》，2000（4）。

轻就熟地选择，有的放矢地进行查找。①

5.1 图书与知识型信息检索

5.1.1 书目、索引、文摘

书目、索引、文摘是人们查检书刊文章等信息的重要检索工具。书目是图书或报刊目录的简称。它著录和揭示一批相关的文献，按一定的次序编排而成，是一种登记、报道和宣传文献的检索工具。索引是把一种或多种文献中具有检索意义的内容，如字、词、句、人名、地名、书名、篇名、主题等摘录下来，按一定顺序加以编排并注明出处以供查检的工具。文摘是以简明扼要的文字对文献原文的摘述。它将论文或书籍的主要观点、论据、数据等摘录出来，并按一定方式编排，是当代报道学术动态的简捷明快的方法，也是一种文献检索和阅读的工具。

书目、索引、文摘同作为线索性工具书，有一些共同点。三者都是揭示和检索原始文献的工具，它们的编制离不开原始文献，其内容受制于被揭示的文献。它们的利用也是以原始文献为依托的，在一般情况下，它们必须与原始文献配合使用，才能够给读者提供完整的信息。三者都是对原始文献的描述和揭示，其作用在于方便读者检索原始文献。它们详细完整地著录了原始文献（单元）的外部特征及部分内容特征，以便读者甄别、选择文献。它们提供了原始文献（单元）的各项检索标识，包括文献题名、作者、分类号、主题词、序号、代码等，便于读者利用这些标识检索原始文献。

书目、索引、文摘也有不同之处，主要体现在著录的深度和详略上。书目一般以每一种完整的文献为揭示对象，著录项目强调版本、发行事项的揭示，提要除介绍文献内容外，还强调著者介绍，并且注重评论得失和考证。书目除用于报道文献外，还有指导阅读和指示藏所的作用。索引一般以文献里的事项或单元知识为揭示对象，强调揭示文献的内容特征，比书目具有更强的检索深度。而且索引的著录内容标明出处，主要起到指向和示址作用。文摘的揭示对象既有论文，也有书籍，著录项目包括基本项目和文献摘要，其摘要主要是原始文献的高度浓缩，一般不加评论。它包含有更多的信息量，不但有指向和示址作用，还有报道的功能。

---

① 参见沈固朝：《信息检索（多媒体）教程》，36页，北京，高等教育出版社，2002。

5.1.1.1 书目

1. 书目的产生和发展

中国是最早出现书目的国家之一。在中国历史上，书目曾经有过多种称谓，如“录”、“志”、“略”、“簿”、“考”、“书录”、“提要”等。西汉刘向及其子刘歆先后编成中国第一部解题式书目《别录》和第一部综合的群书目录《七略》。现存最早的史志书目《汉书·艺文志》，距今已有1 900多年的历史了。唐、宋、元、明、清时期，书目层出不穷，是我国古典目录学的发展时期。到了近代，书目的编纂有了较大的发展，不仅数量增多，而且种类丰富，出现了联合目录、馆藏目录、报刊目录、专科专题目录等。除《北京图书馆善本书目》、《中国地方志联合目录》等一批古籍书目外，还有《全国总书目》、《全国新书目》、《全国中文期刊联合目录》、《鲁迅研究资料编目》等一大批综合性和专题性的书刊目录。随着现代科学技术的发展，不仅有书本式、卡片式目录，还出现了缩微目录、机读目录等新的目录形式。

2. 书目的类型

对于书目类型的划分问题，众说纷纭。不同的划分标准构成众多的书目类型。按照书目的编制目的和社会职能，可分为登记书目、通报书目、推荐书目（导读书目）、书目之书目（书目指南）等；按照书目收录文献的内容范围，可分为综合书目、专题（或专科）书目、地方文献书目、个人著述书目等；按书目反映文献收藏情况，可分为馆藏目录、联合目录等；按文献出版与书目编制的时间关系，可分为现行书目、回溯书目、预告书目等；按书目著录文献类型特征，可分为图书目录、期刊目录、档案目录、报纸目录、丛书目录、方志目录、古籍目录、乐谱目录、盲文图书目录等；按时代，可分为古典书目和现代书目，前者又包括官修书目、史志书目、版本书目、私撰书目等。除上述几种划分标准外，还可按检索途径、出版物文种等划分。按不同角度划分书目的种类，是为了从不同角度揭示其功能，以满足读者的不同需要。因此，同一个书目按不同的划分标准可属于几种类型。

下面重点介绍一些常见的书目类型：

（1）登记书目。它是反映一国或一地在一定历史时期或一定范围的出版、收藏情况的登记统计性书目。国家书目是登记书目的主要类型之一，它比较全面地反映了一个国家的科学、文化方面出版的成果。

（2）馆藏书目。它是揭示和报道一个图书馆或其他信息机构所收藏的各种文献的总的书目。它的主要作用是为读者利用图书馆、档案馆文献指引门径，在传统时代以卡片目录的形式为主，在网络时代以公共联机检索目录（OPAC）的形

式提供网上服务。

(3) 推荐书目。它是针对特定的读者或特定的目的，围绕某一主题，选择推荐有关文献，用以指导阅读而编制的书目。我国的推荐书目产生于传统官学、家塾、书院与科举制度的需要，至今依然有比较大的影响。推荐书目具有较强的教育性，如张之洞的《书目答问》作为一本重要的推荐书目，问世以后对中国的学术界产生了广泛的影响。

(4) 书目之书目。它是将书目、索引、文摘汇集在一起而编成的一种特殊的书目类型。读者可通过它了解已出版的书目工具的性质和特点。如《书目举要》、《中国历代书目丛刊》等。

(5) 综合书目。它是将各个学科门类的图书综合编制而成的一种图书目录。其内容广博，包罗万象，既有哲学社会科学方面的图书，又有自然科学和技术科学方面的图书。

(6) 专题书目。它是反映某一学科或某一专题图书文献的书目。如《鲁迅资料研究编目》、《中国古典文学名著题解》、《中国农学书录》等。

(7) 地方文献书目。它是专门收录有关某一地区历史、自然和社会状况的图书文献的书目。如《安徽文献书目》、《绍兴地方文献考录》等。

(8) 个人著述书目。它是专门收录某一作者的全部著述，并兼收别人研究该作者的生平和著述的图书资料的目录。如《鲁迅研究资料汇编》、《郭沫若著译书目》等。

(9) 联合目录。它是反映图书在全国或某地区若干个图书馆的收藏情况的目录。如《中国丛书综录》、《中国地方志联合目录》、《中国革命历史档案目录》等。

3. 书目的作用

书目的作用主要体现在以下两个方面：

(1) 指导读书门径。“人生也有涯，而知也无涯。”当今图书数量猛增，形态复杂，种类繁多。若想在浩瀚的书海中选择所需之书，必须借助于书目。书目提供了文献的基本信息，如题名、责任者、出版发行、载体形态等。有的书目对文献的内容、作者、流传情况加以扼要的介绍并评论得失。一些现代书目还介绍文献的馆藏。书目一般具有篇名、主题、分类、著者等检索途径，起着宣传图书文献、方便检索、指导阅读的作用。书目向读者展示了一个国家、一门或几门学科在一定时期的图书文献出版情况，大都按学科系统编排，分类详密。通过书目，我们可大致了解各类图书的出版状况，每种图书的信息内容，并根据自己的需要和爱好选择取舍。

（2）指导科学研究工作。任何科学研究，在提出问题、确定研究课题之前，都要知晓本学科的发展历史，了解前人对这些问题有哪些研究成果，只有继承前人的成果，才能在此基础上创新与发明。通过书目，我们可了解本学科的研究历史和研究现状。特别是通过各类新书目，可掌握本学科最新研究成果，这对考知学术源流、确定研究课题是非常重要的。此外，在研究过程中，需要收集和查找资料，这也离不开书目，通过古今各类书目，可查寻到与研究课题密切相关的事实和资料。

4. 书目举要

（1）查现存古籍。

查现今存世的古代图书，可利用以下几部书目：

《四库全书总目》，清永瑢、纪昀等奉敕编，中华书局1965年影印出版，1981年重印。这是我国历史上最大的一部解题目录，于清乾隆年间纂修《四库全书》过程中编录。《四库全书总目》著录收入《四库全书》的古籍3 461种，79 309卷；同时将认为价值不高或“词意抵触”的著作随类收入“存目”，有6 793种，93 551卷。本书目收录比较完备，基本上包括了清代乾隆以前我国古代的重要著作，尤其是元代以前的古籍收录更为完备。此书目对我国古籍进行了系统的著录、评价，为我们了解古代各类著作提供了不少方便，在分类体系上，对传统的四部分类法有所调整和增补，因而对后世影响较大。但它毕竟成书较早，当时被禁毁或后来又被发现的古籍，当然不可能从中查到；同时，其内容也有不少错误，后有若干著作或书目订误或补正。

《书目答问补正》，张之洞撰、范希曾补正，中华书局1963年影印，上海古籍出版社1983年出版瞿凤起校点本。这是一部导读书目。张氏原本选录了常见的重要古籍2 200多种。其中百分之三四十是《四库全书总目》未有的，余下《四库全书总目》虽有，但校本、注本晚出者又占大多数，因此对查找清乾隆以后的古籍很有参考价值。

《贩书偶记》，孙殿起编，中华书局上海编辑所1959年重印1936年初刊本，上海古籍出版社1982年出版增订本。这是一部古籍知见书目，是编者经营古籍购销事业的详细记录，收录了作者所见的清代后期的著述，大体可当作《四库全书总目》的补编，同时也兼收了少量乾隆以前而未被《四库全书总目》收录的明人著作，以及辛亥革命至抗日战争以前有关古代文化的著作，共9 000余种。

《中国丛书综录》，上海图书馆编，中华书局1959—1962年出版，上海古籍出版社1982—1983年出新版。这是我国历史上规模最大、收录最广、体例最完备的一部古籍丛书联合目录。初版反映了全国41个图书馆的馆藏，收丛书2 797

种，包括38 891种古籍。新版订正了原版的一些错误，并补录了黑龙江图书馆等6所图书馆的收藏情况。本书收录比较完备，大体上收录了现存古书的近一半，基本上反映了我国历代丛书的现存情况。

《中国古籍善本书目》，它是查考古籍善本方面最重要的工具书。中国古籍善本书目编委会编，上海古籍出版社1986年起陆续出版。该书目是我国历史上第一部比较完备地反映现存古籍善本的联合目录，收录全国782个收藏单位珍藏的古籍善本6万多种，约13万部。

(2) 查近现代图书。

《民国时期总书目》，北京图书馆编，书目文献出版社1986—1997年出版。该书目收录1911年至1949年9月间我国出版的中文图书12万余种，占此期出书总数的90%以上，基本上反映了这个时期图书出版的情况。该书目按学科分类，分册编辑出版，共20册，具有回溯性国家书目的性质。

《中国近代现代丛书目录》，上海图书馆1979年编印。该书收录上海图书馆所藏1902—1949年10月出版的中文丛书（线装古籍除外）5 549种，子目30 940余条。所收丛书包括中国共产党领导下出版的革命丛书，进步社团、进步出版社、著名作家、书商编印的丛书，介绍外国学术思想、世界文学名著的丛书，重要人物的个人丛书或全集以及百科性的丛书等。

(3) 查当代图书。

《全国新书目》，1950年创刊，月刊。原由国家版本图书馆编，中华书局出版，现由新闻出版总署信息中心主办，全国新书目杂志社编辑出版。该书目是根据全国出版单位缴送的样本编成的，是及时报道全国各地每月出版新书情况的权威性总目录。

《全国总书目》，1949年创刊，年刊，原由中国版本图书馆编，中华书局出版，现由新闻出版总署信息中心编辑出版。它是《全国新书目》的年度累积本，基本上按年度收录了全国各正式出版单位的当年出版物，包括公开发行或限国内发行的各种文字的初版和改版图书（中国香港、台湾出版的未收）。该书目是目前我国连续出版时间最长、收录书刊比较齐全的权威性书目，是了解新中国成立后我国书刊出版情况的必备工具。

《中国国家书目》，年刊，1985年开始编辑，1987年首次出版。1987—1994年版均由北京图书馆《中国国家书目》编委会主编，书目文献出版社出版，1995—1998年版由华艺出版社出版。该书目的具体收录范围是：普通图书、连续出版物、博士论文、乐谱、地图、技术标准、专利文献、非书资料、书目索引、少数民族语文出版物、盲文出版物、中国领土内出版的各种外国语出版

物等。

（4）查近现代报刊。

《中国近代期刊篇目汇录》，上海图书馆编，上海人民出版社 1965—1984 年出版。它是查找我国近代期刊及其资料篇目的一部大型工具书。全书分 3 卷：第 1 卷（一册），1857—1899 年；第 2 卷（上、中、下三册），1900—1911 年；第 3 卷（上、下两册），1912—1918 年。收录了全国 51 个图书馆所藏的 1857—1918 年出版的比较重要的，侧重于哲学、社会科学方面的中文期刊 495 种，11 000 余期。

《全国中文期刊联合目录（增订本 1833—1949）》，全国图书联合目录编辑组编，书目文献出版社 1981 年版。该书目收录了全国 50 所图书馆 1833 年至 1949 年 9 月入藏的国内外出版的中文期刊（包括中国共产党在各个时期出版的党刊，抗日民主根据地和解放区出版的期刊，以及国民党统治区出版的进步刊物）约 2 万种。

《1833—1949 全国中文期刊联合目录补编本》，北京图书馆、上海图书馆编著，书目文献出版社 1994 年版。该书 1981 年增订本有相当部分期刊未被收录。补编本补收清末至民国时期期刊 16 400 余种，涉及范围有：珍藏革命刊物，国民党之党、政、军刊物，抗日战争时期伪满、伪华北、汪伪政权之军政机关刊物，中小学教育刊物，儿童刊物，文艺刊物。增订本和补编本基本反映了 1833—1949 年我国期刊出版全貌及馆藏情况。

（5）查当代报刊。

《中文核心期刊要目总览》（第 2 版），林被甸、张其苏主编，北京大学出版社 1996 年出版。编者用文献计量学方法，从 10 331 种中文期刊中，经过统计、分析，筛选出 1 578 种作为核心期刊，分属 131 个学科，覆盖文、理、医、农、工学科。

5.1.1.2 索引

1. 索引概述

索引与书目一样具有悠久的历史。索引在我国旧称“通检”、“备检”、“玉键”、“韵编”、“针线”，又有根据英文“Index”一词音译为“引得”。在我国古代，虽然没有明确的“索引”一词，但编纂索引的实践从汉代就开始了。司马迁在《史记·太史公自序》中，详列了《史记》130 篇的篇名、序列号，并按本纪、表、书、世家、列传五大部分排序，条理分明，具有一定的检索作用，可以看作是我国索引的萌芽。有人认为，《宋史·艺文志》所著的《群书备检》是我国最早的篇名索引，但原书已佚，无法定论。一般认为，唐代唐林宝的《元和姓

纂》是我国真正意义上的第一部人名索引。到了近现代，随着科学文化事业的发展及报刊的出现，索引大量涌现，索引的作用日益突显，尤其是在20世纪50年代，电子计算机开始进入索引编制工作领域，使这项工作产生了质的飞跃，并产生了新的索引理论与方法。目前国内外索引的编制大部分实行了自动化，使索引的适时性问题得到较好的解决，许多过去用手工方式难以完成的工程巨大的索引大量问世。

索引能够提高文献检索的深度和检索效率，并且可以满足多途径检索的要求，有人将索引的作用喻为“书海雷达”，可以帮助人们迅速地查检到所需的文献资料。

索引的类型多种多样，可以从不同的角度来划分。按照编排方法或检索途径，索引可分为书目索引、篇目索引、字句索引、主题索引、分类索引、专名索引和引文索引。

(1) 书目索引。专以群书（如总集、丛书、类书等）中的图书目录为对象，以一定方式编排而成的索引。如《四库全书目录索引》、《艺文志二十种综合引得》等。

(2) 篇目索引。将各类文献中包含的单篇文章进行分析著录后，再按一定规则排列起来的索引。其作用是供读者从篇名入手查找图书、期刊、报纸等文献中的有关文章。篇目索引一般包含论文标题、责任者及该文文献出处等著录事项，因此又称为“题录”。如《全国报刊索引》就是收录了报刊单篇文献的篇目索引。

(3) 字句索引。以书中摘出的字、词语、句子为著录单位编成的索引。主要用于检索古籍、经典著作等的内容。有逐字索引，如《毛诗引得》、《尚书通检》等；词语索引，如《水浒全传词汇索引》等；句子索引，如《十三经索引》等。

(4) 主题索引。将书中的全部资料按主题集中而编成的索引。其功能是可供读者从文献论述的主题角度来检索特定的文献，比如《马克思恩格斯全集主题索引》等。

(5) 分类索引。以代表某一知识分类体系或逻辑体系的分类号或类名作为标目，按相应的分类系统编排而成的索引。其功能主要是可供读者从文献内容的学科属性角度来查检所需的文献。如《全国报刊索引》就是按学科分类编辑而成的索引。

(6) 专名索引。将文献中的人名、地名、事物名等专有名词加以摘录整理而成的索引。如《三国志地名索引》、《二十四史纪传人名索引》等。

(7) 引文索引。是一种以文献间的引证关系为基础编制的、供人们从被引证文献角度检索引证文献的索引，又称“引证索引”。如《科学引文索引》。

2. 索引举要

《艺文志二十种综合引得》，哈佛燕京学社引得编纂处编，1933 年初版，中华书局 1960 年订正影印，上海古籍出版社 1986 年重印。引得编纂处共编过 64 种“引得”，本引得是其中之一。本书是根据 15 种“正史”艺文志、经籍志和其他 5 部禁毁书目编成的书名、作者综合索引。中国史书的 20 种艺文志中，原有艺文志的有 7 部：《汉书》、《隋书》、《旧唐书》、《新唐书》、《宋史》、《明史》、《清史稿》。后人补艺文志的有 8 部：《后汉书》、《三国志》、《晋书》、《五代史》、《宋史》、《辽金元史》、《元史》以及《补三史艺文志》。5 部禁毁书目是：清代《禁书总目》、《全毁书目》、《抽毁书目》、《违碍书目》、《征访明季遗书目》。所收图书自先秦至清末共 4 万多种。

《十三经索引》，叶绍钧编，开明书店 1934 年初版，中华书局 1983 年出版重订本。这是一部群书文句索引。所谓《十三经》，是古代十三部儒家经典著作的总称，即《周易》、《尚书》、《诗经》、《周礼》、《仪礼》、《礼记》、《春秋左传》、《春秋公羊传》、《春秋穀梁传》、《论语》、《孝经》、《尔雅》、《孟子》等。它们是研究古代历史和学术思想的重要资料。

《十通索引》，商务印书馆编，该馆 1937 年出版。本索引专供寻检商务印书馆印行的《十通》中的文章细目和词句之用。

《全国报刊索引》，原名《全国主要期刊资料索引》，创刊于 1955 年 3 月，由上海图书馆编印，双月刊。1956 年增收报纸资料后，改刊名为《全国报刊索引》，从 7 月起改为月刊。1966 年停刊，1973 年 10 月复刊，仍命名为《全国报刊索引》。自 1980 年开始分哲社版、科技版编辑出版，收录国内公开出版和内部发行的中文报刊资料。

《复印报刊资料索引》，中国人民大学书报资料中心编印。该刊创办于 1958 年。现设专题 122 个，每期都定期出版。其资料主要来源于国内公开出版和内部发行的报刊约 3 000 多种。每期分复印和索引两部分，复印部分选择该专题中水平较高的予以复印；索引部分则以题录形式报道未复印的论文。其优点在于：设置的专业专题齐全，因长期积累，具有系统性和完整性；由于可直接查得某些专题的文献，节省了大量检索文献的时间。现已依据纸本的《复印报刊资料索引》，推出了系列光盘数据库，方便了用户的查检。

5.1.1.3 文摘

1. 文摘概述

文摘作为一种独立的工具书形式，最早产生于欧洲。最初是在期刊杂志中设置文摘专栏，对已出版的各种科学图书正文进行摘录。例如，1665 年法国科学

院创办的《科学家杂志》，以及同年英国出版的《皇家学会哲学汇刊》，均设有文摘专栏。后来，随着科学技术的进步，原始文献相应增多，科学期刊大量涌现，文摘在其中所占的篇幅也日益增大，从19世纪开始逐渐产生了专门的文摘杂志。如德国1830年创刊的《化学总览》，美国1884年创刊的《工程索引》，1896年创刊的《科学文摘》等，1928年创刊的《社会科学文摘》。其中，《社会科学文摘》是第一家综合性的社科文摘刊物。

关于我国文摘的活动可追溯到南宋袁枢的《通鉴纪事本末》。1897年5月陈念萱在上海创办了《集成报》，这可以看作是我国最早的文摘杂志。该刊是综合性文摘，对政治、实业、科技、史地掌故等均有摘录。一般认为，我国最早的科技文摘是1934年中国化学会编印的《化学》杂志，其中，设有“中国化学摘要”专栏。新中国成立后，我国的文摘得到了一定的发展。1956年，翻译了苏联的《文摘杂志》，1958年起创办了《现代外国哲学社会科学文摘》、《中国机械文摘》、《中国化学化工文摘》等。1966—1977年，文摘刊物几乎全部停办。自20世纪80年代起，我国开始定期正规地出版文摘，不仅出版了一大批新闻文摘，学术性检索性文摘也纷纷出现。目前，我国各主要学科基本上都具有自己的文摘刊物。

文摘型检索工具已成为现代重要的检索工具类型。它具有报道、检索、参考、示址、交流等功能，除帮助做出文献相关性判断外，还能避免阅读某些文献的全文，帮助克服语言障碍。

文摘型检索工具中，对应每一文献的文摘，一般由三部分构成：（1）题录。也称书目著录事项，它是对文献的外表特征进行著录，以便识别、索取原文的项目记录。一般由题名、著者及其单位、文献出处等项组成。（2）文摘正文。即表述文摘内容的短文，是文摘的主体部分。（3）补充事项。主要包括参考文献数、插图或表格数、原文所用语种、文摘编写者、文摘员所加的脚注。一般置于文摘正文的末尾。

文摘按其对原文献的揭示程度划分，可分为报道性文摘和指示性文摘。

（1）报道性文摘。它是在对原始文献进行深入的语义和逻辑分析的基础上进行高度浓缩而形成的。它概括地叙述原文献所有或部分的重要信息，包括研究对象和目的、基本观点与方法、主要结论、全部数据、有关资料以及结论的价值和意义。报道性文摘所含信息量大，参考利用价值高，在一定程度上可代替原文，对帮助读者了解某些难得的文献内容和克服语言障碍有突出作用。它适用于那些学术价值高、内容丰富新颖、主题集中专一的文献。

（2）指示性文摘。它是指明原文主题和内容梗概的文摘，又称“简介”。它

一般只指明文献含有何种信息，并不摘录原文中的具体内容。它不能取代原文，只供读者对原文有初步的了解，以决定是否阅读原文。

2. 文摘举要

《新华文摘》，新华文摘社编，人民出版社出版。本文摘创刊于 1979 年，原名《新华月报（文摘版）》，1981 年 1 月改为现名，月刊。它是一种综合性的文摘杂志，选用的报刊近 240 种，绝大多数是中央和省一级的刊物。摘录政治、哲学、经济、历史、文学、文化教育、科技等方面的学术理论文章及科技研究成果，摘登最新文艺作品。并设“学术动态”、“论点摘编”和“读书与出版”等专栏，报道学术理论及出版动态。每期还附有“报刊文章要目辑览”，所收内容，多数是文选，部分为摘要。从中可查阅当前发表在国内主要报刊上较有学术水平或参考价值的论文资料。

《社会科学文摘》，河南省社会科学院情报研究所编辑出版。1983 年创刊，原名《学术资料（文摘版）》，内部刊物。1985 年改名为《学术文摘》，1985 年 7 月起改为现名，月刊，公开发行。它摘编全国报刊上社会科学方面的重要文章，迅速报道国内外社会科学研究的最新动态，全面反映各学科的不同特点，及时提供学术研究资料。

5.1.2 字典、词典

5.1.2.1 字典、词典概述

字典是以字条为单元，对字的形体、声音、意义以及用法或其他属性做出说明的工具书。词典是以词条为单元，对词目的概念、意义及用法做出说明或提供信息的工具书。事实上，中国的字典、词典，都源于古代的字书。在大量的中国古代字书中，有相当的部分，既解释单字，也解释复词，统称为“字书”。此外，对一般的语文性字典、词典来说，不仅解字，而且释词，并无严格的区别。字典是以释字为主，词典是以释词为主，只是侧重点不同。

我国古代最早的字书是《史籀传》（15 篇），见于《汉书·艺文志》。汉字主要有形、音、义三个方面，与此相对应，按注释的侧重点不同，字书形成了三大系统：以讲字形为主的字书（以《说文解字》为代表）、以讲字（词）义为主的字书（以《尔雅》为代表）、以讲字音为主的字书（以《广韵》为代表）。现在我们所使用的字典、词典，都是在这三大流派的基础上发展演变而来的。

1915 年出版的《中华大字典》，吸取了西方的编纂方法，成为现代字典、词典的开创之作。此后，各种字典、词典日渐增多。中华人民共和国成立以后，编制了各种字典、词典，具有代表性的有：《新华字典》、《现代汉语词典》、《汉语

大字典》、《汉语大词典》、《辞海》、《辞源》等。

现代字典、词典种类很多。一般按收录内容归类，可分为语文字（词）典和知识词典两大类（注：字典一般来说都是语文性的，词典可分为语文性的和知识性的两大类）。语文字（词）典是用于解释字（词）的形、音、义问题的，包括综合性字（词）典、专门性字（词）典、字（词）表。综合性字（词）典收字（词）范围较广，对字（词）的形、音、义及用法进行全面的解释，如《汉语大字典》和《中华大字典》、《汉语大词典》、《辞海》等。专门性字（词）典只收一定范围的字（词），或侧重解释字（词）的形、音、义某个侧面，如《古汉语常用字字典》、《中国成语大辞典》、《同义词词典》、《汉语外来词词典》等。字（词）表只汇集字、词，一般不加解释。如《2000常用字表》、《常用构词字典》等。知识词典是为学习学科基本知识和为研究某一专门学科、某一专门问题而编的，可分为百科词典、专科词典、专名词典。百科词典汇集各学科重要的术语和概念，加以概括解释，提供最基本的知识，一般不收普通词语，如《文史哲百科辞典》、《辞海》（兼具语文词典和百科词典的功能）。专科词典收集一个学科或专门领域的术语、概念并加以解释，系统地反映专业知识的概要，所提供的知识往往比百科词典更为详细，如《中国经济大辞典》、《哲学大辞典》等。专名词典以人名、地名、书名等为收录对象，只介绍有关专名的概况，提供事实和资料。人名词典如《世界人物大辞典》、《中国人名大辞典》等。地名词典如《世界地名词典》、《中国古今地名大辞典》等。书名辞典如《二十六史大辞典》、《中国当代文学作品辞典》等。

5.1.2.2 字典、词典举要

1. 查古代汉语字词

《康熙字典》，由张玉书、陈廷敬等30人奉康熙皇帝之命，在明代梅膺祚《字汇》和张自烈《正字通》两书的基础上增订而成，于1716年完稿，后多次重印。它是我国古代收字最多的一部字典，共收字47 035个。全书共分12卷，从子集到亥集，按地支顺序排列，每集又分上、中、下三卷。全书前有《总目》、《检字》、《辨似》、《等韵》，后有《补遗》一卷，收冷僻字，《备考》一卷，收不通用的字。按214个部首编排，每字先注音后释义。该字典收字丰富，释义广泛，引证详尽，注音全面，体例严谨，影响极大。但其错误和缺点也不少，后有多部著作对此考异订误。

《中华大字典》，徐元诰、欧阳溥存等编，中华书局1915年初版，后多次重印。该字典是以《康熙字典》为基础而编纂的，并在注音释义方面加以改进，校正错误4 000多条，增收了近代方言和科技方面的常用字，共收48 000余字，是

新中国成立前收字最多的一部字典。

《辞源（修订本）》，广东、广西、湖南、河南辞源修订组与商务印书馆编辑部合编，商务印书馆出版。它是我国第一部现代意义上的较大规模的语文词典，是查找古代汉语词汇的重要辞书。《辞源》1915 年初版，1931 年出版续编，1939 年出版正续编增修合订本，1979 年开始出版修订本，1983 年 12 月出齐，共四册。

2. 查现代汉语字词

《新华字典》，新华辞书社编，人民教育出版社 1953 年出版注音字母音序本，1954 年出版部首本，共有部首 189 个。商务印书馆 1957 年出版新 1 版，后多次修订、再版，主要供中、小学师生以及中等文化程度以上的读者使用。目前的最新版本是 2004 年的第 10 次修订版，分普通本、大字本、双色本三个不同版式。收单字（包括繁体字、异体字）1 万多个，带注释的词语 3 500 多个。正文按汉语拼音字母音序排列，附部首检字表。该字典在正字、注音、释义各方面均有独到之处，是目前国内流行最广、影响最大的一部小型字典，对促进语言规范化和标准化有一定贡献。

《现代汉语词典》，中国社会科学院语言研究所词典编辑室编。这是一部为推广普通话、促进汉语规范化服务的现代汉语中型词典。所收条目，包括字、词、词组、熟语、成语等。商务印书馆 1965 年印行“试用本”，1973 年重印（内部发行），修订后 1978 年正式出版，1983 年出版第 2 版，1989 年出版补编，1996 年出版修订本，2002 年出版增补本。目前的最新版本是 2005 年的第 5 版，收词约 65 000 条，基本上反映了现代汉语词汇的面貌，对所收的现代汉语的词和文言虚词都作了全面的词类标注。

《辞海（1999 年版普及本）》，辞海编辑委员会编辑，上海辞书出版社出版。1936 年初版，1965 年出版未定稿，后又出版 1979 年版、1989 年版、1999 年版。与 1989 年版相比，1999 年版在条目上有大量的修订，反映了国内外形势的变化和文化科学技术的发展，弥补缺陷，纠正差错，精简了少量词目和释文。收录单字 19 485 个，字头及其下所列词目 122 835 条。收录范围涉及古今中外自然科学、社会科学，所收词目包括普通词语和百科词语两部分。

3. 查古今汉语字词

《汉语大字典》，徐中舒主编，四川辞书出版社、湖北辞书出版社 1986—1990 年出版。全书共 8 卷。1993 年出版合订缩印本。收单字约 56 000 个，按 200 部首排列，是我国收字多、解释最全面的大型历史性详解汉语字典。

《汉语大词典》，罗竹风主编，上海辞书出版社 1986 年出版第 1 卷，从第 2

卷始，改由汉语大词典出版社出版。全书正文12卷，另有《附录·索引》1卷，至1994年出齐。1997年出版3卷缩印本。本书收词包括古今词语、熟语、成语、典故及古今著作中进入一般语词范围和比较常见的百科词语等，全书收词约37万条，是当今世界上收录汉语语词数量最多的汉语语文词典。汉语大词典出版社与香港商务印书馆联合推出了《汉语大词典（光盘版）》。

5.2 数据与事实型信息检索

5.2.1 年鉴

5.2.1.1 年鉴的定义及特点

年鉴是系统概述一年内各个方面或某一方面的进展情况，汇集有关重要文献及统计资料等，按年度编辑出版的工具书。从总体上说，年鉴有如下特点：

1. 时限性

年鉴，顾名思义是一年一鉴。它一般以年为限，逐年出版，以记事为主，汇集最近一年或截至出版年为止的各方面或某一方面的情况、统计资料等，故有人称之为“年度百科全书”或“微型百科全书”。

2. 新颖性

由于年鉴按年出版，能及时反映上一年的最新信息，其主要内容不断更新，其间虽有回溯性资料，但所占比重很小。因此，年鉴时效性强，信息价值高。

3. 资料性

年鉴“集万卷为一册，缩一年为一瞬”，收录资料广泛且集中。人们称年鉴为“知识密集、信息密集、时间密集、人才密集型的资料性工具书”。它一般包括大事记、专文或综述、事实概览、统计资料（图表）、附录及目录索引等基本内容，能给读者提供各方面的资料。

4. 准确性

年鉴选材严格，其学术性条目多由专家学者撰写或审定。其文献、资料、数据主要依据政府公报和文件、有关部门的统计、重要报刊的报道及专业工作者的撰述。

5.2.1.2 年鉴的产生与发展

年鉴的编纂始于欧洲，英文称之为Almanac、Yearbook、Annual。世界上

第一部以“年鉴”命名的书出版于1457年，最初不定期出版，16世纪后逐年出版，逐渐演变为提供一年内事件和统计资料的年度型工具书。据统计，美国在17、18世纪就有近2 000种年鉴出版。英国1758年创刊的《年鉴摘录》、法国1818年创刊的《世界历史年鉴》等，都是历史悠久、颇具影响的年鉴。19世纪后期，美国和英国分别出现了一些重要的，至今仍是最有影响的综合性年鉴。如美国1868年出版的《世界年鉴》和英国1869年出版的《惠特克年鉴》，现在仍是畅销不衰的著名年鉴。目前，世界各国年鉴的种类和数量仍在急剧增长。

我国“年鉴”一词最早出自《宋史·艺文志》，该书曾著录刘玄所撰《年鉴》一卷，原书已亡佚，其内容无从查证。我国现代形式的年鉴是从西方传入的。第一部年鉴是清同治三年（1864）创刊、1948年停刊的《海关中外贸易年刊》，由外国人创办。接着外国人用外文出版了关于中国的年鉴，如日文《第一回支那年鉴》。后来我国知识分子翻译和摘编了外国年鉴，如沈阳出版的《新译世界统计年鉴》。之后才出现了由我国出版机构和个人编纂的各种年鉴。

新中国成立前，除《申报年鉴》连续出版了近十年，其他均为短期，有的仅出了一年。尽管如此，这些年鉴保存了许多历史资料，对了解和研究民国时期的政治、经济、文化，有着重要的参考价值。

新中国成立后，我国连续出版过几部有影响的年鉴，如《人民手册》和《世界年鉴》等，但都于“文化大革命”中停刊。20世纪80年代起，我国年鉴事业得到了蓬勃的发展，不仅品种数量直线上升，内容质量上也有了长足的进步，出现了所谓的“年鉴热”。1980年我国只有6种年鉴，1991年初全国已有400多种年鉴，1996年达到1 300多种，到2000年我国出版的年鉴约2 000种，内容涉及各门学科。

5.2.1.3 年鉴的类型

按不同的分类标准，可将年鉴划分为不同的类型。一般说来，年鉴主要有以下几种类型：

（1）综合性年鉴。综合性年鉴系统反映社会各方面进展情况、各学科研究信息、基本知识和相关资料，涉及的内容广泛，信息丰富。如《中国百科年鉴》、《中国年鉴》、《人民手册》、《世界知识年鉴》等。

（2）专门性年鉴。专门性年鉴集中反映某一专门范围的年度进展情况及有关的资料，多半围绕一定的学科、专业、专题、部门、行业收集和提供有关的情况和资料。如《世界经济年鉴》、《中国出版年鉴》、《中国人物年鉴》、《中国对外经济贸易年鉴》等。

（3）地方性年鉴。地方性年鉴反映一省、一市、一地的基本情况，如《广州

年鉴》、《北京文艺年鉴》等。

(4) 统计性年鉴。统计性年鉴主要以表格和数字来说明有关领域或部门的进展情况，如《中国统计年鉴》、《中国人口统计年鉴》等。

5.2.1.4 年鉴的功能

年鉴主要通过栏目反映各种信息。各类年鉴均有其稳定的基本栏目，并根据需要设置具有本学科、本部门、本行业特色的栏目。每个栏目都具有不同的职能，提供不同的信息，可以从不同角度满足读者的各种需求。年鉴的主要功能有：

1. 提供时事动态信息

年鉴可以帮助读者系统、全面地了解国内外大事、时事动态及有关重要文件。例如，大型综合性年鉴均设有专栏全面介绍一年中的国内外重大事件，记载政府重要文献及党政领导人重要讲话。地方综合性年鉴反映本地区一年度的大事。专门性年鉴提供该领域一年度的要闻。

2. 提供各学科研究信息

专业性年鉴是系统掌握某一学科研究动态、研究成果和发展趋势的重要途径。各专业性年鉴大都设有“科研与进展”、“学科动态”、“研究观点介绍”、“论著选介”等栏目，读者可据此概要地了解某一学科、某一专题或某一学派的相关信息。

3. 提供统计数据资料

统计性年鉴专门汇集各类统计数字，其他年鉴也往往设有“统计数据”等栏目，因此，年鉴是很重要的数据来源。如《中国统计年鉴》和各省市统计年鉴全面详细地反映了全国各地经济和社会发展状况，数据完整、翔实、具体。

4. 提供实用的指南性资料

年鉴一般设有人物传记、机构名录、报刊简介、新学科介绍、新词语浅释等栏目，这些栏目所提供的信息，常是学习和研究中必需的资料。

5. 提供综述及回溯性资料

年鉴中有一些类别不同、长短不一的综述性文章或条目。它们由熟悉本专业、本地区、本领域情况的专家在占有大量的事实、文献数据的基础上，经过归纳、综合、研究之后写出的，可使读者对该学科、该地区、该领域的发展概况有较完整、系统的了解。另外，各类年鉴的创刊号一般都收集一些历史性的资料。因此，通过创刊号可查到有关回溯性大事和数据，很有参考价值。

6. 提供书刊论文线索

提供文献线索是专业性年鉴的一项特殊功用。好的专业性年鉴设有“书

目”、“索引”、“文摘”栏目。这些栏目所反映的著作和论文，都经过了有关专家的认真筛选，是本学科年度研究的重要成果，给学习和科研提供了丰富的资料线索。

5.2.1.5 年鉴举要

《中国百科年鉴》，中国大百科全书出版社中国百科年鉴编辑部编，中国大百科全书出版社 1980 年首次出版。这是新中国成立以来第一部大型综合性年鉴，为配合《中国大百科全书》而编辑出版。它从 1979 年起，逐年收集和记录国内外重大事件和各个领域、各个学科的新情况、新成果、新知识、新资料，为了解国内外政治、经济、文化、科学、艺术等各个领域、各部门的情况，以及查找有关的新材料，提供了很大的方便。

《人民手册》，名为手册，实质是年鉴。大公报社人民手册编辑委员会编辑。大公报社 1950—1966 年出版。此间，除 1954、1955 年合刊外，均逐年出版，1966 年停刊。它按年度汇编了我国政治、经济、文化等各方面的重要文件和资料，对查考 1949 至 1966 年间我国的基本情况，有一定参考价值。1980 年，人民日报社《人民手册》编辑组编辑出版了《人民手册》(1979)，性质与以前相同。

《世界知识年鉴》，世界知识年鉴编辑委员会编，世界知识出版社出版。1953—1957 年名为《手册》，1958 年后改为《年鉴》。1966 年停刊，1982 年复刊，此后逐年出版该书是一部介绍世界各国和国际形势的大型综合性年鉴。

《世界经济年鉴》，中国社会科学院世界经济与政治研究所世界经济年鉴编辑部编，社会科学出版社 1979 年起逐年出版，是一部介绍收录世界及各国地区基本经济动态的大型工具书。该书所收集资料多来源于联合国及各国官方公报，具有较高的准确性和可靠性。

《中国经济年鉴》，中国经济年鉴编辑委员会编，先后由经济管理出版社、经济年鉴出版社出版，1981 年创刊，以后逐年出版，它比较全面地反映了 1980 年以来有关我国经济的发展情况及统计资料。

《中国统计年鉴》，国家统计局编，中国统计出版社 1982 年起出版。这是一部反映我国国民经济和社会发展情况的统计资料性年刊，内容大致分为综合、行政区划和自然资源、人口、劳动力和职工工资、农业、工业、运输和邮电、固定资产投资、商业、对外经济贸易和旅游、财政、金融、保险、物价、人民生活、教育、科学、文化、体育、卫生、民政、司法等部分，是社会各界进行宏观经济和社会问题研究不可缺少的工具。

5.2.2 百科全书

5.2.2.1 百科全书的定义

百科全书是汇集各学科或某一学科的专门术语、重要名词，以词典的方式进行编排，对每一词目都加以全面系统而又客观简明的阐述，并对新的研究成果加以反映的大型工具书。百科全书涉及各个领域，其内容之丰富、规模之宏大、检索功能之完备是其他工具书所不及的。在各类工具书中，百科全书堪称“工具书之王”。

“百科全书”一词最早出自古希腊文“enkyklios”和“paideia”，前者意为“循环的，周期性的，平常的”，后者意为“教育”，合起来即为“普通教育”或“全面教育”，从字面上说就是一个想接受通才教育的人所应该学习的艺术和科学知识。在抄录过程中谬传为新拉丁语词 encyclopaedia，随后又进入英语，最先记载于 1531 年。在新拉丁语中，该词被选中作为一本覆盖各科知识的参考著作的书名。民国时的著名学者李煜瀛是我国最早进行百科全书研究的人，首次将“encyclopaedia”译为“百科类典”。后来，因受《四库全书》命名的影响，改译为“百科全书”。

5.2.2.2 百科全书的产生及发展

西方百科全书的编纂可以追溯到古希腊、古罗马时期。古希腊哲学家斯珀西波斯（Speusippus）和古希腊哲学家、科学家亚里士多德，编纂过概述各种学问的百科全书式的著作，被认为是百科全书的先驱者。亚里士多德还是最早对科学进行分类的人。他的思想和实践对后世百科全书的编纂有较大影响，人们称他为“百科全书之父”。

古罗马学者及作家瓦罗（M. T. Varro，公元前 116—前 27）编写过 9 卷本巨著《科学要义》（*Disciplinarum*），分述“自由七艺”及医学和建筑，成为百科全书的雏形。其后，古罗马学者老普林尼（Pliny，the Elder，23—79），编纂了现存第一部自然科学“百科全书”——《自然史》（*Historia Naturalis*，一译《博物志》）。该书的编纂方法及内容对后世影响很大。

到了中世纪，出现了一批具有代表性的百科著作。1559 年，法国学者斯卡里兹（P. Scalich）编纂了《百科全书，或神与世俗学科知识》，第一次正式使用“百科全书”（encyclopaedia）这个名称。1728 年，英国学者、百科全书编纂家钱伯斯（E. Chambers，约 1680—1740），编纂了《百科全书，或艺术与科学综合大辞典》。该书影响较大，钱伯斯因而被称为“现代百科全书之父”。我国于清朝末年开始编译外国百科全书，如 1903 年范迪吉等编译了日本《编译普通教育百科

全书》。后来，编纂过几部小型百科全书，如1919年王言纶等编纂了《日用百科全书》，1929年王昌汉等编译了《少年百科全书》，1936年杨家骆编著了《中国文学百科全书》等。1980年，我国编纂出版了第一部大型综合性百科全书——《中国大百科全书》。此后，多部百科全书相继编纂或编译出版。

5.2.2.3 百科全书的特点

1. 概括性

百科全书取材广泛，是百科知识的总汇，它用准确精练的语言，系统概述了人类各个知识领域或某个知识领域的基本事实、基本概念和基本理论，提供了各种事物的基本知识、历史和现状。

2. 权威性

著名的百科全书通常设立阵容强大的编辑机构，各条目的撰写，都是由各个领域、各门学科的著名专家学者完成的，以保证其质量。如《中国大百科全书》总编辑委员会及其下设的各学科卷的编辑委员会，都由我国各学术领域的权威人士组成。

3. 易用性

百科全书集中了日臻完善的编排方式、索引和参见系统，重要条目后都附有参考书目，或在文中注明征引资料的出处等，读者能从多种角度、用最短的时间检索到所需的知识。如《中国大百科全书》就设有七种检索渠道，以供检索与查考，是迄今为止我国出版的所有印刷型工具书中最完备的。

5.2.2.4 百科全书的作用

百科全书包罗万象，能为人们提供人类各个知识领域的基本知识，是学习和工作中最常用的、必备的工具书之一。人们往往称之为“没有围墙的大学”、“知识的小宇宙”、“精简的图书馆”。百科全书的主要作用可概括为两个方面：

1. 提供各种资料

百科全书收录资料广泛，是人们解疑释难的好工具。无论是解决学习工作中遇到的疑难问题，还是查找各种问题的基本资料等，都可以利用百科全书。

2. 帮助系统求知

百科全书对知识领域的覆盖面广，对各学科知识不畸轻畸重，客观、系统、完备、翔实地介绍各门学科的基本概况和基本理论。同时，它介绍的知识是不断更新的。目前，外国百科全书把5年以上的资料都看作过时的资料，因而很重视修订工作。利用百科全书，可以系统学到某一学科的基本知识，了解各学科的发展水平。

5.2.2.5 百科全书的类型

百科全书按内容范围分，有综合性百科全书和专科专题性百科全书。前者如《中国大百科全书》、《简明不列颠百科全书》等；后者如《社会科学百科全书》、《中国企业管理百科全书》、《中国农业百科全书》、《中国医学百科全书》、《中国旅游百科全书》、《集邮百科全书》等。

按地区范围分，有国际性百科全书和地域性百科全书。前者如英、美、法、德等国有名的大百科全书，力图反映世界文化遗产和现代成就，具有国际性；后者侧重反映某一地域、某一国家、某一省的各种情况，如《亚洲百科全书》、《加拿大百科全书》、《北京百科全书》等。

按读者对象，有成人学术性百科全书、成人普及性百科全书和青少年通俗性百科全书。成人学术性百科全书，如《中国大百科全书》、《社会科学百科全书》、《科学技术百科全书》、《世界经济百科全书》、《中国医学百科全书》等；成人普及性百科全书，如《环华百科全书》、《中华常识百科全书》等；青少年通俗性百科全书，如《少年百科全书》等。

5.2.2.6 百科全书举要

《中国大百科全书》，中国大百科全书编辑委员会、中国大百科全书出版社编辑部编，中国大百科全书出版社1980年起陆续出版，1993年出齐。这是我国第一部具有权威性的大型综合性百科全书。全书内容包括66个学科和知识门类，共74卷（包括总索引1卷），收录条目77 859个，总字数约1.26亿字，插图约5万幅。全书按学科分类分卷出版，各学科分卷一般由前言、凡例、学科（或知识门类）概观性文章、条目分类目录、正文、彩色或黑白插图插页、大事年表、条目汉字笔画索引、条目外文索引和内容分析索引等构成。正文条目按汉语拼音字顺编排。

该书内容上最大的特点是新、精和实用。它既关注基础，又偏重前沿；既兼顾过去，又重视现代；既侧重中国，又涵盖世界。它阐述的基本知识和提供的学术资料，其广度、深度和质量使之成为一个比较完整的知识体系。

该书的检索系统十分完备，设有多种检索途径，其中主要的检索途径有音序检索、笔画检索、分类检索、内容检索等。

中国大百科全书出版社每年还出版该书的补编《中国百科年鉴》，及时提供新的知识信息，应注意利用。

《中国大百科全书》已有光盘版1.1版和1.2版，每套四张盘，以《中国大百科全书》和中国百科数据库为基础。

《简明不列颠百科全书》，中国大百科全书出版社和美国不列颠百科全书公司合作编译，主要根据英文版《不列颠百科全书》第15版的《百科简编》编译而成，其中有关中国的条目由我国专家学者重新撰写，中国大百科全书出版社出版，1985年9月起出版，1986年9月出齐，共10卷。共收条目71 000余条，插图5 000余幅，内容涉及社会科学、文学艺术、自然科学、工程技术等各学科的概述和专名、术语、各国人物、史、地、团体、机构等，侧重西方文化、科学成就和当代知识。1991年，中国大百科全书出版社又出版了该书的增补卷（第11卷）。

《简明中华百科全书》，中国大百科全书出版社1994年出版。它是我国一部有代表性的小型百科全书。该书共3卷，收录8 000多个条目，概述文章约15万字，插图1 700幅，全书共约500万字，分正文、附录和索引三大部分。本书的内容以全面、系统、简明地介绍中国古今文化为主，内容构成有鲜明的特色。内容侧重社会科学，涵盖了历史、地理、哲学、宗教、人类、社会、政治、法律、军事、经济、文化、艺术、文学、教育等广泛的社会科学知识领域。科学技术方面的内容只着重介绍中国古代较大的技术发明及科学理论和现代在科技领域取得的突出成就。该书贯彻详今略古的原则。在全书的整个知识体系中，1948年以后的近现代中国是介绍的重点，特别是对1949年以后中国各方面的基本情况作了全面介绍，对1978年以后中国在改革开放过程中出现的新事物、新情况、新人物、新知识作了突出介绍。

5.2.3 手册

5.2.3.1 手册的定义和特点

手册是汇集某一方面经常需要查考的基本知识和数据资料，系统地加以编排，以供读者随时翻检的一种工具书。手册的名称很多，有指南、便览、要览、一览、必读、必备、大全、宝鉴等。手册主要有如下特点：

1. 实用性

手册可以说是一种面向实际的工具书，它一般是根据人们在学习、工作和生活中经常碰到、急需解决的知识性问题而编制，提供有关的基本知识和基本资料，如各种事实和数据，科技手册还注重技艺、操作方法、基本公式、图表、规格和条例的收录。

2. 灵活性

在工具书家族中，手册是一种最模糊最不确定的类型。手册不像其他工具书有确切的内容对象，它既可以收录字词文句、书籍篇目，也可以收录人名、地

名、机构名或其他资料；它好像是一种资料汇编，围绕一定的专题汇集有关的基本知识和资料；又类似于辞典，汇释有关学科、专业的知识性条目。其编纂体例，多数是以知识性条目出现，分类编排，也有按字顺编排，还有按篇章、类目汇集知识性材料，有的干脆以表格形式反映知识材料和数据等。总之，手册是一种比较灵活的工具书。

3. 资料性

手册通常是简明扼要地概述某一学科、专业、专题的基本知识和基本资料，注重图表和数据，具有主题明确、资料翔实具体等特点，即具有较强的资料性。所收资料侧重基础知识，通常偏重于已成为现实的、成功的具体专业知识，而不是定义、概念、历史的叙述和当前的发展状况。当代手册也开始注重反映科学文化发展水平和趋势。

5.2.3.2 手册的产生和发展

我国类似手册的书籍产生较早。在敦煌石窟发现的公元 9 至 10 世纪的《随身宝》，就是一部按类编排、供古人随时查阅的书，可作为我国较早具有综合性手册性质的工具书。流传在世近于手册的书有元代阴时夫的《居家必备》、清代石天基的《万宝全书》，都是古人手边常查的工具书。此外，建筑用的《营造法式》、航海用的《针经》、木工用的《鲁班经》、兽医用的《牛马经》、养鸟用的《禽经》等，都是世代师徒相传的具有专门手册性质的书。

近现代以来，由于科学技术的发展，各种知识在社会实践活动中的运用越来越广，作用日益重要，手册也因此得到迅速发展，出版数量日益增大。如艾芜的《文学手册》、舒新城的《中国教育指南》、陆费执的《农业宝鉴》等。新中国成立后，特别是 20 世纪 80 年代以来，手册的种类和数量是惊人的，尤其是科技手册出版数量相当可观，社科各学科各专业的手册也大量涌现。

5.2.3.3 手册的类型

手册灵活多样、资料稳定、实践性强，它能简明扼要地为人们提供各学科专业基础知识和各行业实用知识，是工作学习中不可缺少的方便实用的工具书。按编纂目的和内容范围，手册可分为综合性手册和专门性手册两类。

综合性手册即一般的常识性手册，面向的是广大读者，主要提供学习、生活的基本知识和资料。它又可以分为两种：一种是为各学科专业提供基本知识和资料，如《新社会科学知识手册》、《当代外国社会科学手册》等；另一种是为日常学习、工作提供常识性知识，如《新编读报手册》、《大学生常用手册》等。

专门性手册的服务对象是专业工作者或专门人员，主要提供专门知识或资

料。它又可分为三种：一是侧重为某一学科专业提供基础知识、基本事实，包括数据、图表、条例等，并反映该学科专业新的研究成果，如《哲学基本知识》、《当代世界经济实用大全》等；二是侧重为某项具体工作或某一具体活动提供特定的实用性知识，如《编辑工作手册》、《图书馆管理工作指南》等；三是介绍生活实用知识，如《家庭生活手册》、《妇女实用大全》等。

5.2.3.4 手册举要

《世界新学科总览》，金哲等主编，重庆出版社 1987 年出版。本书介绍了 470 多门哲学、社会科学及与新技术革命有关的自然科学和技术科学的新学科。全面介绍每门新学科的定义或界定、产生的时代背景与社会环境、奠基性著作与学科创始人、研究内容、学科发展与现状、研究机构和组织等项内容。

《当代中国社会科学手册》，汝信、易克信主编，社会科学文献出版社 1988 年出版。这是一部学术情报性的资料书，比较系统地介绍了新中国成立以来特别是党的十一届三中全会以来我国社会科学事业的发展状况、研究成果及有关的资料。

《当代国外社会科学手册》，中国社会科学院社会科学情报研究所等编，杨承芳主编，江苏人民出版社 1985 年出版。这是一部情报资料性的参考工具书，比较全面地介绍了当代国外社会科学发展的概况与动向，以及一些国家社会科学研究的组织与管理情况。

5.2.4 名录

5.2.4.1 名录及其发展

名录是将机构名、人名、地名等汇集在一起，按分类或字顺加以排列，并对相关事项予以简要揭示和介绍的工具书。

我国的名录起源很早，从有文字记载的时期起，人们便感到有必要将人员、机构和事物的名称有次序地记录下来。古代名录，见于著录的多为人名录，有专录历代名人简要事迹的《尚友录》（明代廖用贤），专录同姓名的《历代同姓名录》（清代刘长华），专录小名或别号的《小名录》（唐代陆龟蒙）、《别号录》（清代葛万里），专录少数民族人名的《西夏姓氏录》（清代张澍），等等。民国时期，也出版了几种专门名人录，如宋景祈的《中国图书馆名人录》和金警钟的《中国国术名人录》。

20 世纪初，我国的机构名录发展很快，较好的有《全国图书馆一览》（浙江图书馆）、《图书馆指南》（顾实）、《全国文化机关一览》（庄文亚）、《北平学术机

关指南》(李文绮)、《交易所便览》(伊兰)、《上海金融组织概要》(杨荫长)、《世界百大商部要览》(周世华)等。1949年至70年代末，由于多种原因，我国出版的名录寥寥无几。

80年代后，随着改革开放的深入，信息交流的扩大，经济实体和学术机构纷纷成立，从而促使人名录和机构名录大量出版。同时，随着国际交流的频繁，地名作为地理实体的标记和符号，与人们各种社会活动的关系日益密切，从而使地名信息显得十分重要。1967年，第一届联合国地名标准化会议要求各国都要建立地名管理机构，制定地名标准化原则，指导本国地名标准化工作。我国于1977年成立地名委员会，逐步实行地名管理科学化，由此推动了地名录的编制和出版。

5.2.4.2 名录的特点

1. 资料性

名录是一种比较典型的事实便览型的工具书。它为人们提供了有关机构、人名和地名的基本情况，包括机构的地址、电话、传真、电子邮箱、人员、组织、业务范围及产品，人物的生卒年、籍贯、学历、经历及著作，地点的正确名称及地理位置等。

2. 简明性

名录是提供专名简要资料的工具书，好比是专名基本信息一览表。每一专名的介绍只由最基本的具体资料组成，并形成格式化。它力求在有限的空间内，提供最基本的信息，没有过多的描述，更没有文字的铺陈。

3. 新颖性

名录注重提供有关专名的最新基本信息资料。人名录多半收录在世人物，地名录一般收录当代现行地名，机构名录尽量反映机构的最新情况。名录的及时性是除年鉴以外的其他工具书所不能比拟的。为了及时反映变化了的情况，名录特别是机构名录很注意修订再版。

5.2.4.3 名录的类型

按收录内容，名录大体可分为机构名录、人名录和地名录3类。

(1) 机构名录。它是汇集机构实体的名称并对该机构做概要介绍的工具书，大都由机构实体、出版单位及人事部门协作编辑。

(2) 人名录。它是汇集人的本名和别名并对人物予以简要介绍(诸如生卒年、籍贯、学历、经历、著作等)的工具书。人名录又可分为综合性名录和专门性名录。

(3) 地名录。它是著录地名及相关资料的工具书，它可提供地名的标准名称

（或加上译名）、所在地域（国别、省别）、地理位置（经纬度），有的地名录还简要说明地名变迁、人口状况、特殊事迹等。

5.2.4.4 名录举要

《中国政府机构名录》，新华通讯社中国政府机构名录编辑部编，中央文献出版社2002年出版。是继《中国政府机构名录》1989年版、1992/1993年版、1996年版问世以来，第四次出版的关于中国政府机构有关资料的权威性大型工具书，分为中央卷和地方卷。中央卷收集范围包括国务院、国务院各部委、国务院直属机构、国务院办事机构、国务院直属事业单位、国务院由部委归口管理的国家局、原国务院直属部分全国性公司、人民团体、事业单位和北京市人民政府、天津市人民政府、上海市人民政府、重庆市人民政府以及上述单位所属局、局（厅）机构和处（室）机构。在上述4个直辖市中还收录了乡镇一级政府机构。地方卷收集的范围包括我国各省、自治区、副省级市、省会城市及直属厅局级单位以及下属处室。资料截至2002年10月底。

《中国工商企业名录大全》（1—8卷），该书编写组编，中国国际广播出版社1992年版。这是目前收录商务业务信息最新最完整的大型名录，主要编入截至1991年7月底之前成立的具有法人资格的我国各行业工商企事业单位的名称、邮政编码、通信地址、电话、电报挂号及部分企业的经营范围、产品简介。

《世界工商企业大全》，该书编委会编，中国友谊出版公司1993年出版。反映世界182个国家和地区的1.4万家公司企业的基本情况。

《中国当代名人录》，中外名人研究中心编，上海人民出版社1991年版。本书收录当代中国各界有杰出贡献、地位重要以及知名度较高的人物7 564人，逐一介绍其生平、籍贯或出生地、主要经历、主要贡献。

《中国地名录——中华人民共和国地图集地名索引》和《世界地名录》，这是我国目前出版的两部较权威的国家性地名录。前者由国家测绘科学研究所地名研究室编辑，地图出版社1983年出版，既可与《中华人民共和国地图集》配合使用，也可单独使用。后者由萧德荣主编，中国大百科全书出版社1984年出版，收中外地名近30万个。

【案例】

《十三经索引》

1934年初版的《十三经索引》，是关于《十三经》的索引。它将《十三经》中的文字，按诵句为单位，依每句首字笔画编排，句子下面用简称注明所在的经

名和篇目，如果一篇再分若干节，则标明节次。书前有篇目简称、笔画检字。

1983年重订本索引，随同《十三经注疏》影印出版。重订本订正了初版遗漏、误入、断错句、文字讹误等错误近千条，并加注在《十三经注疏》中的页数和栏次，另增四角号码检字。

举例："君子耻其言而过其行　论宪27·二五一二下"，表明此句出自《论语·宪问》第27节，在《十三经注疏》第2512页下栏可以查到。

关键术语

书目	索引	文摘	字典	词典
年鉴	百科全书	手册	名录	

思考题

1. 书目、索引、文摘有何异同？
2. 怎样查找清代以前出版的图书？
3. 怎样查新旧期刊的基本情况？
4. 利用《民国时期总书目》、《全国总书目》（或《中国国家书目》、《全国新书目》）查找出若干部与某个研究课题有关的书籍。
5. 查新中国成立后报刊论文资料主要利用哪些工具书？
6. 利用《复印报刊资料索引》和《全国报刊索引》查出若干篇与某个研究课题有关的论文资料。
7. 名录有哪几种类型？其主要作用是什么？

CHAPTER SIX

第6章 计算机信息检索概述

【本章要点】

- ◇ 阐述计算机信息检索的含义和类型
- ◇ 介绍计算机信息检索的发展简史
- ◇ 总结计算机信息检索的特点
- ◇ 论述计算机信息检索策略
- ◇ 探讨提高计算机信息检索效率
- ◇ 讨论计算机检索技术

引子

计算机信息检索是随着计算机的出现而发展起来的。计算机检索经历了脱机批处理检索、联机检索、光盘检索及网络信息检索等阶段。计算机检索以其检索效率高、检索效果好而在信息检索中得到了广泛的使用。

6.1 计算机信息检索的含义和特点

6.1.1 计算机信息检索的含义

计算机信息检索指人们根据特定的信息需求，按照一定的方法，利用计算机从相关的信息检索系统中识别并获取所需的信息。计算机信息检索的过程包括信

息存储过程和信息检索过程，其本质是信息用户的提问标识和信息集合数据库特征标识匹配的过程（参见图 6—1）。

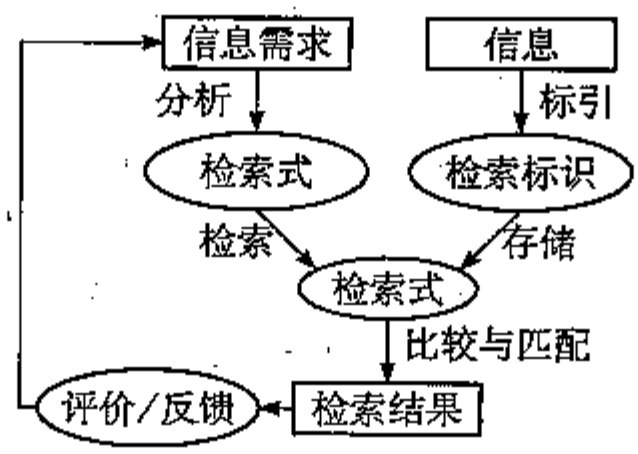

图 6—1 计算机信息检索

6.1.2 计算机信息检索发展简史

计算机信息检索是计算机技术、通信技术、数据传输技术不断发展的产物，同时也是为了满足文献快速增长、信息需求日益增长的需要。计算机检索经历了脱机批处理检索、联机检索、光盘检索及网络化联机检索等阶段。计算机检索以其检索效率高、检索效果好而在信息检索中得到了广泛的使用。

1. 脱机批处理检索

20 世纪五六十年代是脱机检索的试验和实用化阶段，其特征是检索时利用计算机作批处理。计算机信息检索工作开始于 20 世纪 50 年代初期。1954 年，美国海军军械试验站图书馆利用 IBM—701 电子计算机建立了世界上第一个计算机情报检索系统。1959 年美国人卢恩（H. P. Luhn）利用 IBM—650 电子计算机建成世界上第一个定题检索系统，为科研机构提供一定主题的新到文献提供服务。1961 年，美国化学文摘社用计算机编制《化学题录》（*Chemical Titles*），首次利用计算机来处理书目信息。此外，还有 1962 年美国国家航天局开设的 NASA 系统，1964 年美国国家医学图书馆的医学文献分析与检索系统 MEDLARS 等。在这一时期，计算机还没有连接通信网，也没有远程终端装置，主要是利用计算机进行现刊文献的定题检索和回溯性检索。当时的信息检索是脱机批处理检索，即由用户向计算机操作人员提问，操作人员对提问内容进行主题分析，编写提问式，输入计算机，建立用户提问档，按提问档定期对新到的文献进行批量检索，并将结果及时通知用户。这种检索方式，用户不与检索系统发生直接联系，只需把检索要求送往检索中心，由检索人员在计算机主机旁进行文献检索。同时，这一阶段开始利用计算机编辑出版检索性刊物。

2. 联机检索阶段

20 世纪 60 至 80 年代是联机检索试验和实用化阶段。1965 年以后，第三代集成电路计算机进入实用化阶段，存储介质发展为磁盘和磁盘机，存储容量大幅增加，数据库管理和通信技术都有深入发展，信息检索从脱机阶段进入联机信息检索时期。1965 年系统发展公司进行了首次全国性的联机检索表演。1967 年以后，许多联机检索系统相继出现。第一个大规模联机检索系统是 1969 年全面投入运行的 NASA 的 RECON 系统。1970 年美国洛克希德（Lockheed）公司的 DIALOG 系统和系统发展公司的 ORBIT 系统相继建成，美国 MEDLARS 也于 1970 年发展了联机检索系统 MEDLINE。此后不久，欧洲宇航局的 ESA-IRS 系统和美国纽约时报联机检索系统投入运行。随着国际联机检索系统的发展，信息检索在这一阶段实现了远程实时检索。

3. 光盘检索阶段

20 世纪 80 年代以来，一种新型的信息载体激光光盘在信息检索系统中得到越来越广泛的应用。特别是自 1985 年第一张商品化的 CD-ROM 数据库 Biliofile（即美国国会图书馆的 MARC）推出以来，大量以 CD-ROM 为主载体的数据库和电子出版物不断涌现，从而使得光盘检索以其操作方便，不受通信线路的影响等特点异军突起，大有与联机检索平分秋色之势。早期的光盘检索系统是单机驱动器和单用户，为解决多用户同时检索的要求，即同一数据库多张盘同时检索的要求，出现了复合式驱动器、自动换盘机及光盘网络技术。

4. 网络化联机检索阶段

进入 80 年代，随着 TCP/IP 通信协议的普遍采用，以及美国国家科学基金会（National Science Foundation，简称 NSF）的介入，计算机检索发展成了今天的互联网。在互联网网络检索的冲击下，传统的联机检索业纷纷采取措施，改进自己的信息系统与服务方式，在新的环境中寻求新的生长点，以获得新的发展。由于互联网的广泛性、方便性等特征，许多联机系统纷纷上网，把自己的系统安装在互联网的服务器上，成为互联网一个有机组成部分。如 DIALOG、STN、ORBIT 等世界著名的联机系统都建立了自己的万维网服务器，将其服务对象从原来的有限用户扩大到世界各地，大大增加了用户的人数。同时，除保证原来的信息服务项目和内容外，还增加了许多新的动态信息服务，如消息、网络新闻组等。而且，以搜索引擎为核心的网上搜索技术也日益发展，成为网络时代最具有普遍意义的信息检索形式，互联网集成了多种信息检索方式，已成为用户进行信息检索的一个广阔平台。

6.1.3 计算机信息检索的分类

计算机信息检索的对象是计算机检索系统，针对数据库进行，检索过程是在人与计算机的协同作用下完成的，匹配是由机器完成的。计算机信息检索包括许多类型，依据不同的划分标准，可以分为不同的类型：

1. 根据所检索数据库的形式，可以分为书目检索、数据检索、事实检索和全文检索

书目检索（Bibliographic Retrieval）指查出某一主题的文献条目的检索。按检索结果分，包括题录检索、文摘检索、图书与期刊等目录检索。数据检索（Data Retrieval）是利用相关的检索系统查询有关数据，以获得某一问题量化的准确数值，包括统计数据和科学数据等。事实检索（Fact Retrieval）是指在计算机检索系统中查询有关事件或实在情报，以求得对某一问题的解答。全文检索（Text Retrieval）是指直接利用原始文献建库进行的检索。

2. 根据计算机检索服务方式，可以分为定题检索、回溯检索和日常检索

定题检索（Selective Dissemination of Information，简称 SDI）是根据用户检索课题的内容，定期地从新到资料数据中为特定用户提问进行计算机情报检索的服务方法。一次性输入事先确定好的检索提问式保存在检索系统中，检索系统根据数据库更新周期，定期地对保存的检索提问式进行检索，将检索出的最新文献信息提供给用户。它具有定期性、新颖性和批处理式的特点。回溯检索（Retrospective Searching，简称 RS）指追溯查找过去的信息。进行回溯性检索，也可以查找最新的信息，能适应多数用户的查询需求。可用于申请专利时的新颖性检索、科研课题的立项或鉴定时的查新、撰写综述性论文以及编写教材时信息的收集等。日常检索指用户根据自己的信息需求，直接利用终端检索，检索系统即时提供用户所需的文献信息。

3. 根据检索方式，可以分为脱机检索、联机检索、光盘检索和网络检索

脱机检索（Off-Line Retrieval）是成批处理检索提问的计算机检索方式，是计算机信息检索的初期类型。联机检索（Online Retrieval）是指检索者通过检索终端和通信线路，直接查询检索系统数据库的机检方式。光盘检索（CD-ROM Retrieval）指以光盘数据库为基础的一种独立的计算机检索，包括单机光盘检索和光盘网络检索两种类型。网络检索（Network Retrieval）是利用 E-mail、FTP、Telnet、Archie、WAIS、Gopher、Veronica、WWW 等检索工具，在互联网等网络上进行信息存取的行为，目前主要利用的信息检索系统是搜索引擎。

6.1.4 计算机信息检索的特点

手工检索是人们长期以来采用的文献信息检索的传统方法，人们直接凭头脑进行判断，借助简单的机械工具，对记录在普通载体上的资料来进行相应的检索。具体来说，是由检索者通过书本式目录、卡片式目录，或者各类印刷型工具书，来查找自己所需要的信息。检索过程是人的手工操作完成的，其匹配主要依赖人脑的思考、比较与判断。手工检索的优点在于直观性强、灵活性高、费用比较低等。但随着信息数量的迅速增长，人们信息需求的快速拓展，手工检索的不足也日益明显，比如检索速度慢、时空的限制强、更新周期长、新颖性和时效性低、检索途径少等等。

计算机信息检索产生于20世纪50年代，发展于80年代中期，90年代后随着互联网技术的发展而进入了一个崭新的时期。计算机信息检索的应用和普及对于弥补手工检索的缺陷，提高信息检索效率，具有划时代的意义。其特点主要有：

1. 检索范围大

由于计算机的运算速度高和数据库存储量大，计算机信息检索系统收录了数量巨大、内容全面的信息。仅联机检索系统就能提供成百上千个数据库的检索，涵盖主题十分广泛，几乎覆盖了人类社会生活的各个领域。国际联机检索集成计算机技术、通信技术和高密度存储技术，使得计算机检索具备了实效性、完整性、广泛性和准确性的特点，能在短时间内检索世界范围内的有关文献信息资料，真正达到了人类知识的共享。搜索引擎收录了庞大的网络信息资源，成为人们获取互联网信息资源的最重要的计算机检索系统。

2. 检索速度快

计算机的快速运算能力保证了计算机检索系统的检索速度，手工检索需要数日甚至数周的课题，计算机检索只需要数分钟甚至几秒就可以完成，大大地提高了检索文献信息的检索速度，节约了读者的检索时间，提高了检索效率。

3. 检索功能强，组配灵活

计算机信息检索系统一般都提供布尔逻辑检索、截词检索、词组检索、字段限定检索等，各类检索词之间可以灵活组配，还可对检索之间的位置关系和短语进行全文查找。无论光盘检索系统、联机检索系统还是搜索引擎都可以满足多途径的检索要求。这是传统的手工检索无法做到的。

4. 检索途径多

一般来说，计算机检索系统除具有手工检索中采用的途径外，还能满足多途

径交叉检索的需要，尤其适用于综合性课题的检索。大部分计算机检索系统能够提供题名、分类、主题、作者、关键词、机构、中英文摘要、全文等检索途径。

5. 数据更新及时，时效性强

利用计算机检索的文献信息更新周期短，计算机检索系统根据自身的特点更新周期不同，如光盘多为每月更新一次，网络则每天更新一次。而手工检索工具的更新周期则比较长。

6. 检索结果输出形式多样

检索结果可以选择直接浏览、打印、存盘或 E-mail 传送检索结果，部分计算机检索系统还提供不同字段的输出形式，或者选择简单格式和详细格式两种检索结果显示形式。

不过计算机信息检索也有一些不足，计算机检索系统所收录的数据的回溯时间有限，也就是说计算机检索不能够满足所有的信息查询需求。同时，计算机检索需要检索者具有一定的计算机知识，需要有计算机的环境，因而这种检索方式并不是适合每一个信息需求者。就目前而言，计算机检索日益成为人们获取信息的重要方式，但在很长一段时间内，手工检索和计算机检索仍将共存，互为补充，共同满足人们多元化的信息需求。

6.2 计算机信息检索策略

6.2.1 检索策略的含义和作用

信息需求产生之后，如何在茫茫的信息海洋中查找需要的信息？利用哪些信息检索系统？检索提问怎么设计？怎样获得较高的查全率和查准率？信息检索策略对于解决这些问题具有重要的意义。

所谓检索策略，即在分析检索课题内容实质基础上，选择检索系统、检索途径，确定检索词及其相互间的逻辑关系等的信息检索方案。信息检索策略的实质是对检索过程的科学规划。其中关键在于构造能够确切表达信息需求的检索式。

影响检索效果的因素有很多，但对于已经建成的信息检索系统而言，检索策略的优劣是非常重要的因素。正确的检索策略会优化检索过程，有助于提高查全率和查准率，节约检索时间与费用，取得最佳的检索效果。反之，则会降低检索效率。

6.2.2 检索表达式

检索表达式是检索策略的具体体现，简称检索式。检索式一般由检索词和各种逻辑运算符组成，具体来说，它将检索词之间的逻辑关系、位置关系等用检索系统规定的各种算符连接起来，成为计算机可以识别和执行的命令形式。检索式构造的优劣关系到检索策略的成败。

检索表达式主要有逻辑表达式、加权表达式和其他表达式，其中，最为常用的是逻辑表达式。

6.2.2.1 逻辑表达式

逻辑表达式是指利用布尔算符，对检索词的关系进行表达，又称布尔逻辑表达式。布尔算符是19世纪中叶英国数学家乔治·布尔发明的，以集合论与布尔逻辑为理论基础，是目前计算机检索最简单、最基本的匹配模式，也是计算机检索领域广泛采用的表达方式。布尔运算符有逻辑与“AND”、逻辑或“OR”、逻辑非“NOT”等（参见图6—2）。

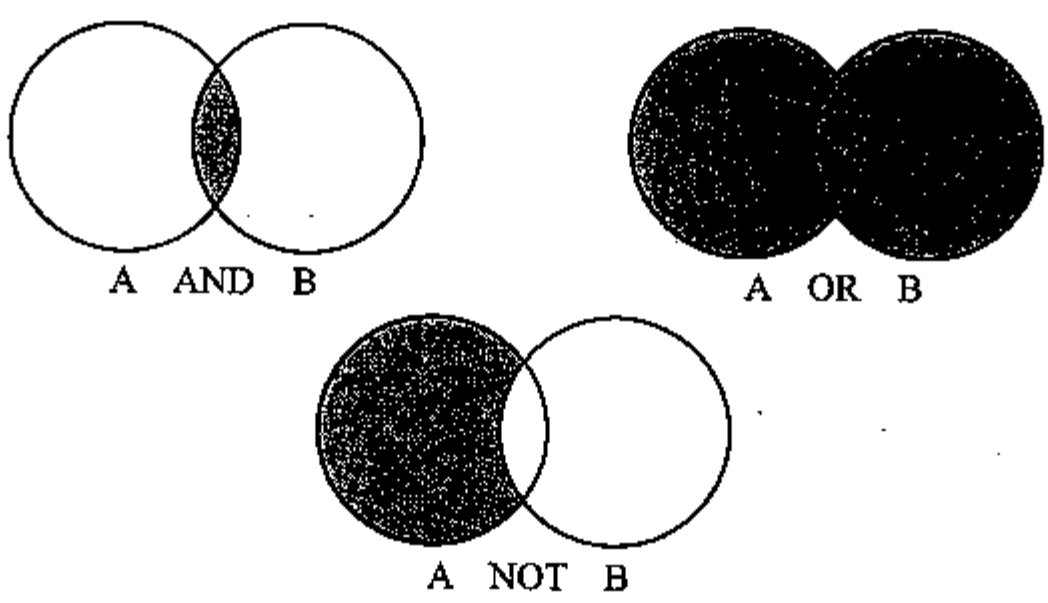

图6—2 布尔算符示意图

逻辑与“AND”，表示它所连接的两个检索词必须同时出现在结果中，检索式可写为：A AND B。含义为A与B重合部分。也有些数据库中用“*”或其他符号表示逻辑与。例如，查找关于“计算机检索”方面的信息，可以表述为：计算机 AND 检索。目前，在一些数据库中提供的二次检索，如中国期刊网，实质上也是逻辑与的运算。

逻辑或“OR”，表示它所连接的两个检索词中任意一个出现在结果中就满足检索条件，检索式可写为：A OR B。它是表示概念并列关系的一种组配，用来扩大检索范围和保证查全率。在一些中文数据库中，用“+”表示逻辑或。例

如，想检索关于“计算机”的信息，可以表达为：计算机＋电脑。逻辑或主要用于表达概念的近义词、同义词、全称和缩写等，以便全面、完整地表达相关的概念，提高信息的查全率。

逻辑非“NOT”，表示它所连接的两个检索词中应从第一个概念中排除第二个概念，检索式可写为：A NOT B。在一些中文数据库中用“－”表示逻辑非。例如，想查找关于“研究生教育”的资料，但要求不包括在职研究生，可以将这一提问的表达式写为：研究生＊教育－在职研究生，也可以写为：（硕士生＋博士生）＊教育－在职研究生。逻辑非表示具有不包含某种概念关系的一组组配，用来缩小检索范围。但在实际检索中要慎重使用。

逻辑表达式在实际检索过程中，易于理解，便于使用。例如，想检索“中国高等教育的发展趋势”，用逻辑表达式可写成：中国＊高等教育＊发展趋势。表示要求查找的文献中同时包含“中国”、“高等教育”、“趋势”这三个词。如果检索有关“文献保护”的资料，逻辑表达式可以写成：（文献＋图书＋档案＋资料）＊（保护＋防潮＋防虫＋防有害气体）。

在逻辑表达式的构造中，根据不同的信息需求、不同的检索策略，其检索式构造也不一致。一般来说，对于以查全为目标的检索课题，在检索式的构造过程中，用“与”连接的概念组面不能太多，应增加用“或”连接的相关检索词。对于以查准为目标的检索课题，其检索式的构造一般可增加逻辑与的使用。

布尔逻辑表达式具有诸多的优点，可以表达与用户思维习惯相一致的查询要求，与计算机逻辑运算功能一致，表达意义比较明显直观。但它同时也存在着一定的缺陷，比如：不能实现检索结果的相关性排序；不能反映表达式中检索词的重要性；如果用户的检索课题中涉及的检索词较多时，可能要写出一个相当复杂的逻辑表达式。

6.2.2.2 加权表达式

为了弥补布尔逻辑表达式的不足，人们提出了加权检索。所谓加权检索，是指在检索提问中，根据每个检索词在检索要求中的重要程度，分别给予一定的数值加以区别，即赋权，这个数值称权值，然后对含有这些检索词的文献进行加权计算，其和在规定的阈值以上的，即确认为命中文献。采用这种方法表达信息需求的称为加权表达式。

例如，用加权表达式来表示查找“中国高等教育的发展趋势”的信息需求，可以写为：

中国（5）高等教育（5）发展趋势（5）阈值 W＝15

括号内的数字 5 即是权值。具体检索时，对同一条记录内包含并且匹配这三个检索词的权值相加，超过阈值 15 时，就作为命中文献输出。逻辑上还是“与”的关系。如《论中国高等教育的发展趋势》这篇文献，各检索词权值相加是 15（中国 5，高等教育 5，发展趋势 5），就是命中文献之一。而《中国高等教育的现状》，检索词权值相加为 10（中国 5，高等教育 5），小于阈值 15，即为非命中文献。

例如，想查找“图书馆学或情报学”方面的文献，用加权的方法可表示成：

图书馆学（15）情报学（15）阈值 W=15

这是一个逻辑“或”的关系，在检索时只要有一个相应的检索词，它的权值之和就可以大于或等于阈值 15，即为命中文献。

从以上可以看出，在采用加权检索时，要对比检索词和标引词，还要统计检索词的权重。还有一种加权检索的形式，就是不直接对检索词进行赋值，而是对文献重点内容的检索词做加权标志，在检索的过程中，如果侧重某一个检索词，可以对该词做加权标志，这样就可以将重点反映该主题的文献查出来。目前许多网络检索工具采用“+”、“-”来表示检索词在检索提问中的分量。在某个检索词前面带上“+”，表示该检索词必须在检索结果中出现，反之，若某个检索词前面带上“-”，则表示该检索词一定不能出现在检索结果中。实质上，网络检索工具的加权检索也仅能控制某个语词是否一定要在检索结果中被包含或被排除，尚不能根据用户的需求来确定某一个具体语词的权值大小，从而确定它对检索结果的影响程度。这种加权的方式还有待于进一步完善。

加权检索可明确各检索词在检索中的重要程度，检索结果按照切题顺序排列，在提高查全率和查准率方面均有一定的作用。但就具体应用来说，加权检索的使用远不及布尔逻辑表达式广泛。

6.2.2.3 位置检索表达式

两个检索词在文献中相隔的距离不同，可能会在一定程度上带来检索结果的差异，单纯依靠布尔逻辑表达式，不能满足多种检索需求。因而，人们又引进了位置检索表达式，也称邻近检索。通过位置算符来表示两个检索词（或短语）之间的距离和位置关系。不同的检索系统可能会采用不同的位置算符，目前应用广泛的主要是“（W）”、“（nW）”和“（N）”、“（nN）”。

（W）表示连接的两个检索词相邻，并且先后顺序不能颠倒，这里的 W 是 with 的缩写，检索式可表达为：A（W）B。（nW）表示连接的两个检索词之间最多可以插入 n 个词（在中文方式下表示 n 个字），而且前后顺序不能颠倒，检

索式可表达为：A（nW）B。例如，如果检索式“文献（2W）检索”，则《文献信息检索》、《文献资源检索》均为命中文献；如果输入检索式“文献（W）检索”，则《文献信息检索》、《文献资源检索》都属于非命中文献。

（N）表示连接的两个检索词相邻，先后顺序可以颠倒，这里的 N 是 near 的缩写，检索式可表达为：A（N）B。（nN）表示连接的两个检索词之间最多可以插入 n 个词（在中文方式下表示 n 个字），前后顺序可以颠倒。例如，检索式 environment（2N）protection 可检索出包含“environment protection”、“protection of the environment”、“protection of water environment”、“protection of forest environment”等内容的结果。

6.2.2.4 截词检索表达式

截词检索表达式指在检索式中用专门符号（截词符号）表示检索词的某一部分允许有一定的词汇变化，也就是说，检索词的不变部分加上由截词符号所代表的任何变化形式所构成的词汇都是合法检索词。截词检索表达式在西方语言检索中应用比较广泛，在中文信息检索中也有一定的应用。采用截词检索表达式，既能防止漏检，又能节省机时，是提高检索效率的有力措施。不同检索系统采用的截词符不完全相同，一般常采用“?”、“*”等。

截词方式有多种，按截断的位置来分，截词有前截断、中间截断、后截断等；按截断的字符数量来分，可分为有限截断和无限截断两种。

后截词，又称右截词、前方一致，允许检索词尾部有若干变化形式。例如，检索式“comput?”将检出包含 computer、computing、computerized、computerization 等词汇的结果。检索式“交际?”，表示检索以“交际”打头的信息，可以检索出“交际艺术”、“交际语言”、“交际行为”等；检索式“中国*银行”可以检索到“中国人民银行”、“中国农业银行”等结果。

中间截词，允许检索词中间有若干变化形式，例如 wom*n 就可同时检索到含有 woman 和 women 的结果。

前截词，又称左截词、后方一致，允许检索词的前端有若干变化形式，例如检索 *physics 就可检得包含 physics、astrophysics、biophysics、chemophysics、geophysics 等词的结果。

截词检索表达式在使用时，一定要合理使用，截断部分要适当，不要截得太短，以免增加检索噪音，查出很多无关的文献。

6.2.2.5 限制检索表达式

在信息检索的实际过程中，有时还需要将检索词限制在标题、文摘等字段

内。限制检索也称字段检索。限制检索表达式指用限制符限定检索词出现范围的检索式。计算机检索系统大部分都支持限制检索，但在不同的检索系统，限制符的表达形式和使用规则有所不同，一般来说，常用的字段限定代码有：标题（TI，Title）、作者（AU，Author）、主题词（SU，Subject）、年代（PY，Publication Year）等。

一些网络检索工具也允许用户采用限制检索表达式，可把检索范围限制在标题、统一资源定位地址（URL）或超链等部分。例如，“TITLE：北京大学”这一检索提问可以查得网页题名中含有“北京大学”的网页。

6.2.3 检索策略的构造步骤

信息检索策略的构造步骤一般包括分析用户信息需求、选择检索系统、确定检索词或检索式、处理检索结果、获取原始文献等。具体步骤如图 6—3 所示：

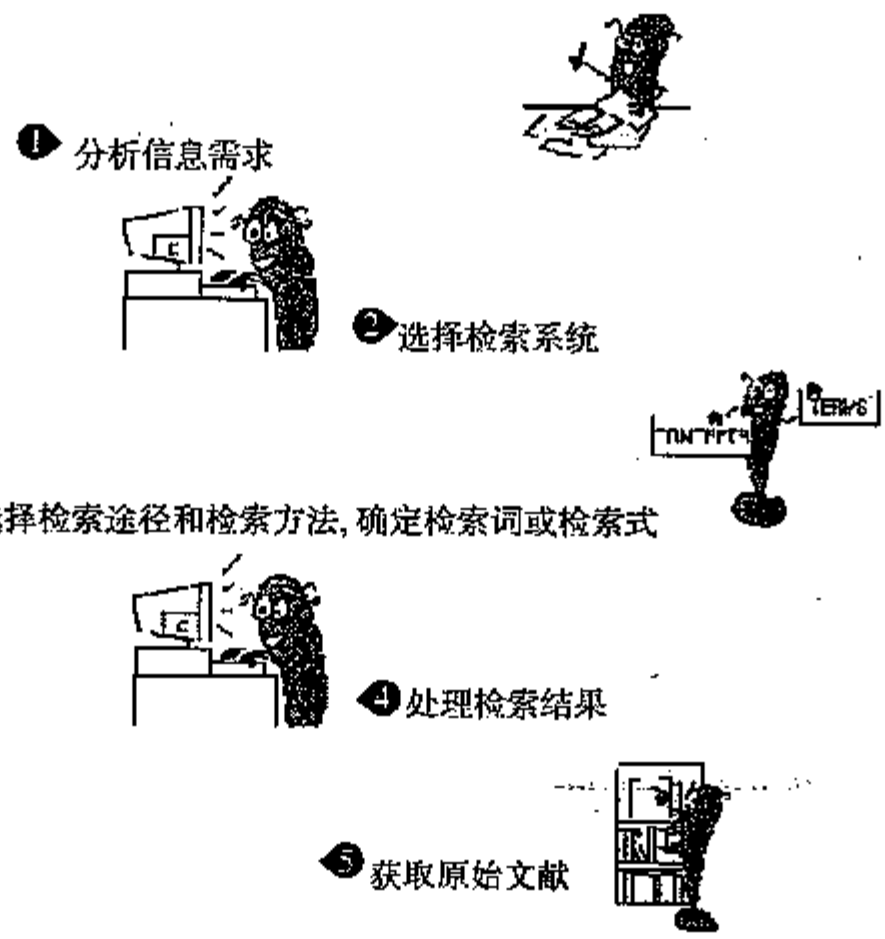

图 6—3 信息检索策略构造步骤

资料来源：http：//www. lib. nus. edu. sg/chz/chilion/chirtopics/strategy. html，2008-05-01。

6.2.3.1 分析信息需求（检索课题），明确检索要求

这是人们进行信息检索的出发点，不同类型的检索课题，信息需求的范围和程度也不尽相同。科技查新需要全面地收集某一主题的文献信息，在查全率上有很高的要求。而对于学习中为解决某一特定问题的检索课题，大部分时候只要求

检出适度的信息即可。在这一环节中，要明确检索目的，明确检索课题内容涉及的主要学科范围和相关概念。在分析课题的基础上，要清楚检索信息的类型，是查文献，查事实，还是查数据，以及要求查找文献信息的时间范围、学科范围等，通过以上分析，对检索需求作出全面的认识。

操作中应尽可能掌握检索课题研究背景，了解检索课题所属的学科领域、学术发展史和现状，借助有关工具书来进一步开拓背景材料，以便于选择正确的检索标识和检索范围。利用掌握的资料背景的相关线索，经过分析、推敲、拓展，发现更多有参考价值的文献线索，再通过这些已知的线索，了解与检索课题有关的学者、科研机构、学术刊物，以增加检索途径，提高检索效率。

6.2.3.2 选择检索系统

在计算机检索中主要是利用数据库，包括搜索引擎。依据对信息需求的分析，选择与检索课题相符、收录信息质量较高、检索功能比较完善的信息检索系统。检索系统的选择要求我们对目前可利用的检索系统有一个大概的了解，如检索系统收录的信息所涉及的学科领域、信息类型、时间范围、检索途径和检索方法、检索费用等等。

6.2.3.3 选择检索途径和检索方法，确定检索词或检索式

检索系统选定后，要对检索途径和方法作出判别和选择。大部分数据库可以提供篇名、作者、主题词、关键词以及全文检索等途径，而且还能利用各种途径的组配进行交叉检索。检索词的确定是建立在检索课题概念分析的基础上，有时，检索课题会包含较复杂的主题内容，应明确组成课题内容的直接概念和相关概念，通过一定的逻辑组配或其他方式形成一定的复合概念或概念关系来表达用户的信息需求。在确定检索词时，应考虑它表达概念的确切性及其与系统存储标识的一致性。

6.2.3.4 处理检索结果

确定了检索词或检索式之后，即可开始实质性检索。在实施检索的过程中，根据检索结果的实际情况，可以调整检索词、检索式、检索途径和检索方法等，也可以充分利用信息检索系统提供的缩检和扩检功能，完善检索结果，直至达到满意的效果。

实施检索之后，将所获得的检索结果加以系统整理，筛选出符合课题要求的相关文献信息，选择检索结果的著录格式，辨认文献类型、文种、著者、篇名、内容、出处等项记录内容，输出检索结果。

6.2.3.5 获取原始文献

使用的信息检索系统类型不同，原始文献的获取方式也不尽相同。比如，利用图书馆公共联机目录查询系统，可以了解图书的基本信息，以借阅或复制的方式获取原始信息；利用联机信息系统，可以用联机传递或脱机邮寄方式获取原始信息；利用有关全文数据库，可以直接打印或下载原始信息；利用网络搜索引擎，除一些收费的信息不可直接得到外，搜索引擎的检索结果大部分都可以在网上直接获取。

6.2.4 检索策略的反馈与调节

信息检索过程是一个比较复杂的过程，受到诸多因素的影响，一次检索的结果往往不能完全满足检索要求，有时会造成一些检索偏差。这就需要我们及时调整检索策略，纠正检索误差，以便获取满意的检索结果。

6.2.4.1 影响查全率和查准率的主要因素

提高信息检索的查全率和查准率，是调节检索策略的主要目标。在检索策略中决定查全率和查准率变化的主要因素有：

（1）主题分析是否准确、全面。对检索课题进行主题分析，是正确选择主题词和构建检索表达式的先决条件，也是提高检索的查全率和查准率的前提。尤其是对于一些由复杂主题构成的检索课题，主题分析具有更为重要的意义。

（2）检索词选择是否准确。选用的检索词的专指度如何，如果检索词过于专指或者过于泛指，都会不恰当地缩小或扩大检索范围。

（3）检索词之间逻辑关系的配置是否合适。一般说来，逻辑与的使用有助于提高查准率，逻辑或的使用有助于提高查全率，截词检索的使用可以提升查全率，限制检索可以将检索词限定在某一范围之内，有利于查准率。但是，如果不合适地使用逻辑算符或其他算符，就会带来一些负面的影响，降低检索的查全率和查准率。

6.2.4.2 提高查全率和查准率的方法

我们知道，在检索策略的构造过程中，检索表达式的形成是最重要的环节。检索策略的调节在一定程度上直接表现为检索表达式的调整。针对提高查全率和查准率的检索需求，可以采用不同的方法。

1. 提高查全率的方法

提高查全率，意味着要扩大检索范围，即扩检。可以通过降低检索词的专指度、增加同义词或近义词或相关词的逻辑或组配、选用截词检索、增加和调整检

索途径等方法。

(1) 降低检索词的专指度。

即选用的检索词范围面要广一些，泛指性要强一些。为了提高信息的查全率，除选择恰当的主题词外，还应该选择比恰当的主题词内容范围更广的上位词。例如，在“中国期刊网”中，检索关于“网络信息资源组织”方面的文章，选择高级检索，从篇名途径，输入“网络信息资源”和“组织”，检索到269篇文献。如果想提高查全率，可以选择降低“网络信息资源”的专指度，输入“信息资源”和“组织”，仍然选择篇名途径，检索到了427篇文献，其中包括《金融信息资源的网络化组织与服务分析》、《浅论因特网信息资源的组织与管理》等一批扩检出来的相关文献。

(2) 增加同义词、近义词或相关词的逻辑或运算。

进行课题检索时，不仅要选择较为规范的主题词，而且要考虑与该主题词相关的同义词或近义词。反映同一概念的检索词越多，则越能保证查全率。比如，一个词语在英文中往往有多个单词与之对应。如“保护”一词在英文中即有conservation、preservation、protection等词与之对应，在构建检索式时，应尽可能考虑到相关的同义词和近义词。如果想检索有关“互联网”方面的资料，而且想获得较高的查全率，检索式可以表达为：互联网 OR Internet OR 因特网。

对于一些表示整体的概念，如果想提高查全率，可以将整体概念进行拆分，并用逻辑或连接。比如，要检索关于“欧洲能源”方面的文献，通过背景知识和课题分析，可以知道欧洲能源也包括英国的天然气、法国的石油等，因而，检索式可以表达为：(欧洲 OR 英国 OR 法国 OR 德国 OR 意大利 OR ……) AND (能源 OR 天然气 OR 石油 OR 煤 OR ……)。

(3) 选用截词检索。

为防止漏检，得到比较全面的结果，可以利用截断的词的一个局部进行检索，利用一组相关词词首一致的特性，进行相关扩检。在一些中文数据库中采用的是前方一致检索。这种方法比较简单易行，通过一个检索词查出许多相关或相近的文献，可避免输入多个词干相同而词缀不同的检索词，从而简化检索过程，节约用户的时间，提高检索速度。

(4) 增加和调整检索途径。

根据检索的需要和检索系统的具体情况，可以增加检索途径，如可以将主题检索途径和分类途径结合起来。也可以调整检索途径，比如，在中国期刊网中，检索关于“计算机检索”方面的文献，选择篇名检索途径，检出215篇文献；选

择关键词检索途径，命中 3 708 篇文献；选择全文途径，得到 206 354 篇文献。①

2. 提高查准率的方法

提高查准率，一般是在有一定查全率的基础上再进行缩检，主要采用的方法包括提高检索词的专指度、增加逻辑与的使用、利用逻辑非、选用限制检索等。

（1）提高检索词的专指度，增加或换用下位词和专指性较强的关键词进行检索。比如，想查找“网络检索工具”的有关资料，为了增加查准率，可以采用提高专指度的方法，增加或换用“搜索引擎”、“网络资源目录”等专指词，提高检索结果的相关性。

（2）用 AND 连接一些进一步限定主题概念的相关检索项，增加相互的制约。在一些搜索引擎和数据库中可以采用“二次检索”（或“在结果中查询”）实现增加逻辑与运算的功能，提高检索的查准率。它要求检索者开始时不要把条件限制得过于严格，如检索结果数量过于庞大，再逐步排除检索结果中不需要的内容。这种逐步缩小检索范围的方法，可实现由查全向查准的逼近。

（3）用 NOT 来排除一些无关的检索项。在第一次检索结果出来之后，根据需要可以采用逻辑非将一些与提问不相关的文献排除，减少检索噪音。但应比较慎重地使用逻辑非，切勿将不该排除的文献去掉。

（4）采用限定检索，缩小检索范围，提高查准率。比如，可以将检索词限定在题名字段、主题字段等，也可以利用文献的外部特征加以限制，如文献类型、出版年代、语种、作者等等。

6.3 信息检索技术

信息检索技术指应用于计算机信息检索过程中的相关技术的总和。本节重点介绍全文检索技术、基于内容的多媒体检索技术等。

6.3.1 全文检索技术

全文检索（Full Text Retrieval），就是以各类数据诸如文字、声音、图像等为主要处理对象，根据数据资料的内容，而不是外在特征来实现的信息检索技术。全文检索技术最早出现在美国匹兹堡大学 1959 年建立的法律情报检索中，进入 20 世纪 80 年代以后，许多商品化联机检索系统都开始大力推行并发展全文

① 检索时间为 2008 年 5 月 26 日。

检索数据库。网络环境下，搜索引擎的发展更大程度地推进了全文技术的发展。与其他检索技术相比，全文检索技术的新颖之处在于，它可以使用原文中任何一个有实际意义的词作为检索入口，而且得到的检索结果是源文献而不是信息线索。

全文检索技术不同于传统数据库的字段检索，它采用特别的索引技术，将相关的文献信息，经过索引产生器的浏览而建立起所谓的索引数据库。当用户进行检索时，系统通过使用者输入的关键词，迅速地从索引数据库中找到用户需要的信息，并且将相关索引显示出来，供用户选择和浏览全文。全文检索技术的显著特点就是提供对海量数据的管理和快速查询。网络搜索引擎就是以全文检索技术作为核心支撑技术，它首先要对 Web 信息进行预处理，包括格式过滤、语词切分、词法分析、语词识别、自动标引等环节，为实现全文检索做好准备。

中文全文检索技术的研发始于 1987 年左右，现已出现了一些商品化的软件，包括 TRS、Quick IMS、南辰、天宇、I-Search 等。其中，最有影响的当属 TRS 全文信息检索系统，它可以广泛地应用于各种信息数据库、信息门户的建设，以及从 Web 站点检索、互联网搜索引擎到电子商务等各种应用中文信息的发布检索。TRS 系统内嵌汉语自动分词系统，支持按词索引、按字索引、按关键词索引、字词混合索引，大大提高了检索的准确性和响应时间。允许使用文中的任意字、词、句和片段进行检索，提供了基于文献内容而不仅仅是文献外部特征的全文检索手段。TRS 所提供的按词和按用户自定义关键词进行索引和检索，以及基于知识词典的扩展检索功能，满足了特殊应用领域的高查准率和高查全率的要求。而且，它检索功能强大，提供了多达 48 种检索运算符。包括外部特征与正文内容的各种逻辑组合检索、位置检索、二次检索、渐进检索、历史检索、词根检索、大小写敏感检索、概念检索、对检索结果按与检索表达式的相关性和重要性程度排序等。

目前的全文检索技术还存在着一些未尽如人意的地方，尤其是在查准方面难以保证，原因是用孤立词和词汇术语作为检索入口，缺乏语义的内在关联，检索的效果不是十分理想。为了解决这一问题，全文检索技术开始和人工智能紧密结合，增加对内容的分析理解、内容表达、知识学习、推理机制。随着智能化技术的发展，全文信息检索技术必将更广泛和高效地应用于网络信息检索领域。

6.3.2 基于内容的多媒体检索技术

数字信息不仅包括文本型文献信息，还包括大量的图形、图像以及声音、视

频、动画等信息。多媒体检索技术指对多媒体信息专有的检索技术，重点是基于内容的多媒体信息检索技术。随着多媒体技术的迅猛发展，网络传输速度的提高，以及新的有效的图像/视频压缩技术的不断出现，对海量多媒体信息的需求日渐增强，在这一背景下，基于内容的多媒体信息检索技术应运而生。它与传统数据库技术相结合，可以方便地实现海量多媒体数据的存储和管理；与网络搜索引擎技术相结合，可以用来检索互联网中丰富的多媒体信息。

影像是一种独具特色的媒体，与文本完全不同。对它的组织、存储、检索、传递与利用，需要一系列新的技术，其中的核心问题是如何表示影像的内容。

20 世纪 70 年代，人们就对图像数据库进行研究，采用的主要方法是利用人工输入图像的各种属性，建立图像的数据库，用户仍然采用传统的文本检索方法，即通过多媒体的外部属性和简单的文字描述进行检索。这些外部属性和文字描述包括：多媒体的创建时间、创建人、创建地点；图像的标题、制作时间、收藏地点、版权状况、出版社；音乐的曲调名称、词曲作家、演奏者；视频的制片人、导演、地点、拍摄时间；相关的文件名、文件格式以及元数据标识等。到 20 世纪 90 年代，随着互联网的迅速普及和多媒体技术的发展，图像和其他多媒体数据日益增多，传统的对影像的文本标引和注释方式逐渐显现出它的不足，如投入的人力过多、难以全面准确地描述出影像的自身特点、描述具有较强的主观性等，因而出现了基于内容的多媒体检索技术。

基于内容的多媒体检索技术突破了传统的基于文本描述和检索的局限，直接对图像、视频、音频内容进行分析，利用媒体对象的语义、媒体的视觉和听觉特征来进行检索。也就是依据图像中的颜色、纹理、形状，视频中的镜头、场景、镜头的运动，声音中的音调、响度、音色等内容特征建立索引并进行检索。基于内容的检索还融合了模式识别、计算机视觉、图像理解等技术，是多种技术的合成。

基于内容的多媒体检索技术具有与传统文本检索不同的特征，它实施的是一种相似性检索，摒弃了传统的精确匹配，采用近似匹配或局部匹配的方法和技术逐步求精，来获得查询和检索结果。它直接对媒体的内容进行分析并抽取内容特征，利用媒体自身的特点进行标引和检索，在很大程度上避免了对影像的主观描述。

根据所检索媒体对象的不同，基于内容的多媒体检索技术又可分为基于内容的图像检索技术、基于内容的视频检索技术和基于内容的音频检索技术等。

6.3.2.1 基于内容的图像检索技术

基于内容特征的图像检索技术 CBIR（Content-based Image Retrieval，简称

CBIR）主要依据图像固有的特征来标引和检索。所谓图像特征包括：图像的画面内容特征，如图像颜色分布、纹理结构、轮廓等；图像描述对象特征，如人、物、景等；图像的相关信息，如作者、时间、地点及其他物理特征；图像的移动和组合特征等。基于内容的图像检索技术通过分析图像的内容，建立特征索引，并存储在特征库中。用户在检索查询时，可以从图像自身的特征将查询需求描述出来，就可以在大容量图像库中找到所需的图像。具体如图 6—4 所示。

基于内容的图像检索技术包括的关键技术有颜色特征提取、纹理特征提取、形状特征提取、相关反馈等等。基于内容的图像检索方式主要有 3 种：

（1）选择颜色的比例、轮廓形状以及纹理图案的图样进行查询。例如用户可以给出红、绿、蓝三种颜色的百分比，或从系统所提供的图例中选择某个作为检索图样。

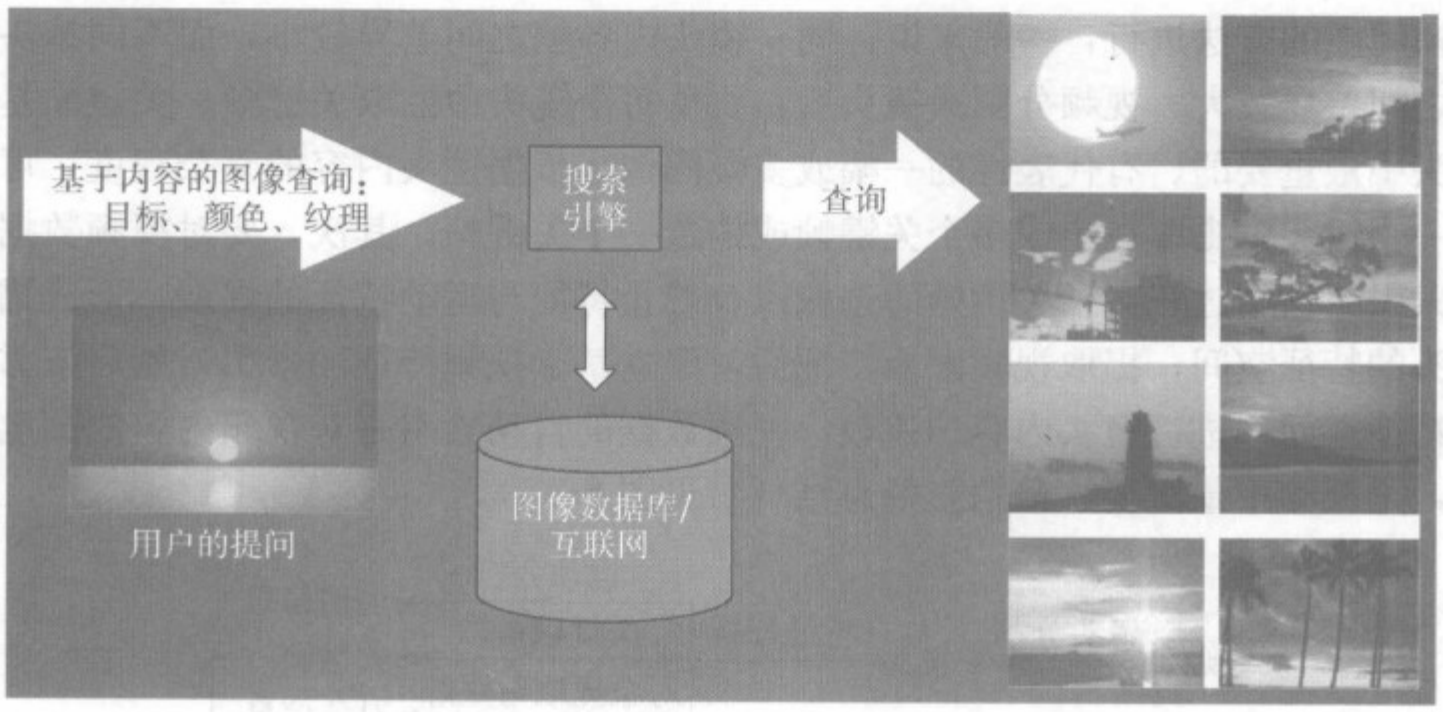

图 6—4　基于内容的图像检索技术

资料来源：http：//www. digiark. com/tian/lecppt/bdlec _ lib07. zip，2003-08-18。

（2）草图查询。用画图工具生成草图，从系统中查询与草图颜色分布、形状或纹理相似的结果。

（3）示例查询。选择系统中的一幅图像，要求系统检索与之类似的图像。用户一般是通过浏览选择系统提供的实例作为查询条件，然后再通过不断修改实例最终找到匹配目标。

目前，比较成功的应用基于内容的图像检索技术的系统有 IBM 公司的 QBIC 系统、MIT 媒体实验室的 Photobook 系统、新加坡国立大学的 CORE 系统、美国哥伦比亚大学的 VisualSEEK 系统等。

6.3.2.2 基于内容的视频检索技术

视频又称动态图像，是一组图像按时间顺序连续表现，它的表示与图像序列、时间关系有关。视频数据可用幕、场景、镜头、帧等描述。视频序列主要由镜头组成；镜头由一系列连续的帧组成；帧是一幅静态的图像，是组成视频的最小单位；场景含有多个镜头；幕是由一系列相关的场景组成，表达一个完整的事件。视频检索实际上是对动态图像进行检索，视频检索的实质就是在大量的视频数据中找到所需要的视频片段。

动态视频检索需要对视频信息进行视频分割和处理，包括视频结构的分析、视频数据的自动索引等。首先，要进行视频结构的分析。通过镜头边界的检测，即把视频分割成基本的组成单元——镜头，镜头就是由一系列帧组成的一段视频，镜头边界检测的核心处理是识别镜头的切换。目前镜头边界检测通常采用计算帧间差的方法进行，一般来说，同一镜头内各帧之间差异较小，而不同镜头的帧之间差异较大。视频分割成镜头后，要从每个镜头中抽取关键帧。关键帧是指镜头中最重要的、有代表性的一幅或多幅图像。依据镜头内容的复杂程度，可以从一个镜头中提取一个或多个关键帧或构造一个关键帧。其次，要对视频数据自动索引。这个过程包括关键帧的选取以及静止特征与运动特征的提取，形成描述镜头的特征空间；提取视频图像特征后，建立基于视频特征的索引，然后依靠这个特征空间来进行镜头内容的比较。视频数据的自动索引是对视频内容的高度概括，是视频中最重要、最精彩的总结（参见图 6—5）。

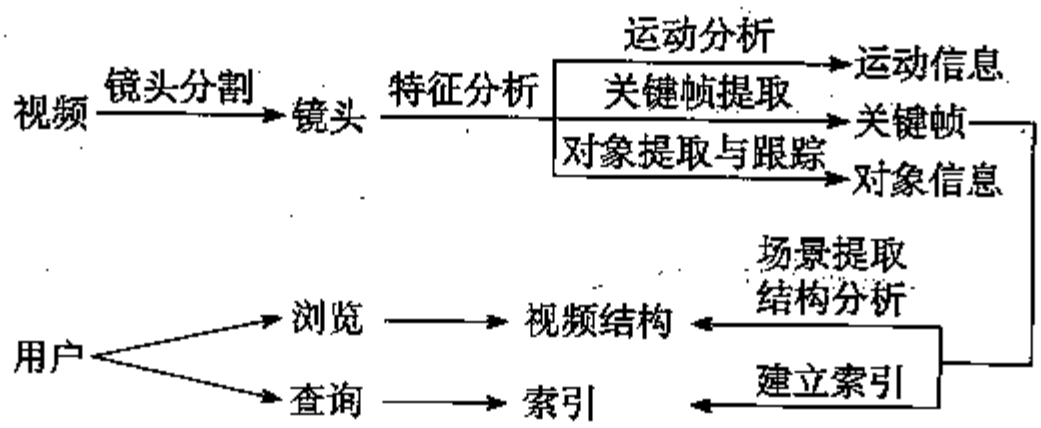

图 6—5 基于内容的视频检索技术

资料来源：http://media.ccidnet.com/media/ciw/1113/d1201.htm，2008-05-26。

这种新型的基于内容的视频检索技术，彻底改变了传统的通过快进或快退等顺序的方法进行人工查找的视频检索方式，满足了用户对大量视频多角度检索的需求。基于内容的视频检索的方法主要有基于关键帧的检索、基于运动的检索和浏览等。

6.3.2.3 基于内容的音频检索技术

音频是对声音进行数字化处理得到的结果。音频数据一般用音量、音调、音强、带宽、音长和音色等属性来描述，其中音量、音调、音强、带宽和音长属性易于通过技术手段进行信息化建模，而对音色的处理较为复杂。在检索前，首先对音频数据建立索引，索引可以基于韵律、旋律以及其他的感知或声学特征。

基于内容的音频检索就是将输入的字符序列和音频数据库中的字符序列相匹配。最简单的音频检索是用准确的序号查找出一段声音，较高级别的检索是匹配任何包含给定样值的声音的检索，更高级别的查询可以涉及频域信息或其他声学属性，最高级别的查询中可以包含声音的概念（主观）特性。基于内容的音频检索主要关心的是上述最后两级的声学和主观特性的查询。声音的一些声学属性，如音调、响度、音色，与音频信息的测量属性非常接近，因此可以在音频数据库中存储这些特性，以供检索。常用的音频检索方法有赋值查询、示例查询和分类浏览等。

目前有代表性的音频检索系统有美国加利福尼亚有限责任公司开发的Muscle Fish系统。我国上海交通大学图书馆也创建了一个音频数据库，允许非音乐专业人员方便地采用传统的检索途径，如音乐家名、曲名、作曲家、生平介绍等进行全文检索，获得相关的曲子；也允许音乐专业人员用乐句进行全曲检索。

基于内容的多媒体检索技术作为一种先进的检索技术，广泛地应用于多媒体数据库、知识产权保护、网络多媒体搜索引擎、数字图书馆、交互电视、艺术收藏和博物馆管理、遥感和地球资源管理、远程医疗、天气预报以及军事指挥系统等等。它与数据库技术相结合，可以方便地实现海量多媒体数据的存储和管理。与Web搜索引擎技术相结合，可以用来检索HTML网页中丰富的多媒体信息，具有广阔的发展前景。

【案例】

Proquest系列数据库检索实例

一、布尔逻辑算符的应用

AND：查找包含所有词语。

实例：Internet AND education

AND NOT：查找包含第一个单词而不包含第二个单词的文章。

实例：Internet AND NOT html

OR：查找任何单词。

实例：Internet OR intranet

二、截词符的应用

符号 * ：仅用作右截词符，它将查找所有形式的单词。

例如，检索 **econom*** 将查找“economy”、“economics”、“economical”等。

符号？：用于替换单词内部或末尾的单个字符。？不能用于单词的开头。

例如，检索“**wom? n**”将查找“woman”和“women”；检索“t? re”将查找“tire”、“tyre”、“tore”等。

关键术语

计算机信息检索	脱机检索	联机检索	光盘检索
网络检索	检索策略	检索表达式	逻辑表达式
位置检索表达式	截词检索表达式	限制检索表达式	信息检索技术
全文检索技术	基于内容的图像检索技术		

思考题

1. 什么是计算机信息检索？
2. 概述计算机信息检索发展简史。
3. 简述计算机信息检索的类型。
4. 与传统的手工检索相比，计算机信息检索有什么特点？
5. 简述计算机检索策略的含义和作用。
6. 什么是检索表达式？
7. 检索表达式的构成可以采用哪几种方法？
8. 提高查全率的方法有哪些？请举例说明。
9. 结合自己的检索实践，说明如何提高查准率。
10. 概述全文检索技术。
11. 概述基于内容的图像检索技术。

CHAPTER SEVEN

第7章
联机检索

【本章要点】

- ◇ 分析联机检索系统的组成结构及其特点
- ◇ 介绍联机检索系统的服务方式和功能
- ◇ 论述联机检索系统的选择和发展
- ◇ 详细介绍三种国外联机检索系统

引子

在信息检索领域，联机检索是计算机信息检索的重要组成部分，有许多信息机构都开展了这项服务。如，中国科学院文献情报中心连通了STN情报检索系统和DIALOG系统，对用户提供定题检索、科技查新和情报服务等。[①] 北京大学图书馆也提供DIALOG系统的联机检索服务。传统的联机检索系统尽管有很多优势，但由于其采用指令检索，并且检索费用昂贵，在很大程度上限制了其用户范围，特别是互联网的产生和普及，使得传统的联机检索受到了新的挑战。因此，国际上著名的联机检索系统都相继推出了基于互联网平台的检索服务。

① 参见 http：//www.las.ac.cn/index.jsp，2008-03-20。

7.1 联机检索系统概述

7.1.1 联机检索系统的含义

联机检索系统是专门提供联机检索服务的信息检索系统。联机检索是指用户利用终端设备，通过国内或国际（卫星）通信网络，与大型计算机检索系统的主机联接，从而检索世界各国存储在计算机数据库中的信息资料的过程。联机检索系统允许用户用人机对话的交互方式直接访问系统及数据库，检索以实时在线的方式进行。用户按照联机检索系统的要求和规定输入相应的检索提问，计算机执行操作，并在用户终端显示屏上输出检索结果。用户可随时修改检索提问，以得到满意结果，系统通过打印或传输方式将结果提交给用户（图 7—1）。

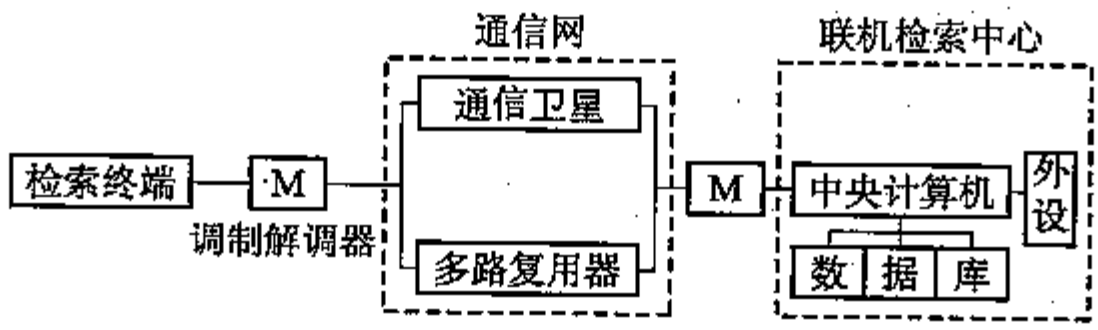

图 7—1　联机检索系统示意图

资料来源：邱宏、付琼：《联机检索与网络信息检索的比较研究》，载《东北电力学院学报》，2001（2）。

联机检索系统通常由检索终端、通信网和联机检索中心 3 个部分组成。

检索终端是联机检索系统与用户联系的接口，它可以是由显示器、键盘和打印机构成的标准终端，也可以是电传终端或微机终端。目前，主要采用的是微机终端。它的作用主要是向联机检索中心发送检索请求和接收信息。

通信网是联机终端与联机检索中心联系的桥梁。联机检索中的通信网有这样一些类型：一是公用电话网，用户通过拨号和租用专线与中央计算机连接，检索结果也通过电话线传送，按时计费，通信质量较差；二是专用数据通信网，但由于专用网的机线设备利用率低，不能实现公用的数据交换，正逐渐被公用数据网取代。三是公用数据网，由于采用分组交换技术（或称分组交换网），线路利用率高、时延小、可靠性好、灵活性强，因而得到了最广泛的应用。著名的有 Telenet（美国）、Tymnet（美国）、Datapac（加拿大）、EPSS（英国）、Tramspac

(法国)、Euronet(欧盟)等。

联机检索中心是系统的中枢，由中央计算机、联机数据库、检索与管理软件及相应的检索服务体制组成。中央计算机是联机系统硬件的核心部分，它在很大程度上决定着系统的检索速度和存储容量，包括中央处理机、中央储存器、通信部件、控制部件和连接外围设备的通道输入和输出子系统，中央计算机的主要功能是在系统软件和检索软件的支持下，有效地进行信息的存储、处理和检索，管理和控制整个系统的运行。联机数据库是联机检索系统的信息源，是系统各种数据库的总称，由系统本身自建或由数据库生产者提供。如，DIALOG 有 600 多个联机数据库，STN 有 220 多个联机数据库。

联机检索系统是计算机技术、数据库技术和通信技术发展的产物。在 1946 年计算机产生之后，人们就开始了将计算机应用于信息检索领域的探索，1965 年，美国系统发展公司进行了关于联机检索网络的第一次全国规模的演示。此后，美国洛克希德公司的 DIALOG、系统发展公司的 ORBIT、美国国家医学图书馆的 MEDLARS 系统、美国书目检索服务公司的 BRS 等联机检索系统纷纷建成，并投入商业运行，成为重要的国际联机检索系统，实现了跨国界跨洲的信息检索，使人工检索需要几天、几个月、几年才能完成的工作，在几分钟、几十分钟内完成。1983 年美国化学文摘社(CAS)、德国卡尔斯鲁厄专业信息中心(FIZ-Karlsruhe)和日本科技情报中心共同推出了 STN 联机信息检索系统。20 世纪 80 年代，发达国家的国际联机网络和检索终端覆盖和深入范围进一步扩展。90 年代，以计算机网络通信技术为基础，光缆为骨干的大容量高速度电子数据传输系统——“信息高速公路”首次在美国提出，将联机检索又推向了一个新的发展。目前大约有 200 个左右联机检索系统，著名的联机检索系统有：DIALOG 系统、ORBIT 系统、ESA-IRS 系统和 STN 系统等。

7.1.2 联机检索的特点

与光盘检索和网络检索相比，联机检索发展比较早，具有一些非常明显的特点：

1. 检索范围广，信息数据量大，数据质量高

联机检索开始于 20 世纪 60 年代，发展至今经历了约 40 年的历程，积累了非常丰富的高质量的信息资源，联机检索系统所拥有的绝大多数数据库都是从国际上权威的数据库生产商那里收购或租借的，信息资源非常可靠，更新也很及时，这是其他任何计算机检索系统所无法比拟的。从联机检索系统查找信息，不仅可以获取最新资料，而且能追溯查询历史性资料。据统计，世界上公开发行文

献一半以上都可以通过国际联机检索系统查到。仅世界上最大的国际联机检索系统 DIALOG 就拥有 600 多个联机数据库，其内容涉及 40 多个语种以及占世界发行总量 60%的 6 万多种期刊。学科覆盖面极广，几乎涉及全部学科范围，包括综合性科学、自然科学、应用科学和工艺学、社会科学和人文科学、时事报道和商业经济等。其数据来源于各种不同的图书、报纸、杂志期刊、技术报告、会议论文、专著、专利、标准、报表、目录、手册等上的信息。与互联网普通信息相比，联机数据库都经过了严格的加工、标引，信息的附加值高，可靠性好，不用担心出现互联网那种良莠不分、加工粗糙的信息。DIALOG 检索系统中有许多世界上非常著名而且经常使用的数据库，包括 CA（《化学文摘》）、INSPEC（《英国科学文摘》）、MEDLINE（《医学文献数据库》）、MATHSCI（《数学文献数据库》）、BA（《生物学文摘》）、NTIS（《美国政府报告》）、SCI（《科学引文索引》）、EI（《工程索引》）、ISTP（《国际科技会议录索引》）、SSCI（《社会科学引文索引》）等等。

2. 检索速度快

联机检索和网络检索不同，主要是由专业的检索人员来完成，这些专业人员一般都具有联机检索系统熟练的操作技能、一定的专业知识和外语水平，而且联机检索是以实时方式进行，从检索提问的输入、调整、修改到获得最终结果的整个过程一般只需几分钟至十几分钟。就检索时间花费和检索结果有效性的比值来说，在单位时间里，通过联机检索所获得的有效结果远大于其他形式的计算机检索，从这个意义上来说，联机检索的速度远远超过了光盘检索和网络检索，更是手工检索所无法比拟的。光盘检索速度虽然也不慢，但在单位时间内得到的数据量不大。网络检索虽然输入关键词后，也会非常快地得到检索结果，但是，对网络检索用户而言，常常是花费了很多时间利用搜索引擎查找网络信息，而最终得到的有效结果或者自己真正需要的信息却很少。因此，目前的科技查新服务基本上仍由联机检索系统完成。

3. 查全率和查准率高，系统检索功能丰富，检索结果输出形式多样

联机检索系统拥有庞大的数据库资源，所收录的信息都是经过人工处理的，同时，它提供多种途径的检索方式，能够从不同的角度满足用户对信息检索的需求。联机检索系统是一种比较成熟的信息检索系统，系统的建设和完善都是围绕着提高查全率和查准率来进行的。同时，联机检索所能提供的检索方法很全面，能够非常有效地提供布尔检索、字段限定检索、截词检索等，一些联机检索系统还为用户提供了同步查询多个数据库的机会，如 DIALOG 系统的跨文档检索功能等。网络检索虽然从理论上也支持布尔检索、字段限定检索、截词检索等检索

功能，但实际检索效果并不十分理想。联机检索系统可以提供多种输出形式，而网络检索的输出格式相对来讲，比较单一。

4. 安全性能高

联机检索系统都是固定地属于某一个机构或公司，集中管理的模式，在很大程度上保证了检索系统的安全性能，确保了数据的稳定性和可靠性。联机检索系统有它自己的通信网络、专用通信软件以及较为完备的安全认证技术，从而保证了系统的安全。而互联网信息量大而无序，存在着大量难以检测的病毒，防火墙屡屡被攻破，引发泄密等一系列问题，安全问题成为困扰网络发展的重要障碍之一。

5. 检索费用较高

国际联机检索的费用不仅包括显示（打印）费、字符费以及计算机检索的机时费，还包括国内国际通信费，使得联机检索费用一般远远高于光盘检索和网络检索。联机检索一个课题总的费用一般在 1 000 元以上，普通用户难以接受。

Susan Feldman 的一项研究比较了传统联机检索和搜索引擎检索，为这两类检索勾画了一个基本的轮廓，反映了它们在检索效果上的优越性及互补性。这项研究由 DIALOG、DJI（Dow Jones Interactive）以及 Information Today Inc. 资助。研究调查了多名有经验的专业检索人员，他们按照特定信息要求检索 DIALOG 或 DJI，之后用 Web 搜索引擎检索同样的信息提问，最后对两者的检索结果进行分析比较。这项研究发现：（1）总体看，DIALOG、DJI 比搜索引擎检索到更多相关的文献。（2）检索时间上，利用搜索引擎比 DIALOG、DJI 联机检索要花多一半以上的时间。如果加上下载、整理信息等过程，则利用搜索引擎所花费时间就更多。（3）查准率上，搜索引擎误检的文献是 DIALOG、DJI 误检的文献的两倍。此外，DIALOG、DJI 检索文献的相关程度得分多居中等偏高的范围（3、4、5），搜索引擎检索的结果得分则相对偏低（分散于 1～5 之间）。这是由于传统的联机数据库收录了高质量的信息，这些信息又经过专门的组织，而 Web 信息的质量变化大，难以预测，也难以确定何时应该停止检索。（4）从总体来看，传统联机系统能够比搜索引擎检索到更高价值的信息，如果同时使用这两种检索设施，检索到的信息源能满足不同的用途，而且可在某些方面起到互补的作用。[①]

7.1.3 联机检索系统服务方式

联机检索是一种重要的现代化检索手段，数千个数据库几乎覆盖了所有学科

① 参见伍宪：《Web 检索与联机检索》，载《图书馆论坛》，2001（1）。

领域，迅速高效地提供信息服务。目前，大多数国际联机检索系统都提供以下一些服务方式：

1. 回溯检索

这是联机用户使用最多的一项服务内容，适用于项目查新、文献调研、课题立项、申报专利、了解市场动态和竞争对手、新产品开发、公司的背景情况调查、经济预测等信息检索需求。回溯检索是根据用户的要求，从现在追溯到过去某个时间，一次性提供若干年内的有关信息，利用它既可以查找过去一段时间的信息，也可以查找最新的信息，能适应大多数用户的信息需求。

2. 定题服务

这是联机检索系统一项重要的服务功能，它能够及时提供有关主题领域的最新文献信息，主要用于研究进程中跟踪同类专题的动态和进展。定题服务指针对用户特定的信息需求，对储存到数据库中的最新文献信息进行检索，并将结果提交给有关用户。实施这项服务的时候，需要将用户的信息需求转化为检索提问式，并将其长期保存在联机检索系统中，每当数据库增加或更新记录时，系统会自动将最新信息检出，定期、连续、主动地提供给用户。

3. 联机订购原文

联机检索系统主要提供的是原始文献的索引、题录或文摘，即所谓的二次信息，只有少部分是原始文献，因而，用户通过国际联机检索到的结果，如果需要原始文献，但在国内馆藏中又无法索取时，可通过国际联机检索终端订购，联机检索系统会根据用户的需要，提供原始文献的传递服务。

4. 电子邮件服务

联机检索系统还提供有电子邮件服务，便于用户与用户、用户与系统之间互相交流检索经验。电子邮件服务是一项成本较低，但却方便快捷的服务，许多大型联机检索系统在不断增加数据库和完善检索软件的同时，增加了电子邮件服务。电子邮件服务也可用来传递系统管理人员对用户提出的检索咨询的回答等。

5. 光盘服务

联机检索系统在提供联机检索的同时，部分联机检索系统也提供光盘检索服务，以达到方便用户的目的。例如，DIALOG 系统就提供有 40 多种数据库的光盘，其中，一些来自于联机数据库，在这些光盘数据库中有很大一部分能直接提供全文。这些光盘数据库的结构、指令与远程联机检索系统一样，利用联机检索系统的光盘服务，可以大大降低检索费用，而且还可以用于自建数据库时数据的套录以及联机检索的实习培训。

7.1.4 联机检索系统的功能

联机检索系统发展得比较成熟，数据质量可靠，检索功能齐全。一般的联机检索系统都提供如下的检索功能：

1. 单词检索

这是利用联机检索系统查找信息的最简单方法，检索词是一个单词，系统在接到检索指令后，检出与该词匹配的相关文献。

2. 词组检索

词组检索分为两种情况，一是固定词组检索，另一种是单元词组配检索。

3. 布尔逻辑检索

联机检索系统都支持布尔逻辑检索（AND，OR，NOT）。布尔逻辑检索表现力较强，可以将复杂的提问进行概念分解，然后通过这些算符把分散的概念连接在一起。但它也存在着一定的局限，在布尔逻辑状态下，所有与 AND 连接的概念必须同在一条记录中出现，难以反映主题概念的重要程度。而且检索结果一般是按照编年的逆顺序或用户专门选定的顺序显示，不说明所有或任何检索词之间的可能关系。近年来，一些联机检索系统为了突破布尔逻辑检索的局限，推出了新的检索技术 Target、Freestyle 与 WIN 技术。它们允许用户选择确定那些被布尔检索漏掉的相关标题，起到辅助性和相关性检索的作用。

4. 截词检索

截词检索是一种灵活性强，简便易用的方法，联机检索系统都采用了截词检索，但各个检索系统采用的截词符略有不同。如 STN 系统中，“?”为无限右截词符，有些数据库支持左截断，“#”为有限截词符，“!”为中间截词符。DIALOG 系统允许右截断和中间截断，截词符统一为“?”。在使用联机检索的截词方式时，输入的词干不应太短，以避免明显的误检。

5. 位置检索

在联机检索系统中位置检索的使用也比较频繁，它有利于提高检索的全面性，联机检索系统均支持位置检索。常用的位置算符有：（W）、（nW）、（N）、（nN）等。

6. 限制检索

为了提高检索的准确性，联机检索系统允许将输入的检索词限定在某一范围（字段）内。联机检索系统的可检字段（也称检索项）通常分为表示文献内部主题特征的基本索引字段和表示文献外部形式特征的辅助索引字段两大类。基本索引字段又称主题字段，含有所有与主题内容相关的词，包括题名（TI）、文摘

（AB）、叙词（DE）、注释（NT）等等。辅助索引字段也称非主题字段，含有记录中除基本索引字段外的那部分信息，如作者（AU）、刊名（JN）、文献类型（DT）、语种（LA）、出版年份（PY）等等。这些字段随数据库不同而有所不同。在检索时，如果没有限制，系统将自动在所有的基本索引字段中进行检索。每个具体的联机检索系统所采用的限制方法略有差异，一般用后缀进行限制，如information/DE等，但也不完全一样。

7.1.5 联机检索系统的选择

联机检索系统经过数十年的积累和发展，汇集了大量的数据库，具备了一定的规模，其中，许多著名的国际联机检索系统或多或少重复收集了一些数据库，也就是说，一条相同的信息往往在不同的系统中都可以查到。因而，在使用联机检索系统前，有必要对检索系统进行选择。选择时应考虑以下几方面：

1. 数据库的信息覆盖和时间范围

同一数据库的记录在不同的联机检索系统中，包含的信息不完全一样，有的检索系统包含有更多的标引信息或文摘。同时，同样一个数据库常常被多个检索系统收录，但有的系统仅收录数据库的一部分，有的收录其全部内容，如CA（化学文摘）数据库在ESA/IRS系统中从1969年开始，而在DIALOG中则从1967年开始。

2. 检索功能和打印格式

联机检索系统都提供有丰富的检索功能，基本上都支持布尔检索、限制检索、位置检索、截词检索等，但各个检索系统所采用的算符和具体的方法会有所不同，尤其是位置算符和截词功能，因而，应选择自己比较熟悉的系统，提高检索效率。此外，同一数据库的可检字段在不同的联机系统中也可能不一样，比较可检字段在不同系统中的数量，也是联机检索系统选择标准之一。还要考虑打印格式的因素，比如，系统能否满足按字段打印检索结果的要求（如著者、题名、文摘等），还是仅给出几个标准格式。

3. 原文联机订购

用户进行联机检索的最终目的是获取原始文献，国内用户利用国际联机检索，常常会碰到找不到原文的问题。因而，联机检索系统能否直接提供原文联机订购服务，能否以合理的价格提供，都是用户考虑选择联机检索系统的重要因素。

4. 检索费用

相对于其他计算机检索系统来说，联机检索的费用是比较昂贵的。不过，不

同系统还是存在着价格差异，这是选择系统要考虑的最重要的因素之一。检索前应该考虑这样一些费用因素：终端与主机之间的通信费；数据库使用费是否包括在数据库连接费中，或者两者分开计算；每个记录的脱机、联机打印费是多少；各系统有无优惠条件，例如一年中联机时间超过预定数，是否可从系统中得到折扣费；系统的响应时间。检索费用一直是影响我国用户使用国际联机检索系统的一个重要因素。

7.1.6 网络环境下联机检索系统的发展

回顾计算机信息检索的发展历程，可以看出，光盘检索的出现曾对信息检索领域霸主地位的联机检索形成了强烈的冲击，光盘检索在检索费用、易用性以及对环境的要求等各方面都呈现出明显的优势。互联网的出现对联机检索系统的发展，更是产生了深刻的影响。尤其是万维网作为互联网发展史上的里程碑，极大地推进了网络检索的发展。20 世纪 90 年代，搜索引擎层出不穷，越来越多的用户开始利用互联网这个全球性的信息资源宝库，来获得需要的信息资源。联机检索在这样的网络背景下，再次积极主动地调整自己，以适应生存和发展的需要。联机检索系统采取的措施有：

1. 建立网站，推出网上服务

面对网络检索发展的强大压力，许多商业联机服务机构纷纷调整政策，做出反应，显现出联机检索系统融入互联网的趋势。它们在互联网上建立自己的网站，利用互联网直接为终端用户服务。DIALOG 在 1997 年左右开始了万维网检索服务，有选择地收集了联机数据库中的部分数据库，采用对话框式的检索方式，以友好的界面为用户提供服务。用户不需要掌握 DIALOG 的检索命令，只要输入检索词就可以进行检索。其他一些著名的商业联机服务系统也纷纷上网，为自己在互联网上开设用户存取节点。用户只需通过互联网，远程登录其主机，就可以使用这些联机系统的网上资源。

2. 调整收费制度，吸引更多用户

联机检索高昂的费用一直是阻碍用户利用联机检索系统的一个主要原因，这与几乎免费的网络检索相比，可以说有天壤之别。在网络检索的强大冲击下，联机检索被迫调整收费制度。DIALOG 系统于 1998 年 5 月机构重组合并后，终止了基于链接时间收费的方式，推行按照用户利用系统资源多少的收费 DISLUNIT 政策。此外，DIALOG 系统还每月提供一两个免费文档，供用户利用。1988 年意大利的 ESA 开始实行新的收费政策，即每次检索只收取数据库使用费和机时费。

7.2 主要联机检索系统简介

大型国际联机检索系统主要有：DIALOG、OCLC、MEDLARS、STN、ORBIT、BRS等，国内联机系统中较大的4个系统是：中国科技信息研究所的ISTIC系统、北京文献服务处的BDSIRS系统、化工信息研究所的CHOICE系统和机电信息研究所的MEIRS系统。

7.2.1 DIALOG系统

7.2.1.1 概况

美国DIALOG系统是目前世界上最强大的国际联机检索系统，也是运作最成功的联机商业数据库系统之一。它始建于1966年，于1972年开始商业运营。最初由美国洛克希德导弹航空公司所属的一个情报科学实验室负责建立。1981年6月，成为该公司的一个子公司，并开始独立经营。其中心设在美国加利福尼亚州的帕洛阿尔托市。1985年，Knight-Ridder公司收购了DIALOG系统，2000年，Thomson公司并购了DIALOG国际联机检索系统。

DIALOG拥有900多种数据库，15 TB的信息总容量，14亿条记录，方便、灵活、快捷、准确、全面地提供各种科技、商业、社科高质量的信息。专业范围涉及：新闻与媒体、商业与金融、知识产权、政府和法规、科学技术、能源和环境、医学、制药、化学、食品与农业、社会科学、电子信息行业等几乎所有的专业。其中，有许多著名的数据库如CA、INSPEC、MEDLINE、MATHSCI、BA、NTIS等都加入到DIALOG系统中；还有著名的几大检索数据库，如SCI、EI、ISTP、SSCI、A & HCI（《艺术与人文科学引文索引》）等也都可从DIALOG系统中检索。世界著名的DERWENT专利数据库以及美国专利、欧洲专利、日本专利等数据库也都可在DIALOG中查询。

DIALOG系统遍布6大洲，共有约25 000个既有客户遍及103个国家并拥有总数量超过200万的最终用户，至今仍是全球最大的专业信息提供商。DIALOG系统的联机数据库主要有四种类型：（1）题录文摘型数据库；（2）名录手册型数据库；（3）全文型数据库；（4）数值型数据库。

DIALOG系统在互联网上设立了网站，用户可以通过互联网检索DIALOG系统，其检索平台主要有远程登录方式、Dialogclassic、Dialogselect、Dialogweb等四种方式。

7.2.1.2 DIALOG 系统的检索步骤

DIALOG 是一个非常典型的联机检索系统，应用 DIALOG 系统时，有这样一些步骤：

1. 联机准备

进行 DIALOG 系统联机检索的基本前提是要拥有该系统的使用权，也就是说要向 DIALOG 系统申请账号，并交纳应付费用。

在联机准备阶段，首先，要做好所需的软、硬件准备工作，包括终端、通信软件和调制解调器等。这是与 DIALOG 联机的物质基础。其次，要制定周全的检索策略，将用户的检索提问转化为 DIALOG 系统所能处理的检索语句，其核心是编制检索式。

2. 选择联机方式

目前可以采用两种方式与 DIALOG 联机：

一种是通过 CHINAPAC（专线）与 DIALOG 联机。过程依次为：输入主机所属的分组交换网址、输入系统标识符、输入用户号、用户密码。

另一种是通过互联网与 DIALOG 联机。有两种具体方法：可以使用远程登录命令 TELNET DIALOG. COM 或通信软件 DIALOGLINK，登录到 DIALOG 联机系统。更常用的方法是直接用浏览器以 Web 方式检索 DIALOG 系统。DIALOG 系统的 Internet Web 界面的检索方法，主要包括四种：

（1）利用 Web 直接上网检索，网址为 http：//www. dialogweb. com/。

这种方式检索费用最低，仅在运行和调用数据的一刹那计算上网费用，缺点是如果用户需要将所有检索过程存盘，就要一屏一屏地存盘，否则随着检索指令的变化，不及时存盘，数据容易丢失。

（2）为专业人员推出的 Web 界面，网址为 http：//www. dialogclassic. com/。

这是最新推出的界面，速度快，检索过程每一屏均保留，不丢失数据，便于存盘，界面是专业人员熟悉的界面，能很快地从旧检索方式转入适应新的 Web 界面。

（3）非专业检索人员 Web 界面，网址为 http：//www. dialogselect. com/。

DialogSelect 作为在互联网上的傻瓜界面主要针对最终用户，而非专业人员。对于初学者、最终信息用户和不愿学习 DIALOG 检索指令的人可使用其傻瓜界面。

（4）数据库蓝页，网址为：http：//library. dialog. com/bluesheets/。

可以按数据库名称、文档号、主题浏览蓝页。数据库蓝页是 DIALOG 系统为了让用户了解每一个数据库的特征、可检字段、字段性质、输出格式等内容提

供的一个检索指南，具体提供有每一种数据库的收录范围、可供检索的字段、打印格式、记录样式及收费情况等。

3. 选择检索方式，熟悉检索指令和操作方式，进行检索

DIALOG 系统提供两种检索方式：命令式和菜单式。菜单式检索方式采用图形界面，简单易用，用户使用起来非常方便，无需经过非常专业的检索培训，DialogWeb（guided search）和 DialogSelect 就采用了菜单式检索方式，这种方式适合于非专业人员检索 DIALOG 系统。命令式检索方式指按照 DIALOG 系统规定的各种指令和格式进行检索，是 DIALOG 系统传统的联机检索方式，远程登录的 DIALOG 系统、DialogClassic 和 DialogWeb（command search）采用的就是这种命令式检索方式。命令式检索方式要求用户掌握和熟悉 DIALOG 系统的指令，比较适用于专业检索人员利用。

目前，用命令式检索方式对 DIALOG 系统实施检索依然是一种非常重要的方式。系统规定了大量的指令、字段限制符、逻辑算符和位置算符，可配合使用，检索功能十分强大，能够进行深度和广度的检索，确保查到非常切题的信息。检索者在保证检索质量的前提下，巧妙地运用一些检索指令，可以有效地降低检索费用，获得更高的检索效率。

系统提供了 60 多种检索指令，DIALOG 系统使用的命令，其一般格式为：c item [/x]。其中 c 是指令，采用单字母的较多；item 是实质项；/x 是约束项；[ ] 内的项目在特定条件下可以缺省。DIALOG 对大小写不敏感，大小写字母等效。其中，基本检索指令有：

（1）b 命令。

b 是 begin 的缩写，用于查询时打开特定的数据库（在 DIALOG 中又称文档 File），使用格式为：b fileno 或 b filegroup。

fileno 是文档号，DIALOG 一共有 600 多数据库，每个数据库编定一个或几个文档号。filegroup 是文档群代码，DIALOG 按学科专业划分了近 200 个文档群，每个文档群用几个字母缩写来表示，包括若干个具体文档，查询时只要指定文档群代码，就同时搜索其包括的所有文档。例如，计算机科学的文档群代码为 COMPSCI，包括一系列文档。不同文档群包括的具体文档相互有交叉，同一文档可以被包含在多个文档群中。还有包含大学科群的大文档群，如 ALLSCIENCE 或 ALLTECHNOLOGY 指科技，ALLSOCIAL 或 ALLHUMANITIES 指文科，ALLMEDICINE 指医学，ALLBUSINESS 指商业和金融，ALL 指所有学科，等等。详尽的文档和文档群目录可以在提供 DIALOG 国际联机检索服务的机构查到。

例如：

b 348（打开 348 号文档）；

b COMPSCI（打开计算机科学的文档群）；

b 2，4，6，8，14，35，347，348，653（同时打开多个文档，文档号之间用逗号隔开）。

(2) s 命令。

s 是 select 的缩写，是基本查询的主要命令，使用格式为：s words。words 是提请查询的词、词组或用逻辑组配结合成的检索式。这是在 DIALOG 系统检索中经常使用的命令。

例如：

s information（查找检索词 information）；

s information/TI（查找检索词 information 出现在题名字段的信息）。

(3) t 命令。

t 是 type 的缩写，用于显示查询结果，使用格式为：t si/fo/no。

si 是集号，系 s 命令所产生，打开一个特定数据库后，第一个 s 命令产生 s1，第二个 s 命令产生 s2，以此类推，t si 就是将第 i 个 s 命令的查询结果显示出来。fo 是显示格式，系统规定了 10 种预定义打印格式，也可以直接用 ab、au、so、ti 等指定显示文摘、作者、来源、题目等信息。no 是指查询结果的记录序号，3 代表第三篇，1—5 代表第一至第五篇，all 表示所有记录。

例如：

t s1/7/all（表示用 7 号格式显示第一次查询所有结果）。

(4) rd 命令。

rd 是 remove duplicates 的缩写，用于对来自不同数据库的文献进行“去重”，使同一篇文献只出现一次（同一篇文献可能被同时收入多个数据库），使用格式为：rd [si]。si 缺省时约定为前一次查询结果。

(5) logoff 命令。

系统脱机的命令，使用格式为：logoff。执行 logoff 后，系统就断开用户与主机的连接，并显示检索时间和费用。类似的脱机命令还有：bye、disc、log、logout、off、quit、stop 等。

此外，还有一些其他的指令：

add <files>：添加查询文档；

cost：显示联机费用；

ds（display sets）：显示查询过的集号；

e（expand）<xx>：扩展查询项目（xx 系字段，包括 au、ti 等）；

edit email address：编辑脱机打印送达电子邮件地址，与 pr 命令配套；

exs（execute steps）<stg>：用于执行存储下来的搜索策略；

idpat <si>：专利分组排序专用命令，有专利去重功能；

map <xx> [<temp><steps>] <si>：按字段和查询步骤存储搜索策略；

pr（print）<si/fo/no><addr/via emailaddr.>：脱机打印查询结果；

print cancel：取消打印；

rank <xx>：将查询结果按字段频次统计排序；

repeat：与 add 配对的命令，对新增文档重复已作过的查询；

report <si/xx/all>：报告查询结果；

save <file>：存储全部搜索策略；

sort <si/all/xx>：将查询结果按指定范围分类排序；

set...：set 命令系列主要用于设置屏显格式，如 set select short 等。①

7.2.2 STN 系统

7.2.2.1 概况

STN（The Scientific and Technical Information Network International，简称 STN）系统创建于 1983 年，由德国卡尔斯鲁厄专业信息中心（Fach infomations zentrum Karlsruhe，简称 FIZ）、美国化学文摘社（Chemical Abstracts Service，简称 CAS）和日本科技信息中心（Japan Science and Technology Corporation，简称 JST）合作开发，是当今世界著名的国际联机检索系统之一。三个服务中心分别位于德国卡尔斯鲁厄、美国哥伦比亚和日本东京，由海底电缆连接。用户只要与其中一个服务中心的主机联机，就可实现对三家主机的同时访问。STN 系统目前收录了 220 多个世界著名的数据库，涉及 55 个专业领域，如化学、工程、生命科学、生物技术、专利、数学、物理、商业等各基础学科领域和综合技术应用领域。STN 中的数据库类型有：书目型、全文型、名录型、数值型和混合型。

STN 系统是一个比较有特色的联机检索系统。主要以科技信息为主，其中

---

① 参见 http://libweb.zju.edu.cn/02/lesson/Teach/SearchYan/Ch5/CH5.htm，2007-08-20。

化学化工信息和专利信息是该系统的特色。STN 系统包括有许多非常权威的专业数据库，如生物学文摘 BIOSIS、化学文摘 CA、英国科学文摘 INSPEC、美国医学文摘 MEDLINE、美国政府四大报告 NTIS、科学引文索引 SCI、世界专利索引 DERWENT、有机化学物质手册数据库 BEILSTEIN 等等。STN 是世界上第一个实现图形检索的系统，能够实现化学物质的结构检索，是检索化学化工信息的最佳系统。

STN 系统的经营机构不是纯商业性机构，每年都得到德国政府和日本政府的资助。它不以营利为目的，联机检索费用明显低于 DIALOG 系统，这些都是优于其他系统的地方。

7.2.2.2 联机方式

用户如果想检索 STN 系统里的信息，必须建立 STN 账号，有自己的用户名和密码。用户可以通过三种方式与 STN 联机：

1. STN Easy（http：//stneasy. fiz-karlsruhe. de、http：//stneasy. cas. org、http：//stneasy-japan. cas. org）

STN Easy 是基于图形的 Web 界面，检索方法简便，无需掌握检索指令，主要是针对没有检索经验的普通用户的。在 STN 系统中，STN Easy 可检索 90 多个数据库，收费也是最低的。STN Easy 具有四种检索选择：EASY SEARCH、ADVANCED SEARCH、CAS NUMBER SEARCH、PATENT SEARCH。STN Easy 可直接在 STN 网上免费申请账号，30 分钟后就可发送至 E-mail 中。

2. STN on the Web（http：//stnweb. fiz-karlsruhe. de、http：//stnweb. cas. org、http：//stnweb-japan. cas. org）

STN on the Web 是基于文本的 Web 界面，它结合了 STN 的命令检索和浏览器的强大功能，用这种方式可以检索 STN 系统的所有数据库，使用所有检索指令。特别是化学物质结构图形的检索，还可直接进入原文库。该方式适用于有经验的检索者，非常方便，检索结果可以保存成 HTML、PDF、RTF、ZIP 等格式，FREE SEARCH PREVIEW 指免费预扫描功能。

3. STN Express with Discover! 6. 0

STN Express with Discover! 6. 0 是 TELNET 界面，适用于专业检索人员，与传统的联机检索界面相似。它是一个非常完整的联机检索经典软件包，允许在脱机状态下轻松编辑检索策略（包括多库检索），在联机状态下迅速检索。该检索方式提供两种检索选择：即 STN 命令语言和 Discover! Wizards。Discover! Wizards 可以帮助不太熟悉 STN 检索方式和检索指令的用户全面地检索 STN 各个数据库。检索结果可以保存为 HTML、PDF、RTF、EXCEL 等格式。也可以

通过化学物质登记号、专利号链接到物质信息及专利信息，并可获取全文。

7.2.2.3 检索指令

STN 系统的检索指令有些与 DIALOG 系统用法类似，也有一些是独有的。STN 系统常用的指令有：

1. begin 命令

表示打开一个或多个数据库。例如：

b ca，inspec（同时打开这 2 个数据库）。

2. search 命令

表示检索相关的词、词组。例如：

s heat/ti（查找检索词出现在题名字段中的信息）。

每使用一次检索指令，系统响应时会给出一个组号 L，它是数据库中包含该数据的所有记录的集合。STN 系统的组号为 L1、L2……每打开一个新的数据库，组号连续给出。

3. expand 命令

查看某检索词在某索引中是如何标引的，不能用于总索引文档检索。扩词指令主要是用来核实拼写、缩写及标引方式等，如公司名称、刊名、产品名称、作者姓名。

例如：

e apple/co（查找 apple 公司）。

4. display 命令

显示指令：显示检索的结果、检索策略、历史、所保存的检索策略及费用。

5. save temp ln 名/q（a）

暂时保存命令，免费保存 7 天，需指定组号（L 号），并自己起名字，/q 保留检索策略，/a 保留某一组号的结果。

6. act 名/q

调用保存的检索策略，在任意一个数据库中执行已保存的检索策略。

7. dup rem 命令

去重命令，用于多库检索，去掉检索结果中重复的文献。

8. log y

关机指令。

Log off：系统提示是中断还是继续；

Log h：关机后1小时内重新联机，直接进入上次关机前所停留的文档中，原检索过程还存在，用d his可显示检索历史，重新开机时需通过同一网络接通主机。

9. help命令

help messages：查看怎样使用命令和一些数据库特征的信息；

help file names：查看可检索的数据库名；

news file：该数据库最后更新的日期；

help directory：该数据库可使用的help命令；

help content：查看某数据库内容范围；

help cost：某数据库价格；

help sfields：某数据库可检字段；

help format：显示某数据库规定的打印格式；

help print：显示某命令的格式；

help commands：某一数据库中可使用的命令。

STN对所有命令、检索词不区分大小写；命令可用前3个字母或第一个字母代替；命令可一行输入，STN允许一行最多240个字符，各命令间分号隔开。

7.2.3 OCLC的FirstSearch系统

7.2.3.1 概况

OCLC（Online Computer Library Center）即联机计算机图书馆中心，创建于1967年，其前身为美国俄亥俄州大学图书馆中心（Ohio College Library Center，简称OCLC）。1981年公司更名为Online Computer Library Center，Inc。OCLC总部设在美国的俄亥俄州，是世界上最大的提供文献信息服务的非营利性组织机构之一。主要是面向图书馆，其目的是推动更多的人检索世界上的信息，实现资源共享，并减少信息费用。目前使用OCLC产品和服务的用户已有70个国家和地区的38 000多个图书馆和教育科研机构。

OCLC的FirstSearch是1991年推出的世界上使用量最大的交互式的联机信息检索服务系统，是OCLC提供的主要信息产品和信息服务。1999年8月，为满足信息检索的需求，适应高新技术的发展，增强系统的检索功能，OCLC研制出了一个全新的联机检索系统New FirstSearch，并从1999年12月开始替代FirstSearch工作。从2000年8月20日开始，New FirstSearch系统已完全替代了旧的FirstSearch系统。

New FirstSearch是一个综合的、以Web为基础的联机检索系统，比旧的

FirstSearch 系统更易于查找、获取和管理信息，界面更加友好，更加面向用户。它除了保留原 FirstSearch 的绝大多数功能外，还增加了许多新功能，实现了 OCLC FirstSearch 和 OCLC 联机电子出版物（Electronic Collections Online，简称 ECO）的完全整合，增强了对 OCLC World Cat 联机编目数据库的馆藏和 OCLC ECO 的检索，实现了各数据库的联机全文共享。New FirstSearch 系统能够使用户通过互联网直接检索到主题范畴广泛的 86 个数据库，其中包括 7 500 种期刊的文本全文。New FirstSearch 可完成对 OCLC 馆际互借系统的无缝访问，数千种印刷型和电子期刊的全文文献的跨数据库的联机显示。New FirstSearch 系统检索功能灵活多样，每个数据库都有多种检索入口，非常方便用户利用。而且，New FirstSearch 系统能在记录表中显示出用户所在图书馆的馆藏标识，为用户有效快捷地从当地获取到文献提供了方便。

New FirstSearch 的 86 个数据库绝大多数由一些美国的国家机构、联合会、研究院、图书馆和大公司等单位提供。这些数据库所涉及主题范畴包括艺术和人文学科、工商管理和经济、会议和会议录、消费者事务和人物、教育、工程和技术、普通科学、生命科学、医学和健康、社会科学、新闻和时事、公共事务、法律、人物、综合和参考等。

7.2.3.2 OCLC FirstSearch 的 12 个数据库介绍

我们可以把 OCLC FirstSearch 中的 12 个数据库分为综合性数据库和专业性数据库两种类型：

综合性数据库有：

1. ArticleFirst

ArticleFirst 数据库包括 16 000 多种学术期刊的文章及索引，主题覆盖了商业、人文学、医学、科学、技术、社会科学和通俗文化等。虽然大多数期刊是英文资料，但也收录了部分其他语言的期刊。该库覆盖了 1990 年到现在的资料，每天更新，其文献数量目前已经超过 23 000 000 篇。

2. ClasePeriodica

ClasePeriodica 提供 1978 年以来有关科学和人文领域的拉丁美洲期刊索引。由 Clase 和 Periodica 两部分组成。其中 Clase 提供社会科学和人文学科方面的文献索引，Periodica 收录科技方面的期刊。数据库提供对以西班牙文、葡萄牙文、法文和英文出版的 2 600 种（Clase：1 200 种；Periodica：1 400 种）学术期刊中的文献检索。数据库每 3 个月更新一次。

3. ECO

ECO 是一个全部带有联机全文文章的期刊数据库。它的主题范畴广泛，总

共涵盖了20个主题。目前记录来自70家出版社的5 400多种期刊。数据库中的文章都以页映像的格式（PDF、RealPage或HTML）显示，在页映像中包括了文章的全部原始内容和图像。该库收录了1995年至今的资料，每天更新，其文献数量目前已经超过3 600 000篇。

4. WorldCat

WorldCat是OCLC的一个联机的联合目录数据库，其资源来自OCLC的成员馆编目的所有记录。它目前包括6 100多万条记录，这些记录来自400多种语言的文献，覆盖了从公元1000年到现在的资料，基本上反映了世界范围内的图书馆所拥有的图书和其他资料。它的主题范畴广泛，并以每年200万条记录的速度增长。该库每天更新。

5. Ebooks

Ebooks收录了参加WorldCat联合编目的OCLC成员馆收藏的所有联机电子书，共计23万多种，其中也包括OCLC的netLibrary电子书。用户可以检索所有这些电子书的书目，并可链接到已订购且包含在WorldCat数据库中的电子书进行阅读。

专业性数据库有：

1. ERIC

ERIC是由教育资源信息中心生产的教育方面的资料来源的一个指南。它囊括了数千个教育专题，覆盖了从1966年到现在的资料，包括约1 016种期刊，100多万条记录。每月更新记录。ERIC涉及的主题主要有：成人、职业、与职业教育、信息与技术、评估、语言学与语音学、残疾与天才教育、阅读和交流、小儿与幼儿教育、师资教育、教育管理、城市教育、高等教育等。

2. GPO

GPO是有关美国政府出版物的数据库。GPO包含52万多条记录，报道了与美国政府相关的各方面的文件。这些文件的类型有：国会报告、国会意见听证会、国会辩论、国会档案、法院资料以及由美国具体实施部门，如：国防部、内政部、劳动部、总统办公室等出版发行的文件。它覆盖了从1976年7月以来的资料，每月更新记录。

3. MEDLINE

医学期刊的文章摘要的数据库。MEDLINE覆盖了所有医学领域，包括临床医学、实验医学、牙科学、护理、保健服务管理、营养学以及其他学科。它收录了国际上出版的9 580多种期刊，覆盖了从1965年到现在的资料，目前有1 500多万条记录每天更新记录。

4. PapersFirst

PapersFirst 是有关国际学术会议论文索引的数据库。该数据库包括在世界各地学术会议上发表的论文，它主要源自 1993 年 10 月以来在“大英图书馆资料提供中心”的会议录收集的每一个大会、专题讨论会、博览会、讲习班和其他会议上发表的论文，每两周更新一次。

5. Proceedings

Proceedings 是 PapersFirst 的相关库，包括 1993 年以来世界范围内的会议目录的引文。每条记录包含在一次会议上提交的论文列表。该库提供了一条检索“大英图书馆资料提供中心”的会议录的途径。该库每两周更新一次。

6. WilsonSelectPlus

H. W. Wilson 公司的全文库。该数据库是一个包括联机全文、索引和摘要的记录的集合，这些全文文章选自 H. W. Wilson 公司的普通科学文摘、人文学科文摘、读者指南文摘和 Wilson 商业文摘。它覆盖了 1 650 多种期刊的从 1994 年到现在的资料，目前该库中有超过 100 万条的记录，并且每周更新一次。

7. WorldAlmanac

WorldAlmanac 是世界年鉴，主要包括人物传记、百科全书目录、事实和统计数据，涉及的范畴包括：艺术和娱乐、新闻人物、计算机、科学和技术、经济学、体育运动、环境、税收、周年纪念日、美国的城市和州、国防、人口统计、世界上的国家等等。目前收录有 1998 年至今的 32 000 多条记录，每年更新。

7.2.3.3 检索方式

NewFirstSearch 有两种检索方式：

1. TTY（Telnet）方式

以远程登录方式登录 fscat. oclc. org，以命令式的方式进行检索。在 1996 年 2 月以前 FirstSearch 检索使用 TTY 方式，目前在 OCLC 的 WWW 服务器上仍保留此方式。该系统所有的功能、行动以及其他有关信息都通过提示显示在屏幕上，用户可以按屏幕上显示的提示成功地进行联机检索。例如，输入 s su：resource-sharing，然后按回车键。s 是 search 的缩写，su 表示主题字段，resource-sharing 是检索词。这种方式屏幕简单，而且每一屏幕信息的后面都有命令提示，缺点是不能显示图像，每次换屏需要敲入命令等。它比较适合习惯于传统联机检索的专业用户使用。

2. 万维网（WWW）方式

使用浏览器，选择不同的接入方式检索 FirstSearch 系统。万维网方式简单易用，可以显示图像、表格以及字符的上下标识等。万维网检索方式呈现的数据

库的界面分为基本检索、高级检索、专家检索。具体如下：

(1) Basic Search 基本检索。

这是一种简单检索，在检索提问框内，直接输入一个或多个检索词（或检索式），进行检索。

(2) Advanced Search 高级检索。

这是一种组合检索，允许用户在每个提问框内输入一个或多个检索词，并对每个提问框后提供的索引字段进行选择。通过这些输入和选择，系统会据此生成复杂的检索式，实现有效的检索。

(3) Expert Search 专家检索。

专家检索是为输入复杂逻辑检索式而设计的。检索式中可以有字段标识符、检索词、逻辑运算符、通配符、截词符等等。

7.2.4 其他联机检索系统

7.2.4.1 ORBIT 系统

ORBIT 是美国 Online Retrieval of Bibliographic Information Time-Share 的缩写，ORBIT 系统即文献目录信息联机分时检索系统。曾是仅次于 DIALOG 系统的世界上第二大国际联机检索系统，现拥有 100 多个数据库，其数据库类型包含石油、生化、环境、医学、运动及安全科学等学科文献。还拥有 SAE（汽车、飞机交通工具）等数据库。该系统有一小部分数据库与 DIALOG 系统相同，近年来，致力于提供一些 DIALOG 没有的数据库，在专利、能源、电子学领域的信息更为齐全。在专利方面，它常年为用户提供 WPI 和 U. S. Patent 等，又将美国专利数据库 USPA 和 USPB 合并成一个数据库 USPM，使用户避免了跨文档检索。其他商情数据库包括 ACCOUNTANTS（《会计文献索引》）、CHEMQUEST（《化工产品市场信息》）、MMA（《管理与销售学文摘》）、MICROSEARCH（《微机产品信息库》）等。ORBIT 系统提供联机检索、联机订购原文、定题检索、回溯检索和建立私人文档等服务。以每周 125 小时以上向全世界 2 万多终端用户服务，ORBIT 系统的 Web 网站的 URL 为 http：//www. questel. orbit. com。

7.2.4.2 BRS 系统

BRS 是 Bibliographic Retrieval Service 的缩写。BRS 系统创建于 1976 年，总部设在美国的拉塞姆（Latham）。最初时，只有 3 名工作人员，4 种数据库，修改了 IBM 公司研制的 STAIRS 软件，提供联机检索。1977 年初，增加为 9 种

数据库，以其廉价政策和团体签约折扣价的方式，赢得了市场。至1994年，数据库增加为160种。BRS的用户主要为生物医学界及学术团体等。1989年1月，MacMillan出版公司收购了BRS，并入其子公司InfoPro Technologies公司，1994年转售给CD PLUS公司，更名为OVID联机系统。BRS/SEARCH检索软件卖给Dataware Technologies公司，OVID系统另外采用OVID检索软件。两个检索软件略有差异。目前，BRS系统拥有数据库近200个，重点在医学、药物学和生命科学等，在工业标准和技术规范方面也拥有一批独家经营的数据库。

7.2.4.3 ESA/IRS系统

ESA/IRS为European Space Agency/Information Retrieval Service的缩写，即欧洲空间组织信息检索服务系统。该系统建于1966年，总部设在意大利首都罗马附近的费拉斯卡蒂（Frascati），是欧洲最大的联机信息检索系统，目前有数据库100多个，专业范围涉及科学、农业、卫生、管理、专利、报告、社会科学和宇航及技术科学等。它拥有的数据库中，虽有近半数与DIALOG系统相重复，14%与ORBIT重复，10%与BRS重复，25%与DATA-STAR重复，但也有自己所独有的数据库，如DATALINE（《金融数据库》）、报道英国制造业情况的INDUSTRIAL MARKET LOCATIONS（《工业市场信息》）、介绍经济和开发方面情况的INFOMAT BIS（《商业信息》）、提供欧洲国家公司财政信息的NEWSLINE/NEXTLINE（《公司金融文档》）等。ESA-IRS网站主页URL为http://www.esrin.esa.it/htdocs/esairs/esairs.html。

7.2.4.4 LEXIS-NEXIS系统

LEXIS-NEXIS联机检索系统创始于1973年，最初只是LEXIS公司，1979年NEXIS加盟，对用户提供数据库联机检索服务。经过30年的发展，LEXIS-NEXIS目前已成为成熟的联机检索和基于互联网的网络检索系统，收录有大量以法律、新闻、商业经济、政府出版物等内容为主的数据库，尤其注重搜集新闻、政府信息、法律信息及商业信息等，所提供的信息均为原始资料，具有很高的使用价值，其中政府法规法律方面的数据库是LEXIS-NEXIS的特色信息源，在法律业界具有非常大的影响力。1998年，该系统为了吸引学术性用户，从已有的各类数据库中，选出了适合大学和学术研究使用的内容，专门做了一个《学术大全数据库》（Academic Universe），内容仍以法律信息、案例、新闻、商业金融信息、政府规章制度为主，增收了医学保健信息和各类参考资料，包含有期刊、报告、政府出版物、新闻快讯等5 200余种出版物，其中约90%有全文或部分全文。

7.2.4.5 北京文献服务处的 BDSIRS 系统

北京文献服务处（Beijing Document Service，简称 BDS）1978 年由中国国防科技信息中心和北京市科协共同策划联合组建，以联机信息检索服务为其主要任务。北京文献服务处计算机信息检索系统（BDSIDS）建于 1981 年，是目前国内系统配置最大、信息量最多的现代化科技信息检索系统之一，现有各种数据库 20 多种，文献量逾 2 200 万篇，联机终端 200 多个，遍及全国 60 多个城市。数据库内容涉及自然科学的各方面，拥有世界著名的《国外专利文摘数据库》（1963— ，即《世界专利索引》）、《美国政府研究报告文摘数据库》（1963— ）、《国外期刊论文文摘数据库》（1987— ，即《科学文摘》）、《国防科技文献》、《中国经济信息数据库》、《中国化学文摘库》等。现已在“http：//bds. cetin. net. cn/”提供检索服务。提供丰富的检索功能：全文检索、字段检索、关键词检索、邻接运算、布尔查询等。

【案例】

利用 Dialog 系统进行科技项目查新实例①

检索课题：有关灌溉用橡塑多孔管的工艺技术

检索步骤如下：

1. 分析检索课题，明确检索要求，确定检索的主题内容、范围等。该课题的主要技术内容包括灌溉用橡塑多孔管也称为橡塑渗灌管，其主要原料为橡胶粉（由废旧轮胎制得）和塑料（如粉状聚乙烯）。该产品主要用于农林业、园艺等方面的农作物灌溉。

2. 确定检索概念。经过分析，其主要的检索概念是：橡胶、塑料、多孔管、灌溉。

3. 选择表达概念的检索词。

（1）选择所有能够表达检索概念的各种不同的词：

灌溉 irrigation ; watering

多孔管 porous pipes; porous tubes; porous hoses; porous piping; porous tubing; porous drip irrigation tubing

橡胶 rubbers

轮胎 tyres; tires

---

① 贾林：《利用 Dialog 系统进行科技项目查新》，载《科技情报开发与经济》，2007（8）。

塑料 plastics

聚乙烯 polyethylene; Polythene; PE; 9002-88-4

(2) 使用截词技术：

灌溉 irrigat?; watering

多孔管 porous pipe? ?; porous tube? ?; porous hose? ?; porous pip???; porous tub???; porous drip irrigat? tub???

橡胶 rubber? ?

轮胎 tyre? ? ; tire? ?

塑料 plastic? ?

聚乙烯 polyethylene; polythene; PE; 9002-88-4

(3) 使用位置算符指定词组或词间的相对位置关系：

灌溉 irrigat? ; watering

多孔管 porou (s 2w) pip??? ; porou (s 2w) tub???; porou (s 2w) hose? ?

橡胶 rubber? ?

轮胎 tyre? ? ; tire? ?

塑料 plastic? ?

聚乙烯 polyethylene; polythene; PE; 9002-88-4

(4) 使用逻辑算符组配检索概念，拟定检索表达式：

分步输入：

? S irrigat? or watering (S1)

? S porous (2w) pip??? orporou (s 2w) tub??? or porous (2w) hose?? (S2)

? S rubber? ? or tyre? ? or tire? ? (S3)

? S plastic?? orpolyethyleneorpolytheneorpeorrn=9002-88-4 (S4)

? S s1 and s2 and s3 and s4 (S5)

一步输入：

S (irrigat? or watering) and porous (2w) (pip??? or tub??? or hose? ?) and (rubber? ? or tyre? ? or tire? ?) and (plastic? ? or polyethylene or polythene or pe or rn=9002-88-4)

(5) 分析检索结果，与用户沟通并确定最后的检索策略，进入 Dialogweb 界面，输入收费账号和密码，进行正式的检索并打印结果。

关键术语

联机检索	联机检索系统	联机检索系统特点
服务方式	联机检索系统选择	联机检索系统发展
DIALOG 系统	STN 系统	OCLC FirstSearch 系统

思考题

1. 什么是联机检索系统?
2. 试述联机检索的特点。
3. 联机检索系统主要支持哪些检索功能?
4. 联机检索系统的服务方式有哪些?
5. 选择使用联机检索系统时应考虑哪些因素?
6. 简述网络环境下传统联机检索的发展。
7. 试述 DIALOG 系统及其检索。
8. 试述 STN 系统及其检索。
9. 利用 OCLC 的 FirstSearch 系统检索关于电子政务的信息。

CHAPTER EIGHT

第8章 光盘检索

【本章要点】

- ◇ 分析光盘检索系统的类型和特点
- ◇ 讨论我国光盘数据库的发展
- ◇ 介绍主要光盘数据库

引子

20世纪60年代，人们开始研究光盘存储技术；1972年，荷兰飞利浦公司首次研制成功激光唱片，1978年又推出了模拟记录激光视盘；1983年CD-ROM光盘机在日本诞生；1985年美国国会图书馆出产了第一个光盘数据库Biblifile（美国国会图书馆机读目录）；1987年著名联机检索系统DIALOG推出了其数据库的光盘版。光盘技术是集成激光、计算机、数字通信和光电集成电路等现代高新技术的结晶。光盘存储器以其高性能、多功能、多类型等突出优点，广泛应用于存储声音、图像、图表、照片、数据、文字等各种信息，成为继磁存储器之后更为新颖有效的现代化信息存储和传播工具，可以说，它的诞生和发展，在信息领域里掀起了一场革命。

8.1 光盘检索系统

8.1.1 光盘检索系统的含义

光盘检索系统即利用光盘驱动器和光盘数据库及其检索软件，结合计算机建立起来的信息检索系统。

光盘检索系统的构成包括硬件和软件两部分。硬件指计算机、光盘驱动器和光盘。计算机是检索的处理中心，光盘驱动器有单盘式和多盘式之分，单盘式驱动器只能一次放置一张光盘，多盘式驱动器可同时放置多张光盘，而且读取光盘的速度也很快，多盘式驱动器有光盘塔和光盘库等。光盘是指存储有数据的光盘数据库。软件是指检索软件，有的检索软件随光盘数据库存储在同一张光盘上，有的检索软件单独出版或发行。见图 8—1。

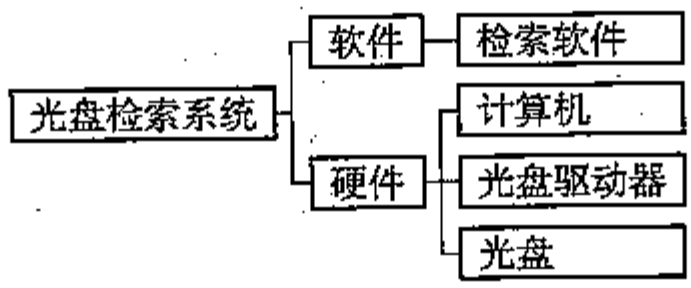

图 8—1 光盘检索系统

8.1.2 光盘检索系统的类型

光盘检索系统依据服务用户的数量，可以分为单机光盘检索系统和网络光盘检索系统。

单机光盘检索系统指由一台计算机、一个或多个光盘驱动器以及光盘数据库构成，只能供一个用户检索的光盘检索系统。在开展光盘检索的初期，使用的都是单机光盘检索系统。随着数据库产业的发展，信息资源数字化进程的加快，光盘数据库日益增多，除了信息机构所提供的单机光盘检索系统，一些个人也开始购买部分光盘数据库，在自己的计算机上使用。比如，《中国大百科全书》电子版作为一个光盘全文数据库走进了千家万户。单机光盘检索系统操作简单，使用方便，但不适合多用户使用，局限性较大。

网络光盘检索系统指能够在局域网上乃至互联网上共享的光盘检索系统，它能够同时为多个用户提供检索服务。这种光盘检索系统的出现是与网络技术发展

紧密联系在一起的。比如，同济大学图书馆的光盘检索系统就是一个非常典型的网络光盘检索系统。该系统于1995年正式建成开通。拥有世界著名的EI（《工程索引》）、NTIS（《美国政府报告》）、INSPEC（《英国科学文摘》）、AST（《应用科学全文》）、ES&PM（《环境科学与污染管理》）、ISTP（《国际科技会议录索引》）、ABI、TRANSPORT、Peterson's Gradline（《Peterson指南》）等国外权威数据库和中国专利等国内光盘数据库。经过1997、1998年的两次扩充和升级，系统现拥有美国MircoMeida CD-NETACCESS Optima 4020光盘塔服务器一台（56驱32速）和美国CYGNET Infinidisc 250C光盘库两台（每台含4驱32速，可容纳250张盘片），并于1998年9月起开始通过校园网提供免费7天×24小时光盘检索服务。

网络光盘检索系统既拥有光盘数据库数据量大、准确权威的特点，又同网络技术的优势很好地结合在了一起，能满足多个用户的信息查询需求，逐渐成为信息机构检索系统的一种重要资源。网络光盘检索系统中，光盘网络共享设备是核心，它主要包括三种：光盘库、光盘塔和光盘镜像服务器。CD-ROM光盘库（CD-ROM Jukebox）是一种带有自动换盘机构（机械手）的光盘网络共享设备，一般由放置光盘的光盘架、自动换盘机构（机械手）和CD-ROM驱动器三部分组成。光盘库一般配置有1～12台CD-ROM驱动器，可容纳50～600片CD-ROM光盘。由于自动换盘机构的换盘时间通常在秒量级，因此光盘库的访问速度较慢。CD-ROM光盘塔（CD-ROM Tower）是由多个SCSI（Small Computer System Interface）接口的CD-ROM驱动器串联而成的，光盘预先放置在CD-ROM驱动器中。用户访问光盘塔时，可以直接访问CD-ROM驱动器中的光盘，因此光盘塔的访问速度较快。光盘镜像服务器将硬盘高速缓存技术和瘦服务器技术相结合，它本身没有通用服务器那样复杂的操作系统和硬件连接，只完成光盘镜像服务器硬盘数据与客户端之间的数据传送，使客户端能以硬盘的访问速度来共享CD-ROM光盘上的信息资源。光盘镜像服务器的外观与光盘塔相似，但它一般只有一台或几台CD-ROM驱动器。网络管理员既可通过光盘镜像服务器上的CD-ROM驱动器将光盘镜像到服务器硬盘中，也可利用网络服务器或客户端上的CD-ROM驱动器将光盘从远程镜像到光盘镜像服务器硬盘中。光盘镜像服务器本身就是一台WWW服务器，客户端可通过WWW浏览器对光盘服务器直接进行远程访问和检索。光盘镜像服务器目前已开始取代光盘库和光盘塔，而成为光盘网络共享的主流产品。

8.1.3 光盘检索系统的特点

光盘检索系统与手工检索、联机检索、网络检索相比，具有以下主要特点：

(1) 光盘检索系统是一个独立的计算机检索系统，受外界的影响较小，在整个检索过程中几乎不涉及远程通信网络问题，因而，光盘检索系统没有国际联机所常有的通信线路不畅、通信费用高等不利因素，而且光盘检索的运行速度一般比网络检索速度快。

(2) 光盘检索系统具有计算机检索的强大优势，软件功能比较齐全，通常具有布尔逻辑检索、截词、字段限定、位置检索等功能，操作简单易学。此外，光盘检索还允许用户方便地将检索结果套录于某一存储介质上，形成某一专题数据库。

(3) 光盘检索系统费用是一次性投入，使用时间一般不受限制，用户可以通过反复实践，充分利用光盘检索系统提供的各种检索功能，提高检索效率，获得满意结果。

(4) 光盘检索系统相对于网络检索来说，虽然没有网络搜索引擎的数据量大，但其数据准确，查全率和查准率远胜于网络检索。

(5) 光盘数据库与联机数据库相比，存在着更新速度慢、数据容量小、专业范围窄、检索时需要不断换盘等缺点。

8.1.4 光盘检索的服务方式

光盘检索系统主要提供以下几种服务：

1. 回溯检索

提供光盘检索服务的信息服务机构所订购的光盘数据库一般是连续、定期更新的，因此可进行回溯检索。

2. 定题服务

光盘数据库一般定期更新，并且光盘检索系统能够长期存储用户的检索需求和重新执行用户的检索策略，因此，可以提供近似的定题服务。

3. 专题检索服务

专题检索服务是光盘检索系统提供的最典型的服务方式，用户可就自己感兴趣的某一专题进行连续多次的检索。

4. 套录子库

光盘检索系统提供了套录检索结果的功能，即允许用户将检索结果套录于某一存储介质上，形成某一专题的数据库，供以后使用。

8.1.5 我国光盘数据库的发展

20 世纪 70 年代末期，我国开始了光盘的研制工作。1986 年，上海市激光研

（3）检索结果显示。

浏览查询结果的标题时，在光标所指的标题上双击鼠标，可以浏览该篇文章的全文，也可以用鼠标单击标题行同时选择多篇文章，然后点击“浏览多篇内容”按钮进行浏览。在改变查询范围或更换光盘前，系统将保留最后100次的查询结果，可以通过此功能随时查看以前的查询结果。

（4）检索结果的处理。

在标题浏览界面点击存盘按钮，保存内容为检索结果的标题。如要保存全文，先在标题页点选，再显示已选文章的全文，然后存盘。点击打印，系统将打印当前显示页。

8.2.1.3 相关数据库

中国人民大学书报资料中心作为人文社科资料的信息中心，还推出了一系列其他光盘数据库产品。其中，重要的有《中国人民大学书报资料中心复印报刊资料索引光盘（A、B）》，该索引盘一套两张，汇集了1978—1997年书报资料中心精心编选的《报刊资料索引》全部题录内容，即：马列主义、哲学、社科总论类，政治、法律类，经济类，文化、教育、体育类，语言、文艺类，历史、地理类，科技、出版类，其他类等。《复印报刊资料专题目录索引》是一个题录型数据库，汇集了自1978年至今的《复印报刊资料》各刊的全部目录。它将《复印报刊资料》系列刊每年所刊登文章的目录按专题和学科体系编排，每条数据包括专题号、专题名、篇名、著者、原载报刊名称及刊期，选印在《复印报刊资料》上的刊期和页次等。用户可以通过从该数据库中检索出文章的题录信息，然后在《复印报刊资料》（纸本型）和1995年以后的光盘数据库中找到全文。该数据库按年更新。《复印报刊资料文摘数据库》汇集了1993—1999年中国人民大学书报资料中心编辑的哲学、政治、法律、经济、教育、文艺、历史、地理等方面的18类专题文摘。文献内容共25 921条记录。每条记录内容包括标题、作者、刊名、刊期，以及1 000字左右的正文文摘，提供二次检索和复合检索功能。《中国法律法规大典》收录了我国自1949年至今颁布的所有法律法规。数据库内容分为国家法、行政法、民法、刑法、经济法、国际法、诉讼法七类，78个子类。

8.2.2 《中文科技期刊篇名数据库》

8.2.2.1 概况

《中文科技期刊篇名数据库（1989— ）》由中国科技信息所重庆分所下属的维普资讯研制，收录了6 000多种国内出版的数学、经济、化学、生物、农业、

据库更新频率更快，检索功能更加全面，因而受到了欢迎。

8.2 主要光盘数据库选介

8.2.1 《复印报刊资料》全文数据库

8.2.1.1 概况

中国人民大学书报资料中心编选的《复印报刊资料》全文，以其涵盖面广，信息量大，分类科学，筛选严谨，结构合理完备，成为国内权威的社会科学、人文科学专题文献资料宝库。它具有大型、集中、系统、连续和灵活五大特点。该数据库是纸本《复印报刊资料》的电子版，《复印报刊资料》一直是我国人文社会科学方面的重要资料汇集，被公认为国家级权威刊物，具有较高的社会价值和学术价值。《复印报刊资料》特聘请著名专家、教授直接进行编辑审校工作，从根本上保证了入选文献的学术质量。同时，该刊全文收录的文章来自于全国范围内众多报刊，入选文章多是学术价值大、质量高、观点新颖、见解独到的重要文献。从1995年开始，100多个专题，每年分马列、哲学、社科总论、政治、法律一张盘，经济一张盘，文化、教育、体育一张盘，语言、文学、艺术、历史、地理及其他一张盘。从1997年开始，按季度汇集100多个专题全文于一张光盘内，全年共4张光盘，可以按专题类别提供服务。年末再将全年内容按类出版四张光盘。光盘在WINDOWS环境下全文检索，可按任意字、词、日期、人物实现秒级检索；图片嵌入正文，可放大，可缩小；文献数据可转存，可拷贝，可编辑，可打印。目前，该数据库也推出了它的Web版，更加方便利用。

8.2.1.2 检索（以年度盘为例）

《复印报刊资料》全文数据库光盘检索步骤如下：

（1）确定查询类别，选择数据库。

（2）输入检索条件，进行检索。

先在屏幕左边的小窗口选择检索字段（如作者、原文出处、标题词等），然后在输入窗口输入检索条件，点击查询按钮或敲回车键。系统根据输入的检索条件从库中取得符合条件的结果。

工具栏中的图标功能从左到右依次为选择检索数据库、浏览数据库内容、检索、二次检索、排序浏览结果、浏览检索结果、浏览多篇内容、清除多篇选择标记等。

（4）检索结果的处理。

系统允许用户对检索结果进行存盘或打印。在浏览状态下，单击视窗顶部的“套录”下拉菜单，选择“标记套录”或“全部套录”，然后在弹出的对话框里修改盘符、文件名称，然后单击保存，就可以实现对检索结果的存盘。点击视窗顶部的“打印”下拉菜单，选择“标记打印”或“全部打印”即可。

8.2.3 《中国科学引文数据库》

8.2.3.1 概况

《中国科学引文数据库》是一个集多种检索功能为一体的文献数据库。由国家自然科学基金委员会和中国科学院共同资助，中国科学院文献情报中心承建开发，该系统全面参照美国 SCI 的编制体系，是我国目前收集被引文献最多的电子出版物。该库目前已积累了 1989—1998 年的数据，共收录我国出版的重要的中英文科技期刊近 600 种，其学科范围涉及数、理、化、天、地、生、农、林、医学、工程技术等领域。《中国科学引文数据库》可查询专著、期刊论文、会议文献、专利和其他非正式出版物的被引用情况；可查询科技期刊被引情况；可查询论文发表情况；可查询专题文献。

该系统提供的数据如实地反映了来源文章的论文题名、著者、著者机构及其所在地区、受基金资助情况以及文章出处，并详细提供被引文献中，中国人在国内外发表及外国人在中国发表的文献的第一著者、被引文献名称、出版年、卷、期、页及文献类型等信息，准确报道来源文章与被引文献之间的关系。

8.2.3.2 检索

《中国科学引文数据库》光盘检索步骤如下：

（1）打开数据库。

（2）选择检索方式，进行检索。

《中国科学引文数据库》提供两种检索方式，即字典检索和命令检索。

第一，字典检索。

在检索视窗中选择字典检索，用户首先选择需要的检索字段，然后从提供的索引字典中选择一个或多个检索词。单击“显示”按钮，则在右侧的显示框内按选定的显示格式显示出查询结果。

第二，命令检索。

可以在检索视窗中直接输入检索表达式，检索表达式由字段名和检索词组成，具体写法为：字段名＝检索词。每个字段名由两个英文字母表示（见“字典

环保、地球、矿业、机械、无线电、轻工、航空、建筑、情报、医学及综合性期刊和我国港台核心期刊，累积数据200余万条。该数据库以“全”为特色，收录的科技期刊（包括港台期刊）能代表国家出版科技期刊的整体水平，而且每种期刊几乎没有人为的选择过程，均逐篇、逐期、逐年收入数据库。该数据库数据量大、数据更新及时、检索方式简单易行、检索结果准确快捷。本库提供《中图法》分类号、主题词、著者及题名检索点，并可进行逻辑组配和年代限定。数据库每季度更新一次，年均文献报道量达28万余条。

《中国科技期刊篇名数据库》是目前国内数据量最大的综合性文献数据库，收录数据数量大，覆盖领域广泛。从2000年起，该数据库将收录范围扩大为包括社会科学在内的各学科期刊，并更名为《中文科技期刊数据库》。同时，开始增加收录数据的全文，从一个题录数据库提升为全文数据库，并开始了基于万维网的检索服务，成为我国目前最重要的三大期刊全文数据库之一。

8.2.2.2 检索

《中文科技期刊篇名数据库》光盘检索步骤如下：

（1）选择光盘数据库。

打开“光盘选择”下拉菜单可对光盘数据库进行选择，有两个选择，即1989—1995年数据和1996—（最新）数据。

（2）选择检索方式，进行检索。

该数据库提供三种检索方式：字段检索、字典检索和组配检索。

字段检索：系统提供主题词、著者、分类、刊名、篇名等五种字段检索途径。用户可以根据自己的需要选择某个检索字段作为检索途径，然后输入相关的检索词，单击确定，进行检索。

字典检索：系统提供主题词、分类号、刊名等三种索引字典。用户可以选择其中的任何一种字典，在弹出的索引视窗中可键入要查询的关键词，单击查找，系统会自动在该索引字典中定位，选中需要的检索词，双击或单击“确定”键即可实现检索。

组配检索：又称复合检索。组配检索是针对已经存在的检索式进行的，所以当使用复合式进行组配检索时，事先应已有两个以上的检索式存在。

（3）检索结果的显示。

系统提供两种显示格式：完整记录和摘要列表。完整记录格式含题录信息及文摘，摘要列表格式只在一行内显示每条记录的分类号、题名、著者、出处等。单击视窗顶部的“显示”下拉菜单，选择一种显示格式，即可完成显示。

（3）检索结果显示。

浏览查询结果的标题时，在光标所指的标题上双击鼠标，可以浏览该篇文章的全文，也可以用鼠标单击标题行同时选择多篇文章，然后点击“浏览多篇内容”按钮进行浏览。在改变查询范围或更换光盘前，系统将保留最后 100 次的查询结果，可以通过此功能随时查看以前的查询结果。

（4）检索结果的处理。

在标题浏览界面点击存盘按钮，保存内容为检索结果的标题。如要保存全文，先在标题页点选，再显示已选文章的全文，然后存盘。点击打印，系统将打印当前显示页。

8.2.1.3 相关数据库

中国人民大学书报资料中心作为人文社科资料的信息中心，还推出了一系列其他光盘数据库产品。其中，重要的有《中国人民大学书报资料中心复印报刊资料索引光盘（A、B）》，该索引盘一套两张，汇集了 1978—1997 年书报资料中心精心编选的《报刊资料索引》全部题录内容，即：马列主义、哲学、社科总论类，政治、法律类，经济类，文化、教育、体育类，语言、文艺类，历史、地理类，科技、出版类，其他类等。《复印报刊资料专题目录索引》是一个题录型数据库，汇集了自 1978 年至今的《复印报刊资料》各刊的全部目录。它将《复印报刊资料》系列刊每年所刊登文章的目录按专题和学科体系编排，每条数据包括专题号、专题名、篇名、著者、原载报刊名称及刊期；选印在《复印报刊资料》上的刊期和页次等。用户可以通过从该数据库中检索出文章的题录信息，然后在《复印报刊资料》（纸本型）和 1995 年以后的光盘数据库中找到全文。该数据库按年更新。《复印报刊资料文摘数据库》汇集了 1993—1999 年中国人民大学书报资料中心编辑的哲学、政治、法律、经济、教育、文艺、历史、地理等方面的 18 类专题文摘。文献内容共 25 921 条记录。每条记录内容包括标题、作者、刊名、刊期，以及 1 000 字左右的正文文摘，提供二次检索和复合检索功能。《中国法律法规大典》收录了我国自 1949 年至今颁布的所有法律法规。数据库内容分为国家法、行政法、民法、刑法、经济法、国际法、诉讼法七类，78 个子类。

8.2.2 《中文科技期刊篇名数据库》

8.2.2.1 概况

《中文科技期刊篇名数据库（1989— ）》由中国科技信息所重庆分所下属的维普资讯研制，收录了 6 000 多种国内出版的数学、经济、化学、生物、农业、

检索”），例如，“CA＝许智宏”表示要查询被引作者是许智宏的所有文章，即查询许智宏所发表的全部文章的被引用情况，如哪一篇被引，被引多少次，都是被哪些人的哪些文章引用的。

系统可以非常方便地实现字典检索和命令检索方式的切换。使用其中任一种检索方式得到检索结果后，屏幕右侧的显示框下方都会出现另外一种检索方式的提示，单击该按钮，则自动切换到另一种检索方式下的相应字段检索。

（3）检索结果的显示和处理。

《中国科学引文数据库》提供四种显示格式，即浏览格式（包括第一著者、文献题名和记录流水号）、题录格式、综合格式（全记录格式）和引文格式（检索被引用情况时，选择该格式可对检索结果进行整理）。

单击显示框下的“保存”按钮，则弹出保存记录的对话框，选择路径，并键入文件名，保存结果将是一个纯文本文件。

8.2.4 其他光盘数据库

8.2.4.1 《中文社科报刊篇名数据库》

《中文社科报刊篇名数据库》是由文化部立项、上海图书馆承建的重大科技项目，由上海图书馆全国报刊索引编辑部负责编辑和研制，具有文献信息量大、检索点多、查检速度快等特点。本数据库收录了全国哲学社会科学期刊6 000多种，报纸200余种，学科范围涉及马列主义、毛泽东思想、邓小平理论、哲学、社会科学、政治、军事、经济、文化、教育、体育、语言文字、文学、艺术、历史、地理等各个学科。条目收录采取核心期刊全收、非核心期刊选收的原则，现年更新量约20余万条，为目前国内特大型文献数据库之一。《中文社科报刊篇名数据库》是国内比较有影响的、较为完整系统的社会科学光盘数据库，是纸本《全国报刊索引》在电子时代新的发展，检索功能比较全面，操作便捷，深受用户欢迎。目前已升级成为网络数据库。

8.2.4.2 《四库全书》数据库

《四库全书》电子版以《景印文渊阁四库全书》为底本，由上海人民出版社、香港迪志文化出版有限公司和书同文数字化技术有限公司联合开发。它分为标题检索版和全文检索版两种，每一种版本又分为网络版和单机版，全文版约181张光盘，标题版为165张数据光盘。全文版的检索功能较为完善，可以从全文、分类、书名和著者等途径进行检索。除了可以帮助用户迅速查到所需的字、词、书名、篇目或作者资料外，还可以随时跳转使用。全文版附有古今纪年换算、八

卦、六十四卦表，同时，从吉林大学出版社出版的《四库大辞典》中移用了有关条目，建立了各种资料大范围的超链接。此外，在上海人民出版社出版的《中华古汉语字典》的基础上补充了汉字的信息，制成了便捷的联机字典。该电子版采用了微软公司的 Single Binary 跨平台技术，使本产品可以在中文（简体/繁体）、英文、日文、韩文多种语言的视窗环境中运行。《文渊阁四库全书》电子版获得了电子出版物国家奖（1999）和莫比斯文化鼓励奖。

8.2.4.3 《四部丛刊》数据库

《四部丛刊》是文史工作者经常使用的一部重要典籍。该书由学者、出版家张元济先生汇集多种中国古籍经典而撰成。本数据库由北京书同文数字化技术有限公司开发，采用扫描技术，重现原书面貌，并在卷首详细记录原版宽窄大小。其制作底本采用上海涵芬楼四部丛刊，其中包括《四部丛刊》初编（1922）、续编（1932）、三编（1936）。

《四部丛刊》电子版保有纸张版本的全部内容，实现了全文检索，特征检索，择要笔记，纪元换算以及简、繁、异体汉字相互关联查询功能。《四部丛刊》的检索途径有：书名检索、著者检索、全文检索和分类检索。

《四部丛刊》电子版分为局域网络版、国际互联网络版以及单机版。《四部丛刊》电子版的光盘全套共计 24 张（不含联机字典）。

8.2.4.4 《中西文期刊联合目录数据库》

《中西文期刊联合目录数据库》是全国性的连续出版物联合目录数据库，依靠先进的信息存取和网络通信技术，同时配合一次信息服务来达到全国范围的连续出版物资源共享，全方位满足不同层次的用户需求；为全国范围的连续出版物订购协调提供依据，促进我国外文连续出版物总引入量的提高；推动图书馆工作的自动化和标准化。

该数据库报道 200 多家图书情报单位收藏的中西文期刊 4.5 万种，收录的单位遍及中国科学院全院和北京地区各大图书情报单位以及国内一些大的图书情报单位。如中国科学院文献情报中心、中国科学技术信息研究所、北京大学图书馆、清华大学图书馆、中华人民共和国国防科学技术委员会情报所、中国医学科学院图书馆、中国农业科学院图书馆等。数据库中数据的著录按照 ISBD（S）及有关国家标准和国际标准，机读格式按照 UNIMARC 格式和 CNMARC 格式。

8.2.4.5 Arts & Humanities Citation Index(A&HCI)

A&HCI 是艺术与人文科学方面期刊文献的多学科的索引光盘，它完整地收录了 25 个学科的 1 100 多种期刊，还包括 ISI 各个数据库中有关艺术与人文科学

方面的其他7 000种期刊的内容，涉及各个艺术领域，如视觉、音乐、表演、文学、工艺、历史、宗教等等，还有人文科学的各个方面，其主题范围包括考古、建筑、艺术、亚洲研究、古典著作、舞蹈、电影、历史、人文、语言学、文学、音乐、哲学、诗歌、广播、宗教、电视和戏剧等。该数据库可按被引作者、被引文献等途径进行检索。该数据库每年增加10万条新记录。

8.2.4.6 Social Science Citation Index(SSCI)

SSCI收录全球1 400种主要的社会科学期刊论文，共涉及50个学科领域，具体包括社会科学及行为科学、人类学、考古学、商业、财政、经济、教育、地理历史、图书馆学与情报学、法律、语言、政治、行销、统计、都市发展等。本数据库每年平均增加12.5万条记录，它除了能检索文章被引用的情况外，同时还可以揭示原文中所有的参考文献，并据此获得一批相关文献。因此，它是人文及社会科学研究领域的最有效并最具权威性的参考工具之一。

【案例】

《中文社科报刊篇名数据库》检索

1. 打开数据库

《中文社科报刊篇名数据库》可以单机运行或在局域网上运行。首先对该光盘数据库进行安装，在弹出的视窗中，单击视窗顶部的“数据库”，选择点击“中文社科报刊篇名（光盘）”或“中文社科报刊篇名（最新数据）”（图8—2）。

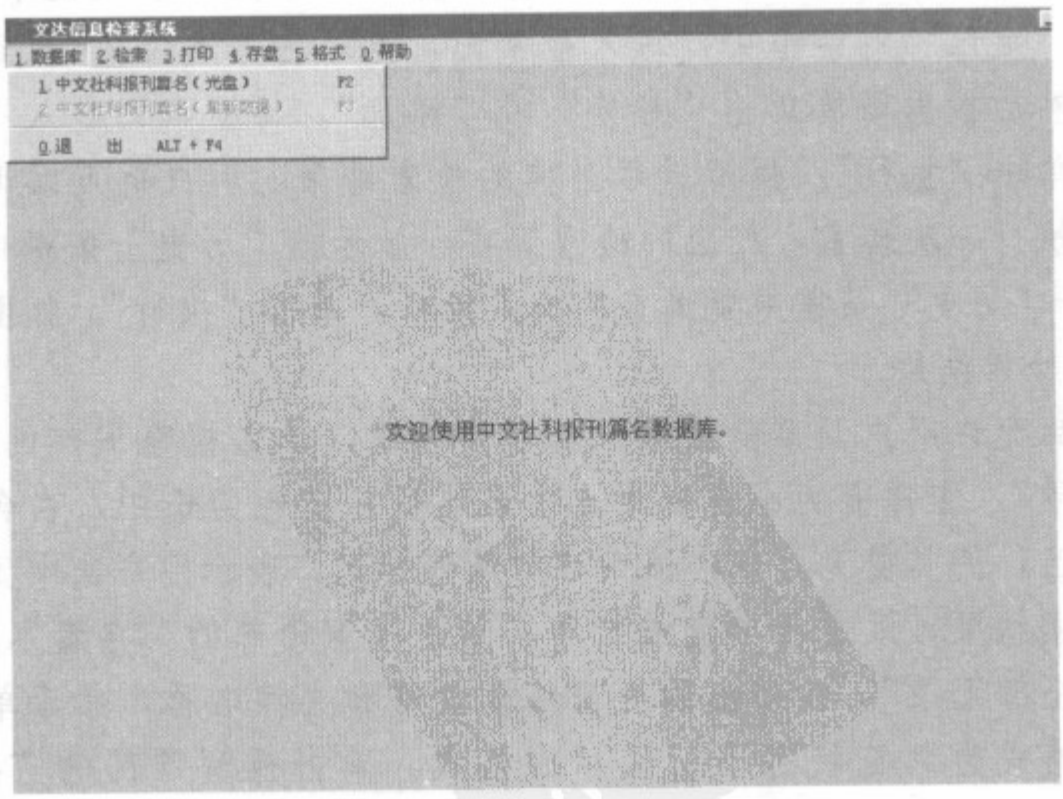

图8—2 打开数据库

2. 输入检索式，进行检索

在检索框中输入检索词或检索式，单击视窗下部的“字段”，打开下拉菜单，单击其中的字段，以选择检索范围，然后单击“执行”，视窗显示的是当前的检索结果。词与词之间的组配可选择视窗中所列的逻辑算符，如图 8—3。

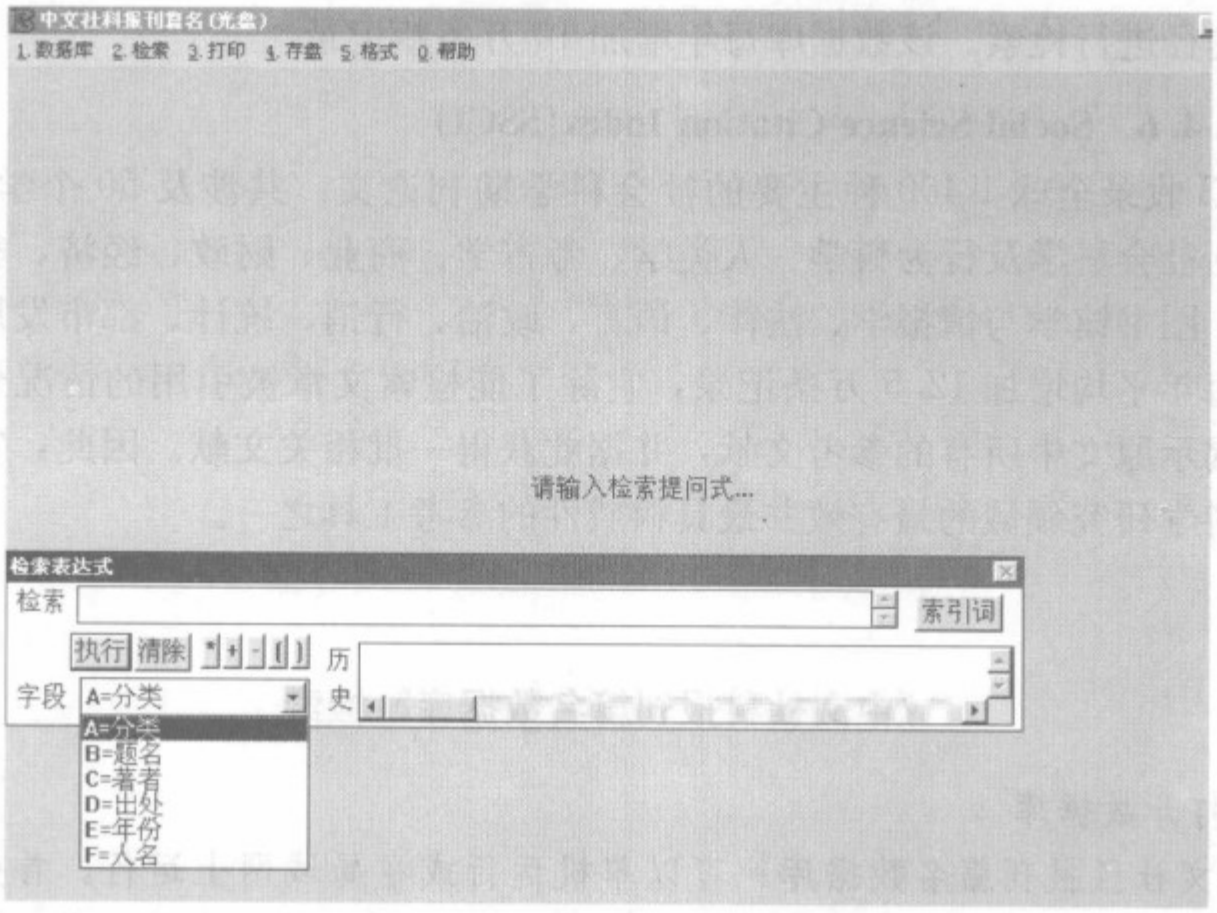

图 8—3　在对话框中输入检索式

也可以使用词库。在“索引词”视窗，打开“字段”下拉菜单，选定字段，将光标放在“定位”对话框中，写入检索词，再单击“定位”，找到所要的检索词并单击，然后根据需要单击“替换”或“添加”。待检索词进入“检索”条状对话框后，单击“执行”，视窗显示当前的检索结果。用户也可以浏览以前的检索结果。每执行一次检索，相应的检索策略会显示在“历史”条框中，单击“清除”，再双击“历史”条框中要浏览的检索策略，单击“执行”。如图 8—4。

3. 处理检索结果

该数据库允许用户调整检索结果的显示格式，单击视窗顶部的“格式”，单击“指定字段”，在弹出的小视窗中点击各字段，蓝色为选中，白色为不选。单击所要的条目，使其变为红色，该条目即做好标记，表示用户选中了该结果。

对于检索结果可以进行存盘或打印。单击视窗顶部的“存盘”，单击“指定记录”或“全部记录”，在弹出的视窗修改盘符和文件名称，然后单击“确定”。所存的记录格式为浏览状态下所显示的格式。单击视窗顶部的“打印”，单击“指定记录”或“全部记录”，单击“确定”，即可以打印检索结果。如图 8—5。

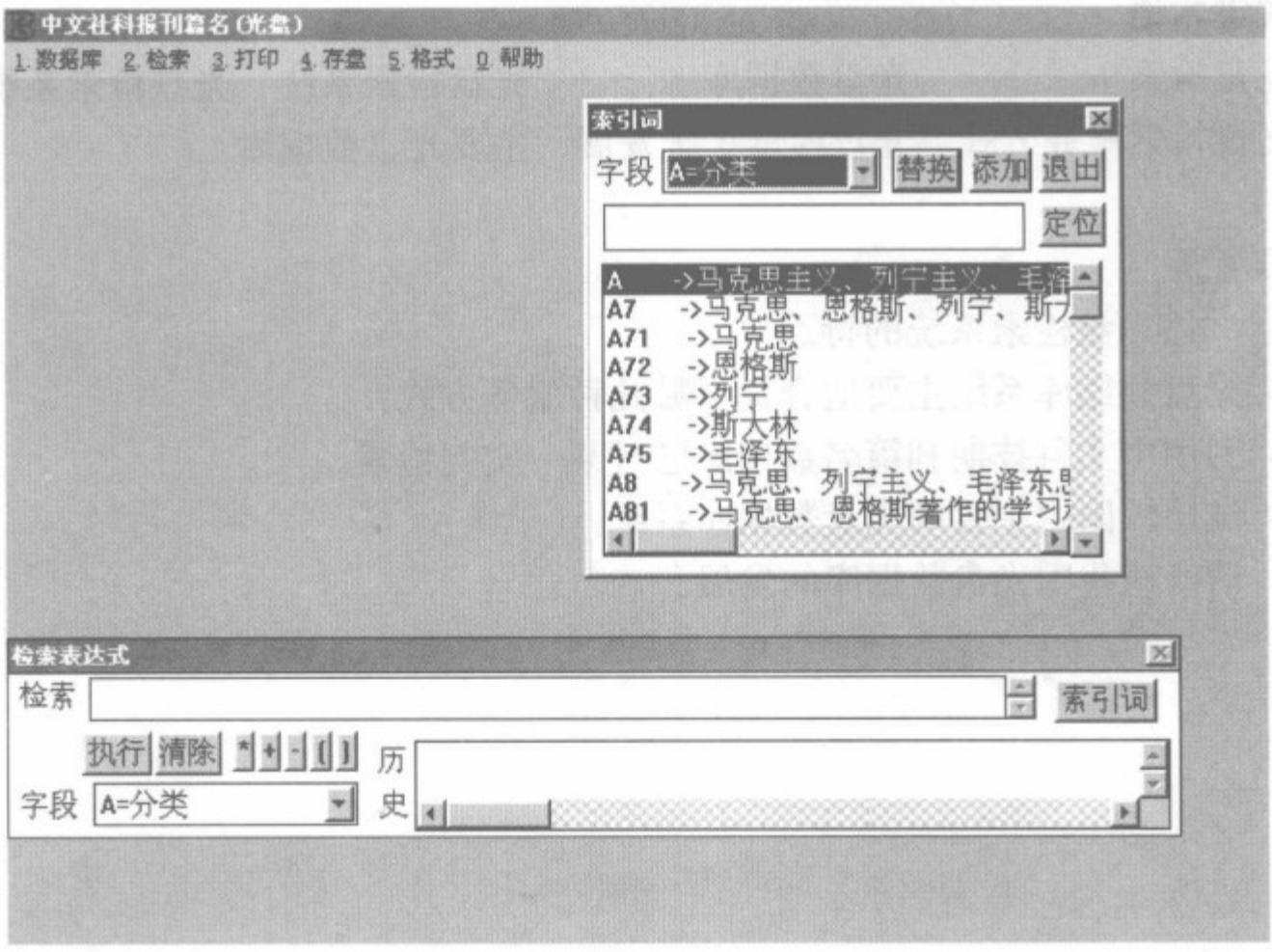

图 8—4　使用词库

图 8—5　处理检索结果

关键术语

光盘检索　　光盘数据库　　光盘检索系统　光盘检索系统类型
光盘检索服务方式　光盘检索系统发展　主要光盘数据库

思考题

1. 试述光盘检索系统的特点。
2. 光盘数据库系统主要适合提供哪几种服务方式?
3. 利用中文科技期刊篇名数据库进行某一实例检索。
4. 试比较国内主要的光盘数据库系统。
5. 请谈谈我国光盘数据库的发展。

CHAPTER NINE

第9章 网络信息检索概述

【本章要点】

◇介绍网络信息资源的概念
◇阐述网络信息资源的特点与类型
◇论述网络信息资源的分布
◇分析网络信息检索的原理及方法
◇简述网络信息检索的相关标准
◇探讨网络信息检索的发展趋势

引子

2008年1月，中国互联网信息中心《第21次中国互联网络发展状况统计报告》称：截至2007年12月，中国网民数已增至2.1亿人。网民数增长迅速，比2007年6月增加4 800万人，2007年一年则增加了7 300万人，年增长率达到53.3%，在过去一年中平均每天增加网民20万人。目前中国的网民人数略低于美国的2.15亿，位于世界第二位。网民平均上网时长是16.2小时/周，互联网已经在网民生活中占据一定的地位。互联网上的信息已是海量，搜索引擎则是网民在汪洋中搜寻信息的工具，是互联网上不可或缺的工具和基础应用之一。目前2.1亿网民中使用搜索引擎的比例是72.4%，即已有1.52亿人从搜索引擎获益，

半年净增加 3 086 万人。[①] 在当前网络环境下，网络信息检索已成为人们获取信息的最重要方式。

9.1 网络信息资源分布

9.1.1 网络信息资源的特点

目前，对于“网络信息资源”尚无统一的定义，出现了一些类似的称谓，如“电子信息资源”（Electronic Information Resources）、“互联网信息资源”（Internet Information Resources）、“联机信息”（On-Line Information）、“万维网资源”（World Wide Web Resources）等。较为流行的两种表述方式是：其一，网络信息资源就是通过计算机网络可以利用的各种信息资源的总和。其二，网络信息资源是指以数字形式记录的，以多媒体形式表达的，存储在网络和计算机磁介质、光介质以及各类通信介质上的，并通过计算机网络通信方式进行传递的信息内容的集合。

我们认为，网络信息资源（Network Information Resources）指可在互联网上发布、查询与存取利用的信息资源的总和。它包括在互联网这个平台上可以获得的一切信息资源，如数据库、电子图书、电子期刊、电子报纸和其他的网站、网页等。这些资源分布式存储在全球互联网的服务器上，与“互联网信息资源”同义。网络信息资源的最重要部分是“万维网信息资源”。

网络信息资源突破了以纸张或其他实物介质为载体的传统信息资源的限制，将大数量、多类型、多媒体、非规范的信息融合为数字化形式，在计算机或计算机网络上方便地存储、检索、处理、传递和加工，进而从根本上改变了原始信息的生产、采集和提供传递的模式，实现了信息表达和传输的质的飞跃。随着网络信息技术在全球范围内的迅速普及和应用，社会公众将从网上获取越来越多的信息资源，而实物形式的各种出版介质，比如图书、期刊、杂志以及光盘等出版介质将成为各种网络信息资源的复制品。

一般而言，网络信息资源是信息资源的一种，它与传统形式的信息资源相比，不同之处表现在其记录载体、表达形式以及传播手段等方面，其中最显著的特征是以数字化形式分布存储于网络节点中。网络信息资源分布在不同的网络节

① 参见 http://www.cnnic.cn/index/0E/00/11/index.htm，2008-02-09。

点上，可以利用现代信息技术进行加工、制作、传输、转换以及进行二次开发。与传统的信息资源一样，网络信息资源涉及人们的生产、生活以及社会生活的其他各个方面，它是随着社会的发展而不断积累起来的，同时它也显现出许多新的特点，主要表现为：

1. 数量巨大，增长迅速

没有人能确切地说清楚互联网上到底分布着多少信息资源。海量是网络信息资源的一个重要特点。互联网是一个基于 TCP/IP 协议联结各国、各机构成千上万计算机网络的通信网，是一个超级巨大的信息资源网，由于政府、机构、企业和个人都可以在网上发布信息，因此它成为无所不有的庞杂信息源。

2. 内容丰富，形式多样

网络信息资源浩如烟海，包罗万象，涵盖了几乎所有的人类社会生活领域，覆盖了不同学科、不同领域、不同语言。有科学技术领域的专业信息，也有与大众日常工作与生活息息相关的信息；有严肃主题的信息，也有体育、娱乐、旅游、消遣和奇闻轶事一类的信息；有历史档案信息，也有现实世界信息；有知识性和教育性的信息，也有消息和新闻的传媒信息。网上还有许多联机馆藏书目数据库及数千个电子图书馆。网络信息资源种类繁多，除了文本信息外，还包括了大量图像、音频、视频、软件、数据库等非文本信息，呈现出多类型、多媒体、非规范、跨地区、跨语种等特征。

3. 结构复杂，分布广泛

网络信息资源本身无统一的标准和规范，信息广泛分布在不同地区的服务器上，服务器有不同的操作系统、数据结构、字符集、处理方式等等。传统信息资源相对结构比较简单，而网络半结构化数据日趋丰富。完全结构化数据有非常良好的数据结构，如关系数据库、面向对象数据库中的数据。完全无结构数据有声音、图像文件等无模式数据。而半结构化数据是介于完全结构化数据和完全无结构数据之间的一种数据类型。半结构化数据虽然有一定的结构，但却是不严格的、多变的和不完整的。为了描述网页半结构化信息资源，产生了元数据的概念。网络信息资源分布在全球互联网的服务器上，从未有过其他任何资源能像网络资源这样有如此广泛的分布，跨越了地理空间的限制。

4. 开放互动，共享性强

开放性是互联网的特征之一，网络具有一个开放的环境，读者可以共享来自全球的各种各样的信息资源，同时可以把自己拥有的信息资源通过网络传输出去，成为网络信息资源的创建者或作者。网络信息资源具有高度共享性，这是它优于物质资源和能源资源的重要特征，同时这也使得它能在更高水平上实现有效

配置。由于信息的数据结构及存储形式具有开放性、通用性和标准化的特点，网络环境下，时间和空间范围得到了最大程度的延伸和扩展。用户不需排队等候就可以共享同一份信息资源。高度共享的网络信息资源有效地缓解了资源配置中“顾此失彼”的尴尬，使得有限的信息资源最大限度地流向网络用户。除享有版权的资料外，不受版权限制的资料，人人可自由取用，为用户创造了获取更多信息的机会。互联网的每个网页可供所有的互联网用户随时登录访问，不存在传统媒体信息由于复本数量的限制所产生的信息不能获取的现象。交互性是网络信息资源的又一特点，具体体现在它具有主动性、参与性、交谈性和操作性。在网络环境下，除了可支持数据库信息检索外，还可采用超文本或超媒体形式进行信息组织和存储，以便用户进行交互式的阅读。

5. 传播快速，利用方便

互联网提供了辐射全球范围的高速信息资源传输通道，它解决了信息传输延迟所导致的信息滞后，使信息资源能更加快捷地分配到各种应用领域中，跨越了时间和空间的限制，传播速度极快，从而实现信息的价值。用户利用网络检索工具，可以瞬间查找到自己所需要的信息资源，并立即访问该资源，把感兴趣的部分通过下载，或者以文件传输，或者以 E-mail 方式将该信息资源传送到自己的主机。同时，任何一个网络信息资源创建者也可以以相同的方式主动把相关的信息资源快速地传送给用户。通过互联网可以实现便捷的信息获取，由于网络信息资源传送与接收的便利，许多原来因为时间或距离而不易取得的资源，都因为网络的存在而成为可能。

6. 更新速度快，动态性强

变化是互联网永恒的主题，网络信息资源也充分体现了这一特色。网络信息资源本身就是一个动态系统，具有很强的时效性，更新频率很快。网络信息资源不仅增长迅速，而且变化也极为频繁。受信息时效性等因素影响，新闻、广告、Web 服务中心等总是不断更新着各自的页面内容，地址、链接信息、访问记录等也时刻处于动态变化中。网络信息具有高度动态性，任何网络信息资源都有可能在短时间内建立、更新、更换地址或者消失，这使得网上的信息资源瞬息万变。

7. 信息使用成本低

在互联网上，大部分信息资源都可免费使用，用户所要支付的主要是网络通信费用。此外，还有一些有偿信息资源，这些资源尽管目前仍然存在较大的成本降低空间及潜力，但与其他非网络信息资源相比，考虑到人力消耗和时间成本，在满足用户相同信息需求的条件下，它仍然需要一定的成本，但费用仍是低廉

的。廉价的网络信息资源有效地刺激了用户对信息的需求，从信息需要的角度也拉动了网络信息资源有效、合理的配置。

网络信息资源与传统信息资源相比，有着明显的优势，但同时也存在一些缺点。

1. 质量参差不齐，良莠不一

由于互联网是一个开放性网络，网络接入者在存储和发布信息时有很大的自由度。在互联网上，任何人都可以不受限制地自由出版、发布自己的网页，分布式存储成为网络环境中信息资源存在的主要形式。这必然导致大量冗余、粗制滥造甚至虚假的信息在网络上迅速传播、膨胀。这样就造成有价值的信息和无价值的信息混在一起，经过深度组织的高质量信息和未经任何过滤处理的低质量信息混为一体，导致了网上信息资源质量的良莠不齐，这给网络用户对有用信息的择取带来很多不便。网络信息资源中，既有有价值的高质量的学术资料或商业信息，也有无价值的劣质的甚至违法的信息，信息内容繁杂、混乱，给用户选择和利用网络资源制造了障碍。目前，互联网上还没有人开发出一种强有力的工具对信息的质量进行选择和过滤。这样，用户会发现大量毫无用途的信息混杂在检索结果中，大大降低了检索的准确性，浪费了用户的时间。

2. 分散无序，缺乏管理

传统信息资源基本上是一种比较集中的存储模式，信息资源进入到信息机构之后，经过加工整理，提供给用户使用；而网络信息资源则不同，它自由地分布在全球各地，伴着互联网的迅速扩张，网络信息资源呈几何数增长。海量的信息资源，没有任何有效管理和控制的机制，显现出明显的分散无序的特征。从宏观上看，网上的信息是无序、分散、不规范的；但从某个局部来看，如某个网站、页面、数据库的信息是有控制、有序、相对集中和相对规范的。这使得网络信息资源有序与无序并存。目前，对网络信息资源的管理主要来自两个方面：一是依赖于人工编制的主题目录，图书馆和信息专业人员通过对互联网的信息进行筛选、组织和评论，编制超文本的主题目录，这些目录虽然质量很高，但编制速度无法适应互联网的增长速度；二是依赖于自动技术，计算机人员设计开发巡视软件和检索软件，对网页进行自动搜集、加工和标引。这种方式省时、省力，加工信息的速度快、范围广，可向用户提供关键词、词组或自然语言的检索。但由于计算机软件在人工智能方面与人脑的思维还有很大差距，在检索的准确性和相关性判断上存在一定的问题。因此，对网络信息资源的控制几乎没有实质性的进展。同时，由于信息动态性和不确定性的特征，互联网上的信息地址、信息链接、信息内容处于经常性变动之中，使得信息资源的更新、消亡无法预测。在网

络上流动着的相当一部分信息资源缺乏稳定性和可靠性，不能满足信息用户安全、重复使用信息的要求。

3. 稳定性差，精确度低，缺乏安全保障

各种网络信息资源具有高度动态性，处在不断更新、淘汰的状态，可以随时发布，也可以及时变更修改，它所连接的网站、网页本身也经常处于变化之中，信息资源的更新、消亡无法预测和控制，缺乏稳定性。由于没有统一的经营管理机构，统一的发布标准，信息编排混乱，针对某一主题的查找结果往往不够精确、不够全面，不能满足信息用户安全、重复使用信息的要求。网络信息生产者并不承担保存网络信息的责任，大量信息生产出来，得不到保存，又很快消失。此外，网络信息产生和传递自由程度很高，因而必然带来诸如信息安全、网络安全等一系列问题，版权保护、隐私保护等缺乏必要的管理和法律制约措施。

9.1.2 网络信息资源的类型

网络信息资源数量庞大，内容繁杂，形式多样，广泛分布在整个网络之中，没有统一的组织管理机构，也没有统一的目录。对网络信息资源进行分类，有助于我们深入地了解网络信息资源。依据不同的划分标准，可以对网络信息资源做出如下的分类：

1. 按网络信息资源的媒体形式分，可分为文本信息、图片信息、音频信息、视频信息、三维虚拟影像信息

文本信息是最为基本的一种媒体存储形式。

图片信息指以 GIF（Graphics Interchange Formats）、JPEG（Joint Photograph Experts Group）等文件格式存储的信息。

音频信息主要指以 WAV（Wave）、AIFF（Audio Interchange File Format）、MIDI（Musical Instrument Digital Interface）、MP3（MPEG—Layer 3）等文件格式存储的信息。

视频信息主要指以 Quick Time、AVI（Audio Video Interleave）以及 MPEG（Moving Picture Expert Group）等形式存储的信息。

三维虚拟影像信息模型是以 VRML（Virtual Expert Modeling Language）组织，以立体三维形式呈现的信息。

2. 按照人类信息交流的方式划分，可以分为正式出版信息、半正式出版信息和非正式出版信息

正式出版信息指通过互联网，用户可以查询到的各种数据库、联机杂志和电子期刊、电子图书、电子报纸等。它们或是传统出版物的数字化，或是有明确创

建者，并且有版权的直接网络出版物。正式出版物具有信息质量可靠、利用率高等特点。

半正式出版信息，如从各种学术团体和教育机构、企业和商业部门、国际组织和政府机构、行业协会等单位的网址或主页上，我们可以查询从正式出版物系统所无法得到的“灰色”信息。它们受到一定的版权保护，但没有纳入正式出版的信息系统中。

非正式出版信息如电子邮件、专题讨论小组和论坛、电子会议、电子布告板新闻等。它们具有动态性强、流动性大、随意性强等特点。

3. 按照信息的加工程度划分，可以分为一次网络信息资源、二次网络信息资源和三次网络信息资源

一次网络信息资源是互联网的原始信息，包括电子图书、电子期刊、电子报纸、电子邮件、网络会议论坛、网络新闻组、企业网站（不包括虚拟的网络型商业网站，如雅虎、搜狐、新浪等）、政府网站、教育科研机构网站等等。

二次网络信息资源是对一次网络信息资源的搜集、加工和处理，主要指搜索引擎、虚拟图书馆等，是网络检索工具的重要组成部分。这类网络信息资源是用户经常利用的工具，是获取一次网络信息资源的门户和入口。

三次网络信息资源是对二次网络信息的搜集和对已搜集二次网络信息的组织，以元搜索引擎为其典型。

4. 按照网络信息的内容和用途划分，可以分为普通型、专门资料型、数据资料型和即时资料型

普通型网络信息资源，主要是反映某个组织或个人相关信息、某类学科知识或者某一方面的信息，一般不具备站内强大的搜索功能，只是通过链接来组织各种内容信息。

专门资料型网络信息资源，主要指以查检为目的，为用户提供全面内容信息的网络信息资源类型，如网络数据库、搜索引擎、专利检索网站等等，它通常具有全文检索的功能，以免费或收费的方式提供服务。这类网络信息资源是我们进行信息检索时经常利用到的信息资源。

数据资料型网络信息资源，通常是按内容、地域、时间、出版所有权或者其他分类组织起来的相关数据集合。如地区或城市介绍，工程实况及记录，企事业机构名录、指南、字典、百科全书、年鉴、手册、产品样本等参考工具等。包括一些统计数据、产品或商品的规格及价格、各种投资行情和分析等。

即时资料型网络信息资源，指在网上论坛、新闻组、留言板等上面实时产生的信息资源。这类网络信息由于发表方便，随意性较大，动态性强。

5. 按照信息的表现形式划分，可以分为全文型、数值型、书目文献型和实时活动型

全文型网络信息资源，如各种报纸、期刊文献的全文，政府出版物、专利、标准以及全文型的其他网站。

数值型网络信息资源，如主要提供统计数据、产品或商品的规格及价格的网站或网页。

书目文献型网络信息资源，如图书馆公共联机检索系统就是典型的这类资源。

实时活动型网络信息资源，如各种投资行情和分析、BBS讨论组、网上商务贸易等。

6. 按照传输协议的不同，可以分为 WWW 信息资源、Telnet 信息资源、FTP 信息资源、网络论坛和 Gopher 信息资源

WWW 信息资源指建立在超文本、超媒体技术的基础上，集文本、图形、图像、声音为一体，并以直观的图形用户界面展现和提供信息的网络资源形式。WWW 代表着互联网信息资源的主流，自 20 世纪 90 年代以来，发展极为迅速，是网络信息资源中最主要、最常见的形式。目前，各类机构纷纷建立 WWW 站点，向社会发布大量信息。我们检索和利用的大部分信息是 WWW 信息资源。WWW 信息资源具有采用超文本传输协议（HTTP）和超文本标记语言（HTML），一般使用浏览器进行浏览等特点。同时，利用 WWW 浏览器还可以轻松地访问 Usenet、FTP、Gopher 等许多其他类型的网络资源。

Telnet 信息资源指借助远程登录（Remote Login）在网络通讯协议的支持下，使自己的计算机暂时成为远程计算机的终端，而进行实时访问和使用的远程计算机中对外开放的资源。这些信息资源包括数据库资源、软件程序等。

通过 FTP 不仅可以从远程计算机上获取、下载文件，也可以将文件从本地机上传到远程计算机上。通过 FTP 可获得电子图书、电子杂志、免费软件等许多类型的信息资源。互联网上有许多提供 FTP 信息资源的服务器，且数量不断增加，检索工具 Archie 专门用于搜索 FTP 服务器的地址及 FTP 文件数据库。

网络论坛是一种最丰富、最自由、最具开放性的网络信息资源，是互联网上最受欢迎的信息交流形式，主要包括：新闻组（Usenet Newsgroups）、邮件列表（Mailing List）、电子公告牌（BBS）、专题讨论组（Discussion Group）等。新闻组是一个巨大的信息集合，它按类别细分成许多小组，每个小组集中了对某类信息感兴趣的人们，可以进行相互的直接交流。新闻组要求用户主动地从新闻服务

器上读取信息，参与讨论。利用邮件列表，许多兴趣相同的人可以互相进行交流。一旦加入了某个电子论坛，就可以收到邮件群其他成员发送的信息，也可以向该论坛发送信息。邮件列表的用户是被动地从邮箱中接收电子邮件。BBS上有许多实时的信息，是重要的网络信息资源之一。

Gopher是一种基于菜单的网络信息系统。利用Gopher服务器，通过选择菜单项，在一级级菜单的指引下，逐级进入子菜单或某一个文件进行浏览。这些文件是以树型结构进行管理的，用户可以穿梭于文件树间寻找所需信息，而不必知道它们的具体IP地址、域名等，像只灵活的信息鼠似的在网上搜寻、漫游，查询所需信息。但随着WWW的发展与普及，以及Gopher不能传送多媒体信息的缺点，Gopher处于面临淘汰的境地，有些Gopher服务器已经关闭。

9.1.3 网络信息资源的分布

随着全球信息化的不断深入，网络信息资源日益丰富。但由于网络信息资源高度分散、数字信息易拷贝等特点导致信息重复率高，无序性更为突出。所发布的信息没有统一的管理和规范，加之现有的检索工具智能化程度较低，导致信息查询困难，大大影响了网络信息资源的利用效率。因此，了解目前网络信息资源的分布情况，加快和扩大网络信息资源开发利用的进度和规模，利用信息网络共享信息资源，弥补目前普遍面临的信息资源短缺的问题，已经迫在眉睫。互联网现已成为全世界最大的信息资源库，网络信息资源可谓浩瀚无边，内容涉及各个方面。

1. 政府信息

政府信息是一切产生于政府内部，或虽然产生于政府外部但对政府活动有影响的信息资源的统称。政府信息是政府活动的原始记录和产物，它的发展与政府机构本身的扩充及其职能的强化有着不可分割的联系；其文献地位和使用价值也随着综合国力的增强和在国际事务中所发挥的作用，而不断上升和升值。网络政府信息的多少，已被视为一个国家民主程度的表征之一。由于政府总以某种方式与人们的工作和生活的每一方面直接或间接相联系，因此，其信息总量常常多得惊人，甚至达到无法计数的地步。

据统计，目前各级政府部门大约集聚了全社会信息资源总量的80%。政府网络信息资源与其他内容的网络信息资源相比，具有权威、可靠、质量高等特点，成为互联网上最重要的网络信息资源之一。政府信息内容非常丰富，包括政府公报、政策法规、政务新闻、机构设置与职责、办事规程和工作动态等等相关信息。其中，除大量正式文件外，也有一系列专业数据库，如政府信息查询的数

据库以及有关企业、产品、科研成果、市场信息等方面的数据库等，还有非常受欢迎的官方的统计数据及分析报告、调查报告、历史档案等。

在互联网上有两种基本类型的政府网站。一类是具体政府部门的网站，即基本网站，这类政府网站主要反映机构本身的信息内容和服务项目。另一类是门户网站，是一个跨机构的入口网，它是提供政府信息和服务的总窗口，这类政府网站是实现电子政务的平台。由于这两种网站的功能不尽相同，它们所发布的信息内容也有所区别。基本网站的内容主要以本政府机构的信息和服务为主，是该政府部门职能实现电子化、网络化的窗口。这类政府网站的信息内容主要包括本政府机构的办事指南、网上咨询、网上查询、网上申报、网上审批、政府网上采购等，有关政务公开的相关政策信息，可公开的资料、档案、数据库等，包括行业新闻、市场动态、产业数据、行业管理政策等。门户政府网站是一个完整的、开放的政府网站体系，它将跨机构、跨行业的政府信息收集起来，通过门户政府网站用统一的方式提供不同种类、不同层次的服务，即“一站式”服务。凭借简单、安全的访问入口，为客户提供全方位的服务和内容。

2. 教育科研信息

教育科研信息主要指各高等学校、科研机构和其他专业学术机构的相关网络信息资源。如中国教育科研网作为综合性的教育科研网络，包含了大量的教育科研信息，而且从这里可以链接到各大学的网站。大学网站的内容也相当广泛，有该校各学院、系、专业的介绍，学位、奖学金的设立，入学申请表，校历以及学校周边环境，生活设施，公共交通，还有各学科专业的教学计划、课程表，及教师的个人网页。在我国，科学院系统的网站上汇集有大量的科研信息和相关数据库，如中国社会科学院及各地社科院有关于社会科学研究的最新动态及其相关信息。其他专业学术机构如学会、协会等所设立的网站上有最新的学术会议安排、学科发展的动态信息和本机构出版的通信杂志。终身教育已经成为当代教育的重要理念，而互联网则成为人们获取知识，了解相关教育信息和科研信息的重要工具。

3. 网上出版物

网上出版物是指在网络环境中编辑、出版、发行的出版物以及印刷型出版物的网络版，主要包括网上图书、网上期刊、网上报纸等。相关技术的发展使网络出版物的数量正急剧增加，内容更是涉及方方面面，其中网上参考工具书更是独树一帜，像一些百科全书、辞典、手册、名录等都进入了互联网，这些网络版参考工具书使用起来方便、快捷。网上期刊在数量上超过网上图书，具有周期短、组织灵活的特点。互联网所具有的交互性，为编辑与作者之间，作者与读者之间

的信息交流和沟通提供了方便。目前互联网上有上万种电子期刊，其中很多是免费提供。网上报纸在近几年也得到了迅速发展，以美国为例，美国共出版 2 200 种左右的日报，据报道，在网上设站的报纸已占到美国全部报纸的一半以上。

4. 网络数据库

网络数据库是网络信息资源中数据质量最高、学术性最强的信息资源，是学术性用户使用最为频繁的网络信息资源。网络上的数据库分为收费数据库和免费数据库两种类型，网络数据库有全文型、文摘型、题录型、事实和数值型、多媒体型等。全文商业数据库大都是收费的，需要通过购买或用户授权才能使用，现在高校图书馆使用的很多数据库就是这种情形。像一些国际上著名的数据库，如：UnCover 数据库，是美国 Carl 公司生产的世界上最大的期刊数据库之一，包括自然科学和社会科学几千种期刊，更新及时，可查最新的期刊文献，并提供原始文献传递服务；UMI 网络数据库，由美国著名的数据库公司 UMI 制作，包括博士、硕士论文数据库、学术期刊图书馆等。在互联网上，也有许多免费的数据库，像一些商业性数据库的题录库、图书馆自建的数据库等，如：中国期刊网的题录数据库、各个图书馆的公共检索服务系统。

5. 电子论坛和电子会议

互联网上设有 USENET 及 Listserv 电子论坛，也称新闻讨论小组。USENET 及 Listserv 都是由成千上万个专题讨论小组构成。每个小组是由某一主题参与的文章所构成。USENET 与 Listserv 类似，但是，前者范围更广泛，几乎无所不包，一般不需订购便可参与；而后者较为严肃，而且更趋学术性，通常还需订购方可参与。互联网上用户通过 E-mail 均可自由参与电子论坛的活动，从中可以获得用任何其他手段都难获得的第一手重要专题信息与资料。因此，它是研究人员及时了解跟踪学科动态与前沿的最有效途径之一，可消除印刷出版物时间滞后的缺点。电子论坛的另一功能是举办国际电子会议。例如，利用 Listserv，一些学术团体与组织已成功举办了多次专业性的国际学术会议。参加这些会议，能了解本专业的最新研究发展动态，获取完整的会议论文与资料。

6. 网上专利信息

网上的专利信息资源主要分布在：(1) 联机检索系统中的专利数据库。一些知名的联机检索系统中都包含与专利有关的数据库。如 DIALOG 系统（http://www.dialog.com）和 STN（www.cas.org/stn.html）系统等。(2) 专利管理机构网站提供的信息。专利管理机构网站主要是指各国（地区）或地方专利局的主页或者由它们及其下属机构开发的网站。这类网站提供的专利信息全面、权威、新颖。例如美国、日本、加拿大等国的专利数据库在 Internet 上均可得到免费使

用。(3) 数据库出版机构提供的信息。主要有 DERWENT 公司、英国 IEE 公司(INSPEC)。

互联网上还有大量的会议信息、学位论文、技术标准、科技政策法规、产品样本目录、科技报告、统计数据、电子论坛、科技新闻、组织机构、通讯讨论组和数据库等，这些信息共同汇成了网络信息资源宝库，其分布特征主要体现在以下几个方面。

1. 离散性

网络信息资源的类型非常庞杂，既有各种类型的数据库、软件资源，也有丰富多彩的电子出版物及动态信息。信息发布内容具有很大的自由性和任意性，由于缺乏必要的过滤、质量控制和管理机制，不仅学术信息、商业信息、政府信息、个人信息混为一体，而且大量不健康信息也得以扩散，引发了许多方面的问题。这些显示了网络信息资源的分散性和无序性，使用户面对眼花缭乱的信息无所适从，不知道如何寻找自己需要的信息资源。

2. 不均衡性

不均衡性主要表现在地区分布和语言分布等方面。信息资源分布基本上反映出该地区经济、文化等方面的发展水平，从全球范围来说，以美国为首的西方发达国家发展水平大大高于发展中国家，而从我国的情况来看，东部地区的发展水平高于西部地区。网络上中文信息的贫乏，也已成为一个现实问题。占世界人口20%的发达国家拥有全世界信息量的 80%，拥有大量信息资源的国家将具有更多的竞争优势，将会更发达、更富有，并在国际事务中占据主导地位。在国际数据库市场上，美国仍然是世界上最大的数据库生产国和输出国，其产品在本国和国际市场上占据主导地位。欧洲的数据库的数量增长很快。其中英国和荷兰的增长尤其突出。韩国的数据库居亚洲榜首。

9.2 网络信息检索原理与方法

9.2.1 网络信息检索原理

9.2.1.1 网络信息检索及其特点

在网络世界这个浩瀚、动荡的信息海洋中，如何能够准确、及时、有效地获取自己所需要的信息，对于我们每一个用户来说都十分重要。从信息检索的发展历程可以看出，无论是联机检索，还是光盘检索，都将转向互联网这个网络平

台，成为网络检索的组成部分。计算机检索也将步入网络检索这样一个新的发展时期，网络信息检索（Networked Information Retrieval，简称 NIR）将代表网络时代获取信息的重要方向。

网络信息检索指通过一定的方法，从已存储的网络信息中查找与用户提问相关的信息的过程。它是计算机检索的发展和延伸，是一种基于互联网的新型的信息检索方式。网络信息检索是对传统信息检索的重大变革，尤其是万维网的出现，打破了传统的线性信息组织方式，创立了超文本超媒体的信息组织方式。网络信息检索与传统信息检索相比，呈现出新的特点。

1. 检索的对象得到了极大的丰富

传统信息检索的主体是文献检索，其中以纸本的图书、期刊、报纸、学术论文、会议文献等为核心。在网络环境下，信息资源组成体系发生了变化，网络资源在内容和形式上均较传统的资源丰富了许多。信息量更大，信息形式更加多样，不仅包括目录、索引和全文等文本型信息，还包括声音、图像、影像等多媒体信息。在单机环境下，由于受硬件资源的影响，文档数据库的数据量受到一定限制，随着互联网，特别是 Web 服务器的出现，可共享的网络信息资源越来越多，如何从数量庞大的文档中查找到每个用户所关心的信息，成为信息共享的关键。联机公共检索目录（OPAC）的发展，使用户可以便捷地查询网上的目录，如，国家图书馆的联机公共目录为我们提供了快速、全面获取书目信息的入口。网络数据库突飞猛进的发展，为我们提供了大量的电子期刊、电子报纸、学术论文等资源，如，《中国期刊网》上的《中国期刊全文数据库》收录了目前国内主要的中文期刊，是目前国内最重要的期刊数据库之一。同时，许多传统的媒体纷纷在网上发行电子版，如《人民日报》、《光明日报》、《中国日报》等各大报纸均设立了自己的网站，为读者提供全天 24 小时的服务。而且，传统的信息检索系统几乎都是基于单语言环境，而网络信息检索面对的是不同的信息资源，互联网信息检索使用不同的自然语言描述各种信息，形成了不同语种的信息检索系统。网上有许多检索工具支持多语种检索，如 Google 等，这也是网络信息检索一个明显的特点。

2. 检索的空间得到了极大的扩展

互联网的发展将全球连在了一起，也将全球的资源汇集成了一个大的资源宝库，网络信息资源的检索将面对全球的资源。传统的信息检索在很大程度上受到了地域空间的限制，信息用户获取相关信息，主要的渠道是图书馆或其他的信息机构，能够得到的资料也仅是该馆图书馆或该机构所存储的信息，虽然有馆际互借，但仅占到很小的比例。现代网络信息检索冲破了传统的空间的局限性，大大

扩展了检索空间。它可以检索互联网上的各类资源，而检索者不必预先知道某种资源的具体地址。其检索范围覆盖了整个互联网这一全球性的网络，为访问和获取广泛分布在世界各地的、成千上万台服务器和主机上的大量信息提供了可能。这一优势是其他任何信息检索方式所不具备的。

3. 检索趋于简单方便

网络信息检索一改以往的信息检索专业性较强的特点，以简单方便的检索方式赢得了广大用户的欢迎。万维网的超文本超媒体技术为用户提供了超链接的浏览方式，用户可以采用直接浏览的方式，获取自己所需要的信息。超文本与用字符串来表达、以线性形式进行组织的传统文本信息的处理有较大的不同。它不是以字符，而是以节点为单位组织各种信息，一个节点是一个“信息块”。节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，节点间通过关系链加以链接，构成表达特定内容的信息网络。它对信息的存储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按顺序线性存取的限制，可以方便灵活地检索信息，表现出较强的关联性。

网络信息检索在用户检索界面、检索结果提供方式等方面都体现了良好的交互性，具有较好的信息反馈功能。用户可以根据自己的需要，方便地调整检索提问，直至得到满意的检索结果。友好的用户界面对用户屏蔽了各局部网络间的物理差异，使用户在使用这些服务时感到明显的系统透明度。用户使用自己所熟悉的方式输入查询提问，就可以实现对网络信息的检索。

此外，自然语言在网络检索中的广泛使用，使得网络检索变得日趋简洁。关键词检索在网络信息检索中的普遍应用，智能信息技术的发展，使得用户的网络信息检索过程变得轻松、随意，无需考虑繁琐的检索规则。同时，与之相关的检索交互性也进一步提高。

网络检索虽然具有以上所提到的诸多优势，但与其他类型的计算机检索形式相比，也存在一些不足，主要表现在以下几个方面：

1. 信息查准率比较低

网络用户表达的需求与获取的检索结果往往相差很大，尤其是学术性信息的查询。尽管不同的搜索引擎涵盖范围不同，检索结果不同，但真正符合用户需要的信息却不多。

2. 检索带有一定的盲目性

超文本一方面使得网络检索独具特色，利用方便，另一方面，也引起了一定的负效应。网络信息以超文本链接，用户从一个检索点入口，整个搜索过程几乎由网络的超链接控制，处于一种失控、无方向的状态。用户信息需求检索的主动

性变为被动性，有“被别人牵着鼻子走”的感觉。一旦进入链接的“死区”（链接点的历史变动或网路堵塞），就会影响检索效率。

3. 各种检索工具的检索方法不统一，造成了用户使用的不便

各种网络检索工具使用的检索符号和检索方式不统一，在检索式的组成上，不同的检索工具也有不同的要求，如，个别检索工具要求用户在写检索的主题时尽可能详细，但有的检索工具则要求用户尽可能以简短的词表示查询主题，有些检索工具要求用户将人名和专有名词都大写，有些则大小写都可以，这也给用户进行网络检索带来了麻烦。

9.2.1.2 网络信息资源检索的原理

当前的网络信息通信多采用客户端/服务器结构。在这种网络通信结构下，用户首先向客户端的应用程序发出数据请求，接着应用程序通过客户端跨越网络向相应的网络服务器传递有关数据请求。网络服务器在接到有关请求后从相应的数据库或其他存储介质中获得有关数据，再把其数据返回到客户端，最后通过相应的用户界面应用程序把有关结果以特定形式呈现给用户。如今通常使用网络浏览器作为网络信息检索客户端工具，它提供良好的用户界面，同时作为通用的基于万维网协议 HTTP 的网络客户端，如图 9—1 所示。

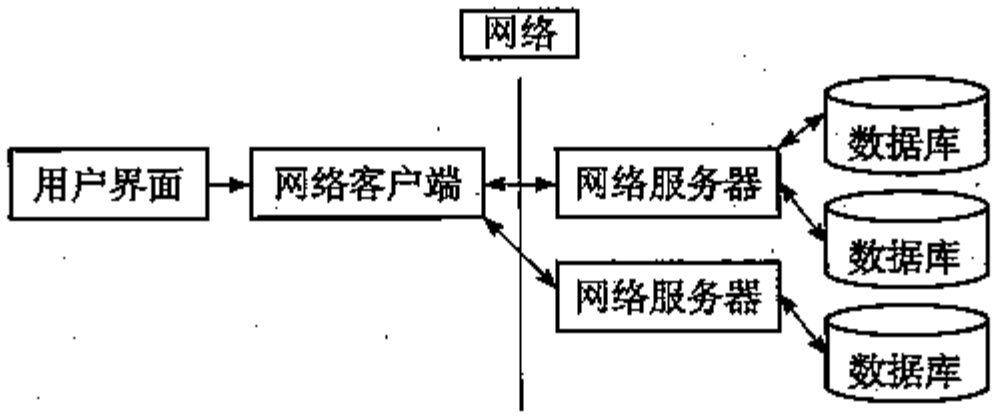

图 9—1 网络信息检索的基本模型

在网页检索的过程中，一般分为两级模式，第一级是通过关键字找到与该关键字相关的网站地址，第二级是在某个具体的网站中通过关键字找到与该关键字相关的网页。目前网络信息检索模型多采用布尔检索模型和向量空间模型。

9.2.2 网络信息检索方法

在互联网上查找信息，主要是要找到提供信息源的服务器。也就是说，首先以找到各个服务器在网上的地址 URL 为目标，然后通过该地址去访问服务器提供的信息。在网上检索信息资源的基本方法如下：

1. 直接浏览——网址查询

如果用户要访问已知地址的信息资源，可以在浏览器地址栏中输入已知的网站或网页地址，直接进行浏览，这是一种最常见最有效的信息资源的获取方式。网络信息资源的用户大都有自己侧重的研究领域或喜爱的主题，会有意识地积累一些与此相关的网址。用户可以充分利用浏览器中的收藏夹功能，保存和管理浏览过的感兴趣的网站或网页，也可以通过与他人的交流征询获取相关的网址。同时，目前在一些刊物上有一些专门介绍某些专业网络资源的文章，也可供我们参考使用。该方法有些类似传统环境下的资料索引收集工作。能否有效地采用这种方法，关键在于用户平时是否能多渠道地收集相关网址。

直接按网址进行查询的方法简单高效，但获得的网络信息资源仅是沧海一粟，能够通过这种方法获得的信息非常有限。

2. 利用网络资源目录

在互联网上存在着一些特殊的门户网站，其主要目的是收集和整理网上提供的各类信息。这些网站是由信息管理专业人员在广泛搜集网络资源并进行加工整理的基础上，按照某种主题分类体系编制的一种可供浏览的等级结构式目录。在每个类目及子类下提供相应的网络资源地址，并给以简单的描述，使用户在目录体系的导引下，逐层点击，直至发现有关的信息。网络资源目录可以分为两大类：一类是网络检索工具的分类目录，如著名的雅虎目录、搜狐目录等等；另一类是各个信息机构整理编制的信息导航，如 CALIS（全国高等教育文献保障系统）的重点学科导航库（http：//www. calis. edu. cn）。

网上除了有许多综合性目录指南外，还有很多专题性的信息资源指南，几乎每一个学科专业、研究领域或热门主题的网络资源指南都可以在互联网上找到。用户可以利用目录浏览，比较全面地获取某个学科或某个主题的网上资源。这种方法比较适合对宽泛性主题的信息进行查询。

这种信息查询方式简便易行，用户不需要经过专门的培训，就可以轻松地利用万维网的超文本技术浏览和获取网上的信息资源，但需要用户平时注重对资源目录的熟悉和了解，其中，包括综合性的目录指南和专业性的资源导航，只有这样，才能有效地利用浏览方式获取丰富的信息资源。

3. 利用以关键词检索为主的搜索引擎

这是获取网络信息资源最为常规和普遍的方式。搜索引擎作为主要的网络检索工具，在网络信息资源的检索中具有重要的地位。搜索引擎使用自动索引软件来发现、收集并标引网页，建立数据库；以 Web 形式提供给用户一个检索界面，供用户输入关键词、词组或短语等检索项；代替用户在数据库中查找出与检索提

问匹配的记录，然后返回结果，并按相关度排序输出。搜索引擎突出的是检索功能。利用搜索引擎进行检索省时省力，简单方便，检索速度快、范围广，能及时获取新增信息。

利用搜索引擎查询获取网络信息资源，比较适合用户从关键词的途径获取信息。但在使用过程中，尤其是利用英文搜索引擎进行查找时，应注意掌握一些检索技巧，如认真分析课题、多键入相关的关键词、使用词组检索和字段限定检索等。

4. 查询网络文献数据库

访问网络数据库是用户获取学术性信息的最有效的方法，尤其是全文数据库的迅速发展，为用户直接获取原始文献提供了便捷的途径。

利用网络数据库查检文献应注重两点：一是要获知网络数据库的登录入口；二是要掌握一些网络数据库的检索技巧。对于免费的数据库，用户可以按其网址直接登录，或通过搜索引擎的搜索找到它的服务器地址；对于收费数据库来说，多为单位购买，然后提供给本局域网上的用户使用，如一些大学的图书馆购买了多种有价值的网络数据库，用户可以在所属校园网的任何一台主机上访问这些数据库，而该校园网外的用户则不能使用这些数据库。网络数据库检索功能较为完善，大都支持布尔检索、字段限定检索、截词检索、位置检索和自然语言检索。在网络数据库的检索过程中，能否构造一个完整的检索式对于提高检索结果的查准率和查全率都有很大影响。这就要求信息机构加大用户培训的力度，使花巨资购买的网络数据库资源得到充分的利用，满足用户查询网络数据库的信息需求。

5. 查询网上图书馆

网上图书馆是查询网络信息资源的又一重要的途径。对于信息资源的用户而言，网上图书馆的主要资源有三种类型：一是联机公共检索目录（OPAC），如我国国家图书馆的联机公共检索目录（http：//opac. nlc. gov. cn/search. htm）。用户通过 OPAC 可以获知相关的书目信息，也可以在网上预约借书。二是图书馆工作人员通过收集整理组织网上资源而形成的学科信息资源导航，如各高校图书馆的信息资源导航。网上图书馆的学科信息导航不同于检索工具的分类目录，它以学科分类为基础，以搜集学术性信息为宗旨，经过人工筛选重新组织形成。这类导航一般具有较强的针对性，主要目的是服务于本校的教学和科研。如北京大学图书馆的“互联网学术资源学科分类导航”[①]，目前建立了哲学、数学、图书馆学情报学、工商管理、历史学、环境科学等学科资源导航。三是图书馆购买

① http：//www. lib. pku. edu. cn/zixun/zixun _ xkdh. htm，2007－10－08。

的各种商业数据库（包括网络数据库和光盘数据库）。网上图书馆是我们利用商业数据库的入口，图书馆提供利用电子资源的各种培训。如北京大学图书馆目前拥有 200 多个数据库、近 2 万种电子期刊和 10 余万种电子图书为读者提供服务。而且，图书馆还对购买的主要网络数据库的收录情况和使用方法做了相关的介绍。长期以来，图书馆被称为文献信息中心，在互联网上，图书馆依然是网络信息的集散地，用户可以充分利用网上图书馆获取自己所需的信息资源。

上面提到的 5 种检索方法，可以归结为两种检索模式，即超文本的浏览模式（browse）和关键词的查找模式（search）。超文本的浏览模式以知识分类为基础，是网络资源目录的具体使用，可以满足用户的族性检索需要。使用浏览模式进行检索时，用户只需以一个节点作为检索入口，根据节点中文本的内容了解嵌入其中的热链指向的主题，然后选择自己感兴趣的节点进一步搜索。在搜索过程中，用户会发现许多相关节点内容根本没有预料到，而是在浏览过程中不断地涌现出来。同时，对信息的搜索可以深入到某一精确主题的信息单元。例如：现在的检索对象是一篇文献，而用户真正感兴趣的只是该文献内集中讨论某一现象的一小段。在互联网上，只要有“热链”连到该段信息，用户就可以从其他节点跳跃到该信息单元进行阅读。关键词的查找模式是网络信息检索中最常用的方法，在很大程度上，主要是针对用户的特性检索需要，即输入检索词以及各检索词之间的逻辑关系，然后检索软件根据输入信息在索引库中搜索，获得检索结果（在互联网上是一系列节点地址）并输出给用户。不同的检索服务可能有不同的界面，不同的侧重内容，但有一点是共同的，就是其庞大的索引数据库，它们收集了互联网上数百万乃至数千万主页信息，包括该主页的主题、地址、包含于其中的被链文档主题，以及每个文档中出现的单词频率、位置等等。

9.3 网络信息检索相关标准

9.3.1 网络信息检索标准 Z39.50

在现实网络信息环境中，我们往往遇到的是用不同元数据描述的多个异构资源系统组成的开放型资源体系，需要有效的基于分布式系统的方法实现跨元数据格式和跨系统的透明检索。这类透明检索的实施主要是基于公共检索协议，例如 Z39.50 协议。

Z39.50 是一种 Client/Server 体系结构下描述客户端检索服务器上数据以及

获得检索结果的数据结构与交互规则的协议，是网络中的应用层协议，定义了客户端与服务器之间数据交换标准。它是美国信息检索方面的国家标准，它的正式名称是 ANSI/NISOZ39.50—1995，全称为“Information Retrieval: Application Service Definition and Protocol Specification for Open System Interconnection”（信息检索：开放系统互联的应用服务定义与协议说明）。

9.3.1.1 Z39.50 概况

Z39.50 起源于 20 世纪 70 年代美国国会图书馆、OCLC（Online Computer Library Center）、研究图书馆信息网络（Research Libraries Information Network）等的书目数据库系统互联项目。在美国国家信息标准化组织（National Information Standards Organization，简称 NISO）的支持下开始了这一方面的标准化研究工作，于 1988 年推出第一个版本 Z39.50—1988，即图书馆应用系统信息检索服务定义与协议规范（Information Retrieval Service Definition and Protocol Specifications for Library Applications）。美国国会图书馆 WWW/Z39.50 网关上对其定义为：“1988 年由美国信息标准化组织（NISO）通过的一个关于计算机与计算机信息检索协议的国家标准。它可以使一个系统中的用户在不知道其他系统的检索语法的情况下，从其他支持 Z39.50 的计算机系统中查找和检索信息。”① 1989 年美国国会图书馆被指定为 Z39.50 标准的维护机构，并逐渐形成了一个非官方的组织 ZIG（Z39.50 Implementors Group）来管理和协调 Z39.50 的发展。经过各方的努力，1992 年出版了第二版：Z39.50—1992。该版是基于 ISO/OSI 标准的网络协议框架（即七层协议，分别为物理层、数据链路层、网络层、传输层、对话层、表示层和应用层），而 ISO/OSI 是一个理论上成功的标准，并没有得到广泛的应用。事实上的网络互联标准是 TCP/IP 协议。因此在推出 Z39.50—1992 后，ZIP 开始研究 Z39.50 第三版，并于 1995 年正式推出，该版功能强大，涵盖了很多新功能，与第二版兼容，并且支持 TCP/IP 协议。目前第四版正在酝酿中。在 1991 年，Z39.50 被国际标准化组织接纳为国际标准 ISO23950，成为一个世界范围的信息检索标准。

9.3.1.2 Z39.50 的内容及特点

Z39.50 是一种基于网络的信息检索标准，主要包括两部分：一部分是信息检索服务的定义，定义了信息检索服务的 11 种机制，包括 Z39.50 协议支持的服务功能说明和服务参数说明；另一部分是 Z39.50 协议的规范，包括协议控制信

---

① 许虹、罗中兴：《Z39.50 远程信息检索标准》，载《情报杂志》，2001（5）。

息定义、信息交换规则和实现协议必备的条件。Z39.50 是国际通用的信息检索协议，已是一个相当成熟的标准。

Z39.50 协议是一种网络协议，它由控制和管理计算机之间通信过程中所涉及的格式和进程的规则所组成，具有以下特点：

（1）与 HTTP、Gopher 等面向传输层的协议不同，Z39.50 是基于会话层的协议，是有状态的。对于面向会话层的协议，当客户端连上一个服务器后，就会建立一个固定的会话，连接在会话完成前不会关闭，前面会话时交换的信息可以被后面的会话所使用。有状态的协议比无状态的协议效率高，因为前者为客户端和服务器之间提供了保持连接的会话机制，而后者要求客户端和服务器之间每次传递消息都要重新建立连接；而且前者允许客户端和服务器约定通信行为（如所需的服务类型），并将这种约定贯穿于整个会话过程，而在无状态的协议中，来自客户端的消息中含有大量对于服务器行为方式的描述，而这种信息在每次传输中会重复出现。

（2）Z39.50 是一种开放网络平台上的应用层协议，利用它可以使不同计算机系统之间实现协同工作。由于它采用 ASN.1（Syntax Notation I，抽象语法标记 1）规则描述协议数据单元，采用 BER（ISO8825，Basic Encoding Rules，基本编码规则）编码规则对 ASN.1 描述的协议数据单元进行编码，因而它支持计算机使用一种标准的、相互可以理解的方式进行通信，并支持不同数据结构、内容、格式的系统之间的数据传输，实现异构平台异构系统之间的互联与查询。

（3）Z39.50 支持分布式 Client/Server（客户端/服务器）模式。当客户端向服务器提交一个检索请求时，服务器在一个或多个数据库里进行检索，并将命中记录返回给客户端。在 Z39.50 协议里，客户端的主要工作是初始化一个查询，发出一个查询请求并要求服务器端做出相应的回答。服务器是远程数据库的一个接口，其主要工作是对客户端的请求做出相应的回应，如对查询请求作出应答，或提供所需的查询记录。在 Z39.50 协议中，客户端为源端，服务器端称为目的端。

（4）Z39.50 既可以采用同步方式，又可以采用异步方式进行通信。采用同步通信方式时，一端发出消息后，就等待另一端作出响应。在异步通信方式下，Z39.50 客户端可以向服务器发出多个请求，服务器既可以处理多个请求，也可以用自己的请求中断这些客户端请求。

9.3.1.3 Z39.50 的运行机制及实现模型

Z39.50 协议是面向连接的应用层协议，它描述了两个信息检索服务系统之间的交互，客户端和服务器分别被称为源端和目的端，源端和目的端的交互是在一个会话里进行的，称为 Z 连接。源端发起 Z 连接并在 Z 连接过程中发起

操作，目的端则接受Z—连接并结束相应的操作，在一个Z—连接中，可能有多个连续的、并行的操作，源端和目的端的角色不能互换，一个Z—连接不能重新开始，就是说，一旦一个Z—连接终结，除了明确特别保留的信息以外，状态信息不会保留。

一个Z—连接主要包括初始化请求（Init Request）、初始化响应（Init Response）、查询请求（Search Request）、查询响应（Search Response）、提交请求（Present Request）、提交响应（Present Response）。在Z39.50协议中，消息的发送和接收以应用协议数据单元（APDU）进行，协议的所有功能和服务由一系列APDU加以描述，不同的APDU完成不同的功能。在Z39.50协议中说明了所有APDU的抽象语法，APDU内容的定义通过ASN.1规则描述协议数据单元，APDU通过BER完成转换，形成与机器无关的字节流。

基于Z39.50协议的信息检索基本过程，如图9—2所示。

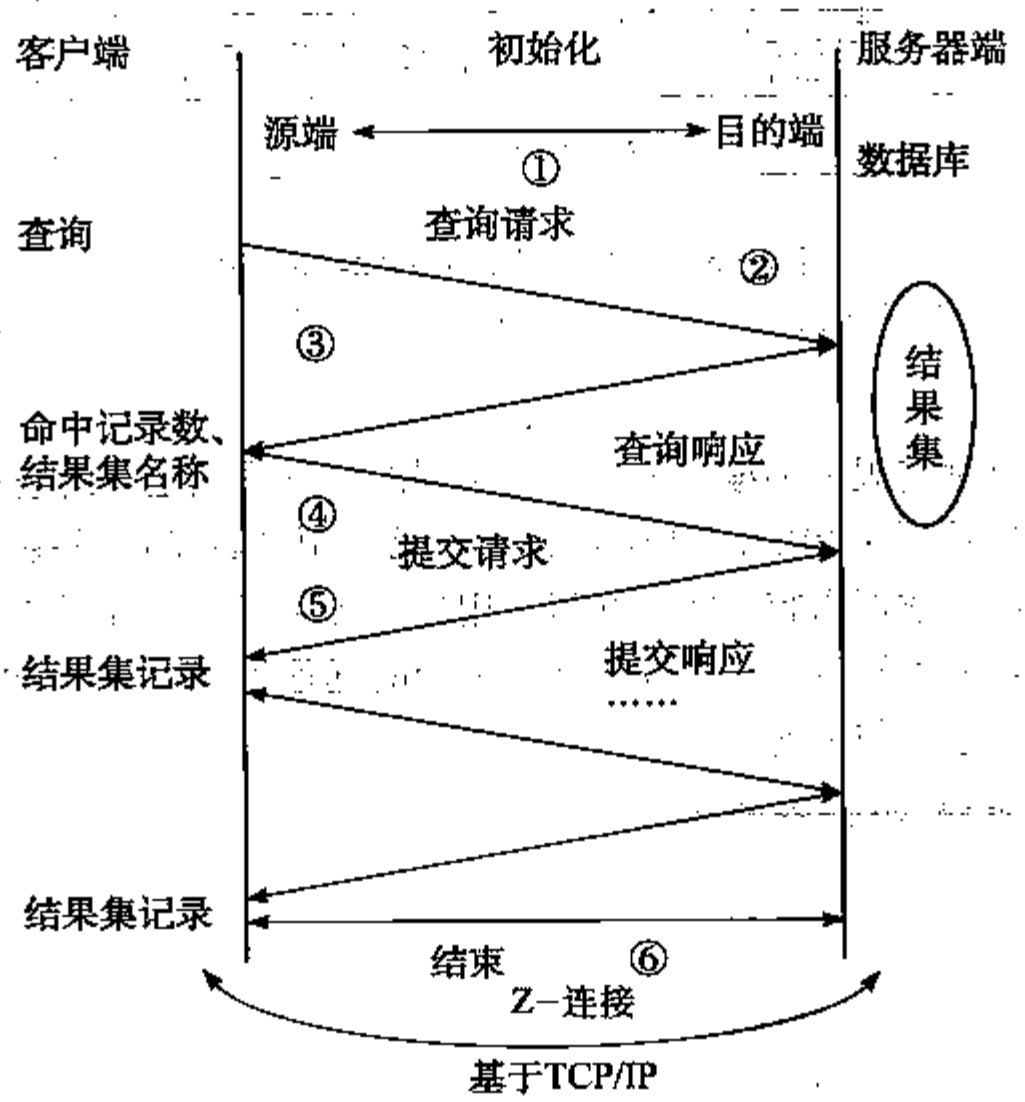

图9—2 Z39.50的基本实现过程

（1）源端向目的端发出初始化请求APDU，目的端发出初始化响应APDU，源端据此判断连接接受与否。

（2）源端发送查询请求APDU，目的端在数据库上执行查询，创建符合查询

请求的结果集并缓存在服务器端。

（3）目的端发送查询响应 APDU，包括命中的记录数和结果集名称。

（4）源端发送提交请求 APDU，指定记录格式和元素定义。

（5）目的端发送提交响应 APDU，返回一个或多个结果集中的记录。

（6）服务结束后，由源端或者目的端关闭连接。

早先的 Z39.50 协议采用了典型的两层 C/S 结构，在客户端程序可以方便地实现与分布于全国各地的 Z39.50 服务系统建立连接、访问和检索其中的数据库系统。两层的 Z39.50 协议的实现模型如图 9—3 所示。

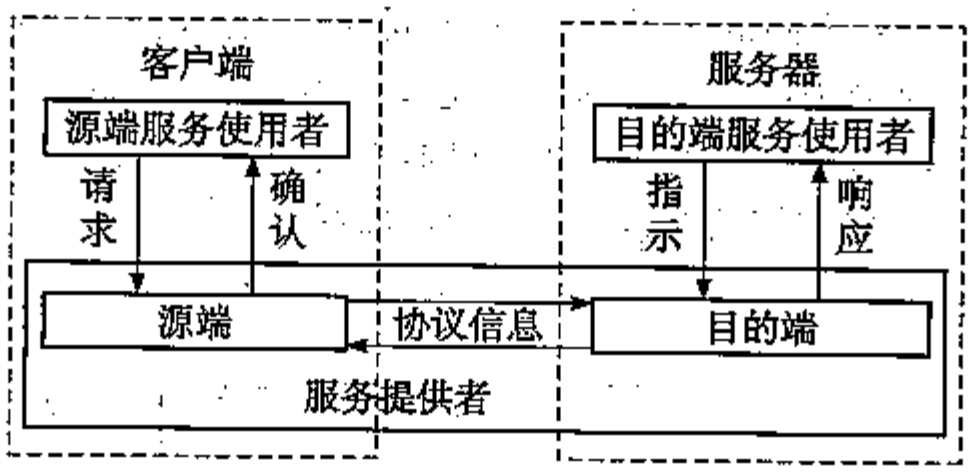

图 9—3　Z39.50 协议的基本实现模型

资料来源：丁峰、马范援：《基于 Z39.50 的分布式 WWW 信息检索》，载《计算机工程》，2001（2）。

该标准通过制定规范和编码来构筑与不同的信息系统之间的连接与通信，是完全独立于下层数据结构的信息检索服务，无需用户具备或掌握远程系统的语法、检索策略以及数据内容等相关知识，用户只要使用其本身所熟知的一个计算机系统的界面，就可检索基于不同软硬件平台的远程系统的信息资源，从而实现网上透明的信息检索与传递。

9.3.1.4　Z39.50 的主要功能及其应用

Z39.50 的主要目的在于定义基于客户端/服务器体系结构的数据库的查询与检索的语法，从而使一套存取标准适用于异构系统的数据。Z39.50 定义的信息检索系统主要包括 11 种功能：

（1）初始化。包括初始化服务，当客户端提出连接时，服务器可以接受或可以改变，客户端必须接受，否则将终止连接。

（2）查询。系统间传递查找信息。

（3）检索。系统间传递检索信息，对目标数据库进行检索。

（4）结果删除。客户端可要求删除在连接期间保留在服务器端所有或部分结

果集。

(5) 浏览。提供对索引或数据库中特定内容的浏览。

(6) 排序。提供对结果集排序的功能。

(7) 存取控制。服务器端在执行开启、查询、展现或删除服务时，皆可对客户端之权限等提出质疑，客户端必须予以回应，否则服务器端可以终止连接。

(8) 会计/资源控制。包括资源控制服务、触发资源控制服务、资源报告服务。当实现使用和预测使用之资源超过协议之范围，则服务器端会通知客户端，只有服务器端同意，客户端才能继续作业。

(9) 解释功能。客户端可以查询 Z39.50 的解释数据库，从而知道服务器端的相关信息，包括可供查询的数据库、检索点及错误信息等。

(10) 扩展服务功能。如客户端可以设定在服务器端上被持续执行或周期性执行的查询命令，客户端也可以向服务器端要求传送某份文件，另外也提供了维护数据库的功能。

(11) 终止。允许服务器端或客户端做两种终止：服务器或客户端可在任何时间送出或接收检索放弃请求，并终止此连接；服务器端在收到开启、查询、展现或删除回应后，可提出检索解除请求，等收到客户端回应后关闭连接。

Z39.50 是国际通用的信息检索协议，是一个相当成熟的标准。自从公布以后，就逐渐被美国和一些发达国家的计算机厂商、数据库中心和图书馆等有关单位所接受。对于缺乏一个信息组织与检索标准的互联网上的大量信息资源而言，Z39.50 的大量应用在一定程度上帮助解决了网上信息的无序和难以检索的问题，为网络中的异构平台和异构系统之间的信息检索和传输提供了条件，实现了与其他具有标准接口的系统之间的数据访问，为信息资源共享提供了新的途径。

Z39.50 对国外（特别是美国等英语国家）信息系统的发展和服务方式产生了重要影响，尤其在图书馆自动化领域中得到了广泛的应用，多数图书馆提供 Z39.50 服务，各软件企业为适应网上检索的需要开发出了大量软件，通过这些软件，可以查询任何一个支持 Z39.50 协议的服务器上的资源。例如，通过美国国会图书馆的 Z39.50 网关，可以查询到几百家图书馆、科研机构和商业数据库的数据信息。

国内出现了不少自行开发的支持 Z39.50 的图书馆集成系统，如北京邮电大学图书馆的 MELINETS、南京大学图书馆的“汇文”图书馆管理集成系统等。同时，也引进了一些国外的大型图书馆管理集成系统，1996 年底上海图书馆引进了 Horizon 系统，1999 年上海交通大学、复旦大学、华南理工大学等五家图书馆也陆续使用。1997 年初，清华大学图书馆引进了 INNOPAC 系统，西安交

通大学也采用了该系统。1998 年，北京大学引进并汉化了 Sirsi 公司的 Unicorn 图书馆自动化集成系统，此后，中国人民大学图书馆、北京航空航天大学图书馆也引进了该系统，这些系统都支持 Z39.50。例如，利用广东省立中山图书馆的 Z39.50 公共查询网关，读者不仅可以查询中山图书馆的馆藏数据库，还可以查询到其他所有支持 Z39.50 协议的各大图书馆及情报信息部门服务器上的数据，为读者提供了一种简便而快捷的检索途径，避免了读者为某一文献针对不同图书馆馆藏数据库系统，进行分别检索操作的麻烦。目前，通过该公共查询网关，可以选择的目标服务器有：北图联合编目中心、中山图书馆、上海图书馆、深圳市图书馆、西安交通大学、清华大学、北京大学、中国人民大学、南开大学、华东师范大学、香港中文大学、贝尔实验室图书馆（Bell Laboratories Library）等。通过 Z39.50 协议传输的记录格式有多种，如 UNIMARC、USMARC、SUTRS、GRS1、XML 等等，该网关目前只支持 UNIMARC 和 USMARC 两种数据格式的显示。具体查询界面如图 9—4：

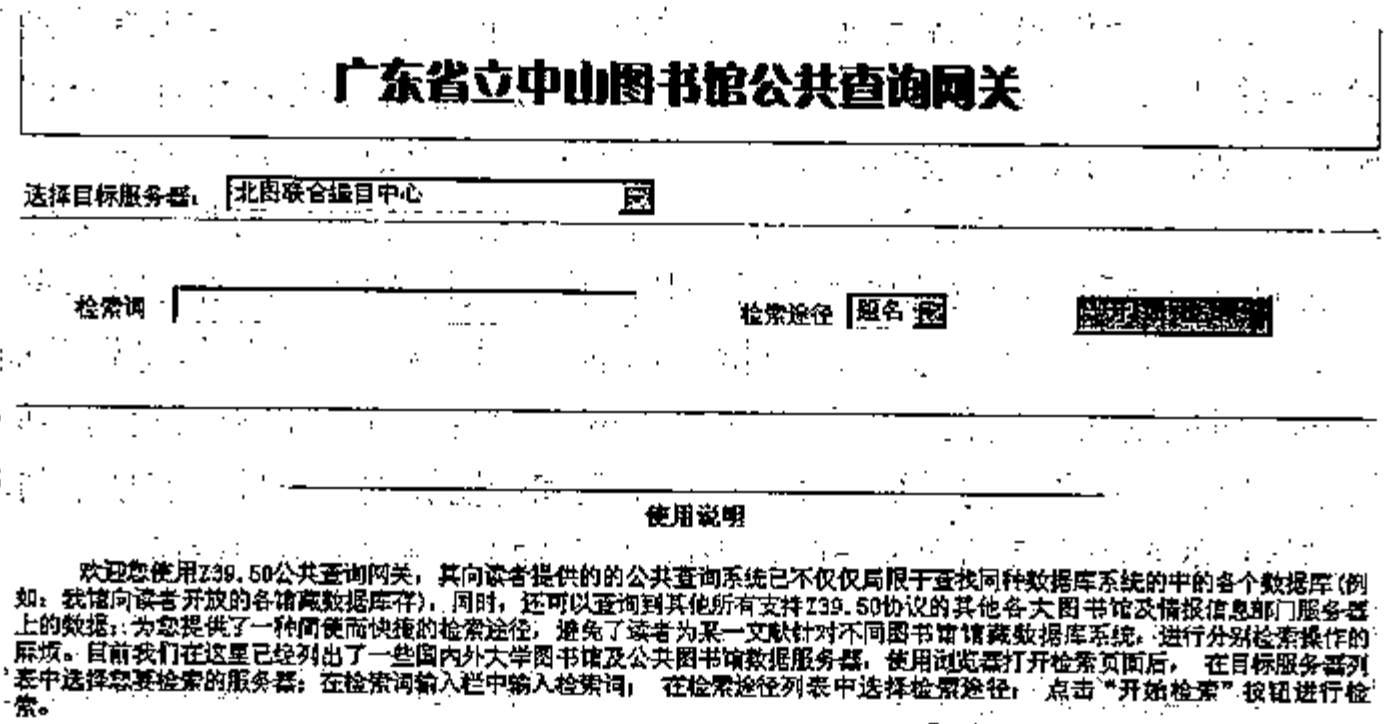

图 9—4　广东省立中山图书馆公共查询网关

资料来源：http：//bmzx.zslib.com.cn/z3950gate/search.php，2008－04－30。

9.3.1.5　Z39.50 标准的不足

虽然 Z39.50 协议从制定之初到后来的几经修订已有三版，但是实际中的实现效果却不甚理想，究其原因，一方面是计算机技术发展的各种因素，另一方面是 Z39.50 协议自身实现的内在原因。这些不足对实际实现的影响表现在两个层次：较低层次的影响是增加系统实现的复杂性，降低了实现效率，但一般不影响实现的通用性和标准性；较高层次是影响了实现的通用性和标准化程度，这是

Z39.50实现的根本障碍。

作为一个开放系统互连的标准集，Z39.50的主要问题和实现的最大障碍是其定义的协议数据结构稍显简单。从信息检索应用的通用性和标准化角度出发，Z39.50既不规定也不限制计算机系统的实现细节，其协议规范仍属于抽象层次。例如其应用协议数据单元APDU，是规定了信息检索应用之间信息传递格式的信息单元，为复合数据类型，而Z39.50仅定义了其数据单元，从Z39.50实现的角度，要和其他Z39.50实现互操作，仅定义组成的数据单元是不够的。其数据类型必须足够具体，使每个有意义的位都可有效的存取，即准确地和其他数据元素相区别。另一方面，如果协议规范制定得过于具体，有可能限制某些现有或将来的计算机系统的实现，使标准集失去普遍性和可扩展性。这种矛盾的权衡和解决是制定一切标准所必须面临的基本问题，一种可行的折衷是制定足够具体的实现细节，以确保可实现性，同时在各层次提供诸如指示符之类冗余信息，精巧的组织和仔细的配置可使实现时的冗余对效率的影响减小，同时保有相当的可扩展性。Z39.50采用了另一方法。它规定了APDU数据类型的转换语法由具体服务提供者确定。换句话说，Z39.50把APDU数据类型精确定义的责任交给了Z39.50的实现者。因此，一个Z39.50的合法实现必须为APDU数据类型的解释做某种假定，也就无法保证和任一其他的Z39.50合法实现互操作。

其次，Z39.50标准所采用的编码标准是在20世纪80年代发展起来的一种成熟标准，主要用于有线通信和无线通信领域。按这种标准对信息编码的过程，几乎是将可读文字转换为机器语言的过程。对很多数据库应用软件设计者来说，在平时的软件产品开发中很少涉及这样的“底层”开发，甚至无法看懂相关标准文献的含义。这意味着Z39.50的实施，使很多图书馆应用软件开发者或提供商面临着技术上的挑战，或面临产品开发项目的投资风险。

再次，Z39.50这个标准是建立在一种比较专门的通信协议的基础上，增加了这个标准的实施中的技术复杂性。这种技术上的复杂性提高了软件实现的门槛，需要软件设计人员和编码人员具有较为丰富的数据结构和网络通信接口规范方面的知识，并掌握较高的代码转换算法技术。然而数据库检索系统软件的开发人员一般很少有人涉及这些专门的技术领域，因此原有Z39.50标准实施与推广的阻力主要不是来自应用软件的使用者，而是来自应用软件的开发者。

9.3.2 其他网络信息检索标准

实现信息源互联互操作的协议主要有两类。一类为重量级协议，如在图书馆界有较大影响的Z39.50协议，这些协议本身较复杂，往往要求在字段级实现互

操作。一类为轻量级协议如 OAI 等，这些协议一般作为一些应用协议的基础。由于 Z39.50 存在着诸多的不足，促使了新的网络信息检索标准的制作和使用，目前正在使用或处于研究探讨的网络信息检索标准还有一些，具体如下：

9.3.2.1 OAI (Open Archives Initiative)

OAI 设想是在 1999 年 10 月在美国新墨西哥州的圣达菲（Santa Fe）举行的开放存取先导的第一次会议上首先提出的，它受到了网络信息联盟（The Coalition for Networked Information，简称 CNI）、数字图书馆联盟（Digital Library Federation，简称 DLF）和美国国家科学基金会（The National Science Foundation，简称 NSF）的支持和资助。OAI 协议自 1999 年提出以后，于 2001 年 1 月推出了第一版（OAI—PMH 110），之后又于 2002 年 6 月推出了第二版（OAI—PMH 210），并于 2005 年 5 月 3 日公布关于 210 协议的最新文档说明。

OAI 是一项简单、灵活的元数据互操作协议，其目标是实现在 Web 上发布信息的不同组织（主要在欧美等地）之间的互操作，形成一个与应用无关的互操作框架，OAI 支持选择性的采集方式，适合于系统间元数据的循环交换。OAI 包含两类角色，即数据提供方和服务提供方。OAI 的核心思想非常简单：在 HTTP 协议的基础上，服务提供方使用 OAI 规定的 6 个命令动词，按照一定的条件采集（Harvest）来自不同数字图书馆（数据提供方）中的元数据，将所收集来的 XML 格式编码的元数据集中保存在一个中心库中，并在此基础上提供增值服务。OAI 的主要目标：(1) 简化数字信息内容有效传播；(2) 提升数字资源的存取；(3) 扩展存取数字资料种数的范围。

OAI 实现了异构、分布元数据资源的采集、集中和交换，并凭借其简单易用、低门槛、低成本以及使用了 HTTP、XML 等通用技术作为基础，在许多数字图书馆项目中得到了广泛的认同和使用。OAI 还制定了相应的元数据采集标准 OAI MHP（Open Archive Initiative Metadata Harvesting Protocol）。该协议是一个元数据采集标准，即从数据提供方采集的只是元数据信息，不包括内容。OAI 的技术框架在设计之初就是本着简单易用的原则进行。目前，采用 OAI 协议框架的项目有：NSDL（National Science Digital library）、Cyclades、the Andrew Mellon Foundation、RLG（the Research Libraries Group）、the Digital Library Federation、the OCLC Office of Research 等。

9.3.2.2 OpenURL

OpenURL（Open Uniform Resource Locators）框架最初是 1999 年在研制 SFX 系统时由赫伯特·范德松佩尔（Herbert Van de Sompel）提出来的，2001 年美国国

家信息标准组织（NISO）成立专责委员会，制定关于 OpenURL 的美国国家标准（ANS）Z39188—200X。OpenURL 是一种开放的信息资源与查询服务之间的通信协议标准，是开放的、上下文相关的链接框架，它提供了一种在信息服务者之间传递对象元数据的格式。OpenURL 的目的是把不同来源和不同通信协议的信息源及相关服务融合在一起，实现不同类型、不同格式和异地分布信息资源的无缝链接。可以使用 OpenURL 来传递上下文相关的分发请求。例如，OpenURL 标准可以让检索到一篇论文引文的用户，通过扩展的连接服务器，直接访问该引文最合适的拷贝。最佳拷贝是根据用户和组织的参数来进行选择，这些参数有拷贝的位置、花费、与信息供应商的协议及其他相似考虑。这种选择不需要用户具备相关知识，它是通过在引文源和“解析者”（连接服务器）之间传递 OpenURL 元数据来实现的，连接服务器存储了以上参数信息和到合适资源的连接。

OpenURL 作为一种开放式链接框架，把信息源、信息服务和用户需求有机地整合在一起，它突破了传统链接框架的局限，可为用户提供上下文相关（Context-sensitive）链接传递服务。以 OpenURL 为基础的 SFX 系统成功地将国外著名数据库和图书馆整合到 OpenURL 框架中。目前基于 OpenURL 框架的参考链接系统产品多达几十种，其中较成熟、应用较广的系统有 SFX、LinkFinderPlus 等。

9.3.2.3 X.500

X.500 是由国际电报电话咨询委员会（CCITT）制定的基于 ISO/OSI 开放系统互连标准的名录服务通信协议。此标准为国际标准化组织（ISO）所接纳，其标准号为 ISO9594。它的目标是向用户提供分布的名录服务，提出了一个将分布在各地的名录服务器连接成为一个全球性的分布式名录服务的体系结构。每个名录服务器都拥有一部分数据库，通过称为名录系统代理（Directory System Agent，简称 DSA）的服务器对用户提供服务。数据库的维护工作是由各 DSA 在本地完成的。全球性的名录服务对用户来说是透明的，似乎就是在本地提供的。X.500 是一个将局部名录服务连接起来，构成全球分布式的名录服务系统的协议。X.500 组织起来的数据就像一个很全的电话号码簿，或者说一个 X.500 系统像是一个分门别类的图书馆。而某一机构建立和维护的名录数据库只是全球名录数据库的一部分。

X.500 不仅提供有关个人和机构信息，它还可以用来提供有关网络资源、应用系统或硬件等方面的信息。X.500 中的信息是依据 DIT（Directory Information Tree）这种体系结构来组织的，从根目录（世界）开始逐层向下：世界→国家→机构→个人（或资源）。X.500 是一种标准化的名录服务，有很好的发展前

景。但由于建立和维护 X.500 的费用较大，目前实施的范围还有限，而且大部分都是用于提供“白页服务”。

9.3.2.4 LDAP

LDAP 的英文全称是 Lightweight Directory Access Protocol，一般都简称为 LDAP。它是基于 X.500 标准的，但是比较简单并且可以根据需要定制。与 X.500 不同，LDAP 支持 TCP/IP，这对访问互联网是必需的。LDAP 的核心规范在 RFC 中都有定义，所有与 LDAP 相关的 RFC 都可以在 LDAPman RFC 网页中找到。现在 LDAP 技术发展得很快，应用前景也很广阔。在企业范围内实现 LDAP，可以让运行在几乎所有计算机平台上的所有应用程序从 LDAP 目录中获取信息。LDAP 目录可以存储各种类型的数据：电子邮件地址、邮件路由信息、人力资源数据、公用密钥、联系人列表等等。通过把 LDAP 目录作为系统集成中的一个重要环节，可以简化员工在企业内部查询信息的步骤，甚至连主要的数据源都可以放在任何地方。

LDAP 最大的优势在于可以在任何计算机平台上，用很容易获得的而且数目不断增加的 LDAP 的客户端程序访问 LDAP 目录，而且也很容易定制应用程序，为它加上 LDAP 的支持。LDAP 协议是跨平台的和标准的协议，因此应用程序就不用为 LDAP 目录放在什么样的服务器上操心了。LDAP 是互联网的标准，厂商都很愿意在产品中加入对 LDAP 的支持，因为他们根本不用考虑另一端（客户端或服务端）是怎么样的。LDAP 服务器可以是任何一个开发源代码或商用的 LDAP 目录服务器（或者还可能是具有 LDAP 界面的关系型数据库），因为可以用同样的协议、客户端连接软件包和查询命令与 LDAP 服务器进行交互。

9.4 网络信息检索发展趋势

由于现代信息通信技术的发展，网络信息检索技术的软硬件环境有了很大改善，信息检索服务功能的不断完善，网络用户对网络信息检索的需求，这些都极大地推动了网络信息检索的发展。网络信息检索的发展主要体现在智能检索技术、知识检索技术、多媒体检索技术、新一代搜索引擎技术、自然语言检索技术和基于内容的检索技术。在以用户为中心的思想指导下，网络信息检索服务呈现出个性化、多样化特点。

9.4.1 网络信息检索技术发展

9.4.1.1 智能检索技术

随着人工智能技术应用于信息检索领域，出现了各种各样智能信息技术方法，智能检索是人工智能与检索技术的有机结合。智能化信息检索是基于自然语言处理的检索形式，它可以模拟人脑的思维方式，分析用户以自然语言表达的检索请求，自动形成检索策略，进行智能、快速、高效的信息检索。智能检索技术主要体现在语义理解、知识管理和知识检索三个方面。它利用语义分析模块自动智能分词，进行用户请求和知识库“数据”的语义理解，最终把知识库中匹配的信息筛选、整序后提供给用户。

基于智能技术基础的智能搜索引擎，拥有机器学习技术、智能代理技术、知识发现技术，基于自然语言理解，拥有智能化的检索、分析和反馈功能。其中，智能代理技术 Agent 是一些智能化的程序，能够学习用户的需求，并利用搜索引擎等系统提供的现有服务来检索用户所需信息。Push 技术采用主动服务新模型，直接向用户推送其感兴趣的信息，而无须用户查找。总之，智能检索以用户信息需求为基点，建立用户检索智能模型，检索过程、检索结果、检索反馈和数据库维护智能化、自动化，还能够实现信息定期和定题检索以及根据用户反馈自动对知识库进行维护和更新。如 Alta Vista 也开始引入自然语言的自动翻译。

9.4.1.2 多媒体检索技术

从基于文本的方式开始，多媒体信息检索至今已发展成熟，但是在大量的多媒体信息检索环境中还是支持不够。多媒体检索技术包括基于描述的多媒体检索和基于内容的多媒体检索。基于描述的多媒体检索其实质依然是基于关键词的检索，基于内容的多媒体检索是在对多媒体对象内容特征分析的基础上，提取多媒体对象的语义信息，构成多媒体语义信息单元数据库，再对用户提交的多媒体的样例进行匹配检索。而多媒体信息检索中的音频检索、图像检索、视频检索将成为独立的研究对象。如 IBM 公司开发的基于 MPEG—7 的多媒体搜索引擎 MARVEL 获得《华尔街日报》2004 年创新大奖。MARVEL 利用多模型机器学习技术，自动对视频文件进行扫描、解析并对概念进行索引，从而自动为每个视频文件添加描述性元数据。

鉴于网络信息多媒体成分越来越多，有人提出了多媒体检索的发展趋势，例如根据人们现有的思维方式和世界潮流，还可以提出一种基于体验的检索方法，

也就是达到根据人的自身体验进行明确的或者模糊的信息检索的目的，将各类数据库合并，可提供综合的各种特征信息，甚至在图像、视频、音频的多媒体基础上，增加诸如气味、口感等多媒体检索。相信计算机科学和技术的发展，将会在新的检索技术、快速算法等方面有新的突破。

9.4.1.3 P2P 检索技术

P2P 是 Peer to Peer 的缩写，一般将 P2P 译为“端对端”或“点对点”。它是一种用于不同 PC 用户之间，不经过中介设备直接交换数据或服务的技术。它允许网络用户直接使用对方的文件。传统检索是以服务器为中心，用户向服务器发送请求。服务器将检索结果发送到用户的浏览器中，即是以服务器为中心的发散型方式，任何检索必须以服务器作为中介。P2P 检索能够共享所有用户硬盘上的文件、目录乃至整个硬盘，也可以不受信息文档格式和设备的限制，达到传统目录式搜索引擎无可比拟的深度。P2P 模式基于分布式共享技术，它使互联网上每台计算机都有可能成为信息资源提供者。

以 P2P 技术发展先锋 Gnutella 进行的搜索为例：一台 PC 上的 Gnutella 软件可将用户的搜索需求同时发给另外 10 台 PC，如果搜索请求未得到满足，这 10 台 PC 继续转发给另外 10 台 PC，这样在几分钟内就可以搜遍几百万台 PC 上的信息资源。美国的 NapsterMP3 网站最早使用这种 P2P 思想，它利用文件共享技术来共享 MP3。基于 P2P 对等搜索理念的搜索技术会为因特网的信息搜索提供全新的解决之道。它使人们在因特网上的共享服务行为被提升到了一个更高的层次，使人们以更主动深刻的方式参与到网络中去。目前基于 P2P 技术开发的软件还有：eDonkey2000、eMule、Jigle 等，Google、Infrasearch 等公司也将这种技术应用到自己的系统中。

9.4.1.4 可视化检索技术

可视化检索是把文献信息、用户提问、各类检索模型以及利用检索模型进行信息检索的过程，展示在一个多维的可视化空间中，并向用户提供信息检索服务。其实质是提供一种可视的语义关系，使提问与检索结果以及检索到的各文献之间的关系可视化。可视化信息检索包括两方面，即检索过程的可视化和检索结果的可视化。检索过程的可视化是指用户在检索过程中各检索对象之间的关系以可视化的形式展现在用户面前，用户顺着可视化检索的画面一步一步地发现检索结果。检索结果的可视化是指用户提交检索词后获得的检索结果不仅仅是现在主要信息检索工具提供的列表这样的一维形式，而是以基于检索结果分析后形成的二维或三维的形式来展示检索结果之间的语义关系。

比如，可视化元搜索引擎 Kartoo 就实现了信息的可视化表达，其检索结果是以图形界面而不是线性列表的形式提供的。Kartoo 采用了 FlashPlayer 软件将来自其他多个 Web 搜索引擎并经整理的检索结果以一种实时交互式镜像图的形式显示出来。镜像图由一组大小不同、形状各异的纸形图标、关键字，以及连线组成。图标用于表示检出的各种资源（如 Web 网站/网页，以及 .pdf、.doc 等各种类型的网络文档），图标的形状反映检出资源的类型，图标的大小反映检出资源与检索式的相关程度，相关程度越大，图标的尺寸则越大。关键字出现在一些图标附近，用于表示其邻近资源所属的主题范畴。连线存在于图标与关键字之间，用以反映资源之间的相互关系。还有 EBSCO 数据库中的可视化搜索也是一种可视化检索环境，其检索结果按主题以交互式的“圆形”或“正方形”可视图的方式排序。“圆形”表示将结果归到哪一类，与较小的圆形相比，较大的圆形含有更多子类和链接。“正方形”表示文章，同一正方形可显示在多个圆形中。

9.4.1.5 语义检索技术

语义检索技术也称为概念检索技术，它不是传统意义上的关键词的字面匹配，而是从词所表达的概念意义层次上来认识和处理用户的检索请求。语义检索主要包括两方面：同义词扩展检索和相关概念联想。同义词扩展检索是指在检索某一关键词时，还能对它的同义词、近义词检索。相关概念联想又分为两方面，一是对概念的上位概念进行联想，称为语义外延扩展，一是对其下位概念进行联想，成为语义蕴含扩展。在采用语义检索技术的信息检索系统中，概念集是其核心，概念集中给出了各个概念的定义和概念之间的相互关系。根据概念集，系统可以对收集到的信息资源进行分类和语义标注；在概念集的协助下，能够使用户对要检索的东西定位得更快、了解得更深入；根据概念集，可以规范用户的查询信息来提高查询效率。因此，这是一个基于知识的检索过程。

美国马里兰大学的 SHOE 项目开发的 SHOE 搜索引擎就是基于语义检索的下一代搜索引擎，它使用类似于 XML 语法的网络本体语言（OWL）和高级人工智能技术构建自己的知识库，它基于上下文检索，通过提供上下文背景知识来补充用户所检索的内容，从而实现语义概念上的检索。基于概念集的信息检索实现了概念语义层次的检索，突破了关键词检索局限于形式的固有缺陷，实现了对于用户检索请求的合理化联想，不仅给出查询结果，还提供了进一步检索的建议，使信息查询变得更加方便、快速和准确，为实现信息的精确语义检索提供了思路和实现的基础。

9.4.2 网络信息检索服务发展

9.4.2.1 多样化信息检索服务

多样化信息检索服务包括检索多样化信息形态、多样化检索语种、服务功能多样化及本地化和一站式服务。

信息检索形态的多样化是指不仅可以检索到不同格式的文本文档，而且也可以检索到声音、图像、动画等多媒体性文件。很多网站都不断增加多种类型的搜索文件。

使用某一种语言直接进行多语种检索；并提供多语种的匹配结果是多语种检索服务的发展方向。由于互联网是一个巨大的数字资源信息库，包含不同语言信息，随着信息查询用户的素质的不断提高，借助网上机器翻译工具，可以直接阅读中外文信息。由于这种需求，在网络版叙词表的基础上，加上新的智能检索技术开发出可以检索不同语言的同类信息的软件。其基本原理是以类似于多语种叙词表的词汇编码，实现不同语种间词汇的转换。在单一检索界面的检索后台有一个多语种词库，对用户提交的某一语种的检索词自动在词库中查找对应的其他语种的检索词，再提交给搜索引擎，以多语种检索结果输出给用户。多语种信息检索需要机器翻译技术的支持，并对多语种检索得出的输出结果相关度或是重要性排序进行研究。这样用户不但可以用自己熟悉的语言进行检索，还可以用一种语言进行提问，获得多种语言的检索结果。许多搜索引擎商纷纷在其他国家设立本地站点，通过增加服务器，分流用户，更加方便用户使用母语检索信息。如Google、Yahoo!、HotBot等都在世界各地设立了分支机构。

一站式服务是信息检索服务多样化的又一体现。一站式信息检索服务是指用户通过一个检索工具能满足自己所有的信息检索需求。一站式信息检索是未来信息检索服务的一种发展模式。网络信息检索已不仅仅是单纯的提供特定信息的检索功能，正在扩大服务范畴，诸如提供站点评论、天气预报、新闻报道、股票点评、各种黄页（如电话号码、航班和列车时刻表、地图等）、免费电子邮箱等，以多种形式满足用户的需求。全球最大的搜索引擎Google正扩展其业务方向，朝着一站式服务的方向发展。目前的Google已经为用户提供了某种程度上的一站式服务了。

网络信息检索服务多样化还表现为信息检索系统应该具有多方面的功能，包括导向功能、评价功能和文化积累功能等。从网络检索工具的实际作用看，它们已经能进行一定程度的评价、导向，有知识“过滤”的作用，有一定的信息资源控制功能。同时，网络文化也如同传统文化一样需要保存。网络将再造人类文明

已是不容争辩。

9.4.2.2 个性化信息检索服务

随着互联网中信息量的增长，互联网信息检索系统的检索效率日益受到关注。为不同的用户提供有针对性的检索结果，也即个性化信息检索，成为一种新的个性化服务形式，也是未来面向用户信息检索的一个发展方向。个性化信息服务，是针对不同用户，采用不同服务策略和方式，提供不同信息内容的服务。它具有以用户为中心、对用户需求进行挖掘、灵活多样和主动将信息推送给用户的特点。个性化服务主要有三种形式：个性化推荐、个性化网站和个性化信息检索。个性化信息检索服务主要体现在：一是允许网络用户的个性化定制，网络用户基本的定制包括选择自己喜欢的检索界面、检索结果的显示格式、检索结果的语言等，而高级定制包括网络用户自己选择检索信息来源、对检索结果进行过滤，检索结果去重等。搜索引擎 AllTheWeb 就提供对冒犯性网页进行过滤的功能，并可对检索结果进行去重显示。二是基于数据挖掘技术对网络用户的检索行为进行分析，挖掘出网络用户的检索需求，利用推送技术主动向用户推送所需的网络信息。

个性化检索定制服务的根本特征就是“一切以用户为中心”，在建立定制服务的过程中，要充分地研究用户的检索行为、目标、习惯方法，同时根据用户的检索行为，能够自动规范用户的检索动作，提高信息检索的效率和效果，推荐更具针对性的信息服务。了解和掌握了个性化定制信息服务的特点，才能够更好地为用户提供针对性更强的服务，从而提高用户对检索服务的满意度。个性化检索定制能够主动地提供一些有针对性的信息内容，同时又能够根据用户检索内容进行关联性分析，为用户进一步的信息检索提供建议。目前，支持定制检索标签功能的搜索引擎有 iBoogie、Clusty、Ask. com（Teoma）和 Yahoo! Search。前两者支持两种形式的检索标签定制，后两者只支持一种形式的检索标签定制。

信息推送服务是基于推送技术发展而出现的一种新型服务，这种信息检索服务最大的特点就是将被动地等待用户提出请求变为有目的地、主动地推荐信息，提供信息的主动性增强。所谓推送技术就是一种按照用户指定的时间间隔或根据发生的事件把用户选定的数据自动推送给用户的计算机数据发布技术。基于网络的信息推送主要有以下几种形式：频道式推送，频道式网播技术是目前网上最普遍采用的一种推送方式，它将某些网页定义为浏览器中的频道，用户可以像选择电视频道那样去选择收看感兴趣的、通过网络播送的信息；邮件式推送，以电子邮件方式主动将有关信息发布给列表中的用户，如《大学图书馆学报》的目次推

送服务；网页式推送，在特定网页内将信息提供给感兴趣的用户；专用式推送，通过机密的点对点通信方式，将指定的信息发送给专门的用户。

9.4.3 网络信息检索标准发展

研究网络环境下异构信息检索的标准体系成为当前信息检索领域的一个研究热点。标准的网络语言、网络符号和网络输出显示方式，会使网络信息的传播更加方便、快捷，有利于被用户接受与利用，实现信息资源共享。可以说，规范化和标准化是网络信息资源共享的必要前提，没有规范的接口和统一的信息检索技术标准，资源共享将成为空谈。

9.4.3.1 Z39.50 网关

随着 Internet 的普及，Web 浏览器已经对 Z39.50 标准的应用范围和应用方式产生了很大的影响，为 Z39.50 的应用提供了另一个应用领域，即作为 Web 网关，为公众提供跨平台、跨服务器的虚拟目录检索服务。Web 网关以双重身份提供信息检索服务，对于众多 Z39.50 服务器而言，它是一个 Z39.50 客户端程序，对于众多万维网浏览器用户而言，它是一个 Web 服务器。用户通过这种网关检索信息，可以将网关所连接的众多数据库视为一个综合信息库，通过统一的检索界面发出一个检索请求后，可以得到很多服务器返回的结果。通过 Web 网关检索信息，用户不必逐一记住各服务器的地址，不必分别进入各服务器进行重复的检索操作。Z39.50 网关在数字图书馆的应用前景十分广阔。其工作流程为：

（1）用户的查询请求由浏览器通过 HTTP 发送给 Web 服务器。

（2）Web 服务器通过 HTTP—Z39.50 转换网关把 HTTP 请求转换为 Z39.50 请求。

（3）客户进程将请求发送给本地或远程的 Z39.50 服务器，进而访问数据库得到查询结果。

（4）以规定的格式将检索结果传送到网关。HTTP—Z39.50 协议转换网关收集由 Z39.50 服务器返回的查询结果，整合后统一以 HTML 页面的形式返回给用户浏览器。

典型的浏览器/网关/服务器的模型如图 9—5 所示。

从图 9—3 和图 9—5 对比来看，三层结构与二层结构的 Z39.50 实现模型差异比较显著，与用户交互的 Z39.50 客户服务转移到了网关，代之以 Web 浏览器为交互界面，无需安装专门的 Z39.50 客户程序，用户通过熟悉的浏览器就可以获得相关服务。

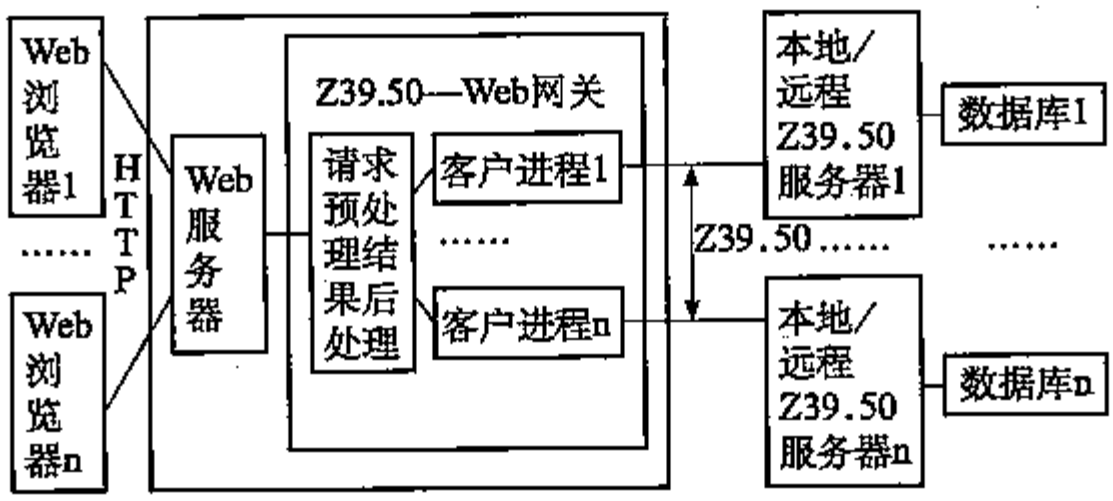

图 9—5 浏览器/网关/服务器模型

9.4.3.2 新一代 Z39.50[①]

为了简化 Z39.50，以欧美为主的一些 Z39.50 开发团体，对 Z39.50—1995 及以前的各个版本，去粗取精，并且做了大量的创新，于 2002 年上半年公布了新一代的 Z39.50（ZING，Z39.50—International：Next Generation），参加开发的机构主要有：AR-TISTE、Pergamum、Tilburg University、Knowledge Intergration Ltd、SIRSI、InQuirion Pty Ltd.、OCLC、EDI-NA、PICAOxford、DBK、Koninlijke Bibliotheek、Library of Congress 等。SRW（Search/Retrieve Web Service）和 SRU（Search/Retrieve URL Service），合称为 SRW/U，是 ZING 的核心所在。SRW/U 集成了 Web 和 URL 技术，对 Z39.50 作了很大的改进，营造了一个崭新的 Z39.50。

SRW/U 从语义学角度提出了搜索数据库的概念。与传统的 Z39.50 相比，创新包括以下几个方面：（1）Z39.50 是面向连接的、会话的、有态的；而 SRW/U 则是面向无连接、无态的。（2）Z39.50 可以把多个服务绑定在一个协议中，而 SRW/U 则认为不同的 Z39.50 的服务就是 Web 服务。（3）Z39.50 的搜索和呈现服务是泾渭分明的，而 SRW/U 的搜索和呈现服务则绑定在一个 Web 服务中。Z39.50 区分数据库和服务器，而 SRW/U 则认为数据库和服务器是没有区别的，都看作是服务器。（4）Z39.50 定义了不同的记录语法，而 SRW/U 只采用一种，即 XML。（5）Z39.50 的提问表达是 RPN（逆波兰），而 SRW/U 则定义了专门的串提问语言——cql。（6）Z39.50 是基于 ASN.1/BER 的，而 SRW/U 则是基于 XM，把 XML 作为 SRW/U 的开发基础。在 ZING 中，SRW 与 SRU 只有细微的不同。SRW 的请求/响应序列的执行是使用 HTTPPOST，通过 XML-

① 参见衡中青、曹翔：《新一代 Z39.50：ZING》，载《情报杂志》，2003（3）。

SOAP/PRC 来执行的，而 SRU 的请求/响应序列是通过 HTTPGET 命令来执行的，请求是通过 HTTPURL 来调用的，SRU 不使用 SOAP。

SRW/U 作为下一代 Z39.50 计划成员之一，它不是对 Z39.50—1995 版的更新和替代，而是一种在继承原有的 Z39.50 标准合理成分的基础上建立的全新的体系。SRW/U 的成熟和发展，最终不会简单地取代原有的 Z39.50 标准，而很可能会与原有的 Z39.50 标准共同发展，在不同的领域发挥作用。SRW/U 实施更简单，更具有市场潜力，更有希望推广到商业信息检索领域。而经典 Z39.50 将继续应用于图书馆等文献信息领域。另外，SRW/U 与原有的 Z39.50 显然是不兼容的体系。具有不同的数据结构和不同的通信方式，SRW 客户不能直接获得 Z39.50 资源。但是，可以利用 SRW 建立与现有 Z39.50 服务器的网关，从而扩展已有 Z39.50 服务器的服务范围。

其他的网络信息检索标准还有 STARTS（Standard Protocol Proposal for Internet Retrieval and Search）、CIP（Common Indexing Protocol）等等，这些标准正在探讨和酝酿之中，有可能成为未来网络信息检索的标准。

9.4.3.3 STARTS

STARTS 全称是“Standard Protocol Proposal for Internet Retrieval and Search”。此标准正在酝酿中，其目标是在 HTTP 协议的基础上建立一个简单而高效的信息检索协议，它直接支持元数据，并允许对结果集排序（用户定义标准）及合并。

9.4.3.4 ZOBRA

ZOBRA 是 Z39.50 和 COBRA 的结合，用 OMG 和 IDL（界面定义语言）来描述“标准可复用信息检索界面”，其功能可覆盖现有的信息检索协议的功能（如 Z39.50）。ZOBRA 致力于设计一个抽象而复杂的数据模型，包括面向记录的、层次的、关系的和面向对象的模型。其目的是隐藏分布环境下信息的不一致性。它还允许服务器支持复杂的信息要求，如关系的连接、并和交。

9.4.3.5 CIP

CIP 全称是公共索引协议（Common Indexing Protocol），是互联网工程任务组（Internet Engineering Task Force，简称 IETF）查询工作小组新的努力成果。CIP 是用于交换索引信息的公共协议，它以先前的协议概念为基础，这些先前的协议包括 whois ＋ ＋ 、X.500（LDAP）和 CCSSO。它还使用 Harvest 的总结对象互换格式（Harvest's Summary Object Interchange Format）的一个版本来构造索引文件格式。此标准尚处于婴儿期，但是很有发展前途，值得研究和遵循。

网络信息检索标准的易用性和灵活性十分重要。网络信息检索标准准则是一个不断发展的、动态的复杂系统。应进一步简化检索标准，使得其更加灵活和可扩展。一个系统是否灵活，最本质的特征就是它是否能够适应变化。对于网络信息检索标准的应用来说，灵活性意味着能够很快并且很容易扩展。当然，这里所说的灵活性是指在规则允许下的灵活，并且存在一定限度。否则，准则无限灵活的后果将是没有准则，如何处理足够灵活与有限灵活之间的关系，也将是未来网络信息检索发展中所要解决的问题。

同时，标准国际化是网络信息检索标准化发展的必然趋势，特别是我国已经加入世贸组织，参与国际经济大体系，进入国际经济大循环，更需要我们遵循国际标准，采用国际标准。采用国际标准是提高标准质量，加快标准制定速度的捷径，我们应认真贯彻执行。与此同时，也应认识到标准的制定不可能一蹴而就，是一个需要不断完善的过程，标准的实施需要各方的共同努力，长期坚持才能达到理想的境界。

【案例】

情报学网络信息资源的分布与检索

随着互联网的普及，网络信息资源的不断增多，有关情报学的网络信息资源成为情报学研究的重要信息源。情报学网络信息资源大致分布在以下几个方面：

(1) 基于网络平台的商用电子信息资源。即正式电子出版物，是情报学研究者从事科研工作利用的主要资源，包括参考数据库、全文数据库、事实数据库、电子期刊、电子图书等。参考数据库如《全国报刊索引数据库》、《中文社会科学引文索引》(CSSCI) 等，全文数据库如《中国期刊网全文数据库》等，事实数据库如万方数据资源系统的《商务信息子系统》等，电子期刊包括收录于数据库的全文期刊和期刊网站等，电子图书如超星数字图书馆、书生之家数字图书馆、方正 Apabi 数字图书馆等。

(2) 重要的学术网站。比如图书情报学科信息门户 (http://www.tsg.cn)，由国家科学数字图书馆项目管理中心资助建设。它对在互联网上可以直接查询到的国内外各学科领域、各类型的重要图书情报系统及其馆藏资源进行搜集、评价、分类、组织和有序化整理、揭示，形成合理的分类组织与浏览体系，为科研人员提供查询各学科领域、各类型信息资源的捷径和方法。

(3) 搜索引擎和网络资源目录。比如 Google、百度、Yahoo! 等。其中，Yahoo! 是典型的网络资源目录，而 Google、百度作为综合性的搜索引擎在专业性方面也具有特色，利用 Google 的学术搜索和图书搜索，以及百度的图书搜索，

都可以查找到与情报学有关的网络信息资源。

关键术语

网络信息资源　　网络信息资源概念　　网络信息资源特点
网络信息资源类型　　网络信息资源分布　　网络信息检索原理
网络信息检索方法　　网络信息检索标准　　网络信息检索发展趋势

思考题

1. 和传统信息资源相比，网络信息资源有何特点？
2. 网络信息资源有哪些类型，其分布特征体现在哪些方面？
3. 简要分析网络信息检索的原理与方法。
4. 与传统信息检索相比，网络信息检索体现出哪些优势？
5. 简述网络信息资源检索服务的发展趋势。

CHAPTER TEN

第10章 网络信息检索工具

【本章要点】

- ◇介绍网络信息检索工具的发展和类型
- ◇介绍搜索引擎的发展历程
- ◇分析搜索引擎的结构及工作原理
- ◇介绍了国内外主要搜索引擎
- ◇探讨搜索引擎的发展趋势
- ◇分析网络资源目录的原理及特点
- ◇介绍国内外主要的网络资源目录
- ◇分析元搜索引擎的原理和特点
- ◇探讨元搜索引擎的技术和评价方法
- ◇介绍国外主要的元搜索引擎

引子

网络信息浩如烟海，并且增长迅速、类型多样，要想从中快速找到自己所需要的信息，必须借助于网络信息检索工具。不同的网络信息检索工具，其特点、原理和作用是不相同的，因而有不同的适用范围，例如，网络资源目录型检索工具更适合于没有明确检索需求的用户，而搜索引擎则更适合于较能明确表达检索需求的用户。即使同一类型的网络信息检索工具，其在收录信息的范围、检索语种、检索原理、检索功能等方面也存在着或多或少的区别，使得即使针对相同的

检索需求，不同的检索工具所达到的效果可能存在着差别。用户要想提高检索效果、降低检索成本、优化检索满意度，首先需要对网络信息检索工具有较为清晰而全面的了解。

10.1 网络信息检索工具的发展和类型

10.1.1 网络信息检索工具的发展

网络信息检索工具是指在互联网上提供信息检索服务的信息检索系统。如搜寻FTP资源的Archie，检索Gopher网站资源的Veronica和Jughead等，近年来广为流行的Yahoo、Alta Vista、Lycos等Web检索工具等。网络信息检索工具的检索对象是存在于互联网信息空间的各种类型的网络信息资源。

在互联网的发展过程中，先后产生了Archie、Gopher、WAIS及Search Engine等检索工具。

1. Archie（文档查询服务）

Archie最早是由加拿大Mcgill大学计算机学院的学生和志愿者开发的一种检索工具。主要是查询FTP资源，是一种基于文件名的信息查询工具。用户可以通过特定文件名或文件说明中出现的字符串进行查询，文档查询服务器定期运行文档索引程序，自动对它所知道的所有FTP服务器进行遍历，以获得这些服务器上更新的全部文件目录，并对全部文件目录编制索引。

2. Gopher（基于菜单的信息检索服务）

1991年，美国明尼苏达大学开发了Gopher，后来又出现了两种配合Gopher的软件工具：Veronica和Jughead。它们已成为Gopher服务器提供的一种标准服务。用户通过逐层展开菜单，对互联网上的远程信息系统进行浏览。

3. WAIS（基于关键词的文档检索服务）

WAIS（Wide Area Information Servers）由思想机器公司（Thinking Machines Corporation）于1991年推出。WAIS以互联网上各种文本数据库作为检索对象，采用自然语言关键词全文检索方法。首先从数据源列表中选择检索对象，然后在选定的数据源范围内进行关键词检索。检索结果按相关度权数大小排序。这样，用户可以选择相关度较高的信息加以浏览，节省时间。

4. Search Engine（基于超文本的搜索引擎服务）

Search Engine是基于万维网（WWW）的检索工具。1992年万维网推出之

后很快成为发展最快和应用最广的信息服务，随着万维网的流行，各种类型的搜索引擎迅速发展起来。1994 年 4 月，互联网上诞生了第一个搜索引擎 WebCrawler。

10.1.2　网络信息检索工具的类型

网络信息检索工具按照不同的划分方法，可分为许多不同的类型。

1. 按索引方式划分，可分为目录型检索工具和索引型检索工具

目录型检索工具，又称网络资源目录或主题指南，主要采用人工或机器搜索信息，由人工对搜集的信息进行甄别、分类、加工，建立分类导航或分类编排网站目录，提供分类浏览的工具。这类检索工具如 Sohu、Yahoo! 等。

索引型检索工具，又称搜索引擎（Search Engine），主要采用搜索软件自动搜索信息，建立网页信息索引库，提供全文检索，用户在检索框中输入关键词或词组进行检索。这类检索工具如 Google、Alta Vista、天网等，目前的搜索引擎从功能上和检索效果上都在努力接近传统大型商业性联机检索系统，已逐渐成为网络信息检索的主要工具。

随着网络检索工具的发展，现在的网络检索工具大多都提供分类检索和关键词检索两种方式，只不过各检索工具的侧重点不同，因此，目录型检索工具和索引型检索工具的界限也越来越模糊，大多数流行的网络检索工具同时提供两种方式的检索，既提供主题指南又有索引功能的混合型检索工具是当今网络检索工具发展的主流和趋势。

2. 按检索时搜索的检索工具数量划分，可以分为独立型检索工具和集合型检索工具

独立型检索工具也称单一检索工具，它拥有自己的索引数据库，提供基于自身索引库的查询服务，如 Sohu、Yahoo!、Google 等。

集合型检索工具是多个独立型检索工具的组合，也称多元搜索引擎、元搜索引擎，通过集合型检索工具，可以同时利用多个网络检索工具进行网络信息查询。如 Dogpile、万纬、Profusion 等。

3. 按检索网络资源的类型划分，可分为 Web 资源检索工具和非 Web 资源检索工具

所谓 Web 检索工具是指利用超文本（或超媒体）技术在互联网上建立的一种提供网上信息资源导航、检索服务的专门 Web 服务器或网站。由于目前以超文本技术建立起来的 Web 已成为互联网信息的主流形式，而且 Web 检索工具既以 Web 形式提供服务，又以 Web 资源为主导检索对象，检索范围还涉及其他网络资源形式，如 Usenet、Gopher、FTP 等。因此，Web 检索工具已成为我们获

取互联网信息资源的主要检索工具和手段，几乎成了网络检索工具的代名词。本课程所涉及的网络检索工具几乎都是这种类型。

非 Web 资源检索工具是查找网上非 Web 资源，主要包括 FTP、Gopher、Usenet、Telnet 等资源，如查找 Usenet 新闻组的 Deja News、查找 FTP 资源的 Archie、检索电子邮件列表的 Liszt 等。这一类检索工具随着万维网的发展，作用有所减弱。

在下面的章节中，我们将介绍搜索引擎、网络资源目录和元搜索引擎 3 种主要的网络信息检索工具。

10.2 搜索引擎

10.2.1 搜索引擎概述

10.2.1.1 搜索引擎的概念

搜索引擎来自英文“Search Engine”，意为信息查找的发动机，是最为常用的网络资源搜索工具之一。关于搜索引擎的定义有广义和狭义之分。

广义的搜索引擎泛指网络上提供信息检索服务的工具和系统，是网络检索工具的统称。广义的搜索引擎包括三种类型：（1）目录式搜索引擎（Directory Search Engine），即网络资源目录，又称目录型检索工具。主要通过人工发现信息，依靠编目员的知识进行甄别和分类，用户在分类结构中进行浏览和查询信息。如 Yahoo!、搜狐等。（2）基于机器人技术的搜索引擎（Robot Search Engine），主要采用自动搜索和标引方式来建立和维护其索引数据库，用户查询时可以用逻辑组合方式输入各种关键词，搜索引擎通过特定的检索软件，查找其索引数据库，给出与检索式相匹配的检索结果，供用户浏览利用。如 Alta Vista、Google 和天网等。（3）元搜索引擎（Meta Search Engine），即集合型检索工具，主要通过调用多个独立搜索引擎的检索功能来实现互联网资源的查询。目前，一些学者采用了广义的搜索引擎定义，将搜索引擎看作是 WWW 检索工具的代名词。

狭义的搜索引擎主要指利用自动搜索技术软件，对互联网（主要是万维网）资源进行搜集、组织并提供检索的信息服务系统。即广义的搜索引擎的第 2 种类型。互联网上最早出现的搜索引擎就是利用 Robot 来建立数据库，“搜索引擎”这个词的原义也只是指这种狭义上的基于 Robot 的搜索引擎。本章采用狭义的搜

索引擎的定义。

10.2.1.2 国内外搜索引擎的发展历程

1990年加拿大蒙特利尔大学的学生艾伦·伊米杰等人开发了一个可以通过文件名来查找文件的程序Archie，它是第一个自动搜索网络上匿名FTP网站的程序，然后对其进行索引，其原理与现在的搜索引擎很相似，但它还不是真正的搜索引擎。因此，Archie被称为现代搜索引擎的雏形。Archie是一个可搜索的FTP文件名列表，用户必须输入精确的文件名搜索，然后Archie会告诉用户哪一个FTP地址可以下载该文件。由于Archie能够在一定程度上满足用户在网络中传输大量文件的需要，在当时大受欢迎。计算机机器人（Computer Robot）是指某个能以人类无法达到的速度不断重复执行某项任务的自动程序。由于专门用于检索信息的Robot程序像蜘蛛（Spider）一样在网络间爬来爬去，因此，搜索引擎的Robot程序被称为Spider程序。1993年6月，美国麻省理工学院的马太·杰瑞开发出了世界上第一个Robot程序：World Wide Web Wanderer，用于追踪互联网发展规模。刚开始它只用来统计互联网上的服务器数量，后来则发展为也能够捕获网址。随后的10月份，马丁·科斯特创建了ALIWEB，它是Archie的HTTP版本，不过ALIWEB不使用“机器人”程序，而是靠网站主动提交信息来建立自己的链接索引，类似于现在大家所熟知的Yahoo!。

随着互联网的迅速发展，检索所有新出现的网页变得越来越困难，因此，在Wanderer基础上，一些编程者将传统的“蜘蛛”程序工作原理作了些改进。其设想是，既然所有网页都可能有连向其他网站的链接，那么从跟踪一个网站的链接开始，就有可能检索整个互联网。

1994年4月，斯坦福大学的两名博士生，美籍华人杨致远和大卫·费罗（David Filo）共同创办了超级目录索引Yahoo!。随着访问量和收录链接数的增长，Yahoo!目录开始支持简单的数据库搜索。因为Yahoo!的数据是手工输入的，所以不能真正被归为搜索引擎，事实上只是一个可搜索的目录。Yahoo!中收录的网站，因为都附有简介信息，所以搜索效率明显提高。不过Wanderer只抓取网址，而网址信息含量太小，很多信息难以单靠网址说清楚，搜索效率很低。

1994年7月，卡内基梅隆大学的迈克尔·莫尔丁（Michael Mauldin）将Spider程序接入到其索引程序中，创建了Lycos。Lycos被称为是真正现代意义上的搜索引擎。Lycos能够对搜索结果进行相关性排序，并且还提供了前缀匹配和字符相近限制。

1998年10月之前，Google只是斯坦福大学的一个小项目BackRub。1999年2月，Google完成了从Alpha版到Beta版的蜕变。Google以网页级别为基

础，判断网页的重要性，使得搜索结果的相关性大大增强。2006 年 4 月，Google 宣布其中文名称“谷歌”，这是 Google 第一个在非英语国家起的名字。

由于汉字处理的特殊性，中文信息资源站点正处于发展时期，中文搜索引擎的出现相对较晚，世界上首个中文万维网搜索引擎 Goyoyo 于 1997 年在香港问世，随后中文搜索引擎才陆续在网上发布。1998 年 2 月，中国人自己的搜索引擎“搜狐”问世，“出门靠地图，上网找搜狐”开始了中国互联网的门户时代。

2000 年 1 月，两位北京大学校友，超链分析专利发明人、前 Infoseek 资深工程师李彦宏与好友徐勇在北京中关村创立了百度（Baidu）公司。2001 年 8 月发布 Baidu. com 搜索引擎 Beta 版（此前 Baidu 只为其他门户网站搜狐、新浪、Tom 等提供搜索引擎），2001 年 10 月 22 日正式发布 Baidu 搜索引擎，专注于中文搜索。2003 年 12 月 23 日，原慧聪搜索正式独立运作，成立了中国搜索。2004 年 2 月，中国搜索发布桌面搜索引擎网络猪 1.0，2006 年 3 月，中搜将网络猪更名为 IG（Internet Gateway）。

根据搜索引擎不同时期的研究重点和服务性能，可以将以上搜索引擎的发展分为三个阶段。

第一阶段起始于 1994 年，以 Yahoo!、Alta Vista 和 Infoseek 为代表。这个时期的搜索引擎一般索引都少于 100 万个网页，一般不重新搜集网页并刷新索引，而且其检索速度非常慢。在实现技术上也基本沿用较为成熟的传统检索技术，相当于利用一些已有的技术实现在互联网上的信息检索。

第二阶段起始于 1998 年，以 Google 为代表。处于这个阶段的搜索引擎大多采用分布式方案来提高数据库规模、响应速度和用户数量，并且只专注于做后台技术的提供者，在服务模式上不断创新，竞价排名和图形图像以及 MP3 的搜索引擎便是这个阶段的产物。

第三阶段是起始于 2000 年左右，也是当前搜索引擎空前繁荣的时期，以 Google、Baidu、Yahoo! 等搜索引擎为代表。这一时期搜索引擎的主要特点是：（1）索引数据库的规模大，一般的商业搜索引擎都保持在几千万甚至上亿个网页。（2）除了一般意义上的搜索外，开始出题主题搜索和地域搜索。（3）能够实现一定程度上的智能化、可视化检索。（4）由于搜索返回数据量过大，检索结果相关度评价成为研究的焦点。这一阶段的发展为搜索引擎拓展了生存空间，同时提高了搜索的质量和效率。

10.2.1.3 搜索引擎的结构

搜索引擎一般主要由搜索器（Crawler）、索引器（Indexer）、检索器（Searcher）和用户接口（User Interface）四部分构成。

1. 搜索器

搜索器本质上是一种计算机爬虫程序，其功能是发现和搜集互联网的信息。一个搜索引擎一般会有多个 Spider 或 Robot，并且会日夜不停地运行，以尽可能多和快地搜集各种类型的信息。搜索器还需要定期更新已经搜集过的信息，以尽量减少甚至避免死链接和无效链接。

搜索器在具体的网页搜集过程中，可以有两种方式。最常见的方式是“抓取”，即将 Web 上的网页集合看成是一个有向图，搜集过程从给定起始 URL 集合开始，沿着网页中的链接，按照深、宽或者某种别的策略遍历，不停地从网页集合中移除 URL，下载相应的网页，解析出网页中的超链接 URL，看是否已经被访问过，将未访问过的那些 URL 加入集合。另外一种可能的方式是在第一次全面网页搜集后，系统维护相应 URL 网页集合，往后的搜集直接基于这个集合。每搜到一个网页，如果它发生变化并含有新 URL，则将它们对应的网页也抓回来，并将这些新 URL 也放到集合中。如果集合中某个 URL 对应的网页不存在了，则将它从集合中删除。这种方式也可以看成是一种极端的宽度优先搜索，即第一层是一个很大的集合，往下最多只延伸一层。还有一种方法是让网站拥有者主动向搜索引擎提交它们的网址，系统在一定时间内（2 天到数月不等）定向向那些网站派出“蜘蛛”程序，扫描该网站的所有网页并将有关信息存入数据库中。大型商业搜索引擎一般都提供这种功能。

2. 索引器

索引器的功能是对搜索器所搜集来的信息进行分析和理解，从中抽取索引项，用于表示文档以及生成文档库的索引表，形成索引数据库。索引数据库中每一条记录基本上对应一个网页，原则上包括 URL、标题、关键词、文档摘要等信息。由于各个搜索引擎中的索引器理解和抽取信息方式的不同，因此其索引表一般也不同。索引表一般使用某种形式的倒排表，即由索引项可以立即查找到相应的网页。

索引器从网页中抽取索引项的基本依据是关键词的词频，即在略去只起语法作用的高频词后，一个关键词在文件中出现的频率越高，则它代表该文件主题的程度就越大，从而作为索引项的准确性也就越高。当然，功能强大的索引器一般还需要利用其他信息来分析关键词，如分析关键词在网页中的位置、所使用的字体大小和型号、关键词之间的相邻位置等。最后，索引器需要对所选取的索引项赋予一个权值，以表示该索引项对文档的重要程度，同时也用来计算查询结果的相关度。

由于互联网上的信息更新速度很快，因此，索引数据库中的索引表需要动态更新，需要对其进行添加、修改、删除等处理，以保证索引数据库能尽可能准确地反映当前网络上的信息状态。索引数据库是用户进行检索的基础，它的质量直接影响到用户的检索效果。

3. 检索器

检索器的功能是对用户的检索请求进行分析，将其分解为一个或多个关键词，并转换成计算机可识别的规范检索式，然后在索引数据库中进行匹配，进行文档与查询的相关度评价，对将要输出的结果按匹配程度的高低进行排序，并实现某种用户相关性反馈机制。

用关键词及其组合来表达用户的查询请求来进行检索，然后检索器输出包含有相关检索词的文档，是目前搜索引擎提供的主流查询模式，而这还停留在字面上的比较和匹配。由于自然语言中的主题词存在着大量的同形异义、异形同义的现象，仅靠关键词进行检索显然会影响到用户的查全率和查准率，并且会浪费用户的检索时间。基于概念的检索是解决这一问题的关键，并且有些搜索引擎已开始在关键词检索的基础上引入了概念检索，如 Excite，它在搜索时不只搜索用户输入的关键字，还可智能性地推断用户要查找的相关内容进行搜索。

4. 用户接口

用户接口的主要作用是输入用户检索请求、显示用户查询检索结果和提供用户的相关性反馈机制。一般搜索引擎的用户接口都提供一般检索和高级检索。在一般检索查询接口里，用户可以输入一个或多个关键词，高级检索接口则能够执行布尔逻辑检索、截词检索、位置检索、相似性检索等。许多搜索引擎还提供一些过滤功能，例如可以对日期、文件格式、区域、语言等进行限制。由于不同的搜索引擎特色不同，提供的功能一般也不太相同，因此，不同的搜索引擎所设计的用户接口也是不同的。不过使用人机交互的理论和方法来设计和实现用户接口，以方便用户使用搜索引擎，并且高效率、多途径地从搜索引擎查询到所需的信息，是所有用户接口设计的原则。

用户查询结果的显示主要需要考虑结果的显示内容和排序方式两个方面。大部分搜索引擎显示的检索结果内容主要有：标题、URL、文档摘要、相关搜索等。比较智能化的搜索引擎则能对搜索结果进行自动聚类，形成若干个簇，每个簇内的信息内容相关性强，而簇与簇之间的相关性弱，这样用户只需要考虑那些与检索需求相关的簇，大大缩小了所要浏览的结果数量。国外有些搜索引擎在此

基础上还提供了检索结果大纲视图或地图视图等，如 iboogie、grokker 等。

搜索引擎返回给用户的是一个和用户查询请求相关的结果列表，这个列表所包含的记录一般都在上百条以上，有时甚至达到上万条，用户无法全部浏览。通过对检索结果进行相关度排序，将相关度高的文档放到返回队列的前面，显然可以方便用户浏览，并且可以提高其检索效果。目前的相关度排序基本上采用的是基于文档内容的方法，即考虑用户的检索请求在文档中出现的情况，如检索请求出现的频率、位置、字体大小等。这种方法有一定的局限性，相关度高的文档有时候不一定是用户想要的信息，例如有些网站为了提高相关度，故意在文档中重复出现某些关键词等。Google 认为一个网页的重要性程度取决于被其他网页链接的数量，采用的是 PageRank 算法来进行检索结果排序。DirectHit 则认为多数人访问的网站是最重要的网站，这种由网络大众集体确认网站重要性的方法具有客观性和公正性。这两种排序方法都能在一定程度上优化基于文档内容的排序方法。不过搜索引擎为了改善检索结果的排序效果，一般都采用了多种排序算法。例如，Google 现在对查询结果进行排序时并不仅仅考虑 PageRank，还考虑到关键词在网页中出现的位置、字体大小和型号、关键词之间的相邻关系等因素。

10.2.1.4 搜索引擎的工作原理

搜索引擎的工作原理如图 10—1。首先，搜索器根据一定的搜集策略抓取互联网上的网页，然后由索引器对搜集回来的网页信息进行分析，抽取索引项，用于表示文档以及生成文档库的索引表，形成索引数据库。用户通过检索接口输入相关的查询请求，并对用户的查询请求进行分析和转换，由检索器在索引数据库中进行查找和匹配，最后将符合要求的文档按相关性程度的高低进行排序，形成结果列表，并通过用户接口将检索结果列表返回给用户。

由以上搜索引擎的工作原理可以看出，搜索引擎的工作过程构成了一个典型的、双层的 C/S 服务模式。当用户访问搜索引擎时，用户端是客户端，向搜索引擎发送检索请求，搜索引擎充当服务器，将符合用户请求的检索结果以应答的形式返回给用户。因此，搜索引擎的用户检索过程是一层 C/S 模式。当搜索引擎抓取网页时，搜索引擎可以被看作是客户端，向互联网的各 Web 站点发送搜索请求，互联网的各种网络资源则是服务器，将相关网页作为应答返回给搜索引擎。因此，搜索引擎的数据搜集过程也是一层 C/S 服务模式。

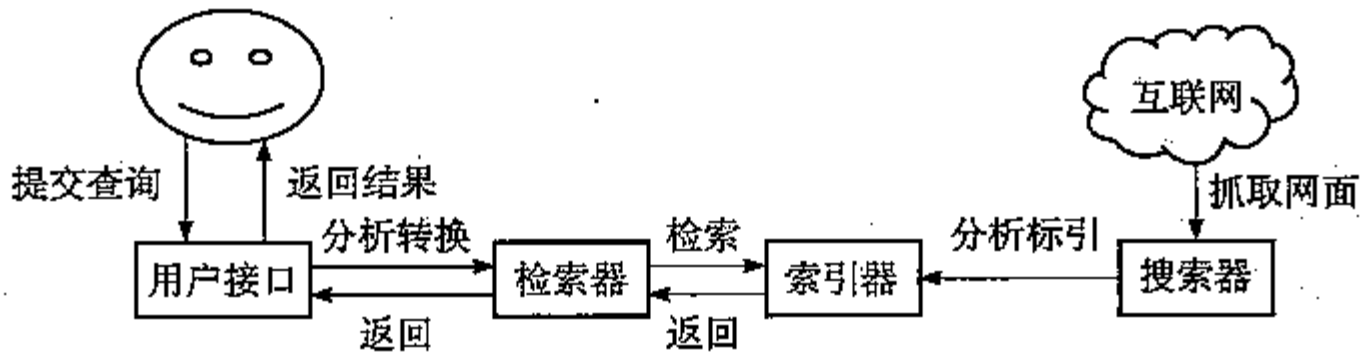

图 10—1　搜索引擎的工作原理

10.2.2　搜索引擎的特点及功能

10.2.2.1　搜索引擎的特点

互联网上的搜索引擎有很多，各有各的风格，有的以查询速度快见长，有的以数据库容量大占优，与传统信息检索工具和其他类型检索工具相比，搜索引擎具有以下优点：

1. 支持全文检索

搜索引擎的出现大大推动了全文检索技术的发展，全文检索技术是搜索引擎的核心。当全文搜索引擎遇到一个网站时，会将该网站上的文章（网页）获取下来，并收入到引擎的数据库中。只要用户输入查询的“关键字”在引擎库的某篇文章中出现过，则这篇文章就会作为匹配结果返回给用户。从这点上看，全文搜索真正提供了用户对互联网上所有信息资源进行检索的手段，给用户以最全面最广泛的搜索结果。

2. 检索功能较为全面，检索方法多样

多数搜索引擎都支持布尔逻辑检索、截词检索、位置检索、字段限定检索等等。不仅能输入单词、词组或句子进行检索，还能指定多个单词之间的逻辑组配及其位置关系；不仅能以词语查询有关主题的页面信息，也能以特定的域名、主机名、URL 等查找有关信息；此外，还可以对被检索文献发表的语种、日期等进行限制。

3. 检索结果按相关性排序

搜索引擎在按照用户要求找到检索结果以后，都会根据自身系统的设定，对检索结果进行相关性排序，然后输出给用户，并将最相关的检索结果排在最前面。相关度的计算一般依据关键词出现的频率、关键词出现在网页的位置、网页被链接的程度等标准来确定。

4. 查询速度快，维护更新及时

搜索引擎是一种自动搜索技术，数据库的容量虽然很大，但搜索的速度还是比较快，更新也非常及时。搜索引擎都具有对索引数据进行动态维护的功能，例如，针对不断更新内容的网页和不断变更的网页地址，对索引数据进行及时更新、添加、删除等处理，以保证索引数据库准确反映网络信息资源的当前状况。因而，在很大程度上保证了它的查询速度和内容更新。

5. 支持关键词检索和分类目录浏览检索

多数搜索引擎在提供关键词检索的同时，或自己设置分类主题目录，或直接采用其他的网络资源目录，支持关键词检索与目录浏览的并行，实现两者的结合。

搜索引擎虽然具有许多优势，但在信息维护、信息重复、网络及站点负载等方面还存在很多的不足。首先，网络信息覆盖范围有限，目前还没有一种能够覆盖整个国际互联网信息资源的搜索引擎，美国《科学》杂志一份研究报告表明，即使功能最完善的搜索引擎，也只能搜索 Web 上大约 1/3 的网页。第二，搜索引擎虽然索引数据库庞大，但检索效果不尽如人意，检索功能尚待改善，检索的查全率、查准率不高。第三，搜索引擎对网络信息的组织与标引缺乏控制，各搜索引擎都有自己的信息收集方式、检索算法和结果排序方法，使得信息的组织没有统一的规范，加上索引方式也不尽相同，也给用户利用搜索引擎带来了一定的不便。

10.2.2.2　搜索引擎的功能

搜索引擎产生和发展的历史虽然不长，但它的功能却非常强大，搜索引擎的检索实际上也是一种数据库检索，几乎可以提供一般数据库的全部检索功能，如布尔逻辑检索、词组检索、截词检索、字段检索、限制检索、位置检索和自然语言检索等。但是，并非每一种搜索引擎均能提供全部的检索功能。

1. 搜索引擎的基本检索功能

(1) 布尔逻辑检索。

布尔逻辑检索是网络信息资源检索中应用最为广泛的检索功能。但常见的三种布尔逻辑符号 AND、OR 和 NOT，应用于具体的搜索引擎的表现方式有所不同。有的只允许使用大写的“AND”、“OR”、“NOT”运算符，有的大小写通用，有的将逻辑符号用“&”、“|”、“!”符号表示，有的不支持或仅支持其中的某个算符，等等。

从理论上讲，OR 的应用会导致大量的检索结果，但由于搜索引擎都在尽力将最相关的结果排在最前面，因而，也基本上不会因为 OR 的使用，而带来结果

选择上的麻烦。

布尔逻辑检索在搜索引擎中还有其他的表现符号和实现方式，比如，有些检索工具完全省略了任何符号和关系，直接把布尔逻辑关系隐含在菜单之中。以“match all terms”表示布尔关系 AND，以“match any term”表示布尔关系 OR；或者以“必须包括”（MUST contain）表示布尔关系 AND，用“一定不含”（MUST NOT contain）表示布尔关系 NOT。

（2）词组检索。

词组检索，也称为短语检索，或字符串检索。它是将一个词组或短语用双引号（“”）括起作为一个独立运算单元，进行严格匹配，以提高检索准确度的一种方法。几乎所有的搜索引擎都支持词组检索。例如，以“信息检索”作为提问关键词检索时，检索结果则仅反馈与“信息检索”完全匹配的内容。

（3）截词检索。

在搜索引擎中，右截断采用的比较多，在中文里，也称为前方一致检索。截词符多采用通配符“*”，可以用它代表多个字符。如 comput * 代表 computer、computing、computerized、computerization 等。绝大多数搜索引擎都支持截词检索，但对于每个具体的搜索引擎的截断方式，截词符的表示方法也不完全一样。

（4）位置检索。

在各个搜索引擎中，所设置的位置算符的表示方法不尽相同，比如，有的搜索引擎用“（nW）”和“（nN）”这两个关系。（nW）关系要求它所连接的两个检索词在结果中相互距离不超过 n 个词，而且前后顺序不能颠倒。（nN）关系也要求它所连接的两个检索词在结果中相互距离不超过 n 个词，但前后顺序可以变换。

（5）字段检索。

字段检索也是搜索引擎检索过程中经常使用的功能之一。比如，把查询的网络信息的检索范围限制在标题（title）、统一资源定位地址（URL）或超链（link）等部分。搜索引擎常用的字段有：Title/t（表示检索词或检索式要出现在标题中）、Subject（表示查询的信息要包含在主题字段中）、Text（表示查询文本中包含检索式的页面）、Host（表示在指定的服务器上查找网络信息）、URL/u（表示查找 URL 中包含检索提问式的页面）、Link（表示查找含有链接至 URL 的页面）、Domain（表示查找指定域名的页面）等。如，可以在 Google 的检索框中输入“电子商务 filetype：doc”，表示要查找一些和“电子商务”相关的 Word 文档。

（6）区分大小写检索。

主要是针对英文的搜索引擎。大写的英文表示专有名称、人名或地名，小写的英文则表示普通名词。例如，“Web”专指万维网，而“web”表示蜘蛛网。有一些搜索引擎提供了区分大小写的检索，以便于进一步提高查准率，尤其是有助于对专有名词的查询。

2. 搜索引擎的其他检索功能

随着网络技术发展的日新月异，搜索引擎的技术也不断地向前发展。除以上提到的一些搜索引擎的基本检索功能外，还发展了一些其他的检索功能。

（1）自然语言检索。

自然语言检索就是一种直接采用自然语言中的字、词甚至整个句子作为提问式进行检索的方法。也就是说，可以直接用“What is the weather in London?”这样的自然语言表达式作为检索提问式。目前，还出现了自然语言智能答询，比如，输入“How can I kill virus of computer?”搜索引擎在对提问进行结构和内容的分析之后，或直接给出提问的答案，或引导用户从几个可选择的问题中进行再选择。

（2）多语种检索。

现在越来越多的搜索引擎开始具有多语种检索的功能，用户可以选择限制检索结果的语言。在 Alta Vista、HotBot、Excite、Infoseek 中都提供这种检索功能，Yahoo! 在中国建立了“中文雅虎”网站，Google 也可以选择检索多种语言。其中 Alta Vista、Infoseek 等还提供检索结果的自动翻译服务。翻译的语种只有西文，如英文、法文、德文、西班牙文等等，实现了英文和其他几种西文的对译，虽然翻译的质量有待于进一步提高，但毕竟为用户利用网络信息资源提供了方便。

（3）概念检索。

概念检索是指使用某一检索提问词进行检索时，能同时对该词的同义词、近义词等进行检索，以达到扩大检索、避免漏检的目的。概念检索在很大程度上可以提高查全率，不只简单地查找含有要查找的单词的文档，同时还可以搜索出同要查找的概念相关的文档。例如，输入“intellectual property right”，不但能检索出那些包含有上述词组的网页，而且还能检索出包含“copy right”等词组的网页。

（4）过滤检索。

过滤检索是指在检索中自动将一些网络信息过滤掉，比如一些内容不健康的黄色网站信息，影响国家安全的政治反动网站信息等。这种检索服务技术受

到父母们的欢迎，可以避免孩子们上网时受到不健康信息的影响。比如：Lycos就采取了过滤技术来避免垃圾信息进入检索结果，它的“Search guard”功能基本上可以有效地过滤污染信息。信息过滤可以帮助用户处理大量的信息，对动态的信息流进行筛选，着重于排除用户不希望得到的信息，提高网络信息检索的效率。

10.2.3 主要搜索引擎介绍

10.2.3.1 Alta Vista（http：//www. altavista. com）

Alta Vista（图10—2）是由数字设备公司1995年12月开发的，寓意为“高瞻远瞩”（“a view from above” or “high view”），初衷是通过对整个Web做全文索引，来展示数字设备公司的alpha服务器的强大。其出色的性能，令Alta Vista迅速超越Lycos和Excite，1996年便成为Yahoo！的关键词搜索合作伙伴，一跃为当时最大的基于Robot/Spider的自动搜索引擎。Alta Vista有不少技术创新，已经获得61项与检索有关的专利。它是第一个多语种的、支持非拉丁语言的搜索引擎，像汉语、韩语、日语；最早免费提供多语种的机器翻译；最早尝试汉语的BG码和Big5码即时转换。Alta Vista的高级搜索，为有经验的搜索用户，赋予了多种可以灵活控制搜索范围和水平的限制手段。与Virage合作推出的声频、图片、视频搜索，也是最出色的多媒体搜索，并有过滤的选择。尽管Alta Vista现在的地位不如Google，它仍被认为是功能最完善、搜索精度较高的全文搜索引擎之一。2004年，Alta Vista被Yahoo！收购，其数据库被Yahoo！搜索数据库替代。

图10—2 Alta Vista一般查询界面

Alta Vista 是一个功能强大的搜索引擎，它的简单检索是推荐使用的首选方法，支持自然语言检索、截词检索、字段限制检索。高级检索支持包含各种逻辑关系符号和多层次括号的检索式，例如，可以利用 AND、OR、ANDNOT、NEAR、通配符“*”和（）组配而成的检索式，常规及高级检索均允许针对 Title、URL 或特定的域名进行检索。用户还可以在限定的字段（包括 Title、URL、Host、Links、Image 和 Text 等）输入框中填入文字，以此为条件进行搜索，对检索要求进行明确的控制，以提高检索效率。

作为一个优秀的搜索引擎，Alta Vita 具有一系列突出的特点和功能：

1. 检索速度快，搜索的结果比较完全和准确

Alta Vista 的 Web 索引数据库十分巨大，容量超过 200G，目前标引了近 25 亿个 Web 页面，以及 1 万多个新闻组两周内的所有文章，而且运算速度非常快，对于大多数的查询，仅需 1～2 秒钟的响应时间。

2. 检索功能全面

除上面提及的各种检索功能外，还可以进行图片搜索，用关键词找出想要的图片；可以进行多语种检索，它支持 25 种语言的检索；同时，还可以按主题浏览，以目录的方式查询需要的信息，它采用的是 LookSmart 的索引目录；此外，还具有检索结果的翻译功能；Alta Vista 在每条结果后面给出一个“Translate”的链点，允许对检索结果进行即时翻译。

10.2.3.2　Lycos（http：//www.lycos.com）

Lycos（图 10—3）由 Terra Lycos 集团 1995 年开发，是搜索引擎中的元老，是最早提供信息搜索服务的网站之一。Lycos 一词来自拉丁语，意思是“狼蜘蛛”（Wolf Spider），这种蜘蛛的捕食方式不是织网，而是采取主动进攻。该网站也犹如它的名字一般，锐气十足，灵活机智。在全球化的扩张中，Lycos 采取了与当地领先企业合作，迅速打入本地市场的策略。1998 年，Lycos 以日本作为突破口，进入亚洲市场，此后，进军韩国。1999 年 12 月，Lycos 与新加坡电信合资创建 Lycos Asia 公司，通过在新加坡、中国、中国台湾、中国香港、印度以及东南亚等九个国家与地区，11 个 Lycos 各地门户网站的成立与整合，将目标定位为满足每位网络使用者个性化服务需求，进而成为亚洲地区访问人数最多的门户网站之一。2000 年 Lycos 被西班牙网络集团收购，已成为目前西班牙语最大的门户网站，并且 Lycos 现已放弃自己的 Spider 索引数据库，目前搜索结果大部分来自 FAST/AllTheWeb 引擎。

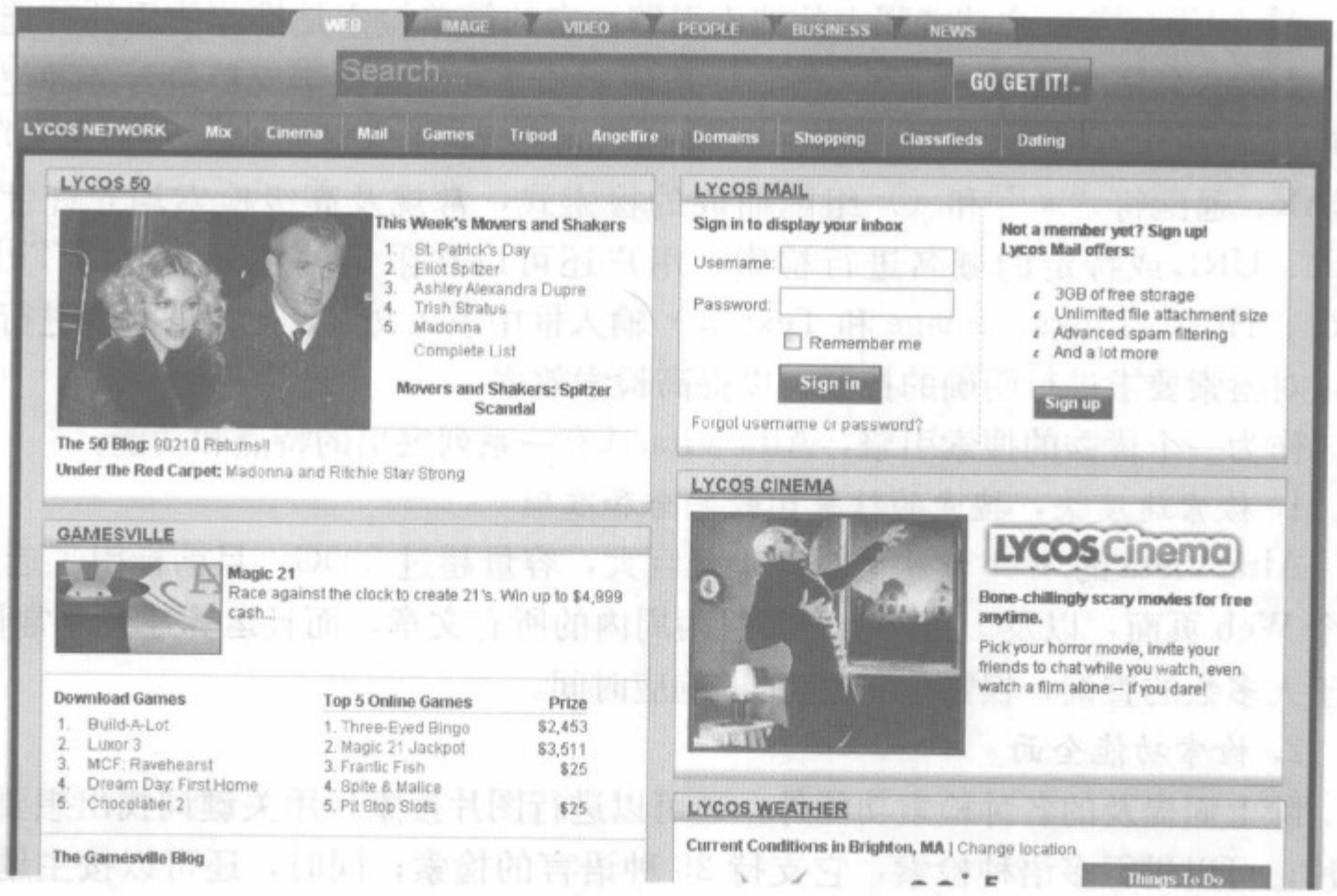

图 10—3　Lycos 主页

Lycos 提供简单检索和高级检索。进行简单检索时，可在检索框中直接输入检索词或检索式。可使用布尔逻辑算符、引号、＋、－、通配符 $ 等。还可以使用连接符（abj），表示两个词相连，但前后顺序不定，例：information abj retrieval，范围符 near：表示两个词之间不能超过 25 个词，如 environment near protection，范围符 far：表示两个词之间可超过 25 个词汇，例：nuclear far agreement，位置符 before：前一词必须出现在第二个词之前，但距离不限。

Lycos 高级检索界面，以“should include”、“must include”和“must not include”来表示布尔逻辑算符“或”、“与”、“非”。并以“in the text”、“in the title”、“in the URL”做了字段限定等等，界面非常友好易读，方便了用户的检索。

检索保护是 Lycos 的特色检索功能，这一功能主要是用来帮助用户在检索结果中过滤掉成人、暴力、仇恨以及与武器相关的内容，使用户可以严格获取特定的领域和链接。可以选择“Disable access to Lycos Chat，Email，Message Board”、“Filter out violent content”、“Filter out hate and racist content”、“Filter out contents about weapons”、“Filter out sexually oriented content”、“Select all of the above”等。

此外，Lycos 也同其他搜索引擎一样，提供目录浏览服务。Lycos 整理了层次分明的分类目录以提供用户检索时参考用，可以更快速地挑选适合的网站内

容。网站的资料来源主要是由 Lycos 搜集，以及网站的自动登录。

10.2.3.3 Google（http：//www.google.com）

Google（图 10—4）是由拉里·佩奇（Larry Page）与塞吉·布林（Sergey Brin）于 1998 年 9 月在美国硅谷创建的高科技公司，他们所设计的 Google 搜索引擎，旨在提供全球最优秀的搜索引擎服务，通过其强大、迅速而方便的搜索引擎，在网上为用户提供准确、详实、符合他们需要的信息。自 2000 年正式开始商业运营以来，目前在全球范围内已拥有了一个正在快速增长的忠实用户群，其中一半以上是国际用户。2000 年 7 月，Google 替代 Inktomi 成为 Yahoo! 公司的搜索引擎。Google 公司不但拥有自身的独立搜索引擎网站，还将其搜索引擎技术售卖给世界上许多公司，1998 年至今，已将其网上搜索技术许可证颁发给 30 多个国家和地区的多家公司。目前有包括 Yahoo!、美国在线、网景和中国的网易等知名网站在内的全球 150 多家公司采用了 Google 搜索引擎技术。Google 非常注重技术创新，并由此获得了多项荣誉，获得了 40 多项世界大奖，如美国《时代》杂志评选的“1999 年度十大网络技术”、《个人电脑》杂志授予的“最佳技术奖”、TheNet 授予的“最佳搜索引擎奖”等等。营销机构 Interbrand 最近公布的用户调查数据显示，Google 已荣登 2002 年最著名品牌排行榜榜首。借助和 America Online、Netscape 和其他公司的合作伙伴关系，它所回应的查询远远多于其他在线服务商，Google 已经成为全球最大的搜索引擎。此外，Google 是阿根廷、澳大利亚、比利时、巴西、加拿大、丹麦、法国、德国、印度、意大利、墨西哥、西班牙、瑞典、瑞士、英国和美国的头号搜索引擎。目前，Google 已有 112 个国际域名。

图 10—4 Google 主页

1. Google 的特点

（1）独树一帜的 PageRank 技术。

PageRank 利用网络自身的超链接结构给所有的网页确定一个重要性的等级，当从网页 A 链接到网页 B 时，Google 就认为“网页 A 投了网页 B 一票”。Google 根据网页的得票数评定其重要性，以此来帮助提高搜索效率。这点和在传统情报检索理论中的引文分析方法颇为相似，即根据引文的数量来确定文献的权威性。除了考虑网页得票数（即链接）的纯数量之外，Google 还要分析为其投票的网页。“重要”网页所投之票自然分量较重，有助于增强其他网页的“重要性”。PageRank 就是要从链接结构中获取网页的重要性，即网页的重要性同时依赖于其他网页的重要性。PageRank 技术根据网页之间的链接结构对网页的重要性进行客观的评价，并将网页的 PageRank 值应用于检索结果的排序。这样，PageRank 技术在很大程度上避免和减少了人为因素，客观地将最恰当的检索结果呈现给用户。随着全球搜索引擎竞争的加剧，Google 对其 PageRank 技术也正在做一些改进。由于某个网页可能不仅仅只有一个主题内容，那么网页的 PageRank 值就不能准确反映网页的所有主题内容。因此，Google 将根据网页的多个主题分别给出几个主题方面的 PageRank 值。在检索结果排序时，将根据检索词的相关主题来参考相应主题的 PageRank 值，这样，网页的 PageRank 值有了“个性化”权值，因此，网页的 PageRank 值可以更准确地服务于检索结果的排序，从而更好地满足用户的检索需求。

（2）强调简单快速，关联性极强。

Google 网站只提供搜索引擎功能，没有花里胡哨的广告。在超过 24 亿网页中搜索问题，而且回复极为相关网页的时间不到 1.5 秒。目前，每天都有数千万用户登录 Google，使用其网上搜索引擎，处理的网页搜索量达到了每秒 2 000 多次，每天超过 1.5 亿次。权威杂志《Wired》的评价很有代表性：“由于简单有效，Google 已成为广大互联网用户的宠儿。”

（3）检索功能全面，易于使用。

Google 提供多种功能的检索，包括目录浏览、简单检索、布尔检索、截词检索、大小写区分、限制检索、词组检索、组合检索、图片搜索、新闻组搜索等。Google 的目录也非常具有特色，它依据世界著名的主题目录“Open Directory”，收录的是网页，收录了来自 150 万个以上网站的网页，同时，Google 根据其专业的“网页级别”（PageRank）技术对目录中登录的网站进行了排序，提高了用户利用该目录的效率。

2. Google 的特色功能

Google 还提供了一些特色的功能，这也成为 Google 吸引用户的重要因素，主要包括网页快照、集成化的工具条、单词的英文解释、手气不错、网页翻译、搜索结果过滤等。

Google 进行网页遍历的时候，会给网页做一份索引快照（Snapshot），并将其存储到 Google 的服务器中。当检索用户并不想访问检索到的网页，只是想大略浏览其内容，或者检索到的网页无法访问或已被删除时，“网页快照”功能可以很好地满足用户要求。用户只需要直接点击“网页快照”，就可以查看 Google 已编入索引的网页的内容，而且经 Google 进行索引处理后，检索词均用不同颜色标明，用户可以对检索出的网页中包含的检索词一目了然。

为了方便搜索者，Google 提供了工具条，集成于 IE 浏览器中，用户无需打开 Google 主页，就可以在工具条内输入关键字进行搜索。此外，工具条还提供了许多其他功能，如显示页面 PageRank 等。最方便的一点在于用户可以快捷地在 Google 主页、目录服务、新闻组搜索、字典、高级搜索和搜索设定之间切换。想安装 Google 的工具条，可以访问 http：//toolbar. google. com，按页面提示可以自动下载并安装。工具条目前只支持 IE5. 0 以上版本。

如果用户想查找某个生词的意思或者想了解某个单词的用法，均可使用在线词典进行查找。例如，想了解“archives”的用法，在检索结果页面会出现“Searched the web of archives”，注意上面句子中，单词 archives 下出现了一个横线，点击这个链接，就跳转到另外一个网站“http：//www. dictionary. com/”，查找到“archives”的解释和说明。

智能化的“手气不错”功能，提供可能最符合要求的网站，省时方便。当用户想浏览一个特定的网站，但是只知道和网站有关的局部信息（如该网站的产品、服务）时，便可通过该局部信息及该网站的其他相关信息来试试“手气不错”功能，Google 将为用户提供一个自认为最准确的检索结果。

Google 提供了网页翻译。Google 正致力于研究网页翻译的强大功能，并已经推出了将意大利语、法语、西班牙语、德语和葡萄牙语等五种语言翻译成英文语言的检索服务测试版本。搜索结果如果是使用以上这五种语言，可以点击“Translate this page”，就可以把该网站翻译成英语。Google 支持多种语言检索。

Google 新设立了成人内容过滤功能，见 Google 的设置页面，http：//www. google. com/preferences，最底下有一个选项 SafeSearch Filtering。不过，中文状态下的 Google 尚没有这个功能。

3. 中文 Google 检索功能分析示例

Google 支持中文搜索，其中文搜索引擎是收集亚洲网站最多的搜索引擎之一，并成为它拓展全球信息市场的重要基础。2000 年 9 月 12 日，Google 公司宣布推出简体及繁体两种中文的版本。Google 的共同创办人兼总裁塞吉·布林表示："每天有上亿的中文使用者在万维网上寻找资讯，我们希望能为他们提供 Google 快速精准的优良服务。"虽然 Google 非中国本土公司，但在国内，使用它的独立搜索引擎的人数正迅猛增长，其搜索引擎技术还受到了中文雅虎、网易等知名门户网站的青睐，采用了其中文互联网服务，这大大提升了 Google 在中国的影响力。中文 Google 已成为用户检索网络信息资源的重要检索工具，具有强大的检索功能。

（1）布尔逻辑组配。Google 可以直接输入关键词进行检索，也可以进行布尔逻辑组配。

逻辑"与"操作，Google 用"＋"或"空格"来表示逻辑"与"操作。例如："电子政务　发展"可以查出同时包含"电子政务"和"发展"两个关键字的全部文档。

逻辑"或"操作，Google 用大写的"OR"表示逻辑"或"操作，且"OR"的两边不得有空格。小写的"or"在查询的时候将被忽略，则操作就变成了一次"与"查询。例如："信息检索 OR 档案检索"可以查找到包括"信息检索"或"档案检索"的网页。

逻辑"非"操作，Google 用"－"来表示逻辑"非"操作。"－"号是英文字符，而不是中文字符"—"。此外，操作符与关键字之间，不能有空格。

混合查询，涉及逻辑操作符的顺序问题。一般而言，搜索引擎按照从左往右的顺序读取操作符号。

（2）限制检索。"site:"表示对搜索的网站进行限制，例如，"档案馆 site:http://www.cctv.com"可以查找出 www.cctv.com 网站上关于"档案馆"的网页；"filetype:"按文件类型搜索文件，可搜索的文件类型包括 Adobe Portable Document Format（PDF）、Microsoft Write（WRI）、Microsoft Word（DOC）、Microsoft Excel（XLS）、Microsoft PowerPoint（PPT）等 12 种；"in url:"和"all in url:"搜索的关键字包含在 URL 链接中；"in title:"和"all in title:"搜索的关键字包含在网页标题中，如"in title: MP3"可以查询到网页标题中含有"MP3"的网页；"link:"搜索所有链接到某个 URL 地址的网页。

（3）图像搜索。Google 还有搜索图像功能，使用 Google 图像搜索可以搜索超过 3.9 亿个图像，因此它获得 2001 Search Engine Watch Awards 的"互联网

上最佳的图像搜索工具”。进入图像搜索界面“http：//images. google. com”，用户可以在搜索框内输入描述图像内容的关键字，如输入“赵薇”，就会搜索到大量的赵薇的图片。同时，还可以采用文件类型限定的方式来搜索图像。即在搜索框中，直接使用“filetype：”来指定文件类型扩展名。例如，如果要查看格式为 .jpg的花朵（flower）的图像，则在搜索框中输入“花 filetype：jpg”。Google 给出的搜索结果是一个个直观的图像缩略图，以及有关该缩略图的简单描述，如图像文件名称以及大小等。然后点击缩略图，就可以看到相关的图像。

（4）新闻组搜索。新闻组是互联网上一种高效的交流方式，在国外，它的使用频率仅次于电子邮件。新闻组有详尽的分类主题，某些主题还有专人管理和编辑，具有大量的有价值的信息。由于新闻组包含的信息实在是海量，因此不利用工具进行检索是不大可能的。DEJA 一直是新闻组搜索引擎中的佼佼者。2001 年 2 月，Google 将 DEJA 收购并提供了所有 DEJA 的功能。现在，除了搜索之外，Google 还支持新闻组的 WEB 方式浏览和张贴功能。

10.2.3.4 Fast/AllTheWeb（http：//www. alltheweb. com）

Fast 总部位于挪威，成立于 1997 年，其技术起源于挪威科技大学（Norwegian University of Science and Technology）的相关研究开发成果。公司全称为 Fast Search & Transfer（FAST）ASA，而 AllTheWeb（图 10—5）是其对外展示技术的窗口网站。AllTheWeb 是当今成长最快的搜索引擎，目前支持 225 种文件格式搜索，其数据库已存有 49 种语言的 31 亿个 Web 文件，与 Google 网页数据库不相上下。而且以其更新速度快，每 7 到 11 天更新一次，搜索精度高而受到广泛关注，被认为是 Google 强有力的竞争对手。

AllTheWeb 属于全文搜索引擎。目前提供一般检索和高级检索功能，还支持新闻检索、FTP 文件搜索、图像检索、视频文件检索等。一般检索支持普通关键词搜索，以及“＋”、“—”、“（ ）”等逻辑命令符号，分别对应 AND、NOT、OR 等布尔逻辑命令，并且可使用引号进行精确匹配搜索（此功能也可通过点选搜索框右侧的“Exact Phrase”实现）。此外，AllTheWeb 引擎还支持以下特殊检索命令：

1. url. tld；domain

表示限定查找顶级域名的网页。比如，“url. tld：cn”意为查找关于中国的网页资料；而“url. tld；com”则表示查找域名后缀为“. com”的商业网站资料。

2. link. all；URL

表示查找链接到某一网页的其他网页。比如，“link. all：www. alltheweb. com”将搜索指向 AllTheWeb 主页的其他网页。

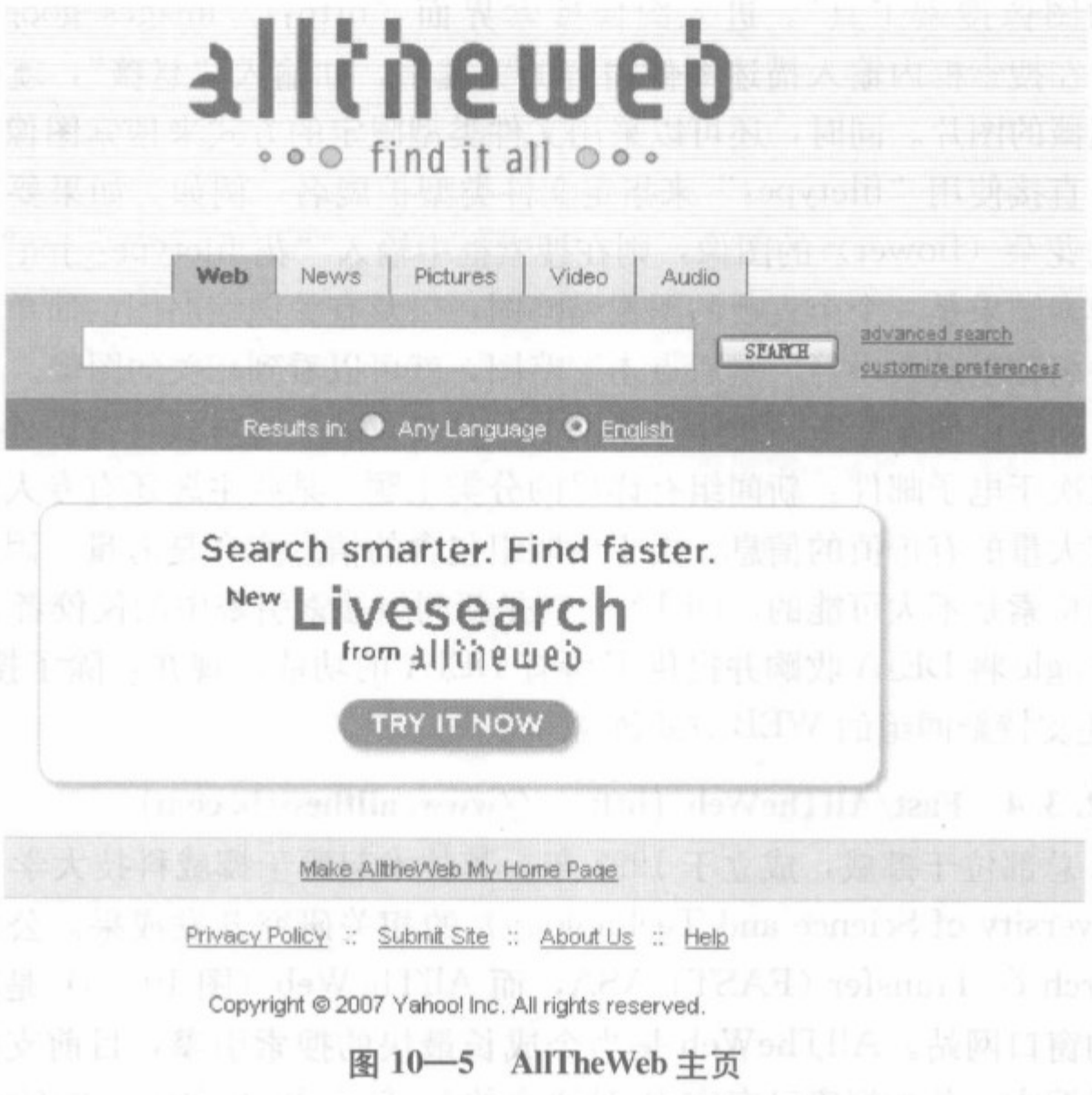

图 10—5　AllTheWeb 主页

3. normal. title：text

搜索网页标题中含有某些特定文字的网页。比如“normal. title：search engine”即为查找所有标题中含“搜索引擎”字样的网页。

4. url. all：text

查询 URL 中含某些特定文字的网页。比如，输入“url. all：alltheweb”时，就可以找到包含“alltheweb”的网页。

AllTheWeb 高级检索提供限定语言、关键词过滤、域名过滤、IP 地址过滤和指定网页大小等高级搜索功能，方便用户进行更精确的查询。

Fast/AllTheWeb 数据库容量大，更新速度快，搜索精度高，是个非常不错的搜索工具。但是它也有不足之处。比如对中文支持不是很好，而且在默认进行任意语言查询时，返回的中文结果有时是乱码，必须手动选择语言才能正常搜索；此外，Fast/AllTheWeb 的网页摘要目前还不是动态生成，造成用户无法方便地根据摘要选择最想要的结果等等，这些方面都还有待改进。

AllTheWeb 于 2003 年 4 月被 Overture 收购，而 Overture 于同年 7 月宣布被

Yahoo! 收购。AllTheWeb 目前使用的是由 Yahoo! 新推出的基于 Inktomi 技术的搜索引擎。

10.2.3.5　Ask（http：//www. ask. com）

Ask（图 10—6）的前身是 AskJeeves。AskJeeves 曾是著名搜索引擎 DirectHit（2002 年 4 月被关闭）的母公司，在 2001 年年末收购了全文搜索引擎 Teoma 并与之进行整合后，其搜索能力得到了进一步的加强。2003 年，搜索引擎界发生了一系列兼并和重组，目前除 Yahoo! 搜索集团和 Google 外，Ask Jeeves 成为硕果仅存的，拥有自主技术的独立一线全文搜索引擎。

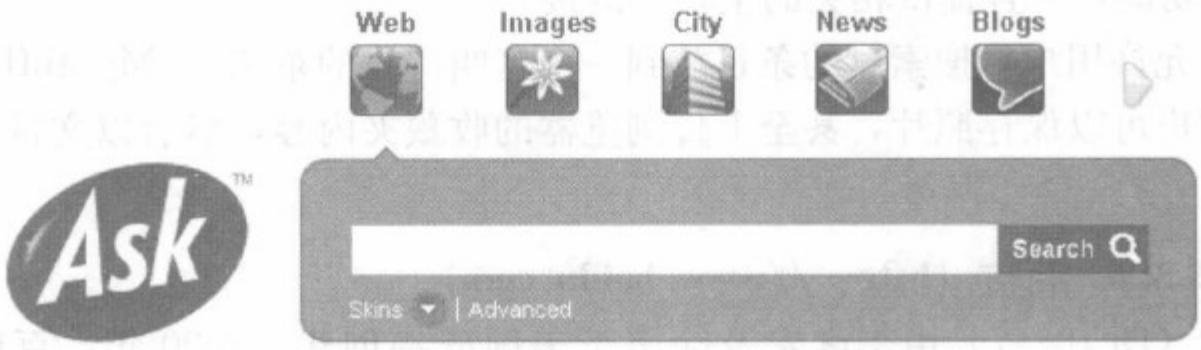

图 10—6　Ask 主页

Ask 主要有以下特色：

（1）不同的“集群”（Clustering）方法，即将类似搜索结果分配到分类目录中。这些集群在搜索结果页面右侧依次排列，吸引了 30%的用户“点击率”。并且，搜索结果中加入了包含搜索条目在内的网页概要，这也令用户在决定点击搜索结果时更为容易，大大减少了用户在找到最终答案之前访问的网站数量。

（2）可以直接使用自然语言来描述搜索请求，系统的自然语言理解引擎会将用户的自然语言问句转换成搜索引擎可以理解的内部搜索请求。支持自然语言检索技术是它的最大特色，而且这种检索方法符合人们日常查找信息的思维方式，这样就抛开了有关关键词和词组的种种限制，也不需要牢记繁琐的检索规则，变成了提问式的检索。

（3）设置了“智能回答”（Smart Answer）。Ask 把搜索结果放置在页面的最上面，位于广告之上。例如：用 earthquakes california 搜索，结果会显示出这个州所有的地震活动。把这个关键词输入 Google，你将会得到很多网页的链接，这些链接不一定就能获得地震的数据，并且你还需要打开网页自己去找。还有很多类型的智能回答，包括体育成绩的记录、电影放映时间、天气预报、字词的定义、翻译、转换、科学与动物信息等等。Google 有时候也提供这些问题的答案，但是它不如 Ask 的搜索结果那么详尽具体。如果用户想对一个事实性问题得到迅速的答案，Ask 比 Google 会更好用一些。虽然 Google 和雅虎都有类似 Ask

“智能回答”的功能，但 Ask 使用得更频繁，效果也更好。且 Ask 提供智能回答的 RSS，比如搜索一个流行博客的名称，将会在搜索结果上面显示这个博客最近的文章。

（4）清楚地区分了广告，且总体上广告很少。首先广告量大幅削减，每个页面只有上方和下方各三个，不像 Google 经常把全部右侧都用来显示搜索结果的赞助商链接。而且广告为彩底，以便与搜索结果区分开来。在 Ask 的搜索结果中广告也很少。

（5）对于各类问题的搜索，都提供扩大或者缩小检索的链接即“缩放”（Zoom）功能，并且提供相关的主题的链接。

（6）允许用户将搜索过的条目放到一个名叫“我的东东”（MyStuff）的特殊页面，用户可以保存照片，甚至上传浏览器的收藏夹内容，然后以文件夹方式进行管理。

10.2.3.6 百度（http://www.baidu.com）

百度（图 10—7）由李彦宏及徐勇于美国硅谷创建。2000 年，百度回国发展。“百度”一词源自辛弃疾的《青玉案》的“众里寻她千百度”，象征着百度对中文信息检索技术执著的追求。百度是目前全球优秀的中文信息检索与传递技术供应商。百度在中国各地和美国均设有服务器，搜索范围涵盖了整个中国和新加坡等华语地区以及北美、欧洲的部分站点。百度主要支持对中文信息的检索，收录范围包括 GBK（汉字内码扩展规范）、GB2312（简体）、BIG5（繁体），并且能够在不同的编码之间转换，是目前更新时间最快、数据量最大的中文搜索引擎。除了收录了中文网站和网页，百度还收录了大量的 Flash，可以对 50 000 个 Flash 进行检索。

百度支持多种检索功能，首先，百度支持“＋”（AND）、“－”（NOT）、“｜”（OR）。如果检索框中的两个关键词之间用空格隔开，则默认为是“＋”连接。其次，百度提供相关检索功能，可以先输入一个简单词语进行搜索，然后百度搜索引擎会提供“相关搜索”作参考，点击任何一个相关搜索词，都能得到那个相关搜索词的搜索结果。第三，提供限定检索，“link：”用于搜索链接到某个 URL 地址的网页，例如，“link：www.ruc.edu.cn”表示检索有链接指向 www.ruc.edu.cn 的网页。用户使用这项功能，可以知道某网页的受欢迎程度。“site：”表示在指定网站内搜索，如“电话 site：www.baidu.com”表示在 www.baidu.com 网站内搜索和“电话”相关的资料。“intitle：”表示在标题中搜索。“inurl：”表示在 URL 中搜索。

此外，百度还有一些特色功能，如 MP3 搜索、Flash 搜索、IE 搜索伴侣、

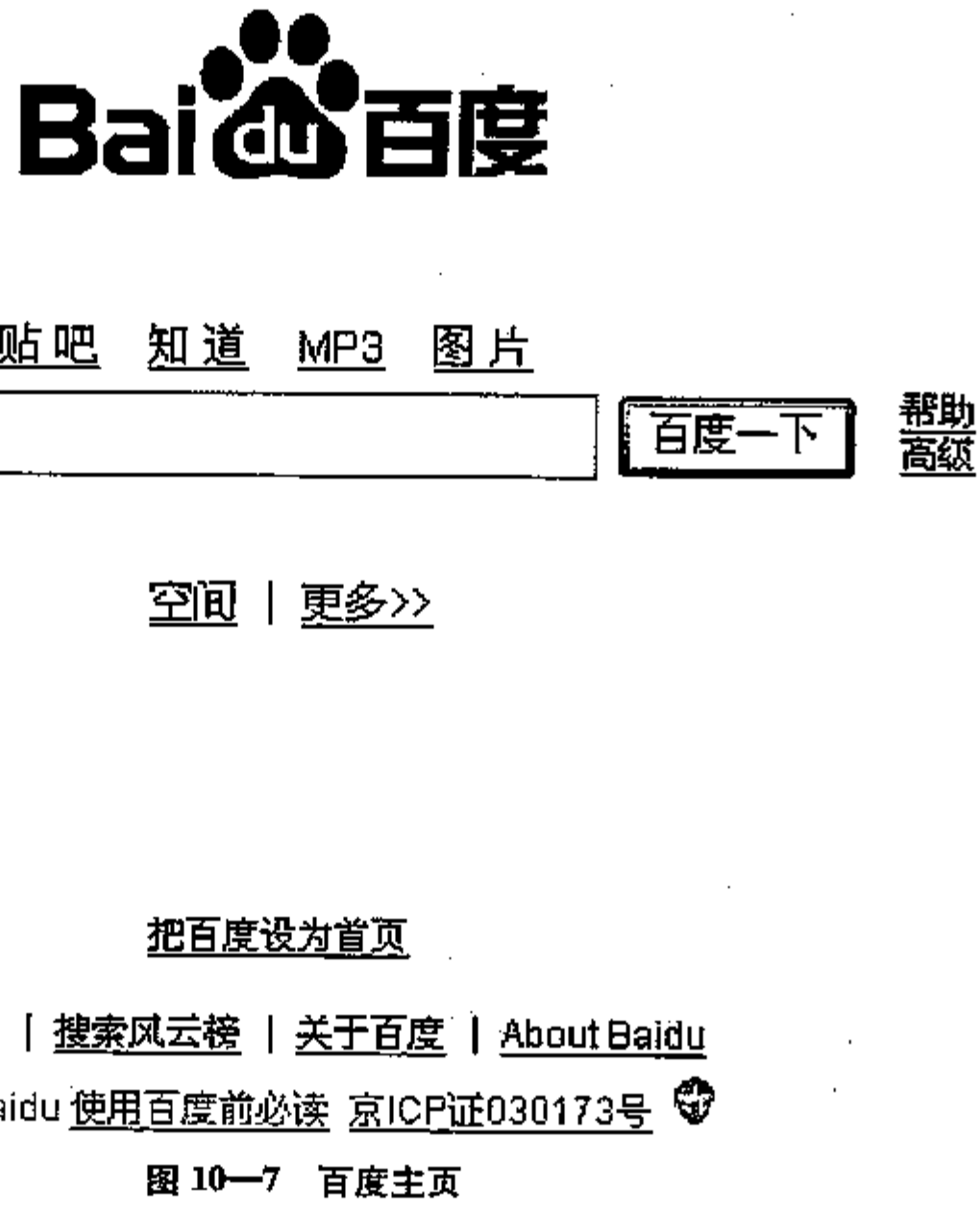

图 10—7　百度主页

百度搜霸、百度快照等。

在众多中文搜索引擎中，百度是一个比较优秀的搜索引擎，它查询速度快，响应时间短，检索结果相关度高。但同 Google 等世界知名英文搜索引擎相比，百度仍有许多不足之处，如数据库资源类型少，信息服务种类比较简单，等等。

10.2.3.7　其他英文搜索引擎

除了以上介绍的 5 种英文搜索引擎外，还有许多比较优秀的英文搜索引擎，如 Excite、AOL、AllExperts、MSN 等。

1. Excite（http：//www. excite. com）

Excite 是由斯坦福大学的 6 个学生 1993 年开发并于 1995 年对外公开服务的搜索引擎。Excite 最初的想法是分析字词关系，以对互联网上的大量信息作更有效的检索。后来 Excite 以概念搜索闻名。它在搜索时不只搜索用户输入的关键字，还可“智能性”地推断用户要查找的相关内容进行搜索。它提供类目、网

站、全文及新闻检索功能。目录分类接近日常生活，细致明晰，网站收录丰富。网站提要清楚完整。搜索结果数量多，精确度较高。2002 年 5 月，被 Infospace 收购的 Excite 停止自己的搜索引擎，改用元搜索引擎 Dogpile。目前 Excite 隶属于 Ask Jeeves 公司。

2. AOL（http：//www. aol. com）

AOL 是“美国在线”（American Online）的缩写，它的搜索引擎服务除了提供一般意义的网页搜索外，同时还提供了天气查询、股市查询、产品和品牌查询等。AOL 同样也支持布尔逻辑检索、词组检索、位置检索、截词检索等功能。AOL 搜索引擎提供的网络、图片、多媒体（音视频）、新闻和地方搜索选项，可提供实时建议的搜索框，搜索结果一目了然。因此，比较适合网络新手和美国在线 AOL 订户。目前，AOL 的搜索技术受到 Google 的支持。

3. AllExperts（http：//www. allexperts. com）

AllExperts 是一个非常有特色的搜索引擎，是以提供专家咨询服务为主的搜索引擎的重要代表。它最大的优势是不需要任何技巧，有人这样形容：“像按门铃一样简单，因为你已经来到了正确的门前。”它是目前互联网上最早的、最大的“提问式”专家服务搜索站点，有来自各个领域的志愿者免费回答用户提出的各种问题。用户向 AllExperts 提问时，首先需要选择一个感兴趣的类目，这个类目是经过层层划分之后不可再分的子类目，然后选择一个专家进行提问。通常AllExperts会提供多个专家给用户选择，并且用户可以查看专家的背景资料，以帮助用户选择恰当的专家。用户在提问时需要填写一个关于问题的表单，其中包括用户的一些背景资料。AllExperts 目前还不支持中文提问。

4. MSN（http：//www. msn. com）

MSN 是由微软公司开发的，其正式版于 2005 年 2 月 1 日推出。目前 MSN 搜索提供 Web、News、Images、Music、Desktop. BETA、Encarta 等的搜索，其特色是可以直接进入 Encarta（微软的百科全书）进行答案搜索。用户也可以通过网络资源目录浏览感兴趣的信息。此外，MSN 还在其检索框下面列出了最近流行信息的搜索。MSN 提供的分类浏览目录是动态变化的，能够及时添加用户非常关注的事件。例如，在其一级分类目录中添加了“Election 2008”，及时详细介绍美国 2008 年大选的情况。目前，微软已经推出了 MSN 中国，提供了针对简体中文的关键词检索和网络资源目录浏览检索。有专家认为，虽然微软如此之晚才进入搜索领域，但凭借它的技术、市场、资金等优势，势必将对现今的搜索格局产生重大影响。

10.2.3.8　其他中文搜索引擎

1. 天网（http：//e. pku. edu. cn）

天网由北京大学计算机系网络与分布式系统研究室研制开发，具有中英文搜索功能，曾被《软件世界》杂志评为最值得关注的中文搜索引擎。它支持简体中文、繁体中文的关键词检索，信息来源是国内 CERNET、CHINANET、CHINAGBN、CSTET 四大网络。支持布尔逻辑检索，逻辑运算符号为“&”（与）、“—”（非）、“|”（或）。检索结果显示格式包括网址、摘要、最后修改时间、长度、相关度、编码类型等。检索结果按关键字串的相关程度来排列。

天网搜索引擎的 FTP 搜索是其一大特色，在天网首页输入框输入用户要查询的文件名，可以包含“*”号（通配所有字符）、“?”号（通配一个字符）、空格（表示逻辑与），点击主页上的“文件”，即得到查询结果。也可以在“FTP 检索”页面进行常用功能的 FTP 搜索，可直接进行一般检索，或按类别搜索文件，输入要检索的文件名后，选择“分类搜索”下的各种类型，如“图片”、“音乐”、“电影”、“压缩”、“文档”、“程序”、“目录”、“源代码”等，则搜索引擎就可以在指定的类型里搜索文件。比如选择“图片”类型，则在所有的图片文件里查找与匹配串相符的文件。

天网搜索引擎还实现了中文网页的自动分类功能，即“天网目录”，测试数据 300 万，这在中文搜索引擎中也是独具特色的，它所搜集的是网页，而不同于其他目录的网站，同时，采用的是自动分类，而非人工搜集整理加工。

此外，天网搜索引擎还提供了“主题搜索”，能够分别实现北京大学校内搜索、西安交通大学校内搜索、新闻搜索、美国 1 000 所大学搜索、Unix 相关搜索。用户可以根据自己的特殊需要，选择使用相关的主题进行检索。

2. 中搜（http：//www. zhongsou. com）

中搜（原慧聪搜索）成立于 2003 年 12 月。目前提供网页、IG 门户、新闻、行业、网站、MP3、图片、论坛、地图、网站等搜索，其中行业搜索较有特色。此外，还提供计算器、量制转换、IP 查询、邮编地区查询、电话区号查询、在线词典等服务。中搜首页的所有内容都是依靠搜索引擎技术自动抓取、聚类而成，所有工作只需要一个编辑就能完成（主要是审核内容的合法性）。在新版本中搜首页上，不再只是单纯的搜索框，而是在保留搜索功能外，增加了互联网热点资讯和热贴，使整个中搜首页猛一望去更像一个门户网站。更值得关注的是，中搜首页的标题栏从“中搜——全球领先的中文搜索引擎”更改为“中搜——个人门户的领导者”，而中搜个人门户产品 IG 在首页也占据了更为突出的位置。中搜首页实际上是把一些质量高的、有特色的、公众关注的搜索内容（新闻、

MP3、论坛、图片搜索）直接推荐给网民。

3. 新浪搜索引擎（http：//search. sina. com. cn）

新浪搜索引擎是面向全球华人的综合型网络信息查询系统。信息资源丰富，索引数据库规范，并有主题分类目录。目前共有 16 大类目录，1 万多个细目和 20 余万个网站，是互联网上规模最大的中文搜索引擎之一。新浪搜索引擎虽然也具备主题目录浏览的功能，但其关键词检索功能在中文信息检索工具中也具有较强的优势。对检索结果的技术处理，新浪搜索同时采用两种技术方案：一是站点类聚，指在检索结果中，如果来自同一站点的网页多于一篇，则除了最相关的一篇外，其余均被隐藏起来，同时会为这个站点提供一个链接，用户如果需要此站点上更多的信息，可点击“此站点上的更多结果”来获得这个站点上其他的相关网页信息；二是内容类聚，指在检索结果中，如果某几个结果的网页内容相同，则只保留一篇，其余被隐藏起来。新浪搜索力图为用户提供最有价值的信息，避免数量过多且重复的检索结果影响用户的使用。

10.2.4 搜索引擎的发展趋势

搜索引擎经过十多年的快速发展，其检索性能不断地得到优化，检索功能和途径越来越多样化，这在用户从海量信息中查找所需信息上发挥了越来越重要的作用。不过，面对数量庞大、增长迅速的信息量和不断多样化的信息类型，用户的检索需求越来越个性化和对检索要求越来越高，这给搜索引擎带来越来越多的挑战。目前搜索引擎自身也存在着一些问题，如对自然语言提问的理解和处理能力差、难以准确地检索多媒体信息、不能基于用户背景进行个性化检索、检索结果常常存在大量重复信息和无用信息等。这些挑战和问题的存在为搜索引擎的发展指明了方向。

1. 集成搜索引擎

当前大型搜索引擎的索引数据库规模都很庞大，一般都能保持几十亿个网页文件以上。例如，目前被公认为全球规模最大的搜索引擎 Google，其目录中收录了高达 80 亿多个网站或网页。不过国外研究却表明搜索引擎的索引能力正在越来越落后于网络的快速增长速度。自 1997 年 12 月以来，搜索引擎的覆盖面相对于网络上可检索的内容实际上是减少了。据专家估计，最好的搜索引擎也只能搜索到 1/3 的网页信息，大部分的网页很难找到。随着互联网规模和信息量的急剧膨胀，仅依赖一家搜索引擎已经无法适应当前互联网的状况。各个搜索引擎的服务特色一般是不同的，其搜集网页的策略和范围也不同，如果把多个搜索引擎有机地集成在一起，提供统一的检索服务，可以很好地解决单一的搜索引擎收录

信息不全的问题。集成搜索引擎正是基于这一思想，将多个独立搜索引擎集成在一起，提供给用户一个统一的检索界面，系统将用户的检索指令发送给各个独立的搜索引擎，并将独立的搜索引擎返回的结果综合整理后反馈给用户。集成搜索引擎是在一个万维网页面上链接若干种独立的搜索引擎，检索时需点选或指定搜索引擎，一次检索输入，多个搜索引擎同时检索，扩展了检索范围，起到了各搜索引擎间取长补短的作用，极大地方便了用户。集成搜索引擎类似元搜索引擎，区别在于它并非同时调用多个搜索引擎进行搜索，而是由用户从提供的若干搜索引擎中选择。

2. 垂直搜索引擎

不同专业领域的用户有不同的专业信息需求，综合性搜索引擎一般能较好地满足普通大众的信息需求，但对于某一专业领域的用户来说，检索结果很可能存在着大量不相关的信息，难以保证有较高的查全率和查准率。垂直搜索引擎是相对综合搜索引擎的信息量大、查询不准确、深度不够等提出来的新的搜索引擎。它通过针对某一特定领域、某一特定人群或某一特定需求提供的有一定价值的信息和相关服务，其特点就是“专、精、深”，且具有专业和行业特色，相比较综合搜索引擎的海量信息无序化，垂直搜索引擎则显得更加专注、具体和深入。垂直搜索引擎在提供专业信息方面有着大型综合搜索引擎所无法比拟的优势，它所采用的原理和技术基本和综合搜索引擎一样，而且基本上都是成熟的技术。因此，基于专业领域的垂直搜索引擎是目前搜索引擎发展的趋势。目前国内外著名的大型综合搜索引擎在自身基础上一般都建立了专门的垂直搜索引擎，如图片搜索、音乐搜索、新闻搜索等。百度除了建立新闻搜索、MP3 搜索和图片搜索等垂直搜索引擎外，在 2006 年还先后开发出了“百度国学”和“百度百科”垂直搜索引擎。

3. 智能搜索引擎

传统的搜索引擎不能很好地理解用户的查询需求，难以基于用户的背景提供个性化的检索服务，缺乏对内容的语义理解、概念推理和自学习功能。智能搜索引擎在传统搜索引擎功能的基础上，还提供用户角色登记、用户兴趣自动识别、内容的语义理解、智能化的信息过滤和推送等功能。智能化的搜索引擎需要用到自然语言理解技术、机器学习技术、知识推理技术、知识发现技术、智能搜索代理技术等。

4. 可视化搜索引擎

传统的搜索引擎在检索结果的显示上一般都采用列表的方式，这种一维的线性排列仅仅是完成了结果的提供任务，对检索结果之间的关系的提示、与用户的

交互等方面则无能为力。尽管检索结果是按照相关度依次排列，但真正符合用户需要的文档有可能排在后面，而用户往往很难有时间和耐心扫描完整个列表。信息可视化是通过一种可视化的表现形式帮助人们理解信息。可视化搜索引擎即是实现检索结果的可视化。将检索结果用可视化方式进行显示不仅可以使人们直接观察到信息，也能实现与用户更直接、直观的交互，还能揭示检索结果中文档之间的关系。它与传统的滚动列表相比，用户不仅能从中快速找到符合要求的文档，也能对所检索的主题获得较为全面的了解。此外，可视化的特征如颜色、位置等信息能帮助用户快速找到感兴趣的区域。搜索引擎检索结果的可视化方法主要有文档簇法和超链接法。文档簇法的主要思想是找出具有相同词的文档，并把包含共同词最多的文档放在同一族中。每个族根据族中文档的主要语义内容给出一个总的标题，以便让用户能找到所需要的信息。在簇的排列上，可以将簇作为节点排列成层次结构，也可以排列成网状结构。超链接法利用文档之间的超链接将检索结果文档之间的关系可视化。这是一种最直接、最省力的方法，可以为用户进一步扩展浏览文档信息提供导航。超链接不仅指明文档的逻辑结构，也具有和用户交互等重要扩充功能。目前 grokker 能够实现一定程度的可视化检索，在检索结果的显示上提供了检索结果大纲视图和地图视图。

5. 多媒体搜索引擎

随着互联网技术与通信技术的发展和人们对图形、图像、声音、视频等多媒体日益增长的需求，互联网中的多媒体数据所占的比例将越来越大，甚至会成为主流，如何帮助用户准确而快速地检索到所需的多媒体信息，是搜索引擎急需解决的问题。目前搜索引擎对多媒体信息的搜索主要是依靠关键词检索，即根据多媒体信息的文件名、路径名、标签以及分析多媒体信息的内容，将其标注为一系列关键词或者别的文本形式的描述，然后通过检索这些描述以达到检索多媒体信息的目的。用关键词对多媒体信息检索存在着很大的不足，最主要的是难以深入揭示多媒体信息的内容特征。国内外大多数学者都认为对多媒体信息的检索应基于多媒体的内容，而不能仅仅停留在文字描述上。基于内容的检索主要是利用媒体对象的语义、媒体的视觉特征或听觉特征进行检索，它利用图像处理、模式识别、计算机视觉、图像理解等学科中的一些方法作为部分基础技术，直接对图像、音频、视频等媒体内容进行分析，抽取特征和语义，利用这些内容特征建立索引并进行检索。基于内容的检索不同于传统的检索手段，它用于检索的是反映媒体内容并与媒体存储在一起的各种量化特征，使用的是基于相似性度量的示例查询方法。它区别于传统的检索手段，融合了图像理解技术、模式识别技术，从而可以为信息用户提供更加有效的检索手段。目前基于内容的多媒体搜索引擎技

术仍然相当不成熟，理论上和实用上均有许多问题亟待解决，尤其在系统模型优化、通用性设计、图像声音特征相关性研究及在互联网上实用化等方面，还是需要着力加强研究的地方。

6. 关联式综合搜索引擎

目前的搜索引擎大都是在甲网站找图片，到乙网站找新闻，再到丙网站找股票资讯等等，十分麻烦且浪费时间。如何将这些图片、新闻、股票等等各种相关联的信息整合到同一界面，让用户一次查询，满足用户全部的查询要求？这就必须引入关联式综合搜索技术。所谓关联式综合搜索技术，就是一种一站式的搜索服务，它使得互联网用户在搜索时只需输入一次查询目标，即可在同一界面得到各种有关联的查询结果。这项服务的关键在于有一架构建在 XML 基础上的整合资讯平台。XML 技术使信息结构化，同时使查询结构化，从而使搜索的准确度和相关性大大提高。

10.3　网络资源目录

10.3.1　网络资源目录的含义和原理

10.3.1.1　网络资源目录的含义

网络资源目录是一种主要采用人工或机器搜索信息，由人工对搜集的信息进行甄别、加工整理、分类，建立分类导航或分类编排网站目录，提供分类浏览的检索工具。也称为“目录式搜索引擎”、“主题指南”、“网络分类目录”等。

1994 年 1 月，由斯坦福大学的研究生杨致远和大卫·费罗共同创建了举世闻名的 Yahoo!。将网站按主题分类收录，分入不同的类目，数据由人工录入，并对收录的每个网站进行了相关的介绍。Yahoo！使网络资源目录的概念深入人心，真正创立了网络资源目录的分类体系，对以后其他网络资源目录产生了深刻和久远的影响。因此 Yahoo! 被尊为网络资源目录的开山鼻祖。

10.3.1.2　网络资源目录的原理

1. 网络资源目录以分类理论为依据

分类作为一种科学方法是人类认识客观世界的重要手段之一。网络资源目录是传统分类法在网络环境下的新的发展，它依然遵循知识分类的原则。分类方法应用于网络信息资源的组织，在很大程度上，可以限定信息资源范围，提高查准

率，分类等级结构在事实上起到了提供上下文检索词的效果，等级结构可以便于用户在查找时进行浏览。

2. 网络资源目录以人工收集信息和组织信息为主

网络资源目录主要采用人工的方式来建立和维护目录。例如，Yahoo! 中国的网站目录就是由编辑们对网站进行访问并作出评判后，把这些资源整理组织起来，按照不同的主题将它们归入 Yahoo! 类目和子类目中，就成为用户现在看见的完整目录。随着信息种类、数量的剧增，以往的手工处理和加工方式不堪重负，传统的人工分类和索引方式虽然保证了质量，但费时费力，面对成指数增长的网络资源，显然力不从心。因而，现在人们开始将更多的目光转向发展自动分类、自动标引、自动编制分类表、词表以及目录、索引、文摘编制等技术，用这些自动技术标引网页内容，并自动归类。网络资源目录除了网站之外，还开始收入大量的网页，如天网的目录就采用了自动分类技术。

3. 网络资源目录将超文本技术融进了分类法

超文本是一种非线性的文本组织模式，它将文本按信息单元存储在节点中，并根据各节点间的概念逻辑关系，用链将其连接成网状结构。从逻辑上看，节点表示信息单元，而链表示各节点间的关系，如等级、并列、引用、补充等。这种方式灵活方便，根据信息间的内在联系提供浏览和查询各类信息的不同角度，用户在查询过程中可以随时转换到自己感兴趣的信息。这与我们所熟悉的分类法有一定的相似性，链相当于分类表中的参照项。分类法的语义关系网络与超文本系统有某种相似之处，正是基于这种相似性，人们在网络资源目录中把分类法和超文本联系在一起，扩展了传统分类法，增强了分类法的扩检和缩检功能及其严格的概念逻辑关系，在网络信息资源的查询过程中起到了指南的作用，对用户的检索过程和检索范围进行控制，为不同专业知识水平的用户提供查询信息的捷径。

10.3.2 网络资源目录的类型和特点

10.3.2.1 网络资源目录的类型

在互联网上有许多网络资源目录，风格各异。按照分类体系建立基础的不同，我们可以将网络资源目录分为两种类型：传统分类法型网络资源目录和创新型网络资源目录。

1. 传统分类法型网络资源目录

传统分类法经历了几百年的发展，积累了丰富的类分文献的经验，拥有着广泛的用户基础。利用传统分类法作为网络资源目录体系，有许多优势。传统的分类法多以学科分类为基础，直接采用传统分类法的网络资源目录提供按学科进行

浏览的功能，比较适应学术性信息需求的用户查询信息。同时，可以依据类号或类目之间的层次隶属关系或平行关系，来扩大或缩小检索范围，类目浏览在某种程度上提供了上下文的环境。而且，绝大多数的分类法采用符号标记，而不是采用专门语言，所以用户可以检索很多语种的文献而不受语言问题的干扰，能够进行多语种检索。还有，类名比较规范，容易理解。

网络资源目录对分类法的应用，并非完全彻底地照搬照抄传统的文献分类法。这类网络资源目录基本上沿用了传统分类法的大类，但个别也略作了调整和取舍：有的是选择分类法中的一部分类目；有的是结合实际情况，去掉了个别的类目；对个别的类名和类号也根据需要作了一些调整；对类目的深度进行了控制，基本上控制在 3 级类目以内，一般不超过 6 级。在类目的排列方式上，有的是完全按照分类法的线性排列方式进行类目的排列，但也有相当一部分进行了调整，融入了多维展开和多元划分。此外，为了让网络用户对分类体系和标记符号等有所熟悉，并能够利用，一些传统分类法型的网络资源目录还提供了类目索引，例如 Internet Public Library（http：//www. ipl. org. ar）就可以进行主题查询，通过主题查询，获得对应的分类号，通过类号的链接进入到相关的类目。

目前，在利用网络信息资源的门户网站上，这种类型的网络资源目录不占据主流。

2. 创新型网络资源目录

创新型网络资源目录指根据网络信息资源的特点，结合网络新环境、新要求创造的新型的网络资源目录。也有人称之为网络信息分类法。它打破了传统的以学科分类为基础的文献分类法的分类方式，采用知识分类的平台，在一定程度上更多地吸纳了主题的因素，成为一种不同于传统分类法的主题分类目录。它作为网络检索工具的主流目录，是我们利用网络信息资源的捷径。如：Yahoo!、蓝帆等都属于这类网络资源目录。

这种类型的网络资源目录的建立不同于传统的文献分类法，设计之初，充分考虑到网络信息数量多、内容庞杂，变化快、稳定性差、类型多样、范围宽、用途广以及网络信息组织特殊、控制性差的特点，还要考虑到网络信息用户的特点和个性差异。因而，实用性原则、自然性原则和针对性原则是编制网络资源目录的基本原则。

（1）实用性原则。网络资源的类目主要是根据收集到的网上资源的实际情况，以及 网上用户的查询习惯来设定，类目设置不仅要讲究全面，更重要的是要有用。它是一个面向用户、面向网络信息资源的一个分类体系，具有很强的实用性。它采用较为宽泛的主题领域建立分类索引，以增加网络分类体系的容纳

性，同时，又具有一定的专指性。在一定程度上结合了传统的分类和主题的优点。对于一些需要从各个角度、各个方面多维反映的类目，利用网状结构，通过多个途径链接相关类目及其信息，以确保将信息有效地组织在相应的类目中，增强整个类目体系的整体性和有序性，同时也可提高检索过程的查全率。

（2）自然性原则。这也是网络资源目录需要遵循的一个重要原则，它摒弃了传统分类法强调科学性和学术性的原则，尽可能少地采用学术性过强、难以理解的专业术语，尽量多地采用通用、规范、内涵外延清晰的自然语言。同时，由于不断地会有新的网络信息及新知识、新事物和新科学的出现，网络信息体系是一个动态体系，动态性也表现得非常明显。

（3）针对性原则。它主要体现在根据检索工具自身的特点以及所面对的用户群体来设定类目，不可能按相同的标准设类，并按相同的标准来划分下面的子类目。有时分类体系为了方便用户查找，专门将使用频率极高的事物和类目单独列出，有针对性地进行类分，并为充分反映该类目，使用有针对性的标准。比如搜狐将“娱乐休闲”放在所有大类中的首位，并将“个人主页”作为一个大类单独列出。

这种类型的网络资源目录没有统一的标准分类体系，彼此之间的类目设置都表现出一定的差异。如，中文雅虎有 14 个基本大类，搜狐有 18 个，新浪有 18 个，网易有 19 个等，而且，具体类目名称也有不同，排列的顺序也不尽相同。分析其原因，大概有这样一些：首先，创新已成为当今网络时代深人人心的观念，每个网络资源目录的建立者都力图在参考他人的基础上有所创新，突出自己的特色。如搜狐单列了“个人主页”、“公司企业”两个大类，突出了“生活服务”一类。一些网络资源目录将商业与经济拆分成几个部分，突出了金融、房地产等热门主题。第二，每个网络资源目录都有自己的覆盖面，都有各自的分类标引人员，分类目录的建立与分类标引人员的思维习惯和收集到的资源情况密切相关等。

创新型的网络资源目录大部分在提供分类浏览的同时，都配有关键词检索与之相结合，加快了系统的反应速度，提高了检索的查全率和查准率。在本章中，我们主要关注和研究的是直接产生于网络环境的创新型网络资源目录的特点及其应用。

10.3.2.2 网络资源目录的特点

无论是综合性的还是专业性的网络资源目录，其目的都是为满足网络资源组织的需要，满足网络用户分类浏览网络信息资源的需要。总结和分析目前网络资源目录的现状，可以看出它具有如下特点：

1. 在体系结构上，以树状结构为主

网络资源目录是一个由类目、子类目等构成的可供浏览的目录等级式树状结构。它根据知识分类原则、收集的信息或网站数量的多少以及用户的需要程度，设置一定数量的基本大类，每一个基本类目下再细分不同层次的次类目或子类目，直至具体主题的网站。每一个网络资源目录都尽力做到结构清晰、内容完整全面、各级详略程度适中。多数网络资源目录的基本大类数量一般在10到20个之间。从基本大类出发，类目体系就像一棵大树的分枝，纵横展开，构成一个多维多级的分类体系。在纵向上，表现为类目的层次，即“基本类目→二级类目→三级类目→四级类目……”。分类的层次决定了网络资源的知识组织的详略程度，网络资源目录为了方便用户使用和节省用户的时间和精力，一般都不会设置过多的层次。在横向上，可以平行地展开各个类目。

2. 在类目设置上，以事物为中心确定类目

网络资源目录不同于传统文献分类法以学科分类为基础的类目设置方式，大多从方便用户的角度出发，以事物为中心来设置类目，更加重视从事物对象和主题的角度来确定类目。网络自身的开放性和互动性，使网络资源表现出明显的灵活性和实用性，这使得网络资源目录在划分事物时，不是简单地采取按其本身的学科属性来划分的所谓的“科学分类”的方法，而是更多地采用了一些能够直接表达事物主题的语词，设置相关的类目，以适应网络环境的需要，便于随时调整。例如搜狐，它的类目都是以一个事物、一个主题为其类目，如一级类目里的娱乐休闲、工商经济、公司企业、文学等，二级类目中的音乐、工业、机械等。而在传统文献分类法中，分散成诸多大类的社会科学的各个学科，在网络资源目录中统一归并为“社会科学”大类。

3. 类目的展开呈现出明显的多维性

传统的分类法的一个重要的功能是解决文献的排架问题，也就是说，文献的放置要按照分类号的顺序来进行排列，一般说来，对同一种文献不会既按照载体形态给定一个分类号，又依据文献内容再给定另外一个分类号（某些特种文献，如档案，可以对实体分类和内容分类采用不同的分类体系），因而，在传统分类法中，在对同一个类目进行下位类（子类目）的划分时，采用的标准是唯一的，类目的展开表现出明显的一维性。

而网络信息资源的分类主要用于网络信息资源的分类浏览，无需考虑资源的搁放问题。虽然，网络信息资源分布在全球不同的服务器上，但超文本技术的采用，轻松地实现了类目和类目之间、类目与资源之间的连接问题。因而网络资源目录类目展开时，无须再过多地考虑一维性问题。比如，在“社会科学”类目展开时，可

以并列设置“期刊”、“信息管理”这样的同时按照载体形态和内容两个标准展开的下位类。如果用户想找“《图书情报工作》”的话，从“期刊”或“信息管理”的相关类目中，通过点击超链的方式，都可以找到网上的《图书情报工作》。

正是有了超文本超媒体技术，网络资源目录可以按照学科之间的交叉与渗透的多元关系，采用多视角、多途径揭示，充分反映学科发展的多维构架和事物的多维属性，用多元划分的方式，构建多维的分类体系，类目的设置可以不局限于单一标准的逻辑划分，而是可以使用若干标准同时对某一上级栏目进行划分，建立若干从属于上级类目的平行的子类目。一个子类可以隶属于多个母类（上位类），一个母类可以重复列举多个子类（下位类）。通过超文本链接，实现有效的跳转，使整个类目体系形成一个多角度，有多重入口的网状结构。当大类为主题对象时，以主题对象为中心设置相关学科类，同时收入地区及各种资源类型；当大类为学科类时，以学科为主题，设置相关主题对象类，同时收入地区及各种资源类型类；当大类为地区或资源类型时，地区下按各个小的区域划分，进一步按资源类型划分，再按主题内容区分。从而构成了与传统分类不同的类目划分与设置方式。

这种多维展开的方式为用户提供了多种的浏览途径，方便用户从不同的角度查询自己所需要的网络信息。如对某一院校的图书馆信息，可以分别从教育、文化、地区等不同角度进行检索，这对于揭示那些交叉学科或总论等具有横向联系的类目，有着其他分类方法所不能取代的优势。

4. 类目直接用语词作为标记

传统分类法主要采用分类号作为类名的标记，对于图书文献的分类主要体现在分类号上。在网络资源目录中，则直接用语词来进行标记。这是由于网络信息资源的分类标记的主要作用是为方便用户检索使用，因而更多地强调标记的直观性、表达性。而最具有表达性和直观性的标记就是语词，语词既是类名又是标记符号，用户在检索网络信息时，直接用语词来检索。在浏览查询中，人们直接从反映查询信息的大类名称开始，通过链接一个个节点，逐步缩小和确定自己的检索范围，完成从上位类到下位类，直至点击打开原始资源的浏览过程。

5. 面向用户的易用性

传统的分类法主要面向专业信息工作人员，必须经过一定的培训，这些工作人员才能熟练地理解和使用分类法。对用户来说，很难全面准确地理解分类法的内涵和分类号的具体含义。与传统分类法不同，网络资源目录面向所有终端用户，为了适应这一要求，网络资源目录在设计和表现方面都体现出明显的易用性。不需要用户懂得很多专业检索知识和高深的检索技巧。无论是老人还是小孩，也无论是中学生还是科学家，只要懂得计算机最基本的操作，就可以按照目录体系的指引，非常

方便地直接通过网络资源目录中不同的类目入口浏览自己所需的信息。

6. 类目体系呈现出动态的特征

由于网络信息资源处于一种动态的环境中，各种信息都在不断更新、淘汰，同时，由于用户的需求也是多方面的，因而网络信息资源的分类也应是一种动态的。网络信息资源分类的动态性，明显地表现在类目的设置上，尤其是子类目的变化。网络信息分类法的每一个子类目，都可以根据网络信息的分布状况进行选择，有信息即可以选用，没有或少有某方面信息的就可以不设这个类目。而且，随着用户关注热点的转变，各个热点类目的位置也将随之调整。这样在网络上所看到的网络信息分类法，将是一个动态的资源目录，即都是依据其网络信息的不同产生的不同偏重的网络信息分类法。整个类目体系是个相对稳定的动态体系，需要适时地更新，从而保证它的适用性。

7. 表现出明显的兼容性

网络资源目录是一个通用性很强的分类体系，具有海纳百川的特征，可以容纳各种类型和各种内容的网络信息。网络信息资源的分类，既要容纳各个专业学术性的资源，又要囊括各种休闲娱乐、生活旅游方面的网上信息；既要能类分官方发布的信息资源，又要归类企业和个人在互联网上创建的内容万千的网站或网页。网络资源目录所采用的以主题和事物为中心的立类思想，在很大程度上保证了网络资源目录的兼容性特征。

网络资源目录的不足之处有：

(1) 类目名称设置欠规范。个别类目名称不规范、不统一、不标准，无规律可循。比如，在网络资源目录中，“商业经济”类出现了几种不同的称谓：“商业与经济”、“工商经济”、“工商财政”、“经济贸易”、“工商产业”、“经济金融”、“经济”、“工商企业”等。类目名称用来表述和概括本类目的信息内容，起着标识、引导的作用，类名应该能够准确地反映该类目的含义。网络资源目录中有些类目名称设置由于过于追求新异或时尚等原因，出现了一些让用户费解的类目，如，FM365 的“银色专题”、网易的“情感绿洲”等。

(2) 类目的设置缺乏逻辑性。逻辑性是传统分类法的核心原则之一，而网络资源目录由于更多强调灵活性和易用性，因而导致了在逻辑规则上出现了一些违背基本逻辑方法的做法，存在归类不当，在类目展开中出现上下位类颠倒，一个类目下包括的子类目范围过广，把不相从属的类目也收入到其下等现象。

(3) 商业化和生活化气息过浓。网络资源目录过分强调商业性和生活化，而忽略了学术性资源的分类问题。互联网的确是我们生活的一种工具，它可以满足我们了解新闻时事、查询事实性信息、网上购物、网上聊天等需求，但同时也应

考虑到网上学术性资源的需求，而且这是一种很大的潜在需求，社会科学和自然科学研究者也希望能从网上获取相关学术信息。

10.3.3 网络资源目录介绍

10.3.3.1 Yahoo!（http://www.yahoo.com）

Yahoo!（图 10—8）是世界上最著名的网络资源目录之一。1994 年 4 月，斯坦福大学的两位电子工程学博士研究生杨致远和大卫·费罗开始编制一个互联网上他们感兴趣的站点目录，这就是最原始的 Yahoo!。1995 年成立 Yahoo! 公司。Yahoo! 以其精心挑选的站点、广泛的内容成为广大用户网上查询的首选工具。目前，Yahoo! 在全球共有 24 个网站，12 种语言版本，其中雅虎中国网站于 1999 年 9 月正式开通，是雅虎在全球的第 20 个网站。

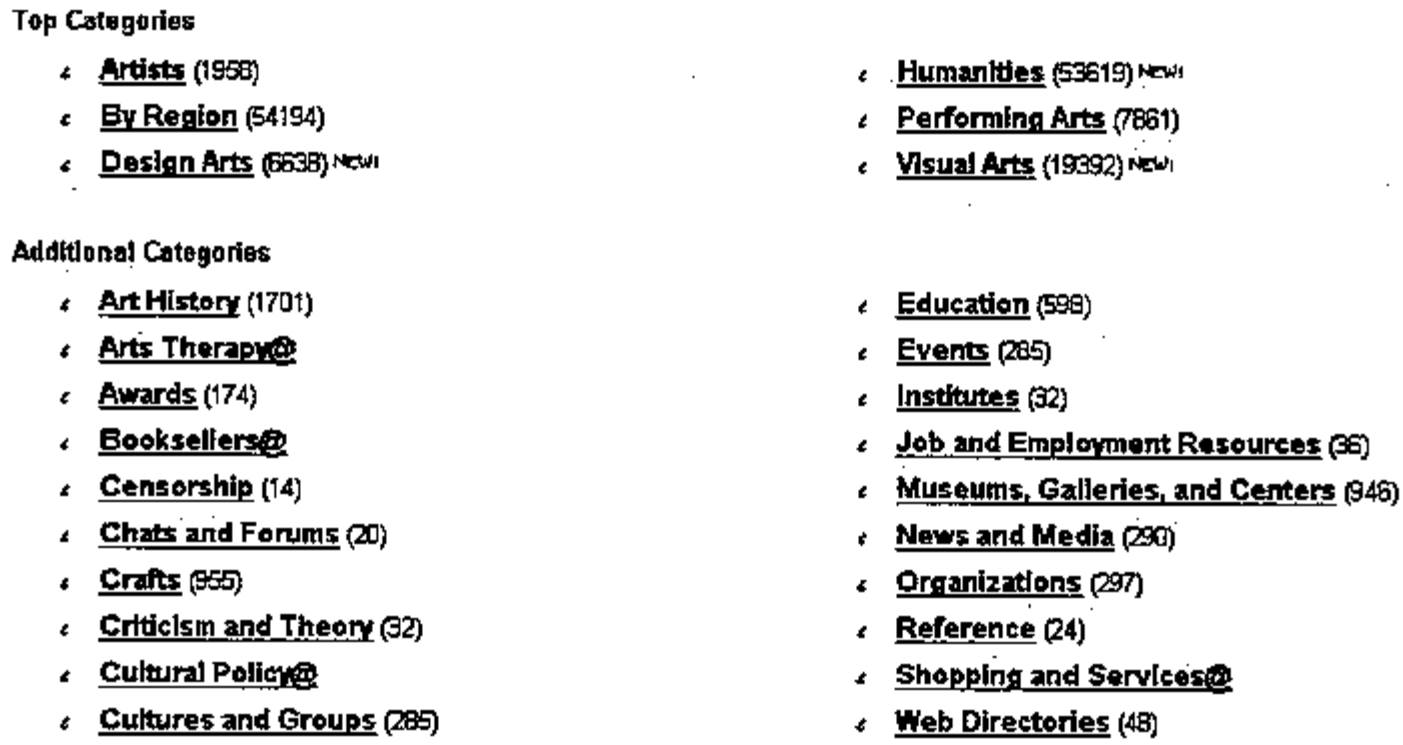

图 10—8 Yahoo! 资源目录中“人文和艺术”子类目的分类目录

Yahoo! 的魅力，就在于它的可浏览式等级主题目录。按照主题建立分类索引，提供全面的分类体系结构，并结合高质量的检索软件，成为网络检索工具的佼佼者和等级式网络资源目录的典型代表。Yahoo! 由 14 个基本大类组成，包括 Art &Humanities（艺术与人文）、Business & Economy（商业与经济）、Computers &Internet（计算机与互联网）、Education（教育）、Entertainment（娱乐）、Government（政府）、Health（健康与医药）、News & Media（新闻与媒体）、Recreation &Sports（休闲与运动）、Reference（参考资料）、Regional（国家与地区）、Science（科学）、Social Science（社会科学）、Society &Culture（社会与文化）。Yahoo! 的这种目录模式成为后来其他网络资源目录效仿的

范例。

关于 Yahoo! 的分类原理，艾梅·格拉塞尔（Aimee Glassel）这样说道："印度著名的分类专家和图书馆专家阮冈纳赞的理论体系（《冒号分类法》，1933）与 Yahoo! 网络信息资源的主题目录之间存在着密切的联系。"这揭示了 Yahoo! 应用分面分析方法进行网络信息资源分类的实质。Yahoo! 采用多标准设类、多维展开的方式，能够为某一信息源在其巨大的分类等级结构中提供不同的路径分支入口，保证了从不同的路径，为检索相同内容的不同用户提供服务。对于交替类目，Yahoo! 利用符号"@"来表示，起到了类似于相关参照的作用，能够指引用户由某一子类目进入 Yahoo! 的浏览性等级结构的其他分支中。例如，想查找"Library and Information Science"，可以从"Social Science"大类入手，也可以从 Reference→Library→Library and Information Science。

Yahoo! 除了提供列表式目录链接浏览外，还提供关键词检索。它的搜索技术目前由 Google 支持，2000 年 6 月 26 日，Yahoo! 公司宣布终止与搜索引擎公司 Inktomi 的合作，而改用 Google 公司的搜索引擎产品，两者的结合堪称珠联璧合：一个提供强大的高质量的主题指南目录，另一个则提供高水平的检索工具。Yahoo! 的关键词检索可以提供简单检索和高级检索。检索时，可以利用双引号、限定检索，等等。

10.3.3.2 Open Directory（http：//dmoz.org）

Open Directory（图 10—9）始于 1998 年 6 月，当时一位程序员里奇·斯克林塔（Rich Skrenta）对 Yahoo! 搜索结果中经常出现老的和死的链接感到不满意，决心要建立一个 Web 上最全面的目录，于是，他便在互联网上发出了倡议，请求位于全球各地的互联网用户都志愿来帮助编辑这个目录。里奇最初的灵感来自于 GNU 计划，这是一项很有名的由志愿者共同努力来编写一个类似 Unix 操作系统的计划，因此他将要建立的目录取名为 GnuHoo，"Hoo"显然是 Yahoo! 名字的一部分，GnuHoo 的含义就是提供与 Yahoo! 相似的目录，但目录的内容由没有数量限制的志愿者来编辑完成。从 1998 年 6 月 5 日站点开通到 6 月 18 日，不到半个月的时间，GnuHoo 就有了 200 名志愿编辑，建立了 2 000 个分类，收集了 27 000 个站点。取得了最初的成功后，GnuHoo 被改名为 NewHoo，更加清楚地表明了这个目录的含义。NewHoo 发展极为迅速，到了 7 月 10 日，编辑人数增加到了1 200名，收集了 40 000 个站点，这一切都发生在短短的五个星期之内。

1998 年 11 月，Netscape 获得了 NewHoo，不久将 NewHoo 改名为 Open Directory Project（ODP），并且决定任何组织和个人都可以在他们自己的网站上使用 ODP 数据库的拷贝。与 Netscape 的联姻，让 ODP 的名气大增，越来越多的

人加入到了编辑的行列中，Open Directory 的内容也越来越全面。到目前为止，大约已有 60 个 Web 站点正在使用 Open Directory 的数据库。Open Directory 收录有超过 38 000 000 个站点，拥有 57 822 个编辑，460 000 个分类。

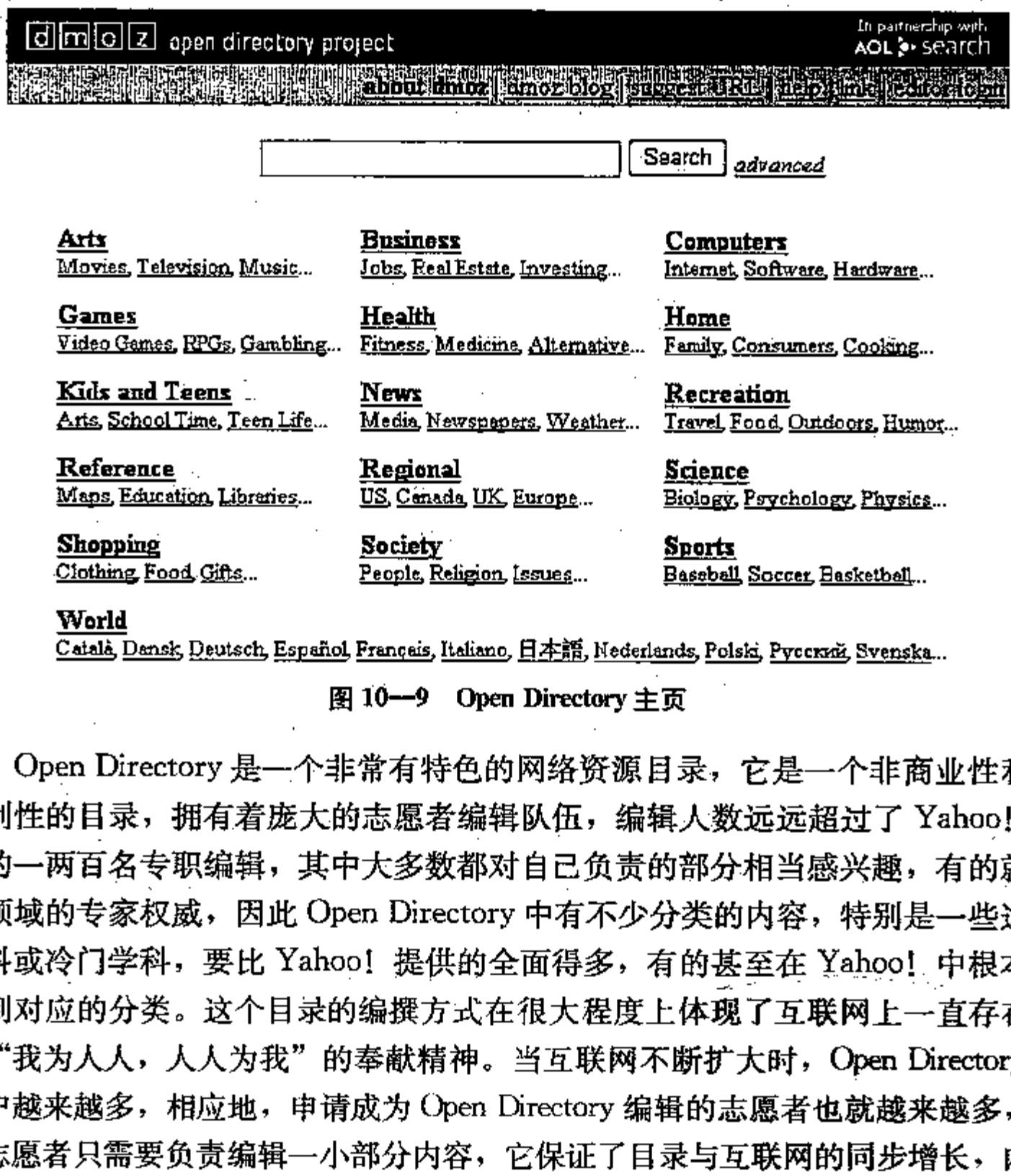

图 10—9　Open Directory 主页

Open Directory 是一个非常有特色的网络资源目录，它是一个非商业性和非营利性的目录，拥有着庞大的志愿者编辑队伍，编辑人数远远超过了 Yahoo! 雇用的一两百名专职编辑，其中大多数都对自己负责的部分相当感兴趣，有的就是该领域的专家权威，因此 Open Directory 中有不少分类的内容，特别是一些边缘学科或冷门学科，要比 Yahoo! 提供的全面得多，有的甚至在 Yahoo! 中根本找不到对应的分类。这个目录的编撰方式在很大程度上体现了互联网上一直存在着的“我为人人，人人为我”的奉献精神。当互联网不断扩大时，Open Directory 的用户越来越多，相应地，申请成为 Open Directory 编辑的志愿者也就越来越多，每个志愿者只需要负责编辑一小部分内容，它保证了目录与互联网的同步增长，内容也可以尽量保持最新。为此，Open Directory 获得了很多殊荣，如“King of the Web”、“Free Center”、“Hot Site”、“Useful Site”、“Cool Site”等。

Open Directory 有 16 个基本大类，包括 Arts（艺术）、Business（商业）、Computers（计算机）、Games（游戏）、Health（健康）、Home（家庭）、Kids and Teens（儿童及青少年）、News（新闻）、Recreation（休闲）、Reference（参

考）、Regional（地区）、Science（科学）、Shopping（购物）、Society（社会）、Sports（体育）、World（世界）。Open Directory 的二级类目相对 Yahoo! 来说，更为详细和专深。

Open Directory 也提供关键词检索，分为简单检索和高级检索。在简单检索中，支持布尔逻辑检索。如果两个检索词之间用空格，则系统作为逻辑与来处理，如在检索框中输入“golf clubs”，表示要查找同时包含这两个词的相关信息；可以用引号“”来表示词组检索，如，“information retrieval”表示查找关于“信息检索”方面的资源；可以使用截词符“*”，如，“comput*”；也可以进行字段限定检索，如，“t：information”；还可以使用“+”、“—”等。它的高级检索提供了相关的选择，包括选择只检索类目、只检索站点等。同时，它还提供了与其他著名搜索引擎的链接，包括 AlltheWeb、AltaVista、Google、HotBot、Netscape、Northern Light、Yahoo!，Open Directory 会直接将检索提问输入所指向的搜索引擎，并获得其他引擎的检索结果。

Open Directory 是一个非常有前景的网络资源目录，它可无限扩展的编辑人员，为其今后的发展注入了极大的活力，它的资源收录的增长速度，内容的更新频次，都是其他网络资源目录所无法比拟的。同时，它详尽的类目体系，开放的管理体制，都形成了它独有的特色，成为用户获取网络信息资源的重要门户网站。

10.3.3.3　Galaxy（http：//www.galaxy.com）

Galaxy（图 10—10）创始于 1994 年 1 月，是互联网上一个老牌的网络资源目录。1997 年，CyberGuard 购买了 Galaxy，1998 年 9 月又卖给了美国健康网络公司（America's Health Network）。2000 年 5 月，Galaxy 成立了自己的公司，即 Galaxy 公司。2001 年 6 月，Logika 公司购买了 Galaxy，并将 Galaxy 与它拥有的 First-Search 的目录和技术结合在了一起，力求推出更好的主题目录和搜索引擎。

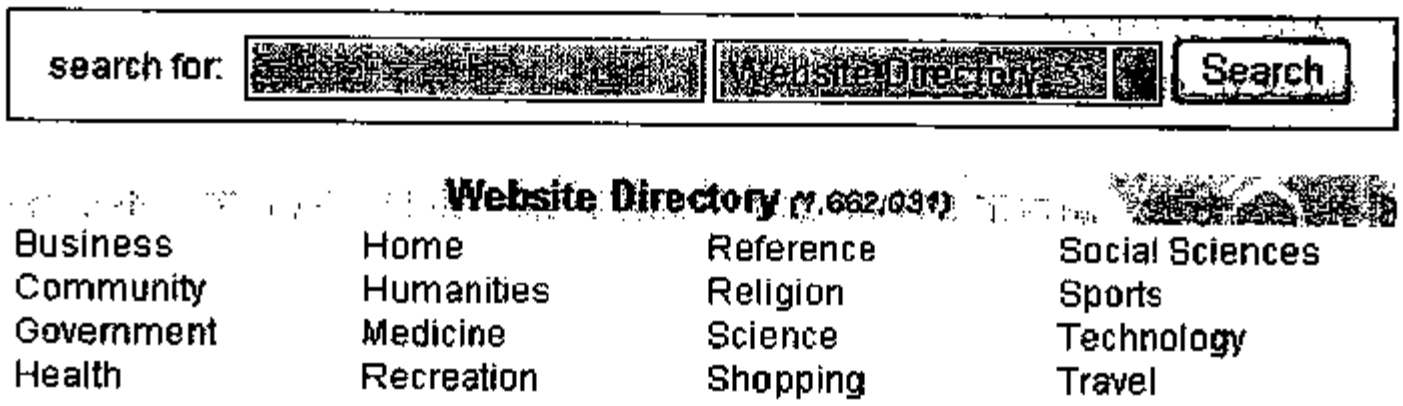

图 10—10　Galaxy 目录

Galaxy 将基本大类分为 16 个，有 Shopping（购物）、Community（公众）、Business and Commerce（商业与商务）、Travel（旅游）、Humanities（人文）、Leisure and Recreation（休闲与娱乐）、Social Sciences（社会科学）、Science（科学）、Religion（宗教）、Sports（体育）、Engineering & Technology（工程与技术）、Health（健康）、Home（家庭）、Government（政府）、Reference（参考）、Medicine（医药）等。它同样也采用人工编辑目录的方式，保证了收录资源的质量。它是一个建立在垂直基础上的等级式分类目录，可以提供给用户集中的相关的信息。

Galaxy 同时也提供关键词的简单检索和高级检索，支持布尔逻辑检索、字段限定检索、引号词组检索等。Galaxy 有多种检索选择，包括万维网检索、目录浏览、域名检索、多元搜索等多种检索方式。多元搜索的结果来自 AllTheWeb、Alta Vista、Galaxy、Google、Open Directory、Teoma、Thunderstone 和 Yahoo!。

Galaxy 是一个较好的网络资源目录，这些年来，不断改进和发展自身，丰富其信息服务功能。

10.3.3.4 搜狐（http://www.sohu.com）

搜狐公司成立于 1996 年，1998 年推出了我国第一个大型的中文网络资源目录，并且以“出门靠地图，上网找搜狐”的服务理念受到用户的欢迎。搜狐的网络资源目录（图 10—11），经过数年的发展，到现在已经发展成为中国影响力最大的分类式网络检索工具。

搜狐的网络资源目录堪称我国第一部系统的网站分类法，对其他中文网络资源目录的发展起到了积极的促进作用。自 1998 年 2 月诞生以来，搜狐网络资源目录中的网站信息资源的收集与处理一直坚持人工编辑为主，确保分类体系和网站信息的人性化特点以及网络资源目录的精确性、系统性和科学性。搜狐分类搜索经过多年的发展，已经形成了一个比较完善的类目体系，为了使网站分类更加科学和规范，搜狐的分类专家总结提炼出一套实用的网站分类法。

搜狐网络资源目录的分类体系的编制，基本上坚持了在符合科学性原则的前提下，充分考虑网站资源和用户的查询习惯的原则。在搜狐的分类体系结构中包括 16 个基本大类，涵盖了 50 000 多个不同层次的子类目，形成了一个十分庞大的树状结构，几乎涉及所有行业或领域。它采用了“纵向成枝、横向成网”、“主题法与分面组配法结合”的分类方式，根据网站的主题，首先把网站分为 16 大类，再按细分主题层层分下去。然后，再根据不同用户的使用习惯，以及不同的分类标准，把不同类目下“相关”的类目“链接”起来，从而形成搜狐的“网

图 10—11　搜狐的网络资源目录

状”分类体系。搜狐的网站分类法以主题分类为主设立了娱乐休闲、电脑网络、卫生健康、工商经济、教育培训、生活服务、公司企业、艺术、社会文化、文学、新闻媒体、政法军事、体育健身、科学技术、社会科学等 15 个大类，另外结合分面组配的方法设立“国家与地区”类目，把其他 15 大类下的所有网站又按所属地域进行分类，因为大多数网站都具有地域性，也便于用户直接查找。

搜狐网络资源目录收录的网站资源都经过了搜狐分类编辑们严格的审核和筛选，质量比较高。搜狐的网络资源目录的查询同样是按照信息所属的类别，层层点击查找信息，所以用目录时首先要考虑清楚想要查找的信息属于哪个类别。比如，想查找关于“图书馆学情报学档案学的期刊杂志”，首先浏览搜狐的 16 个大类，找到“社会科学”大类，点击进入后，然后点击“信息管理”类目，选择“期刊杂志”，就可以找到相关的期刊杂志，具体路径为：社会科学＞信息管理＞期刊杂志。

除此之外，搜狐作为一个综合性很强的搜索网站，还提供多项的检索功能。搜狐提供强大的关键词检索功能，它采用的是百度的搜索引擎技术，搜狐与百度的合作始于2000年8月，两年之后，再次续签合约，扩大原有的合作范围。它提供简单检索和高级检索两种形式，在简单检索中，用户可以在搜索框中直接输入自己想查找信息的关键词，找到相关信息。这种方法对网站、网页、新闻、类目、黄页、软件等都适用。

在高级检索界面，搜狐通过选择“所有输入的关键词”和“至少其中一个关键词”来分别完成逻辑与和逻辑或的运算，同时，也可直接输入运算符：一、&、|、()、空格，来规定多个关键词之间的关系，并完成高级检索。这些运算符既可以是英文，也可以是中文（全角或半角）。

搜狐的高级检索还提供了检索结果的类聚方式的选择，可以选择所返回的网页是“内容类聚”、“站点类聚”，或是两者都要（“站点类聚及内容类聚”，此为默认选择）。“内容类聚”指同一个内容的网页只出现一次，而不管整个互联网上有多少个不同的URL指向该网页。即“内容类聚”屏蔽掉了同样内容的网页，只提供其中一个URL供用户浏览使用。而“站点类聚”则只给用户提供相关网站的主页的URL，屏蔽掉了同一个网站中各个不同的页面。不过，用户可以通过相关摘要底下的URL访问到该网站所有的页面。

此外，还可以对时限进行选择，指定查询结果中网页的生成时间，它包括四种选择：“任何时间的网页”（默认选择）、“三个月内的网页”、“六个月内的网页”、“一年内的网页”。

10.4 元搜索引擎

10.4.1 元搜索引擎的含义和特征

元搜索引擎（Meta Search Engine）又称多元搜索引擎或集合式搜索引擎。元搜索引擎是一种将多个独立搜索引擎集成在一起，提供统一的检索界面，将用户的检索提问同时提交给多个独立的搜索引擎，并将检索结果一并返回给用户的网络检索工具。

元搜索引擎没有自己的网页数据库。元搜索引擎通过向其他独立搜索引擎发送搜索请求来处理用户的搜索请求，然后把这些搜索结果按照一定的方式集成在一起返回给用户。元搜索引擎集成独立搜索引擎的程度及其所使用的机制不尽相

同。最简单的元搜索引擎只是将用户的检索提问传递给它所集成的独立的搜索引擎，然后将各个独立搜索引擎的检索结果的界面一一呈现出来，就好像用户自己对若干不同的独立搜索引擎进行了检索。复杂的元搜索引擎能在显示结果前使用过滤器和其他算法来处理返还的查询结果，根据多个独立搜索引擎的检索结果进行二次加工，如对检索结果去重、排序等，并标明检索结果的来源搜索引擎，输出给用户。

元搜索引擎是建立在已有的独立搜索引擎服务之上的一种搜索引擎，可以将它理解为工具书的工具书，它并不直接针对一次网络资源本身，而是利用下层多个独立搜索引擎提供的服务向上提供统一的检索服务，自身不采集文档，也没有索引，只是维护它所管理的搜索引擎的参数信息，如每个引擎的查询参数、引擎的内容表示。由于元搜索引擎具有较好的扩展性，可以加入多个独立搜索引擎，在一定程度上可以适应网络资源数量剧增的要求，提高网络检索效率，同时也起到对检索工具的推荐和指南的作用。它最大优点是省时，能同时查询多个搜索引擎的数据库，检索的综合性、完整性较好。因而，元搜索引擎技术现在成为检索工具的发展方向，例如，著名的 Excite 和 HotBot 改版之后都增添了元搜索引擎的功能。

元搜索引擎与普通搜索引擎相比有很大的不同。搜索引擎拥有独立的网络资源采集标引机制和相应的数据库；元搜索引擎一般没有自己独立的数据库，更多地是提供统一链接界面（或进一步地提供统一检索方式和结果整理），形成一个由多个分布的、具有独立功能的搜索引擎构成的虚拟整体，用户通过元搜索引擎的功能，实现对这个虚拟整体中各独立搜索引擎数据库的查询显示等一切操作。元搜索引擎具体表现为这样一些特征：

1. 一次检索可以实现对多个搜索引擎的检索

元搜索引擎定制了调用多个独立搜索引擎的统一界面，将用户递交的提问提交给它可支持和调用的多个独立搜索引擎，因此，用户的一次查询可以同时检索多个独立搜索引擎。这期间，元搜索引擎针对不同的独立搜索引擎，将用户的提问作不同转换，以适应相应搜索引擎索引数据库的调用。也就是说，元搜索引擎需要对用户提问进行分门别类的处理，并根据不同独立搜索引擎的要求，按不同的提问表达形式（检索式）提交同一查询。多数元搜索引擎也可以提供一个简便的独立搜索引擎浏览列表，用户可以从中进行选择。

2. 基于独立搜索引擎结果的二次加工

元搜索引擎的结果基于独立搜索引擎的查询结果。除了一小部分元搜索引擎只能简单地直接调用原始的结果页面外，大部分元搜索引擎都会将各个独立引擎

的结果回收之后进行相应的整合，排除相同的结果，并按照一定的排序标准，把二次加工和整理后的结果以统一的格式提供给用户。

3. 标明结果记录的来源搜索引擎及其相关度

元搜索引擎与独立搜索引擎的很大一个区别在于其检索结果的显示页面。随着元搜索引擎技术的不断发展，一些元搜索引擎在用户提问的页面，与独立搜索引擎几乎没有什么明显的区别。而在检索结果的反馈时，在每个检索结果中都清楚地标明了它的来源搜索引擎。有的元搜索引擎还以百分数（%）、星星（☆）等方式标注了该检索结果的相关度。

元搜索引擎的功能很大程度受独立搜索引擎的限制，而且结构相对比较简单，因此不可避免地存在一定局限性：

1. 检索功能简单

实现检索语法转换的能力是有限的，一般只提供一个公共接口供用户输入查询词，实际查询在各个独立搜索引擎中实现。元搜索引擎由于各个独立搜索引擎的检索语法不统一，以统一的查询入口执行多个数据库查询需要进行准确的提问分析和转换，对于简单的布尔逻辑检索和词组检索，元搜索引擎的检索效果很好，对于复杂的检索功能，效果并不是十分理想。因此，元搜索引擎一般只支持通用的检索句法，多数元搜索引擎不支持指定字段检索等特殊检索，掩盖了独立搜索引擎中效果较好的高级查询功能，抹杀了各个独立搜索引擎的特色功能，也在一定程度上影响了检索效果和质量。

2. 在调用搜索引擎和检索结果的数量上都存在一定的局限

大部分元搜索引擎只支持调用 Alta Vista、Excite、GoTo.com、Yahoo!、Infoseek、Lycos 等主要的搜索引擎，有许多大型搜索引擎被排除在外，影响了信息搜索的覆盖面。检索速度的限制从一个侧面反映出了元搜索引擎在检索结果的数量上的局限性，这也就是意味着只能从各个独立的搜索引擎中检索少量的最符合要求的命中记录，因此必然影响了检索结果的全面性。

3. 在返回结果的精确性方面，元搜索引擎不如独立的搜索引擎

元搜索引擎将一次提问同时检索多个搜索引擎，扩大了检索覆盖的范围，提高了查全率。但其结果主要来自独立搜索引擎查询结果中排名靠前的记录，在一定程度上默认了独立搜索引擎的查准效果，而目前独立搜索引擎自身在查全与查准提高方面存在着各种问题。因此，元搜索引擎在为用户提供更全面、综合的结果的同时，难以控制各独立搜索引擎的无关输出。

但无论如何，由于元搜索引擎能够简便地检索多个独立的搜索引擎，它仍不失为我们进行网络信息检索的重要检索工具。

10.4.2　元搜索引擎的原理和分类

10.4.2.1　元搜索引擎的原理

元搜索引擎一般包括用户查询处理、检索机制、结果加工处理和结果页面定制 4 个部分，具体如图 10—12。

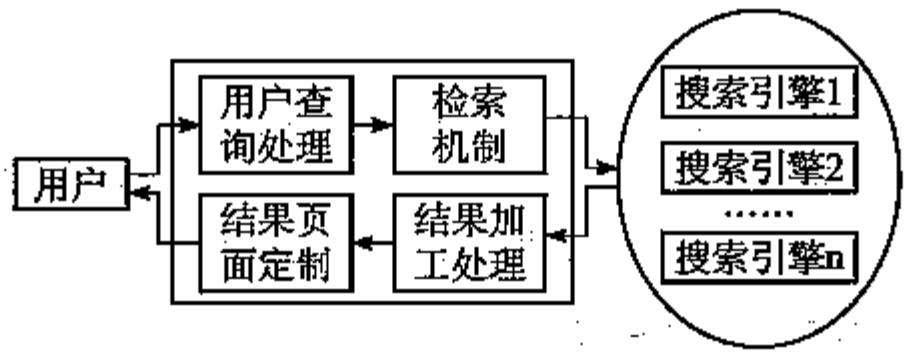

图 10—12　元搜索引擎原理示意图

元搜索引擎在执行查询之前对要调用的搜索引擎列表进行相关的选择，选择方式一般有系统默认和用户选择两种方式。系统默认方式是系统确定了用来检索的搜索引擎集合，用户无权变更；用户选择方式则允许用户自主选定需在哪几个搜索引擎中检索。

用户查询处理机制负责在接收到用户检索提问后，针对不同的搜索引擎进行相应处理，将用户的检索提问转换成能检索不同搜索引擎数据库的提问表达式。元搜索引擎的检索机制是元搜索引擎根据对各成员搜索引擎的检索结果测评分析而制定的一套规则，用于督导检索过程和结果输出过程。作为成员的独立引擎有各自不同检索界面，简单的只采用单个关键词，复杂的可以指定任意的多个关键词之间的布尔条件或词间距。例如采用独立搜索引擎的哪种检索功能（布尔检索、位置检索等）或检索类别（网页搜索、同站检索、类目检索等）。另外，在检索机制中还可以明确对结果反馈的要求，如在检索过程中是选择“结果最好”还是选择“速度最快”的标准或其他限定标准来确定优先显示哪个搜索引擎的查询结果，等等。

结果的加工处理机制负责对从各个搜索引擎得到的结果作综合处理，这一结果处理过程包对结果重复与否、结果之间相关大小等作出判断，最后遴选出满足条件的记录输出。

结果页面定制机制将最终结果以定制的界面呈现给用户。结果输出的页面定制形式在不同的元搜索引擎中有不同的体现，可以直接调用独立搜索引擎原始的反馈页面，也可以由元搜索引擎重新定制一个全新页面。

从理论上来讲，在检索独立搜索引擎时，经常会得到大量返回的结果，那么，元搜索引擎集成了众多的搜索引擎，检索元搜索引擎时返回的结果是否会更

加庞大呢？事实并非如此，元搜索引擎有着自己独特的检索集成方法。目前，元搜索引擎所采用的结果集成的方法主要有以下两种：

1. 直接将不同搜索引擎的结果合并后提供给用户

这种方式使得排在后面的搜索引擎的搜索结果无形中被忽略掉了，它只是提高了搜索结果的完全度，而不能提高用户的满意度，但是这种方式为用户在结果集中再次搜索提供了数据保障。

2. 将响应速度最快的搜索引擎的搜索结果最先返回给用户

由于搜索引擎在将结果显示给用户的时候大多是按照分页方式提供的，这就可以在用户浏览当前页面中搜索结果的时候，同时将其他的搜索结果取到元搜索引擎本地，这种并行处理的方式可以大大提高元搜索引擎的响应速度。

10.4.2.2 元搜索引擎的分类

对元搜索引擎进行科学的分类，有利于我们进一步了解和利用元搜索引擎。元搜索引擎根据不同的标准可以划分为不同的类型：

1. 根据运作平台的不同，可以分为网络型元搜索引擎和桌面型元搜索引擎

网络型元搜索引擎是指提供检索服务的 Web 元搜索引擎站点。这种元搜索引擎必须通过网上调用来在线使用，我们所利用的元搜索引擎主要都是网络型的，它的使用非常普遍。用户通过浏览器就可以方便地访问这些元搜索引擎，并检索需要的网络信息，如 Vivisimo、SavvySearch、Mamma 等都是著名的网络型元搜索引擎。它们具有操作简单的特点，任何一个连入互联网的用户都可以直接利用它们检索自己需要的信息。

桌面型元搜索引擎是一种客户端元搜索软件，它与客户端环境充分结合，代理用户递交提问，一次性检索多个独立搜索引擎，并能获取实际的 Web 页面。这种元搜索引擎可直接在用户计算机上运行，相当于用户自己拥有一个元搜索引擎，因而形象地称之为桌面元搜索引擎。这些桌面元搜索引擎可从网络上下载，多为收费软件，但也有免费的。桌面元搜索引擎是一个包括多个成员搜索引擎的完整系统，它们往往允许用户自定义检索式运行的搜索引擎集合（例如一个或全部成员搜索引擎），甚至可由用户添加新的搜索引擎，这些桌面元搜索引擎不仅可以实现对多个搜索引擎的并行检索，而且也能提供重要的后期处理功能，更容易提供个性化的检索服务。例如用户定义结果排序方式、删除重复记录等功能。目前已经有许多这类成型产品，如 Copernic、BullsEye、Infoseek Express 等。

2. 根据检索机制的不同，可划分为目录式元搜索引擎和统一入口式元搜索引擎

目录式元搜索引擎指按照一定的形式将所有的独立搜索引擎集中罗列在页面

上，有的按分类编排组织成目录，帮助和引导用户根据检索需要来选用搜索引擎，有的虽然提供一个公共的检索入口，但是实际上用户一次只进入一个独立搜索引擎检索，检索的还是某一独立搜索引擎的数据库。这种元搜索引擎并不整合或处理它所支持的独立搜索引擎的检索结果，只是提供了一个到达各个独立搜索引擎的窗口。它的结果反馈页面多直接引用原始搜索引擎的结果页面。主要代表有 All-in-one 等。这类搜索引擎的特点有：只提供一个简单的界面来帮助用户选择和使用各搜索引擎；只能选择一个搜索引擎进行检索；检索结果直接调用原始独立搜索引擎的结果页面；只支持原始独立搜索引擎支持的检索句法，它没有统一的全局外部模式，而是以各搜索引擎的检索模式和数据格式直接面对用户。

统一入口式元搜索引擎指利用统一的检索界面，实现对多个独立搜索引擎索引数据库进行检索，并将检索结果以统一格式显示的元搜索引擎。也就是说，用户发出检索请求后，检索式被分别提交给多个独立搜索引擎，最终反馈的结果是多个独立搜索引擎查询结果的综合。这类元搜索引擎一般具有这样一些特征：(1) 统一检索界面：它将所有成员搜索引擎构成一个逻辑整体，元搜索引擎检索界面构成唯一的全局外部检索模式，用户通过这个全局界面实现对多个或任意一个搜索引擎的检索；(2) 检索指令转换：将用户通过统一界面以统一形式输入的检索指令（全局指令）转换为各个成员检索工具的具体指令（局部指令），以达到用户使用同一指令语言检索不同的搜索引擎的索引数据库的目的；(3) 统一结果集的组织与显示：元搜索引擎对各成员搜索引擎返回的结果进行处理，形成全局结果集，并以统一格式显示，主要涉及数据格式转换、去重、统一排序等。这种元搜索引擎十分活跃，显现出很好的发展前景。如 Metacrawler、Savvysearch、Dogpile、Profusion、Vivisimo 等。我们所利用和探讨的主要是统一入口式搜索引擎。

3. 根据结果显示的不同，元搜索引擎又可分为直接调用原始页面型、分散综合型和混合综合型

直接调用原始页面型元搜索引擎指检索结果直接来自原始搜索引擎站点的结果页面。这种元搜索引擎对所得的检索结果不作任何处理，只是把它所包含的独立搜索引擎的原始检索结果页面返回给用户。

分散综合型元搜索引擎指依次按照每个独立搜索引擎为单位显示检索结果，也就是说，同一个独立搜索引擎所得的检索结果被集中列在该搜索引擎之下。如旧版的 Dogpile 就是一个完全的分散综合型搜索引擎，对于检索结果的显示，新版的 Dogpile 除了保持原有的“View by Search Engine”（按搜索引擎浏览结果）外，增添了“View by Relevance”（按相关性浏览检索结果）功能，用户可以根

据自己的需要来选择检索结果的显示形式。

混合综合型元搜索引擎指将各个独立搜索引擎中查找的结果进行综合，按照元搜索引擎自身设定的排序，将查询各个独立搜索引擎的结果逐个显示给用户，在检索结果的每条记录中显示有该记录的来源。例如，Vivisimo 就是一个比较典型的混合综合型元搜索引擎。

10.4.3 元搜索引擎的技术和评价

10.4.3.1 元搜索引擎的技术

元搜索引擎与搜索引擎的最大区别是它没有自己的索引数据库，不需要解决网络资源的收集问题。元搜索引擎的核心问题是要解决如何调用其他搜索引擎的索引数据库，如何获取检索提问在其他搜索引擎中的查询结果，以及如何评价、排序、呈现结果等。因而，元搜索引擎围绕解决这些问题，主要涉及的技术有用户提问转换、分布式数据库调用、检索机制设计与优化、检索结果输出等技术。

1. 用户提问转换技术

一般元搜索引擎都具有统一的检索界面，供用户输入检索提问，元搜索引擎集成了各个不同的独立搜索引擎，不同的搜索引擎有不同的检索语法和操作符使用技巧，这就需要元搜索引擎将用户输入的检索提问进行处理，根据不同的搜索引擎转换成可以进行检索的检索表达式，递交各个搜索引擎进行检索。同时要对搜索引擎不能处理的检索方式进行排除，并选择一种合适方式来匹配。

2. 分布式数据库调用技术

目录式元搜索引擎不需与独立搜索引擎的索引数据库直接交换数据，只需直接引用独立搜索引擎的结果页面即可。而统一入口式元搜索引擎则需在满足独立搜索引擎的数据库访问权限的情况下，实现对其索引数据库的访问和二次开发。各独立搜索引擎的数据库分布在不同的地域，要实现对异地、异构数据库的访问，需要使用一系列诸如分布对象技术等相关的核心技术。同时，不同数据库调用结果响应时间长短不一，这也会直接影响到结果页面的呈现。

3. 检索机制设计与优化技术

检索机制的设计主要对搜索引擎的初始化方式、各搜索引擎结果平衡处理等问题进行规划。它接影响到用户对元搜索引擎的满意程度。元搜索引擎初始化主要有用户参与、系统默认或自动随机处理等方式，用户可以根据自己的需要作出相应的选择。简单的处理方式是以搜索引擎为单位，在选定的搜索引擎下面显示比较靠前的结果；复杂的处理方式是以记录为单位，综合判定某一记录在多个搜索引擎中被评价的指数，如果多个搜索引擎都检出该结果，则该记录将被排列在

元搜索引擎检索结果的前面，同时，在各个检索结果中注明该检索结果来自哪一个或哪几个独立的搜索引擎。

4. 检索结果输出技术

元搜索引擎的结果输出处理一般有两种形式：直接引用原始结果页面技术与结果页面定制技术。直接引用原始结果页面是通过 CGI 技术，利用表单提交来调用数据库，在自制的页面中，将表单提交的对象修改为独立搜索引擎调用数据库的脚本文件。这种技术比较简单易行，一般无须进行结果去重、再排序等进一步的加工处理，只需完成表单提交的转换即可。结果页面定制技术则要对结果进行更多的加工处理，主要包括：遴选处理各个成员搜索引擎返回的结果，同时对检索结果进行去重；对检索结果进行再度排序，元搜索引擎根据检索机制对相关度判断的标准来比较各个搜索引擎得到的结果，经过算法对结果进行抽取和再度排序之后，将检索结果显示给用户。

在这一过程中，由于不同搜索引擎反馈的结果页面格式相差很大，对于这些页面的处理难度也是相当大，一方面要解析页面找到查询结果，同时还要能够把这些结果的内容抽取出来，目前采用最多的是固定查找和智能判断相结合的策略。而且作为一个元搜索引擎，如何能够将获取的信息按照相关度进行排序也是比较复杂的问题。因为不同搜索引擎在本身查询结果排序过程中采用的算法相差很大，元搜索引擎必须结合这些使用不同排序算法产生的结果，并以统一的结果形式返回给用户。这些都成为成功实现元搜索引擎功能的难点和关键技术。

10.4.3.2　元搜索引擎的评价

元搜索引擎是一种非常有特色的检索工具，选择和评价元搜索引擎可以从以下几个方面着手：

1. 元搜索引擎的初始化方式

指元搜索引擎是否提供明显的多种选择独立搜索引擎的方式，是否允许用户浏览并选择要调用的独立搜索引擎。好的元搜索引擎要能够提供一个一目了然的、可供浏览和选择的引擎列表，并允许用户设置调用方式。但也有许多元搜索引擎将这些信息隐藏在联机帮助或高级检索项中，或根本没有体现。

2. 覆盖的网络资源类型

指元搜索引擎是否覆盖多种网络资源类型。有许多元搜索引擎，除了搜索引擎数据库外，还可以选择搜索 Usenet、MP3 文件、图像文件、声音文件等类型的其他网上资源。如，Dogpile 就可以检索 Web 资源、图像、视频、多媒体、新闻等。

3. 网络信息获取方式

指元搜索引擎是否提供多途径的信息获取方式，是否除了关键词检索之外还

提供主题范畴的目录服务以及其他的专项服务等，让用户可以更加便捷地获取信息资源。元搜索引擎作为一种网络检索工具，能否提供全面综合的信息服务，也是衡量和评价元搜索引擎的一项重要指标。如，Vivisimo 就提供了非常有特色的分类目录，为用户增加了直观的扩大检索范围和缩小检索范围的新途径，并成为自动分类技术应用的典范。

4. 检索功能

元搜索引擎是否可以提供较为丰富的检索功能，是否支持布尔逻辑检索、短语检索、自然语言检索等高级检索特性，能否准确地向各个独立的搜索引擎转换用户的检索请求，都是选择和评判元搜索引擎的重要依据。此外，一个优秀的元搜索引擎必须实现不同搜索引擎间特殊检索语法规则之间的转换。如对于不支持“NEAR”算符的搜索引擎，要自动实现由“NEAR”向“AND”算符的转换等，否则将失去很多重要的高级检索功能。

同时，还要考虑元搜索引擎是否提供了足够多的检索选项和功能设置。如，是否有最长检索时间设置，是否提供高级检索服务，是否可设置每个搜索引擎返回的检索结果数量，是否能够自动检查链接的有效性，是否提供 URL 注册等附加功能等。

5. 检索结果输出格式

利用元搜索引擎进行检索，用户要得到的是检索结果。因而，它的检索结果输出格式如何、检索结果的信息描述是否全面，会在很大程度上影响用户对元搜索引擎的选择。最常见的形式是，将各个独立搜索引擎返回的结果进行集中的去重处理后，以统一的输出格式和相关度指标进行排列输出。常规信息描述主要包括资源名称、摘要信息、URL、来源搜索引擎等。好的元搜索引擎，还要能够显示出该记录结果与用户检索需求的相关度，尽可能降低用户的决策负担。

一个优秀的元搜索引擎应该涵盖了较多的搜索资源，可随意选择和调用源搜索引擎；具备尽可能多的可选择功能，如资源类型选择、返回结果数量控制、结果时段选择、过滤功能选择等；具有强大的检索功能和不同搜索引擎间检索语法规则、字符的转换功能；要有详尽全面的检索结果信息描述；可以支持多种语言检索。我们期待着在互联网上能够出现更多理想的元搜索引擎。

10.4.4 主要元搜索引擎介绍

元搜索引擎以其涵盖较多的搜索资源，能够在尽可能短的时间内提供相对全面的检索结果等诸多优异功能受到用户的青睐，已逐渐成为一种不可或缺的极具潜力的网络检索工具。互联网上有一些比较优秀的元搜索引擎，人们通过访问和

检索它们来获得自己所需要的网络信息。

10.4.4.1　Dogpile（http：//www. dogpile. com）

Dogpile 诞生于 1996 年 1 月 2 日，是一个老牌的非常受欢迎的元搜索引擎，现在属于 InfoSpace 公司，是目前性能较好的统一检索入口式元搜索引擎之一，集成了诸如 Google、Yahoo!、Ask Jeeves、Live 等优秀的独立搜索引擎。Dogpile 提出的口号是“Good Dog，Great Results”。最新改版的 Dogpile（图 10—13）更是增添了许多的功能，加强了 Dogpile 在元搜索引擎中名列前茅的地位。

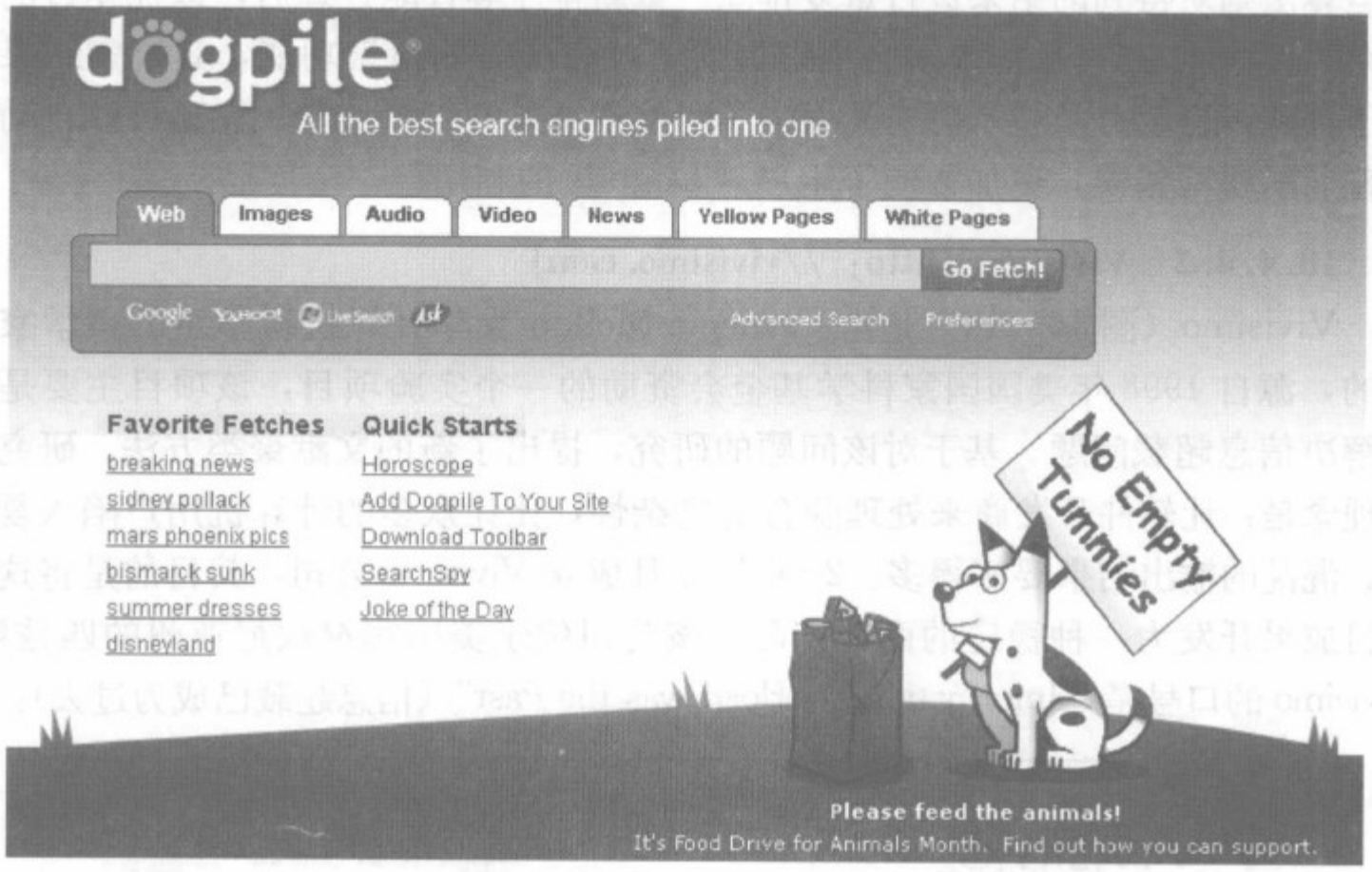

图 10—13　Dogpile 主页

1. Dogplie 简单检索

在 Dogpile 的简单检索界面里，可以选择检索网页、图像、视频、多媒体、新闻、黄页和白页。Dogpile 的搜索技术十分先进，支持布尔逻辑运算等检索技术的使用。检索结果中除了显示包含有检索提问的搜索结果外，还提供了检索结果提示和显示最近的检索历史。

2. Dogpile 高级检索

Dogpile 的高级检索功能比较全面，可以通过输入检索提问，并选择“All of these words”（逻辑与功能）、“The exact phrase”（“词组”检索功能）、“Any of these words”（逻辑或功能）、“None of these words”（逻辑非功能）来形成最终的检索式，或是使用布尔逻辑算符进行组合。

3. 其他检索

新改版的 Dogpile 提供了丰富的检索功能，它的偏好检索（Preferences）可以使用户根据自己的爱好定制个性化的信息检索服务，并可以保留这种定制，直到用户下一次改变定制。同时，它新增加的黄页检索和白页检索可以检索企业和个人的相关信息。此外，还允许用户下载免费的 Dogpile 检索工具集成条。

总之，新版的 Dogpile 是一个非常不错的元搜索引擎，展现了元搜索引擎发展的最新成果。它将用户的查询请求同时向多个搜索引擎递交，按照自定义的关联运算法则对得到的结果进行重复排除、重新排序等智能处理后，以优化过的搜索结果返回给用户。Dogpile 为用户提供了较为全面的检索功能，其检索结果更易于浏览，自动分类技术的应用增强了对检索结果的组织功能。它还可以自动修正普通的拼写错误，更加方便了用户对 Dogpile 的利用。

10.4.4.2 Vivisimo（http://vivisimo.com）

Vivisimo（图 10—14）是由 Carnegie Mellon 大学计算机科学系的科学家建立的，源自 1998 年美国国家科学基金会资助的一个实验项目，该项目主要是为了解决信息超载问题。基于对该问题的研究，提出了新的文献聚类方法。研究者的理念是：让软件开发商来处理隐含的复杂性，比让众多的计算机用户陷入复杂的、混乱的输出结果要好得多。2000 年 6 月成立 Vivisimo 公司，其目的是将这个项目成果开发为一种稳定的商业产品。该公司位于美国宾夕法尼亚州的匹兹堡。Vivisimo 的口号是“Information Overload was the Past”（信息超载已成为过去）。

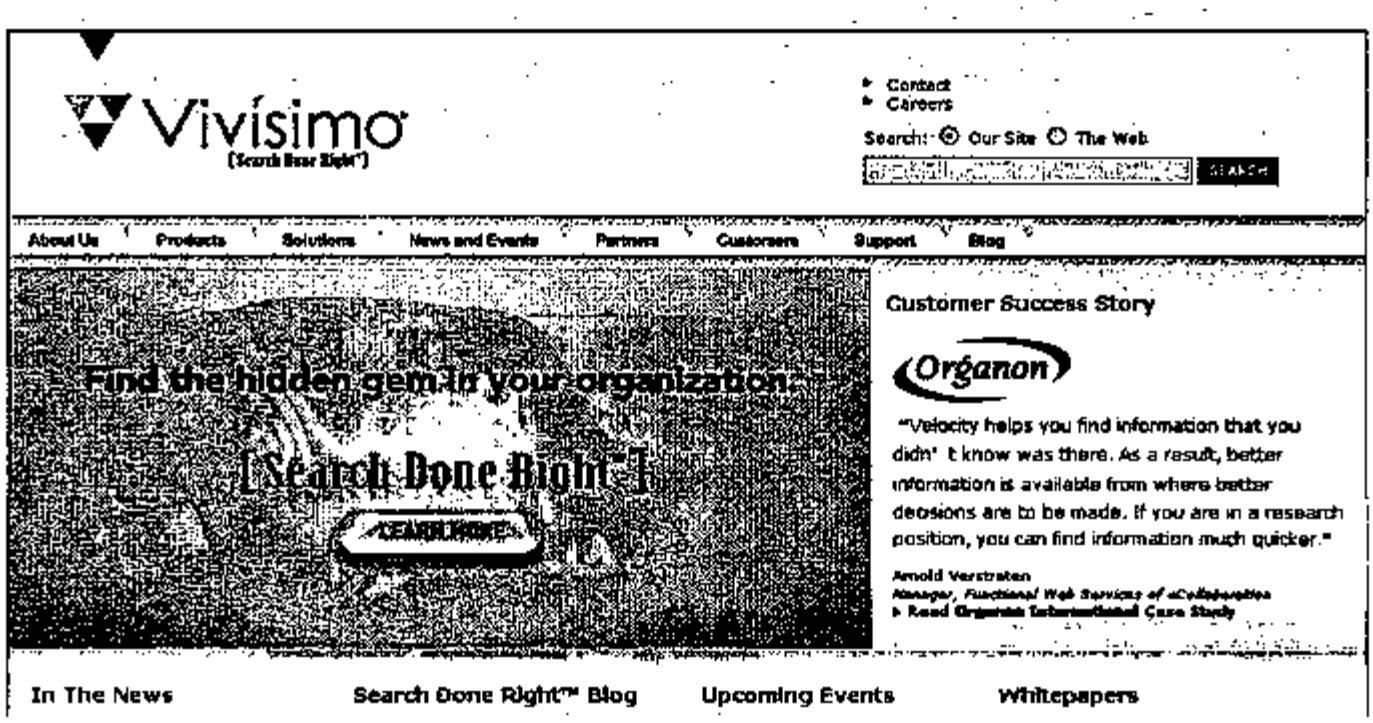

图 10—14 Vivisimo 主页

在拉丁语系里，Vivisimo 表示最高程度的活泼、愉快和聪慧，公司认为他们

的产品可以让死气沉沉的巨量文本信息处理过程变得生动活泼，因此，采用了Vivisimo的名字。

Vivisimo是一个很有特色的元搜索引擎，它采用了一种专门开发的启发式算法来集合或聚类原文文献。这种算法汲取了传统人工智能思想，强调对检索结果拥有更好描述和聚类。它的文献聚类技术将文本信息自动分类，分成了有意义的等级式排列的目录，它是完全自动化的，不需要人为地进一步干预，不需要维护。

1. Vivisimo检索功能

Vivisimo的检索功能在元搜索引擎中是一流的，提供了站内检索和站外检索功能。站内检索是对网站内的资源进行检索。站外具有简单检索和高级检索功能。在简单检索界面中，用户可以直接输入关键词、词组或组合的检索式，它的响应速度也很快。Vivisimo可以提供多种检索功能，能进行布尔逻辑检索，用“and”或“+”表示逻辑与，用“or”表示逻辑或，用“not”或“-”表示逻辑非。还可以进行限制检索，“domain:”表示检索的提问要出现在域名字段里；“host:”表示主机字段限定检索；“link:”表示链接限定，如，link：www.google.com表示检索链接了Google的网页；“title:”表示检索词要出现在题名字段里；“url:”表示网址限定。在高级检索界面里，用户可以自由地选择很多的限定条件，可以选择具体的Web搜索引擎、新闻搜索引擎、返回的结果条数、语言、显示格式、每条记录的打开方式、等待的时间、是否过滤等。

2. Vivisimo显示机制

Vivisimo的显示机制在网络检索工具中显得尤为突出。在检索结果的界面上，分为左右两个部分。界面的左边是类目显示，Vivisimo将一组组文献组成树状的等级式目录，采用“Windows Explorer”风格界面，允许用户轻松地逐层点击，以查询相关主题类目的信息。

站内检索结果所形成的目录是Vivisimo预先组织好的，包括News and Events、About Us、Other、Blog、Partners五个子类目。如图10—15。

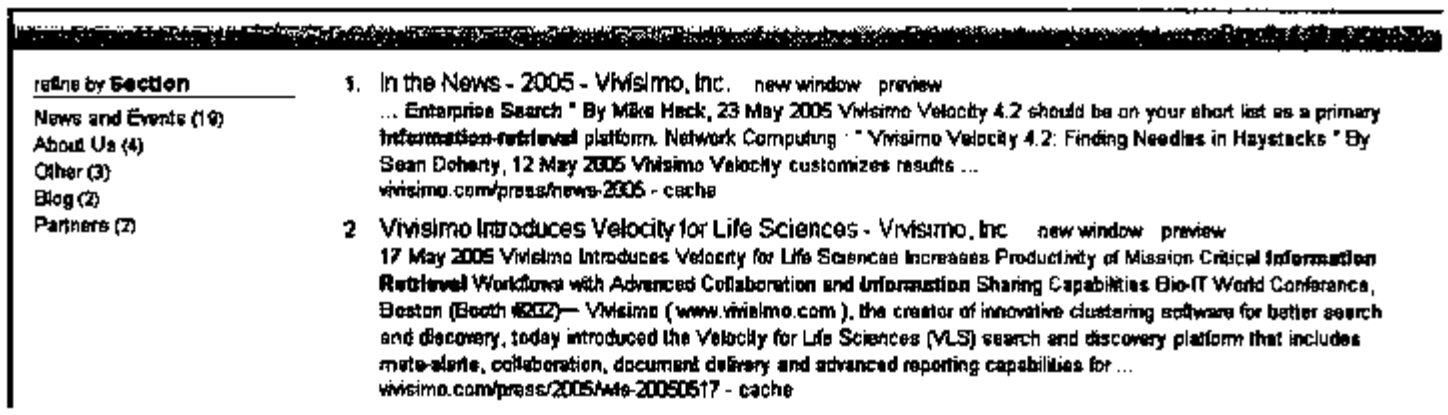

图10—15 Vivisimo站内检索结果

站外检索所形成的等级式的目录（如图 10—16）是 Vivisimo 对检索结果自动聚类的结果，是 Vivisimo 核心技术的直接体现。它与 Yahoo! 人工编制的目录有很大的差别，Vivisimo 的分类目录在体系结构和逻辑上也不是非常的严谨。由于它完全是自动处理的，因而，几乎没有任何的维护成本。而且，目录的显示非常易读，不需要对用户加以任何培训。Vivisimo 的自动分类是对检索结果的过程处理，具有很大的随机性，所以，即使输入相同的检索提问，每次所得的分类结果也不完全一样。

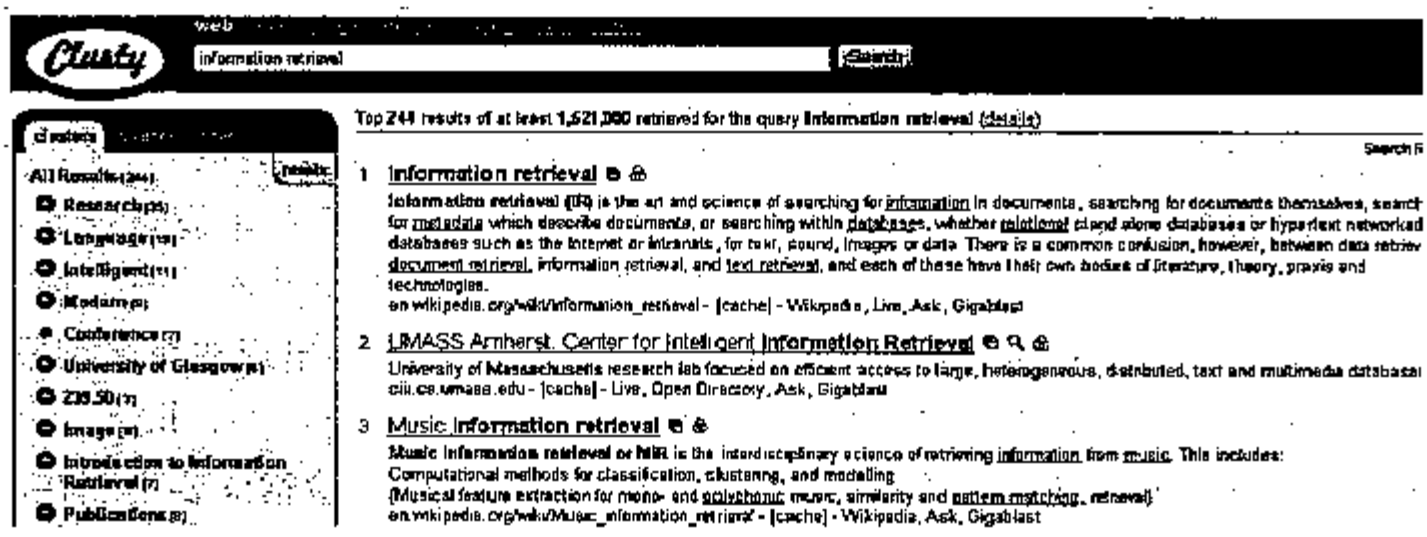

图 10—16　Vivisimo 站外检索结果

界面的右边是具体检索结果的显示，每条记录的格式比较规范，包括题名、摘要、URL、来源搜索引擎等。可以选择检索结果在新的窗口中打开、全屏打开、预览或保存等格式。

总之，Vivisimo 是一个功能比较全面的元搜索引擎。它开发的文献自动聚类技术代表了搜索引擎技术发展的新方向，它细致的检索结果显示机制是目前检索工具中的佼佼者。为此，2001、2002 年 Vivisimo 连续两年被《搜索引擎观察》(Search Engine Watch) 评为“最佳元搜索引擎”。

10.4.4.3　Ixquick (http://www.ixquick.com)

Ixquick（图 10—17）创建于 1998 年，现属于 Surfboard Holding BV 公司（荷兰的一家公司）。Ixquick 的口号是“全球最强大的元搜索引擎”（the world's most powerful metasearch engine）。它以简洁清爽的界面、灵活创新的风格，成为元搜索引擎家族中的最具光芒的新星，在 SearchIQ 的流行搜索引擎智商排行榜上以 145 的高分位居榜首。Ixquick 现支持 18 种语言搜索，其中包括简体中文和繁体中文，搜索的对象有网页、视频、图片和国际电话。它的搜索数据来自如 All the Web、MSN、Yahoo!、Ask、Open Directory 等 13 个搜索引擎。

Ixquick 独创了对检索结果的排序算法，即“星星体系”，用“☆”的多少来

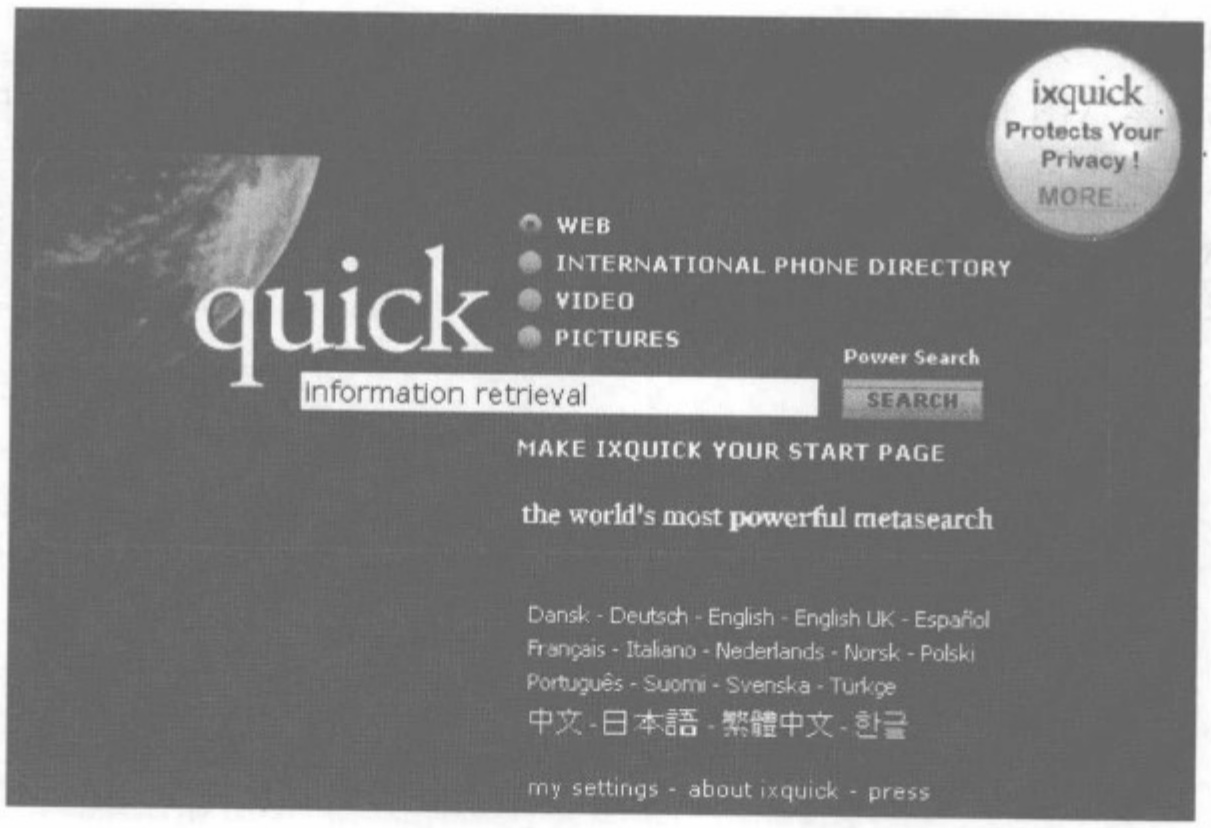

图 10—17　Ixquick 主页

决定检索结果的排序。Ixquick 只获取每个搜索引擎返回的前十条记录，如果一条记录被一个搜索引擎列入前十位了，它将获得一颗“☆”，如果被两个搜索引擎列入前十位了，它将获得两颗“☆”，依此类推。谁获得的“☆”最多，说明它得到的青睐最多，说明它已被更多的人所认可，Ixquick 自然认为它就是最好的，并将它排列在检索集合的首要位置上。正是由于采用了这样一种机制，保证了 Ixquick 有异乎寻常的检索速度和准确率。在传统上，多数搜索引擎开发商为了将最满意的结果最先呈现到用户面前，不遗余力地创建出各种“相关度指标体系”，比如依照检索词的位置、频率、链接等等方法。而 Ixquick 不像大多数元搜索引擎那样，致力于开发自己的指标体系，而是采取充分肯定和接纳的态度，以该记录被多少个搜索引擎所青睐为基本衡量标准，独创了它的“☆”的排序方法。

Ixquick 检索结果的输出格式也十分简单实用，包括：网页名称、文摘描述、URL、源搜索引擎以及该记录在源搜索引擎中的位置信息等等。如果点击了某个源搜索引擎，可打开另外一个窗口，全面了解该搜索引擎的检索结果。同时还在检索结果显示的页面上，进行相关检索，如检索“metadata”，得到了 57 条有“☆”表示的记录（即它们位于源搜索引擎前十条记录之中），这 57 条记录来自 73 999 013 个相匹配的结果。与“metadata”相关的检索主题有：Core、Dublin Core、Metadata Definition 等，用户可以通过它们获取相关的信息。

10.4.4.4 万纬搜索（http://www.widewaysearch.com）

万纬搜索（图 10—18）是上海万纬信息技术有限公司开发的一个中文元搜索引擎，集成了英文搜索引擎如 Google、Yahoo! 等和中文搜索引擎如天网、新浪、搜狐、雅虎（中文）、中文 Google、百度等。用户可根据需要自由选择其中多个引擎进行同步搜索，搜索结果可按相关度、时间、域名和引擎分类。

图 10—18 万纬搜索主页

1. 万纬搜索的检索功能

万纬搜索支持简单检索和高级检索，在检索框中键入关键词，选择好结果显示数量后，用户可以点击“一般查找”或“精确查找”键，引擎立即开始搜索。用户可以自由选择、设定查询结果的数量，如 10 个、20 个、50 个、100 个、150 个、200 个。引擎默认为 20 个查询结果。“一般查找”即为普通的检索，而“精确查找”的功能则指在最短的时间内，将使用网页智能分析和精确网络环境模拟技术，为用户提供最符合关键字的 10 条结果，这在一定程度上可以节省用户筛选检索结果的时间。

在高级检索界面中，除输入关键词外，还可以进行相关的限定。可以选择检索结果的排序方法，有四种选择：相关度、时间、域名分类、引擎等；可以选择具体的源搜索引擎；选择检索时间等。

2. 万纬搜索的特色

万纬搜索为用户提供了多种搜索结果的排列方式，如上所述，包括相关度、时间、域名分类、引擎等选择。选择按相关度方式排列，搜索结果按与关键字的相关程度排列；选择按时间方式排列，搜索结果按信息返回的时间长短排列；选择按域名分类方式排列，搜索结果按信息所属站点排列，共分为商业（.com）、教育（.edu）、政府（.gov）、组织（.org）、ISP（.net）五种，每种中再按相关度排列；选择按“引擎”方式排列，搜索结果按信息所属引擎分类后，再按相关度排列。

万纬搜索是目前一个比较优秀的中文元搜索引擎，在某些方面作了一定的尝试。但与国外先进的元搜索引擎相比，还存在着不小差距，如检索功能比较简单等，有待进一步发展和完善。

10.4.4.5 其他重要的元搜索引擎

除以上介绍的一些网络型元搜索引擎之外，还有一些经常利用的元搜索引擎，主要包括：

1. MetaCrawler（http：//www. metacrawler. com）

MetaCrawler 是世界上最早出现的元搜索引擎之一，于 1994 年由华盛顿大学的研究生 Erik Selberg 和副教授 Oren Etzioni 创建，次年提供 Web 服务，2000 年加入 InfoSpace 网络（the InfoSpace Network）。它曾两次被《PC Magazine》评为最好的搜索引擎。它的检索功能强大，除可以同时检索 About、Ask Jeeves、Fast、FindWhat、Google、Inktomi、LookSmart、Open Directory、Overture、Sprinke 等多个独立的搜索引擎外，本身还提供了包括近 20 个主题目录的检索，提供检索 Web 网页、图像、声频、多媒体、新闻组等信息。

2. Mamma（http：//www. mamma. com）

Mamma 是一个智能元搜索引擎，自称为“搜索引擎之母”（The Mother of All Search Engines），创建于 1996 年。可以调用 8 个独立的万维网搜索引擎（最多同时调用 7 个），可查询万维网、新闻、图像和声音文件等资源。

此外，元搜索引擎还有 InfoGrid（http：//www. infogrid. com）、Infonetware RealTerm Search（http：//www. infonetware. com）、Ithaki（http：//www. ithaki. net/dir. html）、qbSearch（http：//www. qbsearch. com）、Query Server（http：//www. queryserver. com）、Turbo10（http：//turbo10. com/）、Search. com（http：//www. search. com）、1Blink（http：//www. 1blink. com）、Gimenei（http：//gimenei. com/）、IcySpicy（http：//www. icyspicy. com）、Kartoo（http：//www. kartoo. com）、SurfWax（http：//www. surfwax. com）、ByteSearch（http：//www. bytesearch. com）、Fazzle（http：//www. fazzle. com）等等。

【案例】

互联网搜索大赛赛题节选[①]

1. 中国古代四大名镇不包括下列哪个城镇：（单选）

① http：//mail. galiji. cn/html/2/1/20071225/6118. html，2008-05-15。

A. 河南朱仙镇

B. 江西景德镇

C. 江苏周庄镇

D. 广东佛山镇

2. 我国周朝时曾将乐器分为八类，即“金、石、土、草、丝、木、匏、竹”，请选出下列选项中不正确的一项：（单选）

A. 金类：钟，铃

B. 土类：埙，缶

C. 丝类：琴，瑟

D. 竹类：箫，笙

3. 赛迪网中国信息化频道专栏作者王延东在该频道发表过哪些作品？（多选）

A.《企业信息化寓言新解》

B.《遗憾三叹信息化》

C.《ERP 的“紧箍咒”》

D.《CRM 离传统企业有多远》

E.《企业信息化随感：逾淮之橘》

4. 雨果的小说《巴黎圣母院》中漂亮的吉卜赛女郎艾丝米拉达和善良的“钟楼怪人”卡西莫多的生活舞台巴黎圣母院，位于塞纳河中心的“西岱”岛上，是巴黎最负盛誉的名胜之一，它是法国_____建筑的代表？（单选）

A. 哥德（特）式

B. 巴洛克风格

C. 拜占廷式

D. 罗马式

5. 漂亮富饶的青海湖是我国面积最大的咸水湖。古称“西海”，藏语称“错温波”，蒙语称：（单选）

A. 鲜水

B. 鲜海

C. 波拉库

D. 库库诺（淖）尔

6. 美国硅谷，是一个时刻都在制造奇迹的地方，是全世界高科技创业者、投资家的梦想乐园。你知道硅谷位于美国的哪个州吗？（单选）

A. 密苏里州

B. 新泽西州

C. 华盛顿州

D. 加利福尼亚州

7. “纸上谈兵”这一成语出自古代的哪场战争？（单选）

A. 围魏救赵

B. 长平之战

C. 白登山之围

D. 淝水之战

8. 国际标准是指国际标准化组织（ISO）、国际电工委员会（IEC）和_____制定的标准，以及国际标准化组织确认并公布的其他国际组织制定的标准。（单选）

A. OEM

B. UIC

C. WTC

D. ITU

9. 中国首次成功地运用自己的运载火箭完成为国外用户发射商用卫星的服务是在：（单选）

A. 1988 年

B. 1990 年

C. 1992 年

D. 1998 年

10. 古人云：五岳归来不看山，黄山归来不看岳。黄山以其罕见的地质结构，奇异的峰林地貌吸引了无数中外游客。下面选项中不属于黄山四绝的是：（单选）

A. 奇松

B. 怪石

C. 古洞

D. 温泉

关键术语

网络信息检索工具	搜索引擎	搜索引擎发展历程
搜索引擎及结构原理	搜索引擎特点及功能	搜索引擎发展趋势
网络资源目录	网络资源目录类型和特点	元搜索引擎

元搜索引擎原理　　元搜索引擎类型　　元搜索引擎评价

思考题

1. 网络检索工具可以分为哪几种类型，各有什么特点？

2. 简述搜索引擎的发展历程。

3. 简述搜索引擎的结构及工作原理。

4. 试比较中文 Google 和百度。

5. 利用 Google 查找中央电视台网站上关于“伊拉克战争”的网页。

6. 简述网络资源目录的原理和特点。

7. 试分析分类法与超文本技术的关系。

8. 网络资源目录有哪几种类型？

9. 目前在网上使用的传统分类法型的网络目录资源有哪些？试对它们加以评价。

10. Yahoo!、Open Directory、搜狐的目录各有什么特点？试用该目录查询“信息检索”方面的有关网络资源，并说明查询过程。

11. 简述元搜索引擎的原理和特征。

12. 元搜索引擎可分为哪几种类型，各有什么特点？

13. 元搜索引擎主要涉及的技术有哪些？

14. 试分析和评价 Dogpile 和 Vivisimo。

15. 请自拟一个检索课题，利用三个元搜索引擎进行检索，并比较检索结果。

CHAPTER ELEVEN

第11章
网络数据库检索

【本章要点】

◇ 阐述网络数据库的由来、优势及其检索步骤
◇ 介绍国外重要网络数据库
◇ 介绍国内主要网络数据库

引子

2006年3月，中国互联网络信息中心发布的中国互联网络信息资源数量调查报告显示：2005年全国在线数据库的总量为29.54万个。其中，企业网站拥有的在线数据库数量最多，占全部在线数据库的50.4%；其次是个人网站拥有的在线数据库，占全部在线数据库的21.5%；政府网站拥有的在线数据库排第三位，占全部在线数据库的9.4%；其他公益性网站的在线数据库占7.3%；教育科研网站的在线数据库占6.4%；商业网站拥有的在线数据库占4.5%。[①] 在线数据库即网络数据库，是数据库技术与现代网络技术相结合的产物。早期数据库生产商往往以自己生产的数据库提供国际联机服务，或以光盘数据库形式推向市场，随着网络技术的崛起，国内外大型的传统数据库服务提供商都逐渐开始网络数据库的相关服务，如ProQuest、EBSCO、Web of Science、《中国期刊全文数据库》、万方数据资源系统等。网络数据库由于其特有的优势，已逐渐成为数据

① 参见http://www.cnnic.net.cn/index/0E/00/12/index.htm。

库的主流形式。

11.1 网络数据库概述

11.1.1 网络数据库的由来及其优势

网络数据库（Web-database），或称网络版数据库，是指由数据库生产商在互联网上发行，通过计算机网络提供信息检索服务的数据库。

数据库技术是计算机处理与存储海量数据的最有效、最成功的技术，而网络则是共享资源数据最方便、最成功的典范。因而，网络数据库既具有一般数据库的特点，同时又有着明显的网络化特征，成为目前数据库服务方式的主流。

11.1.1.1 网络数据库的由来

随着计算机网络技术和计算机存储技术的发展，基于数据库的信息检索大致经历了三个主要的发展阶段。

第一阶段：20 世纪 70 年代初至 80 年代中期，为专线联机阶段。业内人士称这一阶段为 DIALOG 称雄时代。1972 年 DIALOG 联机检索系统开始投入商业运营，又很快兼并了 Data Star 公司，使其拥有的数据库迅速增加，真正成为了信息检索服务业的龙头公司。专线联机检索使数据库进入了现代化服务阶段。我国通过数据通信专线与 DIALOG 联机的终端 20 年来没有普遍发展，大都设置在研究机构或大学里，不可能个人拥有。而且操作繁琐，只能使用复杂的指令式检索方式，一般限于受过专门训练的情报检索人员使用，检索费用也比较高。因此，妨碍了机器检索的普及。

第二阶段：20 世纪 80 年代中期至 90 年代中后期，为光盘数据库阶段。1983 年，日本索尼公司和荷兰飞利浦公司联合生产出第一张只读光盘（CD-ROM），这种新载体具有存储量大、体积小、便于携带和保存等诸多优点，成为数据库的极好载体。光盘促使数据库进入大发展时期。1988 年 6 月，全世界光盘数据库只有 200 种，到了 1994 年，Mechlermedia 公司的在版 CD-ROM 收录的光盘数据库已有 6 000 多种。很快世界上绝大多数二次信息数据库都有了光盘版。光盘使数据库的用户大大增加，普通用户开始享受到机器检索的方便快捷。

第三阶段：20 世纪 90 年代中后期至今，为网络数据库阶段。美国专业人士称 1998 年是数据库的上网年，认为尽管在 1998 年以前许多数据库已经上网，但从 1998 年起上网的数据库激增，并显示出 Internet 网络数据库技术的进一步发

展，数据库提供商在网络数据库信息传递服务方面趋于成熟。目前几乎所有大型数据库都已建成网络数据库，提供远程信息检索服务。①

11.1.1.2　网络数据库的优势

Internet 在全球的迅速发展，为信息资源的数字化、网络化提供了契机。网络成为现代信息资源存储、交流和利用的主要媒介，信息资源网络化已成为一种趋势。网络数据库作为数据库技术在网络环境下的新发展，表现出一定的优势。

1. 信息容量大、增长迅速、更新及时

目前，数据库生产已形成规模，走向产业化和商业化，这就使得网络数据库的整体发展呈现出以下两个特点。一是数据库规模大、数据量多，增长迅速。二是数据更新速度快、周期短。由于它的存储载体是服务器硬盘，随着硬盘容量的升级，网络数据库的海量存储能力也将提高。例如，网上的 SCI 数据库（《科学引文索引》）收录期刊容量是 5 600 余种，比光盘版 SCI 数据库 3 500 余种的收录范围增加了 2 100 余种。网络数据库在数据数量和更新速度上都显现出一定的优势。大型库一般每周更新，有的甚至通过网络即时更新，不像光盘数据库存在明显的时滞问题。甚至有的电子期刊数据库的更新通常早于其相应的印刷版，为每周或每日更新。

2. 使用方便，界面友好

WWW 浏览器为用户提供了便捷的信息查询方式，由于网络数据库基于 Web 站点和服务器实现用户共享，且大都不间断提供服务，24 小时开放，因此，用户只要拥有一台上网的计算机，并拥有使用数据库的授权，就可随时查检所需的网络数据库，这种服务方式不仅优于单机使用的光盘数据库，而且在查询技巧上也比检索指令复杂的联机数据库简捷方便。

网络数据库面向大众用户，检索界面清晰友好，表现生动形象，易于理解，便于使用。如在有的数据库中，不同的文献类型用不同的图形符号标示，生动直观。同时，允许用户对要查找的信息资源进行选择和限定，如，可在不同的数据库或文档、不同检索方式之间自由切换与选择，可对文献类型、出版时间、出版形式、可检字段等进行限定与选择，用户只需点击相关的选项，即可完成选择与链接操作。

3. 检索功能强大

网络数据库具有较为强大的检索功能，查全率和查准率比较高。可以提供不同层次的检索方式。除提供基本或简易检索模块，供一般用户使用外，还可提供

---

① 参见何小清：《数据库服务方式的发展趋势》，载《情报学报》，2002 (2)。

各种形式的高级检索模块，以方便用户进行各种限定检索，或使用逻辑算符（AND、OR 和 NOT）、括号、位置算符、截词符等构造检索式，进行组配检索，使得检索更为灵活，更为准确。一般网络数据库都提供多途径检索入口，允许用户根据自己的需要选择不同的检索途径，包括关键词（Keyword）、题名（Title）、著者（Author Name）、文摘（Abstract）、全文（Full Text）等等，有的网络数据库甚至可以提供几十个检索入口。

网络数据库的回溯检索能力虽然无法与联机数据库相比，但与网络检索工具（如搜索引擎）和光盘数据库来比，还是比较强的。Web of Science 检索可回溯到 1973 年，比光盘版回溯到 1980 年更早。中国学术期刊网检索目前可以回溯到 1994 年，比数据始于 1996 年的光盘版又早了两年。

4. 检索结果的显示与输出形式灵活、多样

比较成熟的网络数据库一般都提供了灵活多样的检索结果显示形式。用户可以按照自己的需要选择检索结果显示的排序方式，常见的有按照相关度、日期等，还有按照文献标题、著者、来源、语言、出版国等多种方式升序或降序排列。用户还可以自主地选择每页显示的检索结果的数目，如 10 条、20 条、30 条、50 条等。也可以选择检索结果的显示格式，可提供题录（Citation）、题录＋文摘（Citation ＋ Abstract）、全记录（Complete Field）或选择字段（Select Field）等多种格式显示。

大部分网络数据库给用户提供了更灵活的输出方式，用户可以直接对检索结果进行存盘和打印，可利用 E-mail 发送检索结果，抑或直接在网上订购文献全文。

5. 可在异地建立镜像站点

网络数据库不需要本地驱动器（包括光盘塔、光盘库）和相应的服务器等硬件设备，利用 Internet 的服务器和 FTP 功能，网络数据库可以在不同地区建立它的镜像站点，这样不仅使用户获得最佳的检索效果，而且节省时间与传输距离，突破空间限制，实现异地远程检索。比如中国期刊网、万方数据中心等国内著名数据库都已相继成立多个镜像站点。

6. 原文获取功能强

据有关统计资料表明，全世界每年约出版图书 80 万种，期刊 16 万种，而我国每年引进量分别不到 10 万种和 3 万种，再加上经费不足等原因，无法购全各种资料，这就造成了在检索中查到的大量文献源在国内找不到原始文献。利用网上全文数据库在一定程度上可以弥补该缺陷。

全文型的网络数据库直接为用户提供了获取全文的服务，同时，一些书目索引文摘等二次信息数据库也与全文数据库之间建立链接，帮助用户迅速、直接访

问、获取所需原始文献信息，增强数据库的全文提供能力。目前，数据库供应商提供原始文献链接的方式主要有两种，一是链接到出版商的电子期刊全文，二是链接到相应的全文数据库。

7. 较强扩展整合功能

网络数据库除了为用户提供信息查询服务外，还提供有多种整合功能。首先，网络数据库可以与图书馆馆藏进行链接与整合。目前数据库供应商提供的链接方式有两种：一是数据转入或人工直接输入；二是单向式或双向式直接与 OPAC 链接。数据库与图书馆馆藏的整合通常可通过数据的上载和下载实现。其次，网络数据库与其他数据库进行链接与整合。利用网络中的导航功能，能为用户提供网络中其他相关数据库资源的信息检索。如 Web of Science 具有和 DERWENT 专利数据库相对应的链接，可以因此了解到某些科研成果应用于实际生产中的情况以及与此相关的交叉学科，对科学研究、查新检索等具有较高的参考价值。

8. 可提供多种服务形式

许多网络数据库在满足用户查询信息的基本要求的前提下，也开发了一些其他的电子信息服务。主要包括：文献传递服务和定题服务。文献传递服务指当用户从二次信息数据库中查到所需信息并希望得到文献全文时，可以通过电子方式在线订购该文献全文。原文订购若选择电子文献传递方式，一般在 24 小时内即可获得所需文献，方便快捷，可弥补书目索引文摘等二次信息数据库不能提供全文的不足。定题服务，是根据用户需求，定期不断地将符合用户需求的新的信息传送给用户的一种服务模式。数据库供应商提供的 SDI 服务主要有两种。一是指定参考用书（Reserved List）服务，即系统提供图书馆依主题方式整理出类似所谓的指定参考用书功能选项，从而达成专门的 SDI 服务。二是个性化文献报道服务，即由用户创建自己的检索策略，系统定期将符合条件的检索结果传递给用户。例如，Uncover 的最新文献报道服务（Uncover Reveal），由用户选择自己感兴趣的关键词或期刊（最多可选 50 种）建立用户需求文档，系统每周一次，自动地将相关文献及用户所选期刊的最新一期目次信息发送到用户的 E-mail 信箱，用户只需定期查看自己的 E-mail 信箱，即可及时了解最新研究动态。

11.1.1.3　网络数据库的评价

网络数据库作为一种重要的电子资源，已成为人们获取信息的重要来源。根据特定的方法评价网络数据库的优劣对我们选择、开发和优化数据库资源有着非常重要的意义。

1. 数据库内容

内容是数据库的核心，是评价一个数据库的首要标准。数据库内容的评价指

标主要包括文献的收录范围、权威性与连续性、时间跨度、文献总量、更新频率、全文占有量等。

评价数据库文献的收录范围，主要是分析其所覆盖的学科范围是否与服务需要相符，期刊数量是否全面。如果一个经济学的数据库提供一些非经济学的资源，这部分资源的利用率往往达不到预期效果；而如果这个数据库没有收录重要的经济学期刊，那么它的收录也是不全面的。

对数据库收录文献的权威性与连续性进行评价，也是一种对文献外部特征进行的评价。首先，就权威性而言，判断的依据主要看数据库文献是如何进行选择的，其收录的文献的来源机构是什么。学术性、权威性和专业性较强的出版社、学会或专业机构出版的刊物往往具有非常高的权威性，能够得到广泛的认可。期刊第一次被收录的时间、收录期间的期数，是否存在缺年和缺期等不完整的情况，则反映了文献的连续性，一份连续性较强的刊物所具有的学术价值会比较高。

内容的时间跨度是指期刊收录的起止年限，时间跨度越大学术价值越高；文献总量并非越多越好，但是文献总的种类和篇数多的数据库往往具有较高的价值；文献的更新频率会影响到用户对数据库的使用体验，更新越及时越能体现数据库的新颖性和实效性；即使是全文数据库，也并不能保证文献都是全文收录，因而全文占有量也是衡量数据库内容的一个重要指标。

2. 检索功能

数据库的资源量非常庞大，然而，如果没有好的检索功能，即使有浩如烟海的数据，对用户而言也是没有什么意义的，因为他们无法把需要的信息找出来。对检索功能的评价主要有以下四个方面的指标：用户界面、检索手段、检索技术和检索效果。

（1）用户界面的设计应该简洁明了，符合用户的视觉特点和阅读方式，便于用户检索。同时，如果界面能够根据用户的不同需求进行个性化的设定，将更加方便用户的使用。

（2）一个数据库检索功能的优劣还和检索手段有关。一个好的检索系统提供的检索手段不仅有初级检索、高级检索、分类检索等，还应当覆盖文献的各种外部信息，如篇目、作者、关键词、摘要、期刊名等等。著名的数据库 EBSCO 还提供了一种非常特别的检索方式——视觉检索。这种检索方式可以帮助用户更加直观地找到自己需要的文献。

（3）检索技术主要是看数据库在用户检索时是否允许用户使用布尔逻辑运算符、通配符运算，是否支持检索项的扩展和跨库检索等，以及在显示检索结果时是否允许用户对检索结果进行自主设置，如排序方式、每页显示的结果条数、全

文下载的格式等。

（4）检索效果则主要是通过检索结果的查全率、查准率，检索速度等指标来衡量。提高查全率或者查准率，同时缩短检索花费的时间也有助于提高数据库的评价。

3. 数据库的服务

数据库的服务与检索的过程和结果无关，但是却依然会影响到数据库的使用。有的数据库提供电子邮件发送检索结果的功能，有的数据库会有非常详细的在线使用指南，还有的数据库会提供一些特色的额外服务。例如，引文数据库 Web of Science 还提供了对检索结果的分析功能，引入这项服务可以大大节省用户对文献进行人工筛选和分析的时间，也吸引了更多科研人员使用 Web of Science。

11.1.2　网络数据库的检索方式与步骤

网络数据库是一种基于浏览器/服务器（B/S）的数据库，可分为免费数据库（只要连入 Internet 就能使用）和付费数据库（只有付费获得授权才能使用）。其中付费的网络数据库，我们称其为商业数据库。

11.1.2.1　网络数据库的检索方式

网络数据库将数据存放在远程服务器上，用户可通过 Internet 直接访问，也可通过 Web 服务器或中间服务器访问。目前，商业数据库提供商一般都使用单位的 IP 地址控制方式进行授权访问（个别的系统还加上账号和密码）。所以，只要一个单位购买了某些网络数据库，该单位局域网中任何一台计算机都能免费访问该数据库。对用户来说，网络数据库的检索方式有以下几种：

1. 免费检索

对于免费 Web 数据库，用户在选定数据库并输入检索提问式后，就可以进行查找并显示出符合条件的所有记录。网上免费数据库一般多是题录数据库或文摘数据库，只能检索到文献的题录或者文摘，不能看到全文。例如，我们可以直接登录“中国期刊网 CNKI 数字图书馆”的网址（http：//www. chinajournal. net. cn/index. htm），免费检索《中国期刊题录数据库》。

2. 普通用户检索

对于计费数据库，在选定计费数据库并输入检索提问式后，就可以进行检索，并显示出符合条件的记录；但每条记录只显示部分字段内容，用户不能看到记录的全部字段内容。

3. 授权检索

对已经申请注册并拥有合法用户名、口令的用户，在选定计费数据库，输入

用户名和口令，并输入检索提问式后，就可以显示出符合条件的所有记录和记录的全部内容。如各高校、科研院所等购买的网络数据库，就得到了数据库生产商的授权，可以提供给本单位局域网上的用户来利用。

11.1.2.2 网络数据库的检索步骤

网络数据库的检索步骤与一般数据库的检索步骤相似，主要有以下几个步骤：

1. 检索课题的主题分析

实施检索前，首先要对所检主题进行深入研究，确定检索的主题概念。

2. 数据库的选择

应根据所检课题的学科范围或主题概念来选择相关的数据库。网络数据库很多，应首选与该课题有关的最具权威性和数据容量大的网络数据库，在此基础上还可选择一些与其主题概念密切相关的网络数据库作为补充。选择好数据库后，进入载有网络数据库的站点，或可直接进入网络数据库生产商的网址进行检索。

3. 检索策略的选择

依据对课题主题分析的结果，确定检索词和检索式，将检索需求转换为网络数据库认可的检索式。这是网络数据库检索过程中重要的环节之一。

4. 实施检索

输入拟定的检索式或检索词，开始检索。

5. 检索策略的优化

在对检索结果进行分析后，可根据需要改进或改变检索策略，各种不同的网络数据库的优化技巧各不相同。

6. 辅助性检索

可以依据网络数据库所提供的一些辅助检索功能，进行相关的检索，或者进一步精确检索结果。

7. 检索结果的输出

网络数据库检索结果的输出形式各不相同，大致有：存盘、打印或 E-mail。

11.2 国外网络数据库检索示例

11.2.1 ProQuest 系列数据库

ProQuest 系列数据库（http：//proquest. umi. com/pqdweb）是 ProQuest Information & Learning 公司通过 ProQuest 系统提供的一组数据库，内容涉及商

业管理、社会与人文科学、新闻、科学与技术、医药、金融与税务等广泛领域。1985 年，该公司收购了数据收集与生产公司 UMI，并使其成为缩微胶片产品的品牌，包括 18 000 多种外文缩微期刊、7 000 多种缩微报纸、150 多万篇博士/硕士论文、20 多万种绝版书及研究专集。1996 年起公司开始推行数据库的网络信息服务。该公司 Web 版数据库的主要特点是将二次信息与一次信息“捆绑”在一起，为最终用户提供文献获取一体化服务，用户在检索文摘索引时就可以实时获取大部分全文信息。

该系列数据库检索功能完善，检索方法多样，包括基本检索、指南检索、高级检索、自然语言检索、出版物检索等，在一定程度上体现了英文数据库的检索特色。

11.2.1.1　基本检索（Basic Search）

进入数据库后，默认的界面为基本检索界面（图 11—1）。基本检索操作简便，查询速度快，在检索框中输入一个单词、词组或短语就可以进行检索。但检索结果过于宽泛。

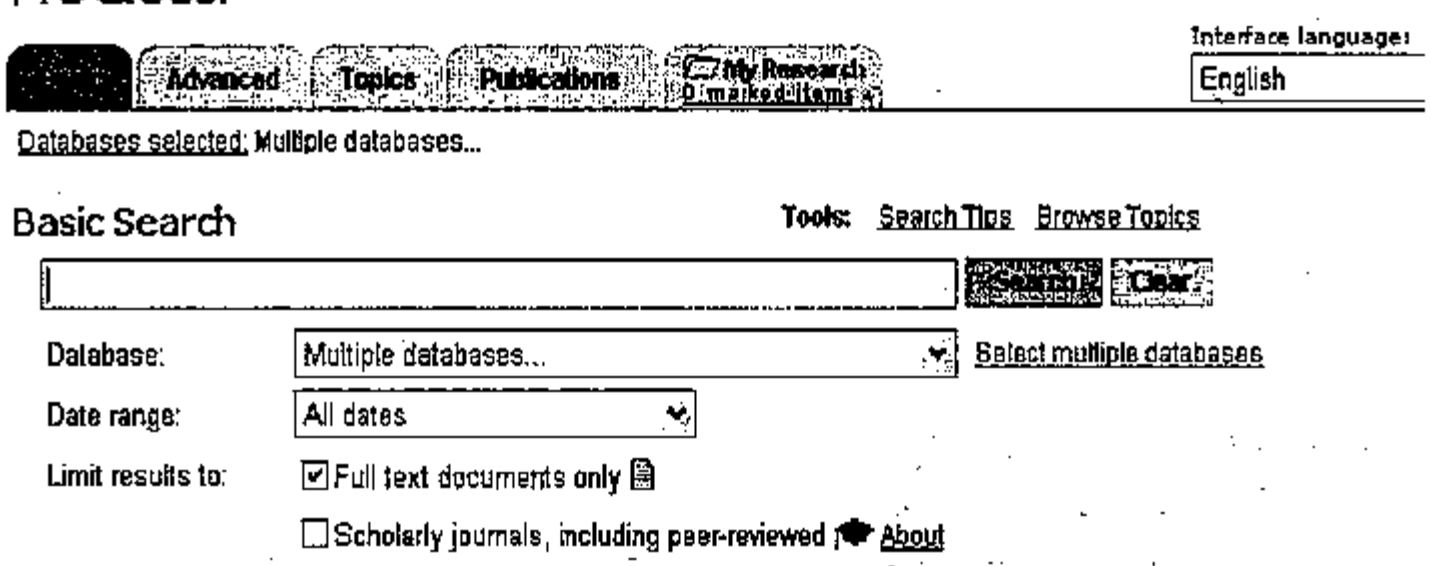

图 11—1　ProQuest 系列数据库基本检索界面

（1）检索词输入框。输入的检索词可以为 a word（单词）、words（词组）、specific phrase（短语）或 a sentence（一句话），检索结果为包含该单词、词组、短语或句子中每一个单词的文章。如输入 information retrieval，检索结果为包含 information 和 retrieval 的文章。

（2）选择数据库。在数据库选项框的下拉列表中选择要查询的数据库。可以选择一个或者多个数据库进行检索。

（3）限定时间范围。用户可以选择最近的一周、一月、一季度、一年等时间范围，用户也可以自由选择日期范围。

（4）限定检索结果的显示。可以只显示带有全文的结果，或者学术期刊等等。

11.2.1.2 高级检索（Advanced Search）

系统支持组合检索，实现检索词间的组配关系。提供的逻辑算符包括：AND、OR、AND NOT 等。同时，ProQuest 还可以限定检索词出现的字段，如题录和文摘、题录和全文、文摘、作者、分类号、公司/机构名称、文献类型、文献代码、文献语种、文章标题、图表标题、出版物名称、主题等。

11.2.1.3 主题检索（Search by Topics）

主题检索可以方便用户浏览和检索某一主题范围的文献。有检索主题和主题浏览两种方式。主题检索在某种意义上说就是通过词表进行的检索。

11.2.1.4 出版物检索（Search by Publications）

检索某一种特定出版物的全文，包括对某一特定卷期内容的检索。用户在已知刊物的情况下，可以通过出版物检索来检索和浏览文章。这种检索方法比较适合对整刊的浏览。

11.2.2 EBSCO 系列数据库

EBSCO 公司是世界上最大的提供期刊、文献订购及出版服务的专业公司之一，Academic Search Premier（《学术期刊数据库》）和 Business Source Premier（《商业资源数据库》）是 EBSCO 公司最重要的网络版数据库。

11.2.2.1 数据库的选择

通过 EBSCO 设在国内的镜像站点，选择“Business Searching Interface”直接进入 Business Source Premier 数据库，而选择“BSCOhostWeb”或“BSCOhost Text Only”（速度较快）则进入数据库选择页（图 11—2），在数据库选择页用户就可以勾选需要检索的数据库了。

11.2.2.2 数据库的检索

EBSCOhost 提供三类检索方法：基本检索（Basic Search）、高级检索（Advanced Search）和视觉搜索（Visual Search）。以 Academic Search Premier 数据库为例，EBSCOhost 又分别提供关键词（Keyword）、出版物（Publications）、主题（Subject Terms）、引文（Cited References）、索引（Indexes）、图片（Images）等多种检索途径。如不特别指定，系统默认关键词（Keyword）方式。

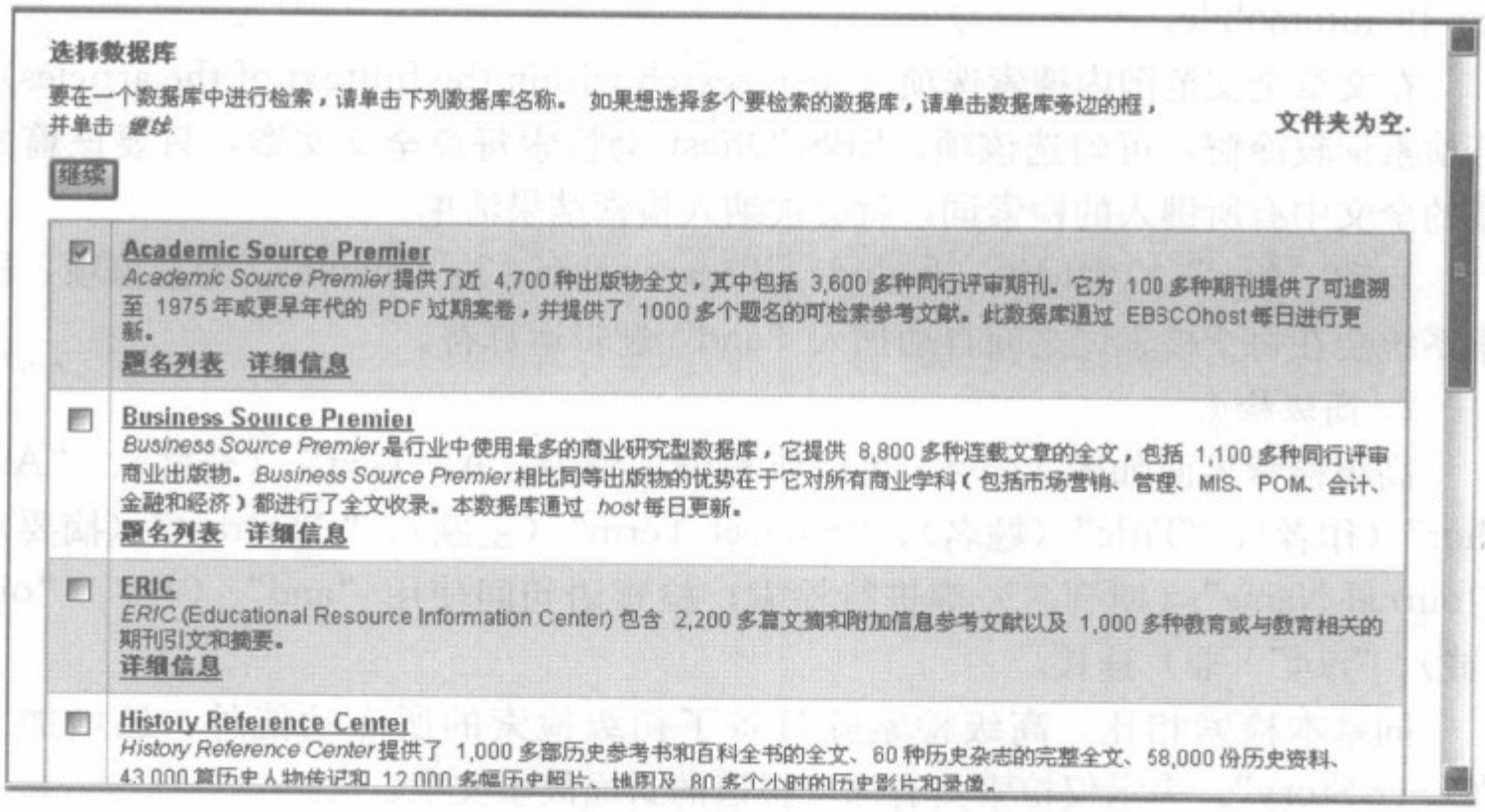

图 11—2　EBSCO 选择数据库界面

1. 基本检索

基本检索界面如图 11—3。用户除了可以选择是否只是全文检索、文献出版时间、出版物名称和类型等限定条件之外，还有扩展条件，如“也可以搜索相关关键词”、“也可以在文章的全文范围内搜索”、“自动‘And’检索词语”等等。

图 11—3　EBSCO 基本检索界面

搜索相关关键词选项（Also search for related words）：勾选该项，会将检索词的同义词或单复数的文献一并检索。如：键入“car”，EBSCOhost 会检索到

car 和 automobile。

在文章全文范围内搜索选项（Also search within the fulltext of the articles）：若检索词较冷僻，可勾选该项，EBSCOhost 会检索每篇全文文章，只要该篇文章的全文中有所键入的检索词，就会被纳入检索结果清单。

自动“And”检索词语（Include all search terms by default）：勾选该项，检索系统会在每个检索词之间自动加入“and”逻辑运算符。

2. 高级检索

高级检索界面如图 11—4。可以将检索项设为“All Text”（全文）、“Author”（作者）、“Title”（题名）、“Subject Term”（主题）、“Abstract”（摘要）、“Journal Name”（期刊名）等进行检索；检索语句间使用“and”（与）、“or”（或）、“not”（非）连接。

同基本检索相比，高级检索除具备了初级检索的所有功能外，还增加了“Cover Story”，表示仅检索具有深度报道的封面故事文章。

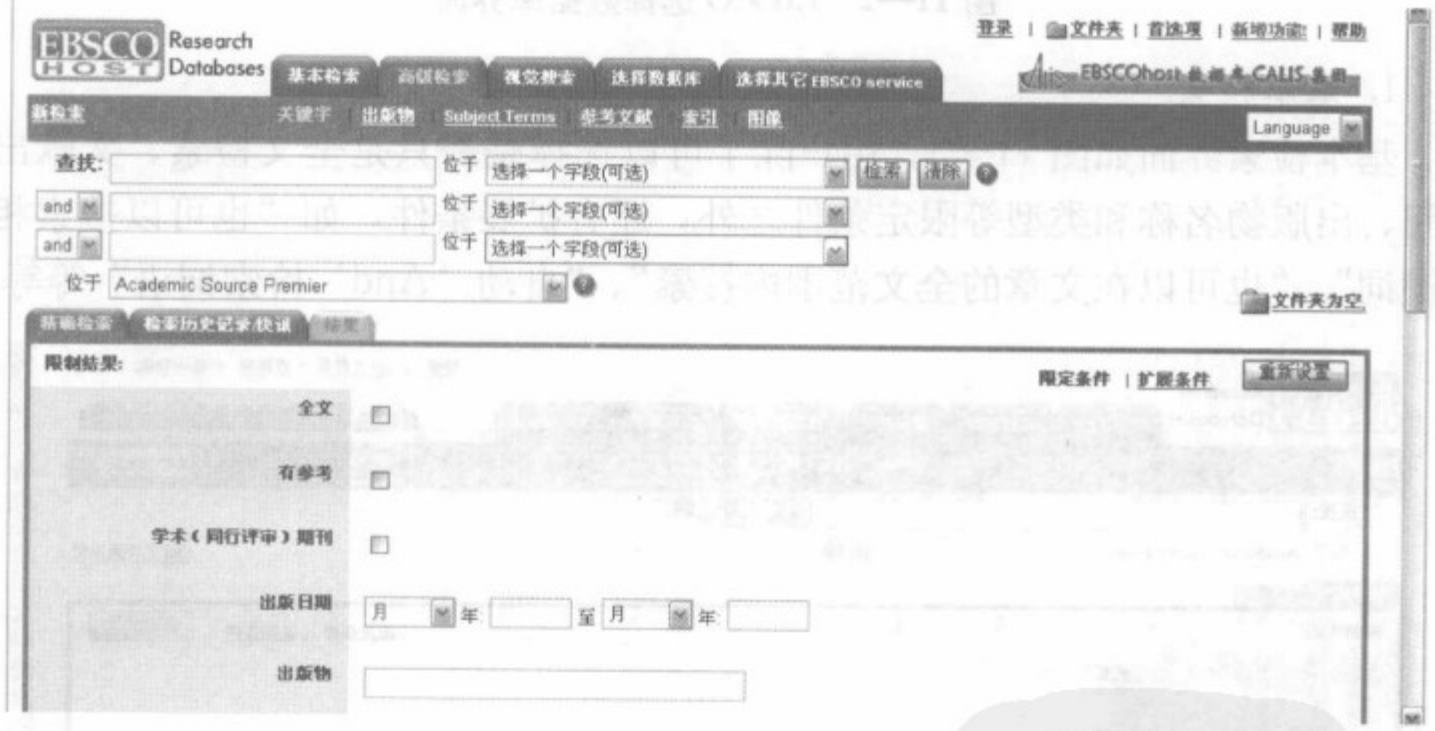

图 11—4　EBSCO 高级检索界面

3. 视觉搜索

EBSCOhost 的视觉搜索是该数据库中特有的图形检索方法，使用简单方便，其界面如图 11—5。使用视觉搜索可在广泛的主题中高效地进行搜索，之后返回结果的视觉导航图，并按主题进行排列，以查找“information retrieval”为例，检索步骤如下：

第一步：点击“视觉搜索”选项卡，在“查找”文本框中输入“information retrieval”，点击“检索”。

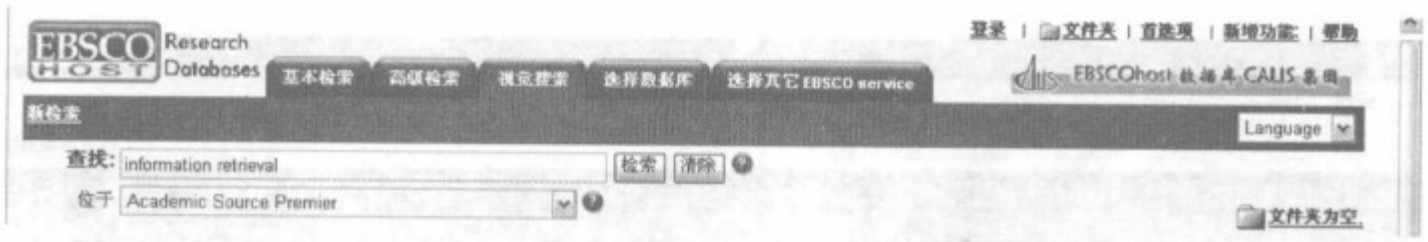

图 11—5　EBSCO 视觉搜索界面

第二步：检索结果根据不同主题组织在一个可以放大缩小的视图中。用“information retrieval”作为主题词进行视觉搜索的检索结果如图 11—6。“information retrieval”在视觉搜索中分为“Computer System”、“Electronic Data Processing”、“Electronic Information Resources”、“Information Science”、“Management Information”、“Query Information Retrieval System”等子类。点击一个类目（圆形图案），可以进入该类的内部；点击类目的外部或者“返回上一视图”，可以返回到视图结构上一层。

图 11—6　以“information retrieval”为主题词进行视觉搜索的结果

第三步：与第二步相似。点击“Query Information Retrieval System”，进入下一层级，进入后点击“Information Services”。在“Information Services”里有 8 个矩形，可以链接到具体的文章。用户如果希望预览文章的基本信息，只需要将鼠标移至矩形上，关于该文章的部分元数据就会显现，如图 11—7。

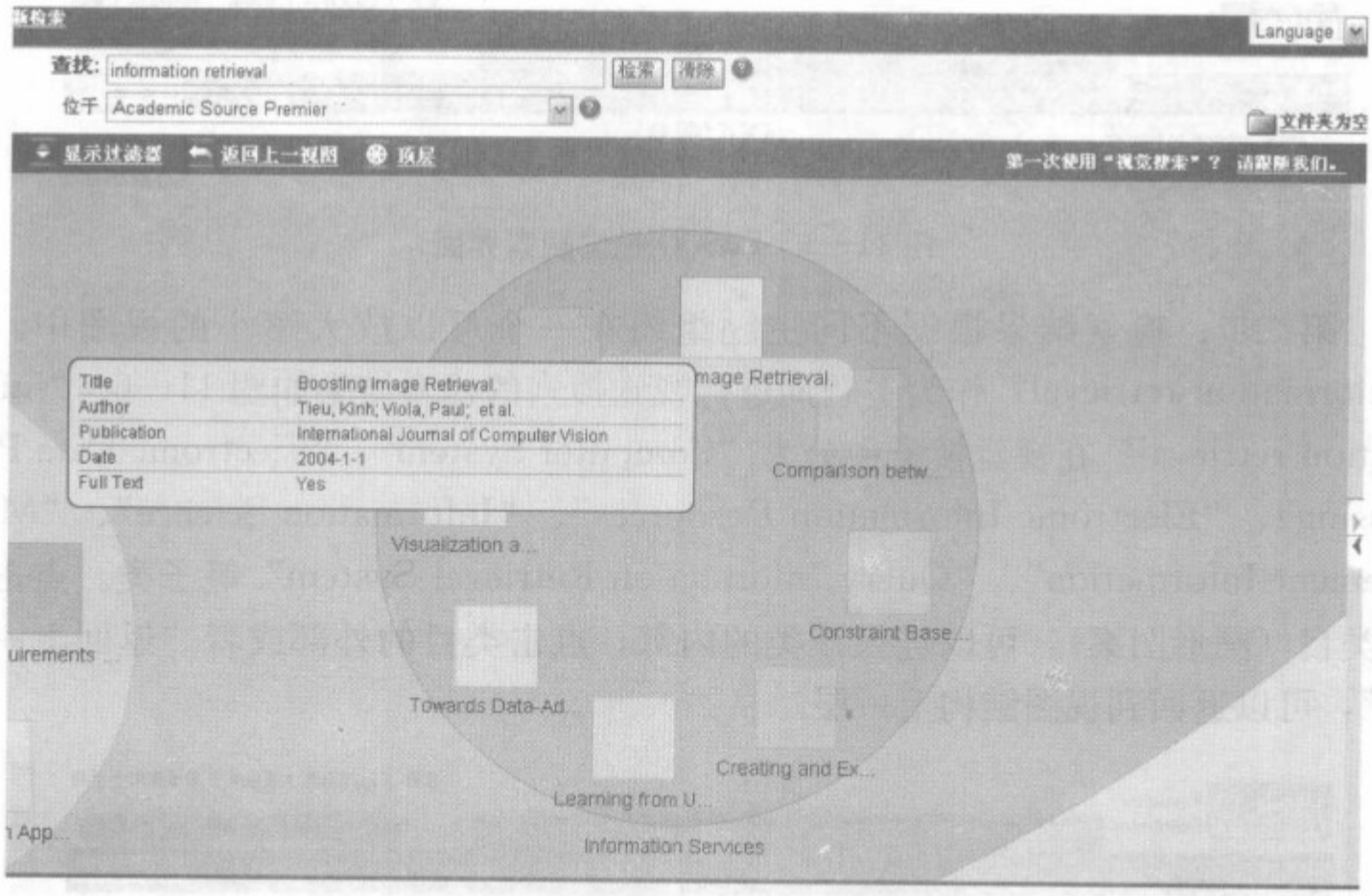

图 11—7　第 2 层检索结果显示界面

第四步：点击矩形“Boosting Image Retrieval”，其著录信息表示在右侧，如图 11—8。用户这时就可以选择下载。

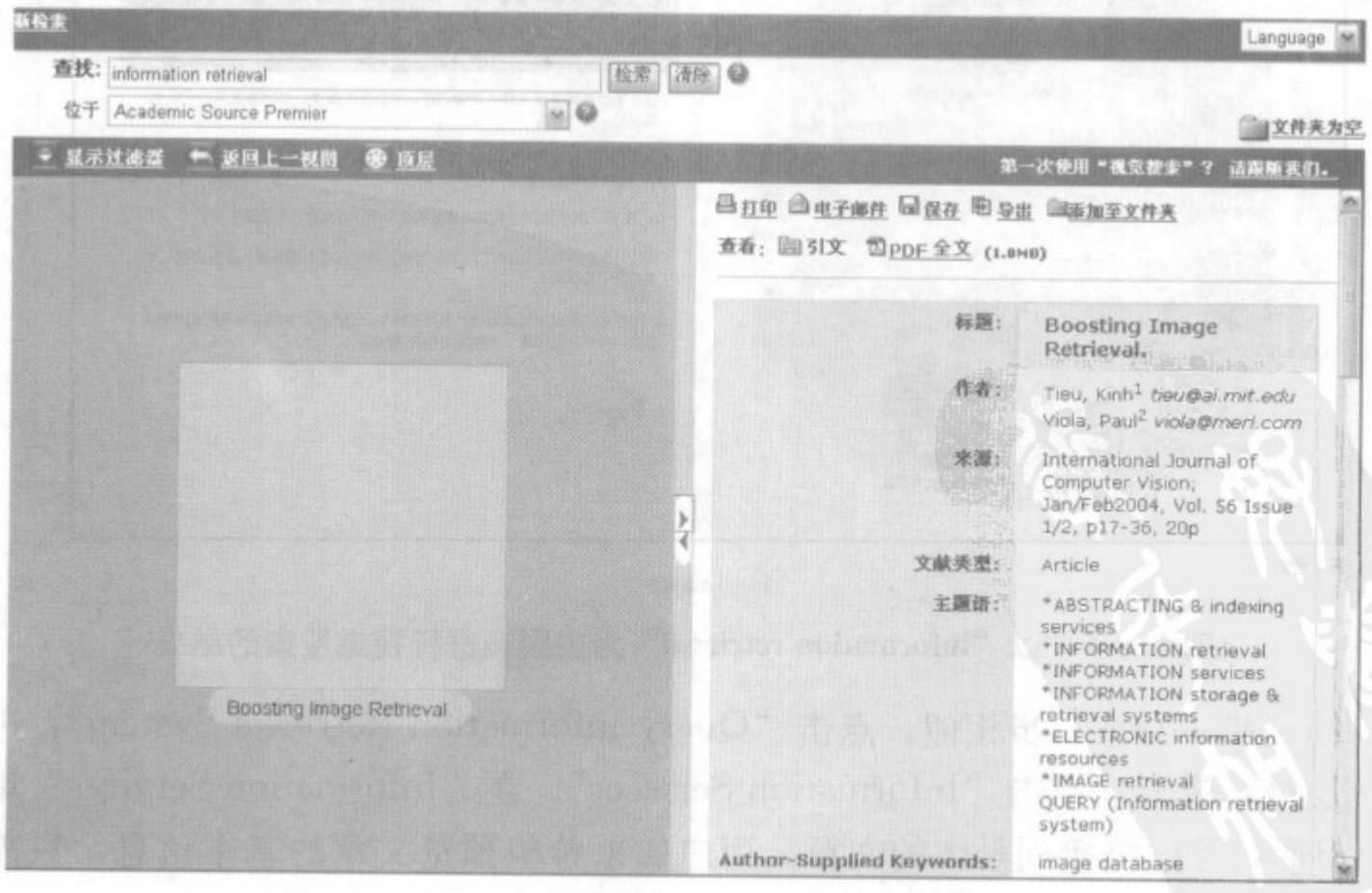

图 11—8　单篇文献详细信息显示界面

第五步：检索结果的过滤。点击“显示过滤器”，会出现检索结果的过滤器，如图 11—9。用户可以限定标题包含的字段、文章发表时间、是否只显示全文收录的文章等等。

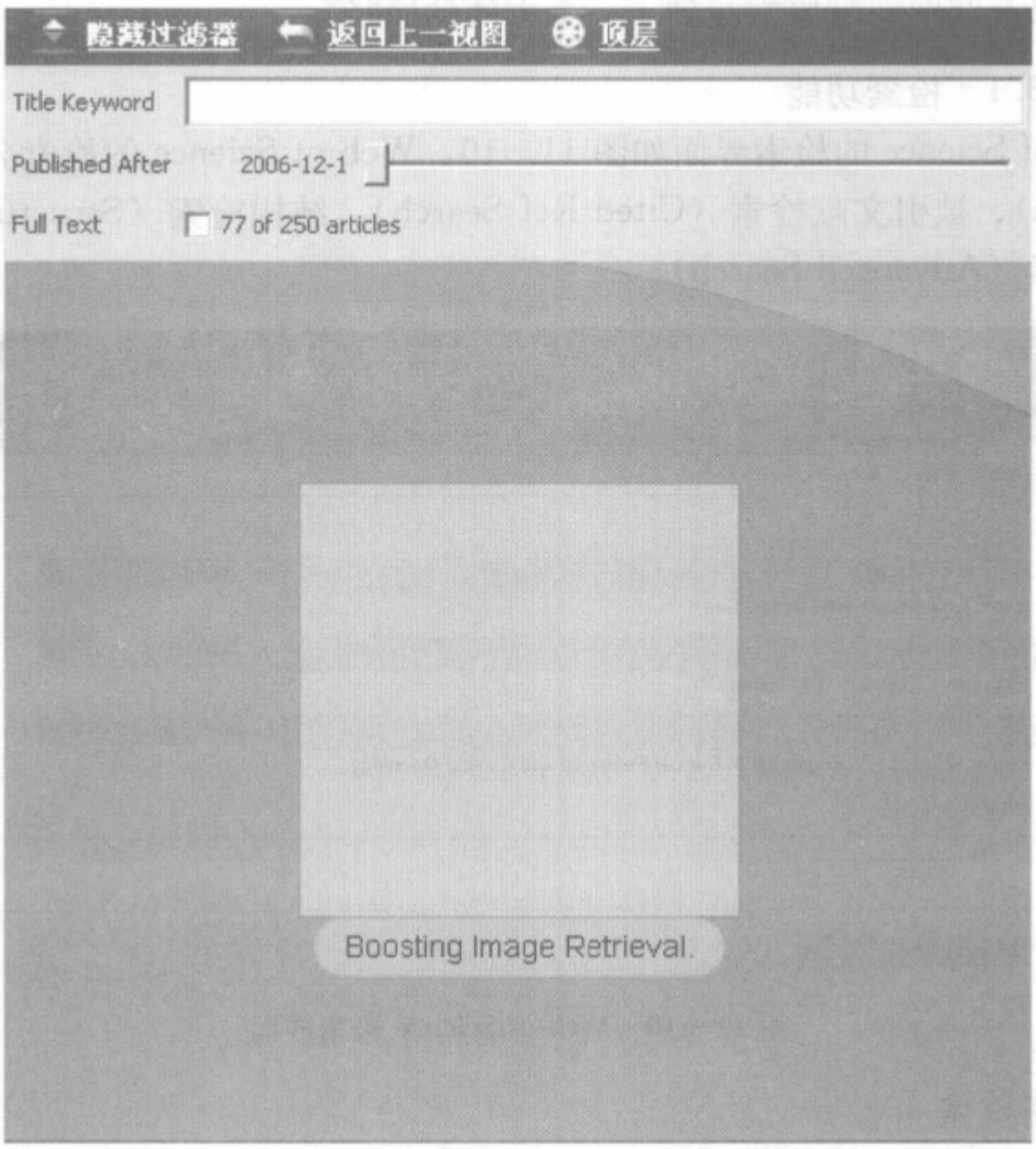

图 11—9　检索结果过滤

11.2.3　Web of Science（三大引文数据库）

Web of Science 是美国 Thomson Scientific 公司基于 WEB 开发的产品，包括三大引文库（SCI、SSCI 和 A&HCI）和两个化学数据库（CCR、IC），以 ISI Web of Knowledge 作为检索平台。三大引文数据库包括：《科学引文索引》（Science Citation Index Expanded，SCI），收录 6 300 多种科学技术期刊；《社会科学引文索引》（Social Sciences Citation Index，SSCI），收录 1 800 多种社会科学期刊；《艺术和人文科学引文索引》（Arts & Humanities Citation Index，A&HCI），收录 1 100 多种艺术与人文类期刊。

Web of Science 可以极大地方便用户查找文献资料。通过引文检索功能，我们不但可以查找相关研究课题各个时期的学术文献，获取论文摘要，而且还可以得到所引用参考文献的记录、被引用的情况及相关文献的记录，等等，这就为文献研究，以及获取文献原文找到了一个方便的途径。

11.2.3.1 检索功能

Web of Science 的检索界面如图 11—10。Web of Science 的检索分为基本检索（Search）、被引文献检索（Cited Ref Search）、结构检索（Structure Search）和高级检索（Advanced Search）。

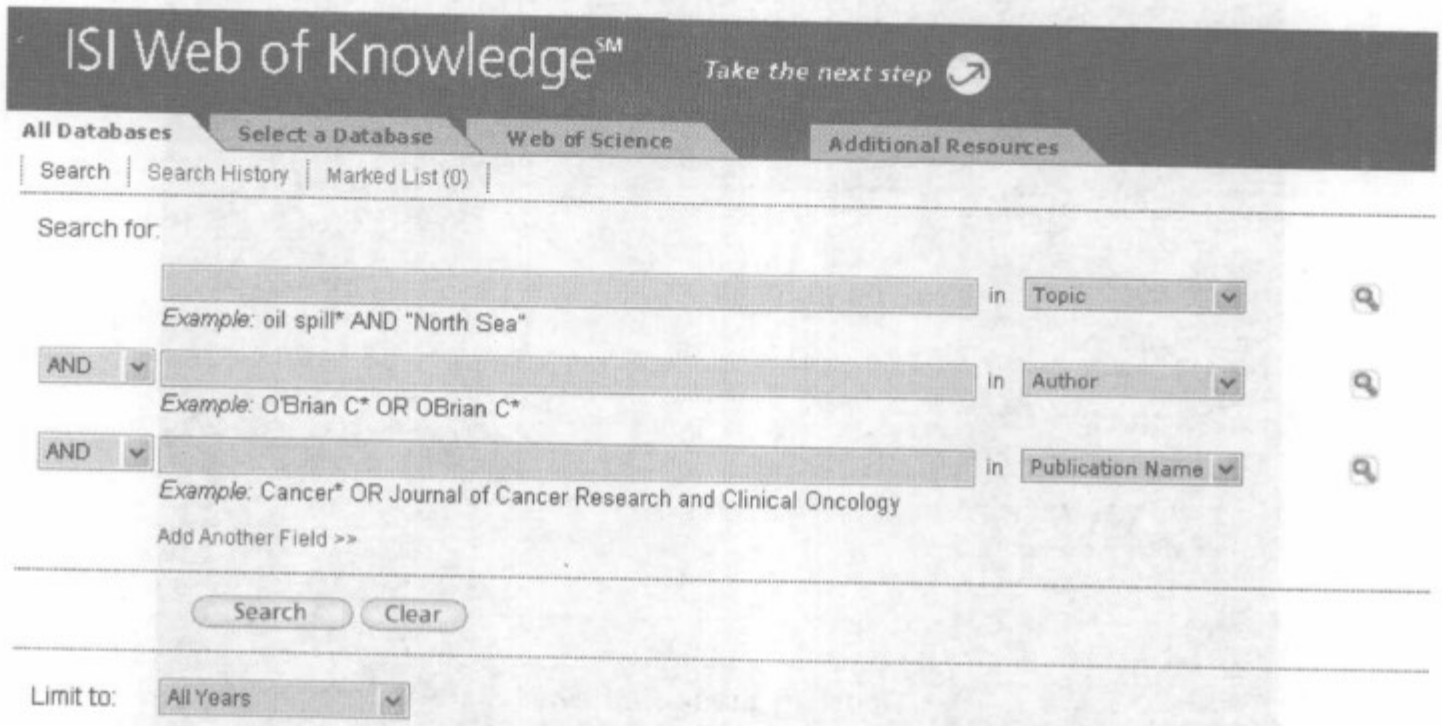

图 11—10 Web of Science 检索界面

1. 基本检索

基本检索如图 11—11。主要按文献的主题（Topic）、篇名（Title）、作者（Author）、期刊名（Publication Name）、作者的地址（Address）、出版年份（Year Published）等进行检索。

基本检索的检索步骤如下：

第一步：确定要检索的数据库和时间区间。选择数据库的范围可以是三大引文数据库和化学数据库中的任意一个，也可以是它们之间的任意组合；时间范围则最早从 1986 年开始，默认的是选择所有年份。

第二步：选择需要的检索条件，并在检索框中输入检索词，点击“Search”。例如，用户要检索作者名，在下拉选项框中选择“Author”，并输入作者名即可。如果用户还有不明白之处，只需要点击检索框右侧的问号就可以得到系统提供的帮助。

第三步：点击选中的文献查看详细信息。

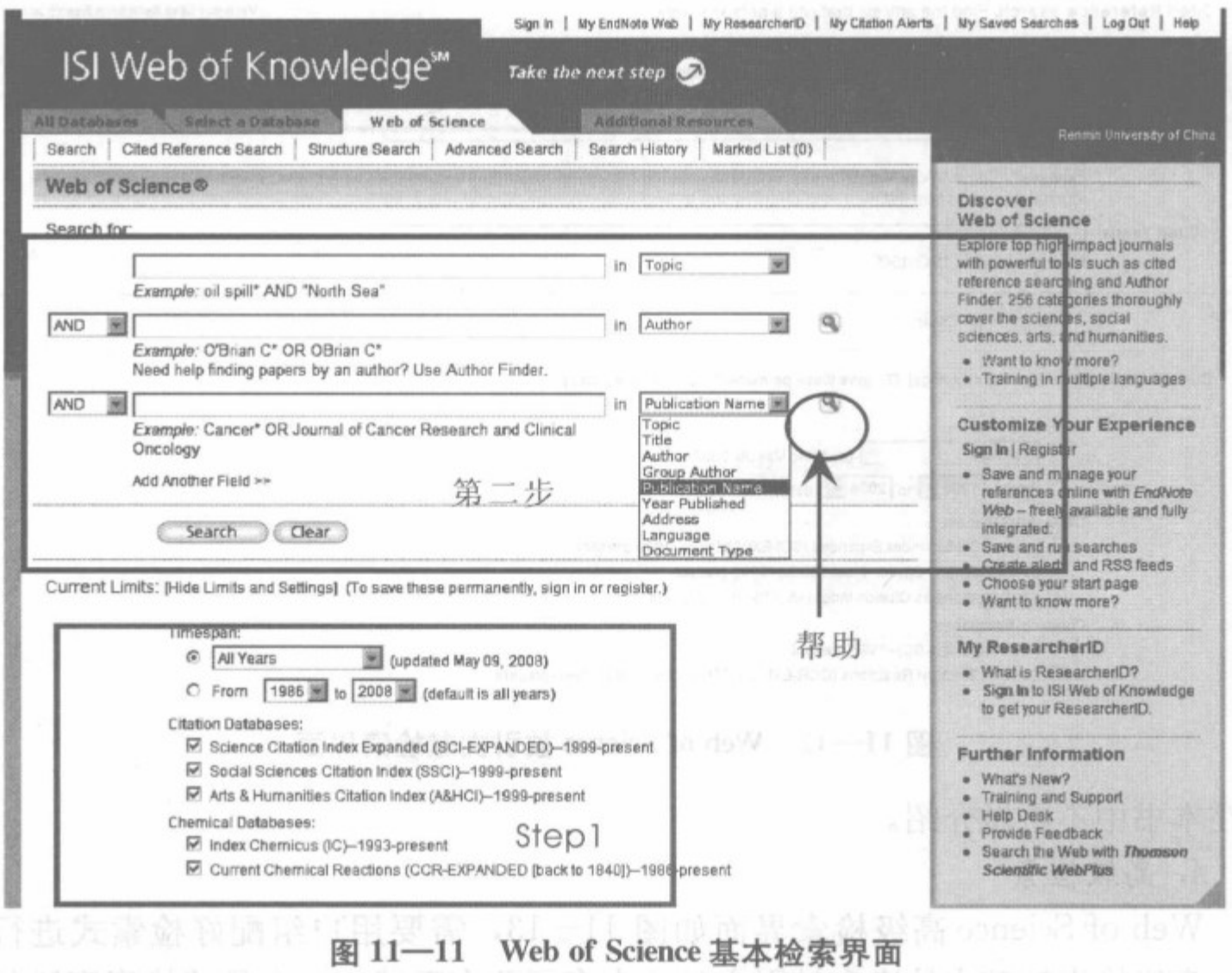

图 11—11　Web of Science 基本检索界面

2. 被引文献检索

Web of Science 被引文献检索检索界面如图 11—12，主要按被引用文献的特征检索，包括被引用的作者（Cited Author）、被引用的期刊名或者书名（Cited Work）、文献发表的年份（Cited Year）。

被引文献检索的基本步骤如下：

第一步：和基本检索类似，被引文献检索的第一步需要确定要检索的数据库和时间区间。

第二步：输入作者名、作品名和（或）发表年份。下图的例子在作者名中输入的是“Stiglitz J”。在这里需要强调的是，作者名要按“姓在前，名在后”的格式输入。

第三步：点击“Search”图标进行检索。

3. 结构检索

结构检索主要针对两个化学数据库，检索化合物、化学反应、化学结构等。这部分数据库内容需要安装 Web of Science 的插件，并且几乎没有文献内容，因

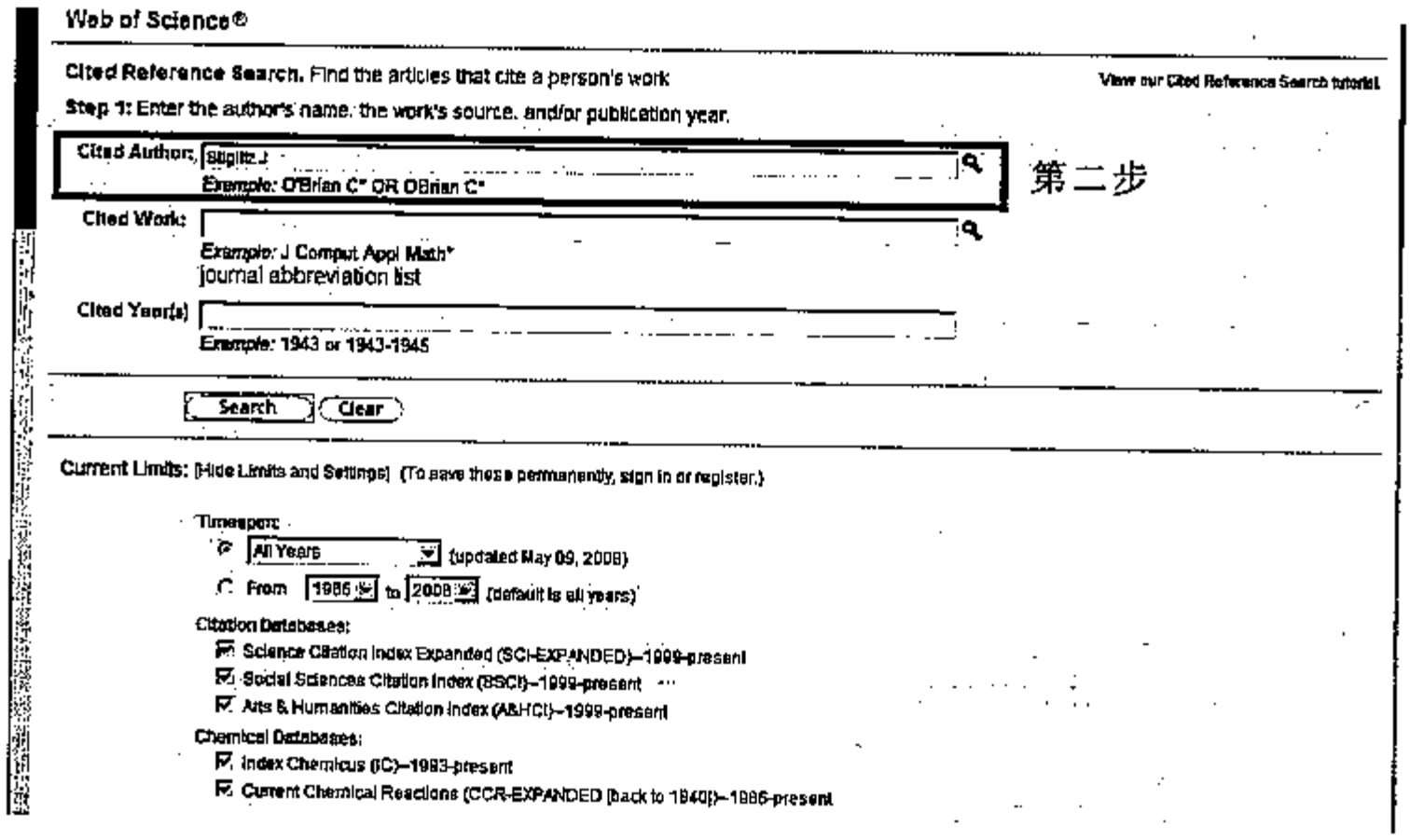

图 11—12　Web of Science 被引文献检索界面

此在本书中不再做介绍。

4. 高级检索

Web of Science 高级检索界面如图 11—13，需要用户组配好检索式进行提问。高级检索较基本检索和被引文献检索多了两个区域。一个是“检索字段代码和布尔逻辑运算符”区（图 11—13 椭圆部分），这部分区域是帮助用户组配检索式的；另一个是“语言和文件格式选项”区，这个区域可以让用户限定检索文献的语言和文件格式。

例如，用户想要检索 2005 年撰写的，题名含有“信息检索”，作者居住在斯坦福大学或者耶鲁大学的文献，在检索框内输入“TI=（information * AND retriev * ）AND PY＝2005 AND AD=（Yale Univ OR Stanford Univ)”，点击“Search”图标就可以了。

11.2.3.2　研究分析功能

Web of Science 不仅是世界著名的检索工具，也是一个评价学术水平的工具。其分析工具使用非常简便，可以帮助研究人员方便地对文献信息进行统计。其使用方法如下：

第一步：用户可以在检索结束之后，点击结果上方的“Analyze Results”，就可以进入分析页面。图 11—14 是以“Information Retrieval”为主题进行检索的检索结果页。

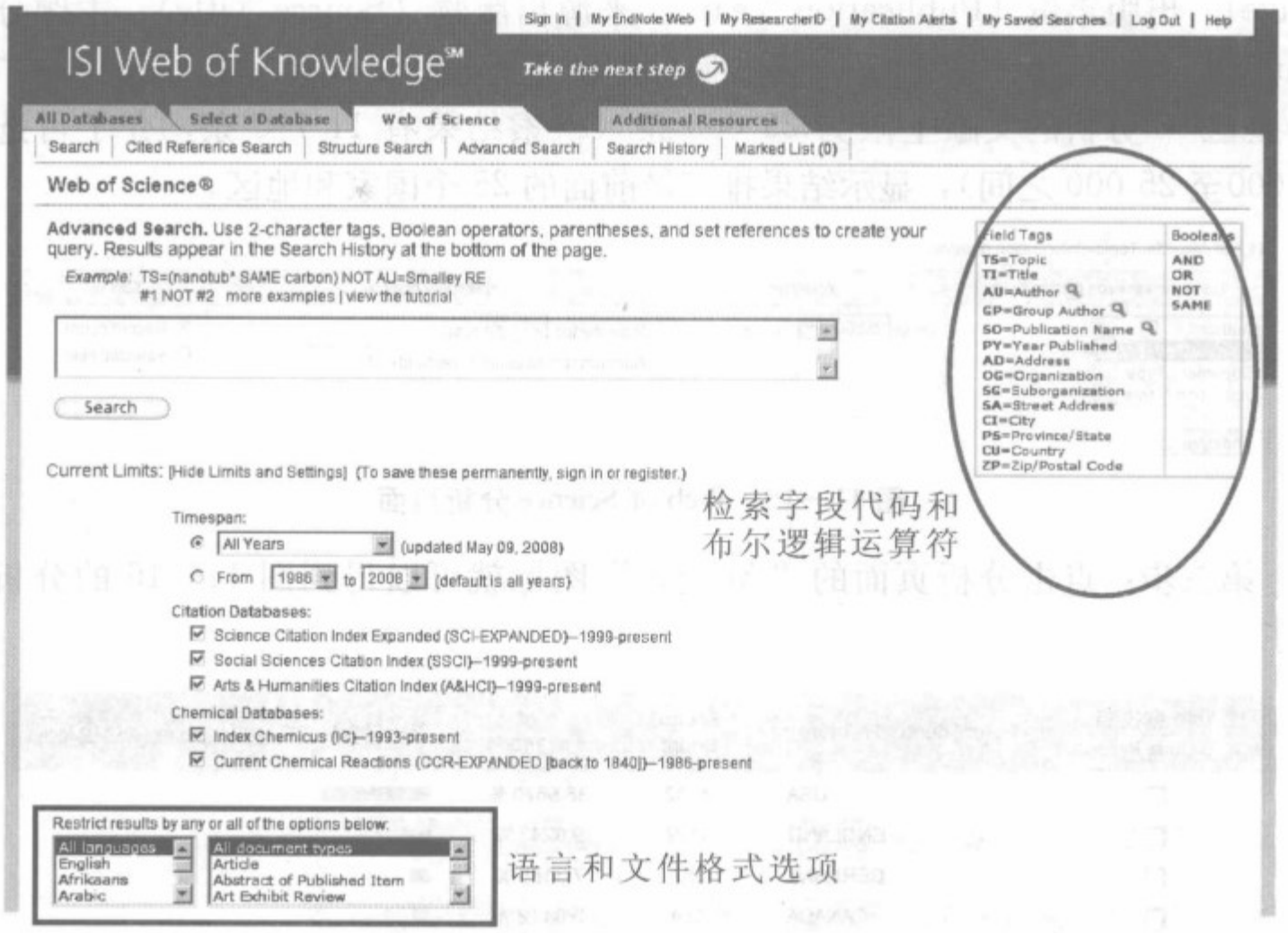

图 11—13　Web of Science 高级检索界面

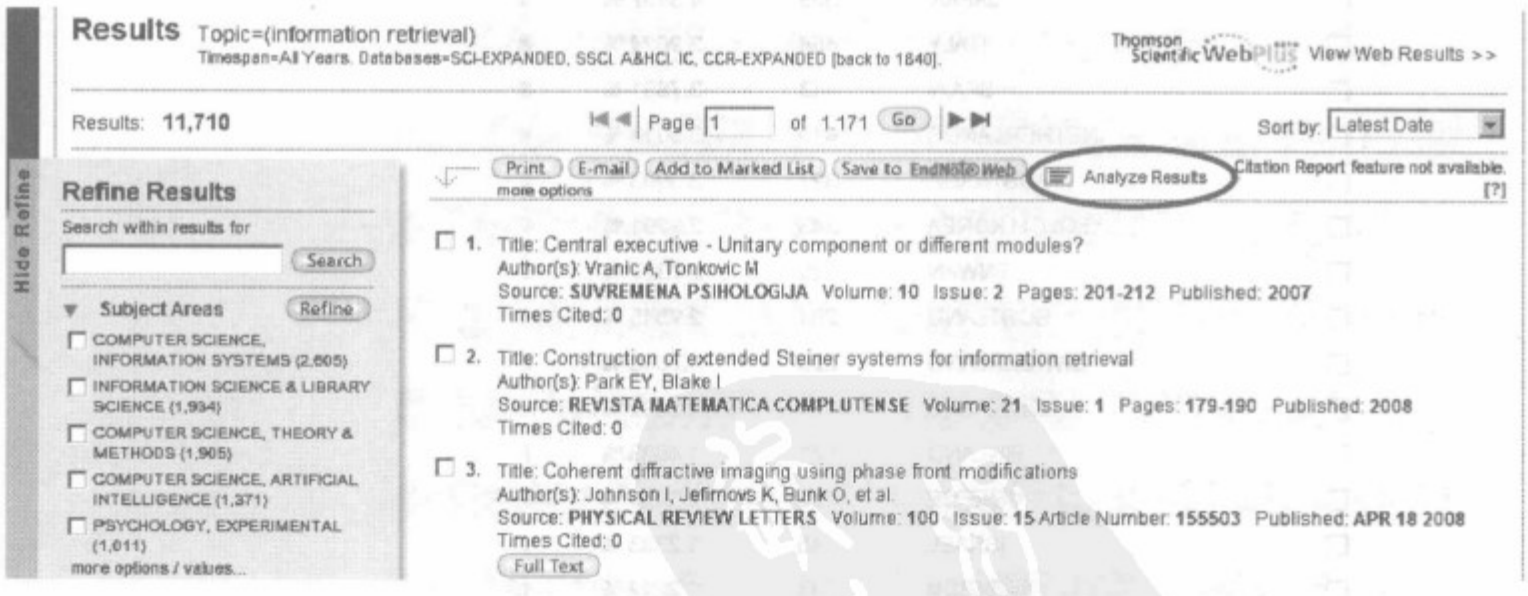

图 11—14　Web of Science 检索结果页面

第二步：在 Web of Science 的分析页面下，用户可以选择排序字段、分析的文献数的上限、分析图显示的数量、分类方法四个参数进行分析的设置。

其中，可供排序的字段有作者（Author）、国家或地区（Country/Territory）、文档类型（Document Type）、机构名称（Institution Name）、语言（Lan-

guage)、出版年份（Publication Year)、来源出版物（Source Title)、主题分类（Subject Area）等。图 11—15 是按照文献数进行排序，选择的排序字段是“国家/地区”，分析的文献上限为 25 000 条（检索结果有 11 710 条，介于可选项 10 000至 25 000 之间），显示结果排序最前面的 25 个国家和地区。

11,710 records. Topic=(information retrieval)

Rank the records by this field:	Analyze:	Set display options:	Sort by:
Author Country/Territory Document Type Institution Name	Up to 25000 records.	Show the top 25 results. Minimum record count (threshold): 2	Record count Selected field

Analyze

图 11—15　Web of Science 分析页面

第三步：点击分析页面的“Analyze”图标就可以得到图 11—16 的分析结果了。

View Records / Exclude Records	Field: Country/Territory	Record Count	% of 11710	Bar Chart	Save Analysis Data to File
☐	USA	4282	36.5670 %		
☐	ENGLAND	1127	9.6243 %		
☐	GERMANY	857	7.3185 %		
☐	CANADA	684	5.8412 %		
☐	FRANCE	623	5.3202 %		
☐	PEOPLES R CHINA	594	5.0726 %		
☐	JAPAN	509	4.3467 %		
☐	ITALY	464	3.9624 %		
☐	SPAIN	443	3.7831 %		
☐	NETHERLANDS	410	3.5013 %		
☐	AUSTRALIA	397	3.3903 %		
☐	SOUTH KOREA	343	2.9291 %		
☐	TAIWAN	320	2.7327 %		
☐	SCOTLAND	264	2.2545 %		
☐	SWITZERLAND	209	1.7848 %		
☐	SINGAPORE	193	1.6482 %		
☐	FINLAND	171	1.4603 %		
☐	GREECE	152	1.2980 %		
☐	ISRAEL	145	1.2383 %		
☐	BELGIUM	143	1.2212 %		
☐	SWEDEN	131	1.1187 %		
☐	INDIA	127	1.0845 %		
☐	IRELAND	103	0.8796 %		
☐	WALES	102	0.8711 %		
☐	DENMARK	96	0.8198 %		
View Records / Exclude Records	Field: Country/Territory	Record Count	% of 11710	Bar Chart	Save Analysis Data to File

图 11—16　Web of Science 分析结果页面

通过这样的分析，用户可以了解到“信息检索”（Information Retrieval）在各个国家的研究情况了，非常便于研究人员进行选题研究和文献调查。

11.2.4　其他国外网络数据库介绍

1. ABI/INFORM

ABI/INFORM 数据库是 UMI 公司出版、在欧美大学普遍应用的著名商业经济类数据库。该数据库涵盖的学科范围有财会、银行、商业、计算机、经济、能源、工程、环境、金融、国际贸易、保险、法律、管理、市场、税收、电信等，涉及这些行业的市场、企业文化、企业案例分析、公司新闻和分析、国际贸易与投资、经济状况和预测等方面。ABI 数据库共收录期刊 3 800 多种，其中收录全文刊 2 800 多种，被 SSCI 和 SCI 收录的期刊有 400 多种。

2. Academic Search Premier

Academic Search Premier，简称 ASP，是全球最大的学术参考全文资料库之一。由 EBSCO 公司提供。收录有关工商经济、资讯科技、人文科学、社会科学、通信传播、教育、艺术、文学、医药、通用科学等领域的期刊 4 286 种，其中 3 288 种为全文刊（最早回溯至 1975 年）。被 SCI 收录的核心期刊为 993 种（含全文刊 350 种）。该库收录图书馆学和信息科学方面的期刊共 85 种（其中全文刊 54 种）。

3. Business Source Premier

Business Source Premier，简称 BSP，由 EBSCO 公司提供。收录 2 800 种全球商业相关活动刊物的索引及摘要，含 2 300 种全文期刊（最早回溯至 1965 年），被 SCI 收录的核心全文期刊 238 种。涉及主题范围有国际商务、经济学、经济管理、金融、会计、劳动人事、银行等。

4. Academic Research Library

学术研究图书馆（Academic Research Library，简称 ARL）是一个综合参考及人文社会科学期刊论文的数据库，涉及商业与经济、教育、历史、传播学、法律、军事、文化、科学、医学、艺术、心理学、宗教与神学、社会学等学科，收录 2 300 多种期刊和报纸，其中，全文刊占三分之二。可检索 1971 年来的文摘和 1986 年来的全文。

5. SpringerLink

德国施普林格（Springer-Verlag）是世界上著名的科技出版集团，由它开发的 SpringerLink 系统可以提供其学术期刊及电子图书的在线服务。SpringerLink 中的期刊及图书等所有资源划分为 12 个学科：建筑学、设计和艺术，行为科学，

生物医学和生命科学，商业和经济，化学和材料科学，计算机科学，地球和环境科学，工程学，人文、社科和法律，数学和统计学，医学，物理和天文学，一共收录了千余种期刊和700多种丛书。

11.3 中文网络数据库

11.3.1 中文网络数据库发展概况

随着当代通信技术、网络技术的飞速发展，国际数据库产业得到突飞猛进的发展，数据库规模不断扩充，采用了商业化的经营模式，许多企业在数据库产品的开发和服务中发挥了重要的作用。如前所述，一些先进国家涌现出了大量的优秀网络数据库。在这样一个国际大环境下，伴随着中文网络资源建设，我国的数据库市场飞速发展，而且表现出更为巨大的增长潜力。此间，越来越多的数据库开始提供基于互联网的数据库服务。

综观中文网络数据库的发展，我们可以看出在国内网络信息服务市场上，形成了3个大型的期刊网络数据库集成化中心，即CNKI中文系列数据库、万方数据资源系统和维普信息资源系统。

（1）CNKI工程即中国知识基础设施工程（China National Knowledge Infrastructure），是采用现代信息技术，建设适合于我国的可以进行知识整合、生产、网络化传播扩散和互动式交流合作的一种社会化知识基础设施的国家级大规模信息化工程，由光盘国家工程研究中心、清华同方光盘股份有限公司、中国学术期刊（光盘版）电子杂志社和清华同方教育技术研究院联合承担。CNKI推出的中文系列数据库有：《中国期刊全文数据库》、《中国学术期刊题录数据库》、《中国重要报纸全文数据库》、《中国图书全文数据库》、《中国专利数据库》、《技术创新数据库》、《中国学位论文全文数据库》、《中国重要会议论文数据库》等。

（2）万方数据资源系统（ChinaInfo），是北京万方数据股份有限公司在中国科技信息研究所数十年积累的全部信息服务资源的基础上建立起来的，形成以科技信息为主，集经济、金融、社会、人文信息为一体，实现网络化服务的信息资源系统。自1997年8月对外开放，以其丰富的信息资源在国内外产生了较大的影响。其中，万方数据——数字化期刊群属国家“九五”重点科技攻关项目——科技期刊网络服务系统。整个系统以刊为单位上网，保留了刊物本身的浏览风格和习惯，方便读者随时阅读和引用，形成了网上期刊的门户特征。

（3）维普信息资源系统，是由重庆维普资讯有限公司研制开发的网络信息资源。维普资讯有限公司是科学技术部西南信息中心下属的一家大型的专业化数据公司，自 1989 年以来，一直致力于报刊等信息资源的深层次开发和推广应用，集数据采集、数据加工、光盘制作发行和网上信息服务于一体。收录有中文报纸约 400 种，中文期刊 8 000 多种，外文期刊约 5 000 种，拥有固定客户 2 000 余家。目前，已成为中国最有影响力的数据库建设者之一。

在中文网络数据库中，全文数据库越来越占据主导地位。《中国期刊全文数据库》、维普的《中文科技期刊数据库》（全文版）和万方资源系统的数字化期刊群是全文网络数据库中的优秀代表。我国信息基础设施的建设和完善，为我国信息资源的共建共享提供了良好的网络环境。网络数据库作为资源数字化的重要形式，作为网络资源共享的重要载体，将会得到更大的发展。

11.3.2　《中国期刊全文数据库》

《中国期刊全文数据库》是我国第一个连续的大规模的集成化、多功能学术期刊全文检索系统，是中国知识基础设施工程 CNKI 中最重要的数据库之一。1999 年 6 月，在原光盘数据库的基础上，正式开通了它的网络版。

《中国期刊全文数据库》收集面广、内容丰富、信息量大。收录有国内 8 200 种期刊全文，其中核心期刊 80%左右，年新增文献达一百多万篇，这些期刊覆盖了自然科学、工程技术、农业、哲学、医学、人文社会科学等各个领域，全文文献总量 2 300 多万篇。全部期刊分为 10 个专辑：理工 A、理工 B、理工 C、农业、医药卫生、文史哲、政治军事与法律、教育与社会科学、电子技术与信息科学、经济与管理。用户可以在线浏览、章节下载、整本下载、分页下载。数据库每日更新。《中国期刊全文数据库》主要提供初级检索、高级检索和专业检索，以及期刊导航等。

11.3.2.1　初级检索

1. 登录《中国期刊全文数据库》

可以直接登录中国期刊网的《中国期刊全文数据库》（http：//www. cnki. net/），输入账号和密码，然后登录；也可以直接从用户所在的校园网上图书馆链接的中国期刊网入口登录该数据库，系统默认的检索方式为初级检索方式。界面如图 11—17。

可以看到，界面的右上方为检索区，主要包括以下内容（表 11—1）。

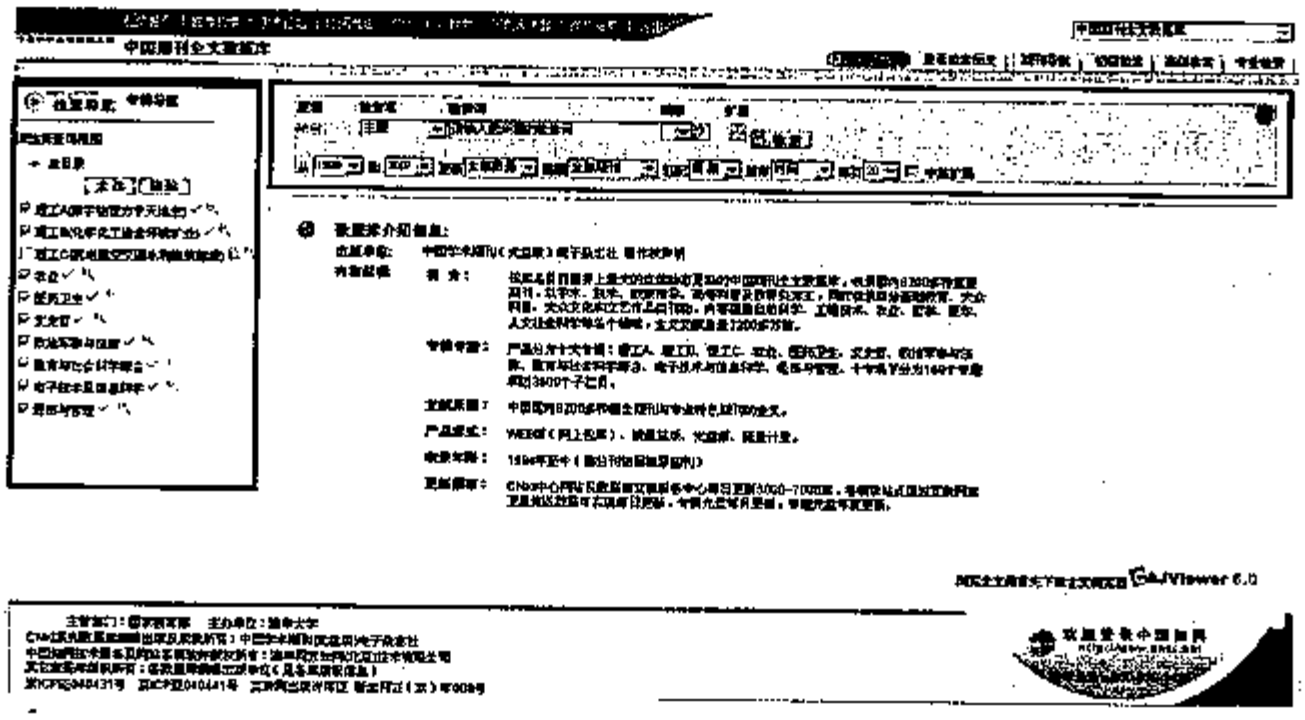

图 11—17 《中国期刊全文数据库》初级检索界面

表 11—1 《中国期刊全文数据库》初级检索界面内容

逻辑	“+”、“—”分别表示增加和减少检索词，最多可扩展到 5 项。
检索项	主题、篇名、关键词、摘要、作者、第一作者、单位、刊名、参考文献、全文、年、期、基金、中图分类号、ISSN、统一刊号。
检索词	检索词为文章检索项中出现的重要词。
词频	指检索词在相应检索项中出现的频次。词频为空，表示至少出现 1 次；如果为数字，例如 3，则表示至少出现 3 次，以此类推。
扩展	显示以输入词为中心词的相似词。
检索时间范围	文献发表的时间范围。
更新	全部数据、最近一周、最近一月、最近三月、最近半年。
范围	全部期刊、EI 来源期刊、SCI 来源期刊、核心期刊。
匹配	精确或者模糊检索。
排序	检索结果的排列顺序。有无序、相关度、时间三个选项。
	无序：检索结果无序排列。
	相关度：以检索词在检索字段内容里出现的命中次数排序，次数越多越靠前。
	时间：以数据更新日期排列，数据更新的日期越新越靠前。
每页显示	选择检索结果页面所要显示的记录条数，提供 5 种值：10、20、30、40、50。
中英扩展	可以用英文查对应的中文内容，用中文查对应的英文内容。
?	查看帮助。
检索	点击“检索”键进行数据检索。

检索界面的左侧为检索导航区。《中国期刊全文数据库》以“专题数据库”

的形式设计分类导航体系，全部期刊按学科共分为 10 个专辑，将各学科、各门类的知识分为 168 个专题，每个专题下再分三级类目，第三级类目下可以按主题看到具体的文章。在分类检索中，可以通过导航逐步缩小范围，最后检索出某一知识单元中的文章。例如：利用学科专业导航，理工 A→数学→数学概论→数学史和数学范畴，双击后自动进行检索，或选择该主题后，点击检索进行查询，可以直接检出其中的文章。这种浏览方式可使用户查询某一学科的所有文献，层次清晰，方便快捷。在初级检索和高级检索中，利用导航选取检索范围，可以节省检索的时间，提高查准率。

2. 选取检索途径

在检索项的下拉框里选取要进行检索的字段。如篇名、作者、关键词、机构、中文摘要、引文、基金、全文、中文刊名、ISSN、年、期、主题词、篇名/关键词/摘要、第一作者等。

3. 输入检索词

根据用户对检索课题主题概念分析的结果，选择检索词，并进行输入。

4. 进行检索

点击“检索”按钮，在页面的检索区下部就会列出检索结果，点击其中一项的题名，可以在新的页面列出详细的信息，包括篇名、作者、刊名、机构、关键词等。

5. 检索结果的优化与处理

对于一些简单查询，建议使用初级检索的功能，该查询的特点是方便快捷、效率高，但查询结果有很大的冗余。如果在检索结果中进行二次检索或配合高级检索则可以大大提高查准率。

在结果中检索（二次检索）：一次检索后可能会有很多用户所不期望的记录，用户可在第一次检索的基础之上进行二次检索。二次检索只是在上次检索结果的范围内进行检索，这样可逐步缩小检索范围，使检索结果越来越靠近自己想要的结果。

6. 相似词显示

点击“检索”按钮，在检索导航区的下方会出现以检索词为中心词的相似词。例如，以“检索”为检索词，可以有以下的显示，方便用户找到精确的检索词（图 11—18）。

7. 检索相关信息的链接

《中国期刊全文数据库》提供相关信息的链接，包括参考文献、共引文献、二级参考文献、相似文献、相关研究机构、相关文献作者、文献分类导航等。通过这些相关信息的链接，可以回溯检索。

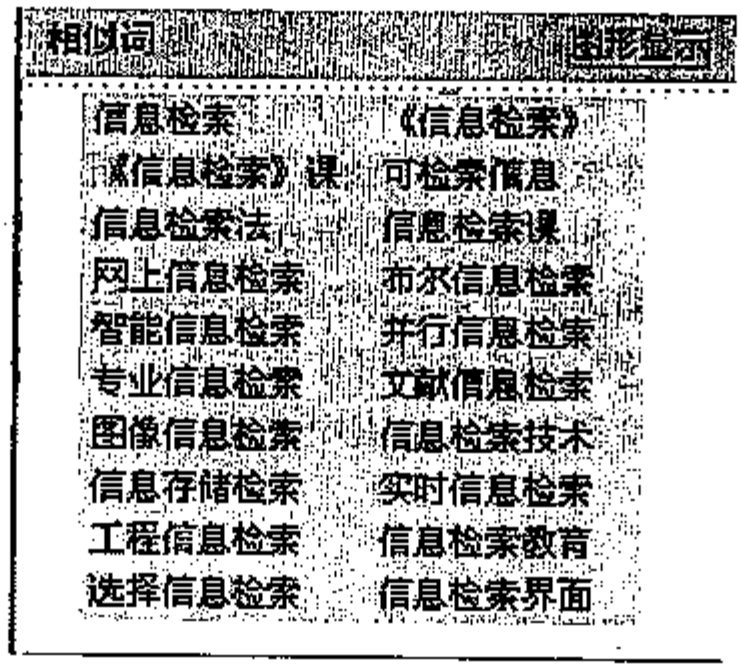

图 11—18 “信息检索”相似词

11.3.2.2 高级检索

利用高级检索系统能进行快速有效的组合查询，优点是查询结果冗余少、命中率高。对于命中率要求较高的查询，建议使用该检索系统。

该系统可组合检索项：最多有 10 个检索项，可以依次输入检索条件，然后选择与（and）、或（or）、非（not）操作，这样就可以进行快速准确的组合查询。

检索结果的处理：检索的结果可以在线浏览，也可以下载。只要点击题名后面的“CAJ 原文下载”即可，然后系统提示“是否在当前位置打开或者保存到磁盘”，文件格式为该系统特定的文件格式“.caj”或“.kdh”格式，必须在中国期刊网首页上下载使用系统特定的全文浏览器。也可以使用“PDF 原文下载”。

11.3.2.3 专业检索

专业检索比高级检索功能更强大，允许用户按自己需要组合逻辑表达式，进行更精确的检索，但需要检索人员根据系统的检索语法编制检索式进行检索。适用于熟练掌握检索技术的专业检索人员。例如，想检索篇名包括“企业”并且关键词为“结构调整”的信息，可以在专业检索界面的检索框中直接输入“TI = 企业 and KY = 结构调整”。

11.3.2.4 检索举例

1. 分类检索举例：查询有关情报事业的期刊文章

先在检索导航里选择“电子技术及信息科学”专辑，然后层层进入，选择下位类“情报学、情报工作”，再点击下一级类目“情报事业”，就可以直接检索出相关文章。

2. 初级检索举例：查询2000年至2007年有关信息服务的期刊文章

第一步：选择相应专辑及年份限制。

分析检索要求，首先估计所检内容应该出现在哪个专辑中，这样就可缩小检索范围，根据各专辑内容分析，“信息服务”应该属于电子技术与信息科学辑，所以我们将检索范围限定在电子技术与信息科学辑中。由于要检索的是2000年至2007年的文章，所以在年份的下拉框里选择从2000年到2007年。

第二步：选择字段并输入检索词。

可以首先选择“篇名”字段。为了提高查全率，也可以尝试一下其他字段的检索情况，如“关键词”、“中文摘要”等。然后，输入检索词“信息服务”。

第三步：点击“检索”按钮，即可进行检索。

第四步：优化检索。如果认为检索结果太多，还可进行二次检索进一步筛选。

3. 高级检索举例：查询陈晓红在《安徽教育学院学报》上发表的文章

第一步：选择相应专辑及年份限制。既然是在《安徽教育学院学报》上发表的文章，应属于教育与社会科学辑。年份没要求，所以不用考虑时间段的选择。

第二步：确定字段，输入检索词及确定各检索词之间的连接关系。作者为“陈晓红”，刊名为“安徽教育学院学报”，二者是逻辑与（and）的关系。

第三步：点击“检索”按钮，即可进行检索，并得到检索结果。

4. 专业检索举例：查询钱伟长在清华大学以外的机构工作期间所发表的，题名中包含“流体”、“力学”的文章

第一步：选择页面上方的专业检索；

第二步：在检索框中输入检索式：题名＝‘流体 # 力学’and（作者＝钱伟长 not 机构＝清华大学）

第三步：点击“检索”，得到检索结果（图11—19）。

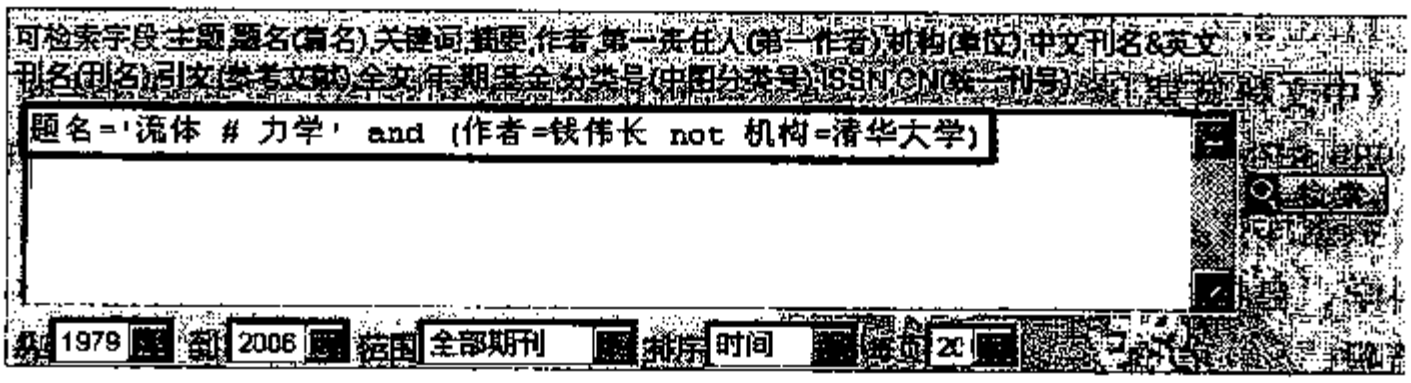

图11—19 专业检索界面

5. 在结果中检索举例：检索2007年有关“信息检索”的期刊文章

第一步：选择在检索项“主题”中检索。

第二步：输入检索词“信息检索”。

第三步：选择从“2007”至“2007”；“匹配”中选择“精确”。

第四步：点击“检索”，进行第一次检索。

检索结果显示，这样的期刊文章一共有209篇，数量较多，需要进行筛选。

二次检索（在结果中检索）：重新选择检索项“篇名”，输入检索词“信息检索”，在检索结果页面上勾选“在结果中检索”，再点击“检索”，检索结果为58条，这样就在某种程度上提高了检索的精确性。

11.3.3 国内其他网络数据库介绍

11.3.3.1 万方数据资源系统

万方数据资源系统（网址：http://www.wanfangdata.cn/）是建立在互联网上的大型综合性信息资源系统，由中国科技信息研究所开发制作。该数据库大多实行有偿服务。收录内容以科技信息为主，同时涵盖经济、文化、教育等相关信息。万方数据资源系统2001年全新改版后，被整合为科技信息子系统、商务信息子系统和数字化期刊子系统3个部分。科技信息子系统面向广大科技工作者提供全方位的科技信息，共有科技文献、名人与机构、中外标准、科技动态、政策法规、成果专利6个栏目，各栏目中包含大量相关数据库资源；商务信息子系统面向企业用户推出工商资讯、经贸信息、成果专利、商贸活动、咨询服务等栏目；数字化期刊子系统共集纳了100多个类目的6 000多种核心期刊全文内容上网。

其中，数字化期刊子系统被广泛应用于各高校，成为我国重要的期刊全文数据库之一。提供分类检索、高级检索和引文检索三种方法。万方数据资源系统主页如图11—20。

11.3.3.2 《中文科技期刊数据库》

《中文科技期刊数据库》由重庆维普资讯有限公司推出，为全文数据库，源于1989年创建的《中文科技期刊篇名数据库》，其全文和题录文摘版一一对应。

数据库按照《中国图书馆分类法》进行分类，所有文献被分为8个专辑：社会科学、自然科学、工程技术、农业科学、医药卫生、经济管理、教育科学和图书情报。总共包含了1989年至今的9 000余种期刊刊载的1 500余万篇文献，并以每年250万篇的速度递增。目前已成为我国图书情报机构、教育机构、科研院所等系统必不可少的基本工具和获取资料的重要来源。

《中文科技期刊数据库》有9种检索入口可供选择。包括：关键词、刊名、作者、第一作者、机构、题名、文摘、分类号、任意字段。其中“任意字段”检

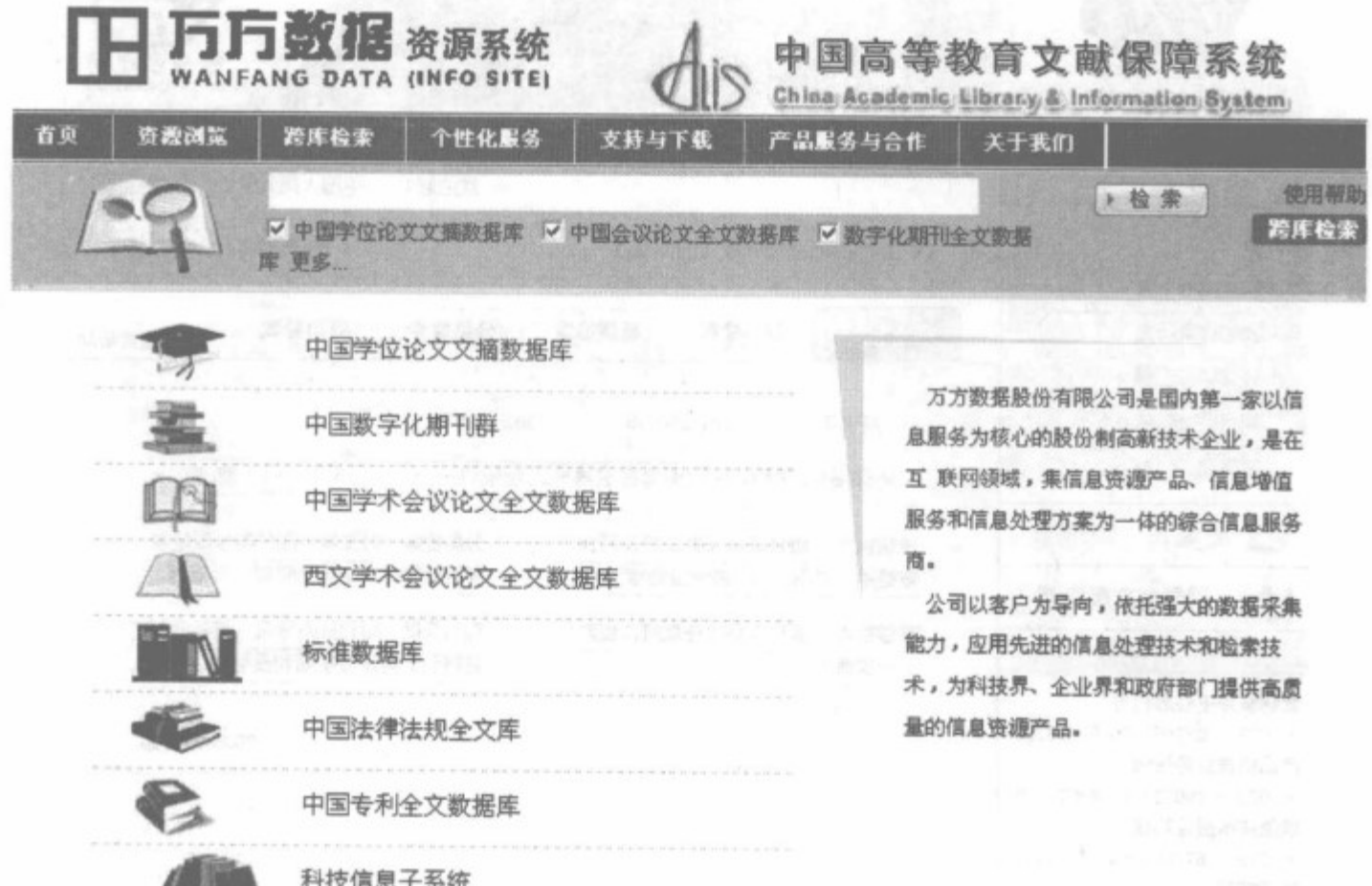

图 11—20　万方数据资源系统主页

索指在所有字段内检索。选定某一检索字段后，可在检索输入框输入检索词，点击“检索”按钮后，就可以进行相应的检索。字段名前的英文字母为检索途径代码。

《中文科技期刊数据库》提供学科分类导航和刊名导航系统。学科分类导航是树形结构的，参考《中国图书资料分类法》进行分类。用户可以首先选择学科，然后层层点击，选择需要的相关类目，直到最后显示出该类别的结果。《中文科技期刊数据库》检索界面如图 11—21。

该数据库还提供二次检索和复合检索，允许用户直接输入复合检索式，例如输入“K＝信息检索＊J＝情报学报”，检索词前面的英文字母是各字段的代码，可在检索选择框中查看。本数据库检索符号的对应关系为“＊”为逻辑与、“＋”为逻辑或、“－”为逻辑非。

11.3.3.3　《中国社会科学引文索引》

《中国社会科学引文索引》（http：//cssci. nju. edu. cn）由南京大学中国社会

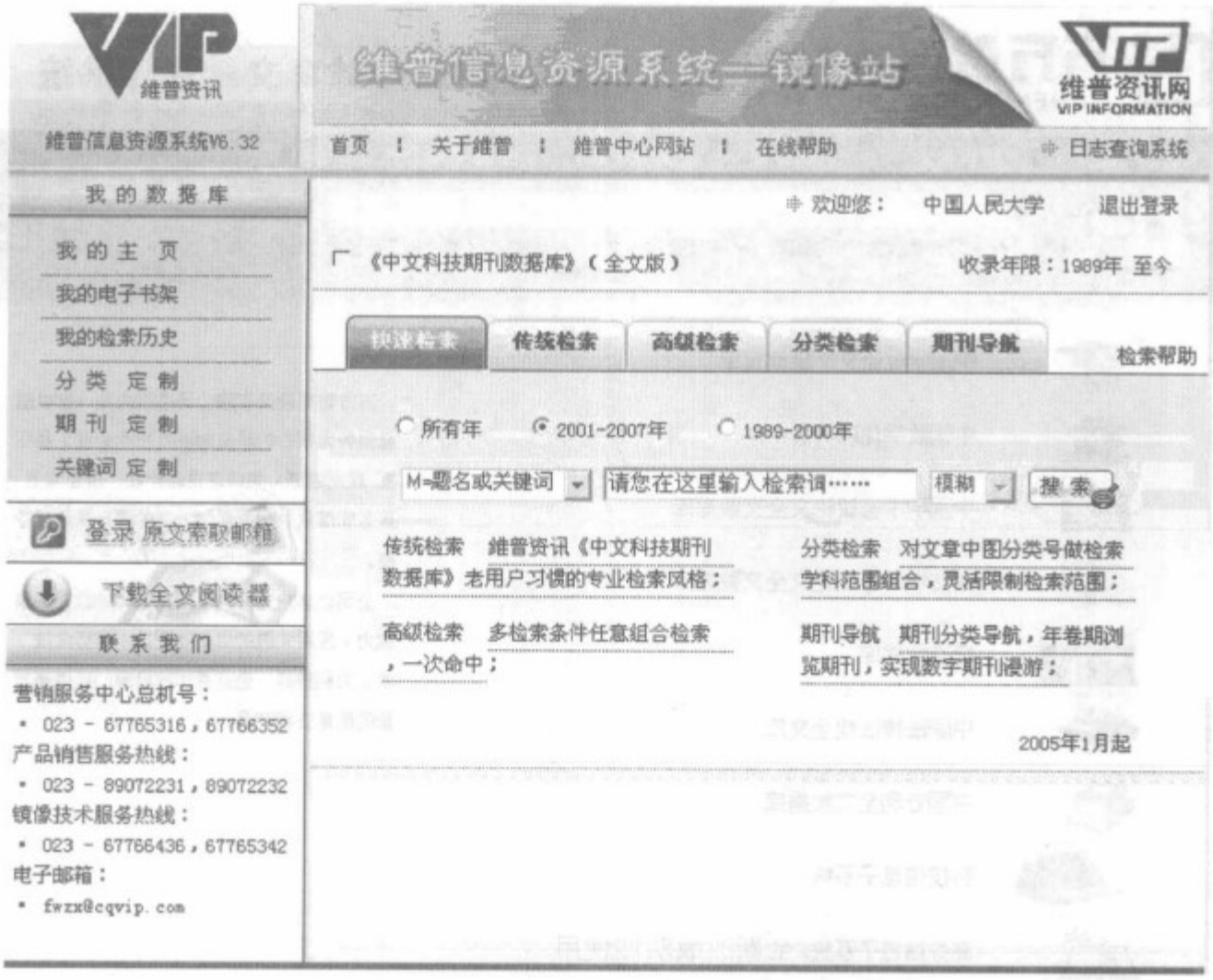

图 11—21 《中文科技期刊数据库》检索界面

科学评价研究中心研制，是教育部人文社会科学重大研究项目，是我国人文社会科学文献信息查询与评价的重要工具。该数据库选用了我国出版的中文人文科学、社会科学学术期刊 496 种，来源文献 50 万余篇，被引用文献 300 余万篇，是我国社会科学研究评价的重要工具。它主要从来源文献和被引文献两个方面向用户提供信息。前者主要用来查询本索引所选用的源刊的文章的作者（所在单位）、篇名、参考文献等，检索途径有：论文作者、篇名（词）、作者机构、作者地区、期刊名称、机构名称、标引词、学科类别、基金项目以及年代等 10 余项；后者主要用来查询作者、论文、期刊等的被引情况。其检索途径有：被引文作者、篇名或书名（词）、期刊名称、出版社、年代、被引文献类型等。

11.3.3.4 中国资讯行

中国资讯行（China InfoBank）是中国香港专门收集、处理及传播中国商业信息的高科技企业，为世界各地各行各业的公司和研究机构提供经济新闻、商业报告、统计数据、科研资料等信息。数据每日更新。http：//www. bjinfobank. com 在线提供 12 个数据库，包括《中国经济新闻库》、《中国商业报告库》、《中国法律法规库》、《中国统计数据库》、《中国上市公司文献库》、《中国香港上市公司资料

库》、《中国医疗健康库》、《中国企业产品库》、《中国中央及地方政府机构库》、《名词解释库》、《中国人物库》等。

11.3.3.5 国务院发展研究中心信息网

国务院发展研究中心信息网，简称国研网（http：//www. drcnet. com. cn/）。通过它可以查询《国务院发展研究中心调查研究报告》（简称《国研报告》）。《国研报告》是国务院发展研究中心专家不定期发布的有关中国经济和社会诸多领域的调查研究报告。每年200期，100万字左右，不定期出版，每天在线更新。通过《国研报告》，用户可获得全面的、具有政策性意义的研究资料。

【案例】

网络数据库检索实例

1. 课题题名：古希腊政治制度史研究

2. 课题分析

(1)“古希腊政治制度史”的所属科目可按如下列出：

历史、地理→世界史→古代史（公元前40世纪—公元476）→古代希腊

政治、法律→政治理论→政治学史、政治思想史→各国政治思想史

政治、法律→中国政治→政治制度与国家机构

政治、法律→世界政治→世界政治概况

(2)“古希腊政治制度史”相关同义词、近义词：

政治制度与政治体制为近义词组。

由于课题本身的准确性和针对性较强，同义词、近义词较少。

3. 选择网络数据库

网络数据库之一：《中国期刊网全文数据库》　　检索年代1990—2007

检索词：古希腊、政治、制度、历史

检索途径：篇名、主题、关键词、摘要

检索式：

(1) 篇名＝古希腊＊篇名＝政治＊篇名＝制度　　1

(2) 主题＝古希腊＊主题＝政治＊主题＝制度＊主题＝历史　　72

(3) 关键词＝古希腊＊关键词＝政治＊关键词＝制度＊关键词＝历史　3

(4) 摘要＝古希腊＊摘要＝政治＊摘要＝制度＊摘要＝历史　　31

网络数据库之二：《中文科技期刊数据库》　　检索年代1990—2007

检索途径：题名或关键词、文摘

检索词：古希腊、政治、制度、历史

检索式：

（1）（题名或关键词＝古希腊）＊（题名或关键词＝政治）＊（题名或关键词＝制度）＊（题名或关键词＝历史）＊全部期刊＊年＝1990—2007　　4

（2）（文摘＝古希腊）＊（文摘＝政治）＊（文摘＝制度）＊（文摘＝历史）＊全部期刊＊年＝1990—2007　　27

网络数据库之三：万方数据资源系统　　检索年代 1990—2007

检索途径：全部字段、跨库检索

检索词：古希腊、政治、制度

检索式：

全部字段＝古希腊＊全部字段＝政治＊全部字段＝制度

《中国学位论文文摘数据库》：80

《数字化期刊数据库》：61

网络数据库之四：Web of Science　　检索年代 1990—2007

检索途径：题名或关键词、文摘

检索词：History，Ancient Greek，Political System

检索式：

（1）topic＝（Ancient Gree＊ and Politic＊ and System）

Databases＝MEDLINE，Web of Science，ISI Proceedings；Timespan＝Latest 5 Years

命中结果：2

（2）topic＝（Gree＊ and Politic＊ and History）

Databases＝MEDLINE，Web of Science，ISI Proceedings；Timespan＝Latest 5 Years

命中结果：10

关键术语

数据库	网络数据库	检索方式
检索步骤	《中国期刊全文数据库》	ProQuest 系列数据库
EBSCO 系列数据库	Web of Science	万方数据资源系统
《中文科技期刊全文数据库》		

思考题

1. 基于数据库的信息检索大致经历了哪几个发展阶段？

2. 网络数据库具有哪些优势?

3. 网络数据库的检索方式有哪几种?

4. 简述网络数据库的检索步骤和检索方法。

5. 试利用《中国期刊网全文数据库》和 ProQuest、Web of Science 进行一些实例检索，掌握其使用方法与技巧。

CHAPTER TWELVE

第12章
特种文献检索

【本章要点】

◇介绍科技报告的检索工具及网络检索方式
◇介绍会议文献的检索工具及网络检索方式
◇介绍学位论文的检索工具及网络检索方式
◇介绍专利文献的检索工具及网络检索方式
◇介绍标准文献的检索工具及网络检索方式
◇介绍档案文献的检索工具及网络检索方式

引子

特种文献是指有特定内容和用途、出版发行渠道特殊的文献资料。它涉及的内容广泛，类型多样，是人类从事生产和科学研究的真实记录，反映了科学技术的发展水平和动态，因此具有重要的参考价值。特种文献通常包括科技报告、会议文献、学位论文、专利文献、标准文献、档案文献等。特种文献由于其特殊的发行方式，使得它有别于一般的图书和期刊的检索。检索特种文献需要首先了解特种文献的检索工具和方式。

12.1 科技报告检索

12.1.1 科技报告概述

科技报告是对科学技术研究结果的报告或研究进展的记录。它可以是研究成果的总结，也可以是科技进展情况的实际记录。许多最新的研究成果，尤其是尖端科学的最新探索往往出现在科技报告中。

目前，美、俄罗斯、英、法、德、日等国每年都发表大量的科技报告。例如：美国政府的四大报告、英国航空委员会（ARC）报告、英国原子能局（UKAEA）报告、法国原子能委员会（CEA）报告、德国航空研究所（DVR）报告，以及一些科研单位和大专院校不定期连续出版的“著作集”、“学术札记”等。

我国科研成果的统一登记和报道工作，是从1963年正式开展的。凡是有了科研成果的单位，都要按照规定及时处理，按照程序上报、登记。中华人民共和国科学技术部根据调查情况发表科技成果公报和出版研究成果报告。截至1965年7月底，《科学技术研究报告》已出至1616号。1971年11月起，这套研究成果报告继续由中国科技情报所出版，报告名称统一改为《科学技术研究成果报告》，分为内部、秘密、绝密3个保密级别，由内部控制使用。我国出版的这套研究成果报告内容十分广泛，是一种较为正规的、代表了我国科技水平的科技报告。

12.1.1.1 科技报告的特点

科技报告的主要特点有四个：(1) 内容新颖、详实专深。科技报告的内容可以是基础理论研究或者工程技术，但涉及的一般是尖端科学的最新研究成果，创新性强，具有前瞻性。并且，科技报告的内容非常详尽、具体，注重详细记载科研记录的全过程，因此，其中既反映了研究的成功经验，又有失败经验教训，一般都附有大量的数据、图表和事实资料等。(2) 出版形式多样。科技报告出版的形式有报告、札记、备忘录、论文和译文等，并且一般无固定出版周期。(3) 质量高。科技报告所报道的内容一般必须经过有关主管部门的审核与鉴定，具有较好的成熟性、可靠性。(4) 主要由政府机构资助。大多数科技报告都与政府的研究活动、国防及尖端科学技术领域有关，提出者一般都是政府的相关机构，而研究主体一般是大学、企业或者政府的附属科研机构，研究过程在资金上受到政府机构的资助。因此，科技报告的发行范围一般受到政府机构的控制，往往只在一定范围内公开或半公开发行。

12.1.1.2 科技报告的类型

科技报告可从不同的角度进行分类：

（1）按技术内容划分，科技报告可分为：技术报告（Technical Report），是指公开发行的出版物，内容比较完整，一般为科研成果的技术总结报告；技术札记（Technical Note），是指科研过程中的临时记录和小结，一般为编写技术报告的素材；技术论文（Technical Paper），是指准备在学术会议或论文期刊上发表的论文，一般用单篇论文形式出版；技术备忘录（Technical Memorandum），是指仅供专业或机构内部人员之间沟通信息所用的资料；技术通报（Technical Bulletin），是指对外公布的内容较为成熟的摘要性文献；合同户报告（Contractor Report），是指合同户与接受资助单位在科研、试制、生产过程中编写的成果资料、进展报告、年度报告、总结报告等。

（2）按报告所反映的研究进度，可划分为：初期报告（Preliminary Report）、进展报告（Progress Report）、状况报告（Status Report）、中间报告（Interim Report）、年度报告（Annual Report）、终结报告（Final Report）等。

（3）按报告的流通范围，可划分为：保密报告（Classified Report），包括绝密（Top Secret）、机密（Secret）、秘密（Confidential）等三种，属于国家机密，公众难于得到；非密限制发行（Restricted）报告，或称内部（Limited）报告，在一定范围内发行，数量有限；解密（Declassified）报告，秘密报告或限制报告经过一段时间后解除限制，成为公开的科技报告，较易获得；公开报告，也称非密（Unclassified）报告，是可直接获得的一种科技报告。

（4）按报告的性质，可划分为：正式报告（Formal Report）、非正式报告（Informal Report）、试验报告（Test Report）、交流报告（Circular Report）、专题报告（Topic Report）、经济报告（Economic Report）、评估报告（Evaluation Report）、生产报告（Production Report）等。

12.1.2 科技报告检索工具

科技报告传播研究成果的速度较快，注重报道进行中的科研工作。大多数科技报告都涉及国家部署、支持的尖端科学技术研究项目，有生产技术方面的，也有基础理论方面的，所报道的研究成果一般必须经过主管部门组织有关单位审查鉴定，可靠性和成熟性较高。所以说，科技报告是一种非常重要的信息来源。据统计，科技人员对科技报告的需要量，约占其全部文献需要量的10%到20%。特别是在那些发展迅速、竞争激烈的高科技领域，人们对科技报告的需要量更大。但由于其机构分散、种类繁多、出版目的不尽相同等，科技报告一般难于收

集，不便掌握。

12.1.2.1 国外科技报告检索工具

世界著名的科技报告是美国的四大报告：PB（Office of Publication Board）报告、AD（ASTIA Documents）报告、NASA（National Aeronautics and Space Administration）报告、DOE（U. S. Department of Energy）报告。

1. PB报告

PB报告最初是由美国商务出版局从德国、日本和意大利夺取的科技资料整理而成。凡是该局出版整理和出版的报告，均依次编号，并在之前冠以PB代号。故这类报告称为PB报告。PB报告10万号以后的主要是美国的科研机构、军事科研单位、高等院校等的科技报告，并由美国商务部下的国家技术情报服务局（NTIS）负责收集、整理、报道和发行。20世纪60年代以后PB报告的内容逐渐转向工程技术、环境污染、城市规划等方面。

2. AD报告

AD报告原为美国武装部队技术情报局（Armed Services Technical Information Agency，ASTIA）收集、出版的科技报告，起始于1951年。AD报告是美国陆海空三军科研机构的报告，其内容侧重于军事技术和工程技术，也广泛涉及许多民用技术，包括航空、军事、电子、通信、数学、化学、地球科学等22个领域。编入AD报告的文献还有期刊、图书、会议录和学位论文等。AD报告的密级包括机密、秘密、内部限制发行、非密公开发行。报告号的编号方法起初采取混排，后在AD后再加上一个字母，以区分不同的密级，如AD—A表示公开报告，AD—B表示内部限制发行报告等。

3. NASA报告

NASA报告由美国航空与航天局（National Aeronautics and Space Administration，NASA）出版，其内容侧重于航空与空间技术领域，同时也包括许多基础学科和技术学科。NASA报告的报告号均采用“NASA—报告出版类型—顺序号”的表示方法。

4. DOE报告

DOE报告由美国能源部出版，其前身是AEC（Atomic Energy Commission）报告和ERDA（Energy Research and Development Administration）报告。DOE报告的内容范围已从核能扩展到整个能源方面。起初DOE报告没有统一的编号方法，而是由各研究机构名称的缩写字母加数字号码构成。由于所属机构较多，编码不统一，因此难以识别。1981年开始，美国能源部发行的报告都采用“DE—年代—顺序号”的形式，所以DOE报告1981年以后又叫DE报告。

美国四大科技报告的检索工具主要有：美国《政府报告通告及索引》（*Government Reports Announcement & Index*，GRA&I）、美国《宇航科技报告》（*Scientific and Technical Aerospace Reports*，STAR）、美国《能源研究文摘》（*Energy Research Abstracts*，ERA）等。

GRA&I，创刊于1946年，现为半月刊，主要以文摘的形式报道美国政府机构及合同户提供的研究报告，也报道政府主管机构出版的科技译文和少量其他国家的科技文献。包括全部PB报告、所有非密级和解密的AD报告、部分NASA报告和DOE报告及其他类型的报告，还报告一些美国专利申请说明书的摘要。GRA&I采用的是NTIS的主题分类表，现共分为38个大类、363个小类，不使用类号，直接按大类和小类的类名字顺排列。大类用数字表示，小类用英文字母表示。GRA&I提供五种索引途径：关键词索引、个人著者索引、团体著者索引、合同号/资助号索引、NTIS订购号/报告号索引。

STAR于1963年创刊，原为半月刊，现为月刊，由美国国家航空与航天局编辑出版。它是查找NASA报告的主要检索工具，也包括一些AD报告、PB报告和DOE报告。STAR的文摘部分按类编排，1975年以来采用新的分类表，包括11个大类和75个小类。它每期由说明、分类表、正在进行的研究计划、文摘和索引部分组成，并且附有五种索引：主题索引、个人著者索引、团体来源索引、合同号索引和报告号/入藏号索引。此外，它还提供有季度、半年度和年度索引等累积索引，可进行追溯检索。

ERA创刊于1976年，半月刊，由美国能源部科技情报局编辑出版，是检索DOE报告的主要工具。它报道的文献主要以能源方面为主，也涉及一些环境科学、生物医学、物理学等方面的文献。ERA报道的文献按分类编排，共分为38个大类和313个小类，并提供五种索引途径：主题索引、个人著者索引、团体来源索引、合同号索引和报告号索引，还提供半年度和年度索引。

除了以上三种检索工具外，我国也有多个部门收藏有美国的四大报告。中国科技信息研究所是我国收藏国外科技报告最主要的单位。中国国防科技信息中心收藏有大量的AD报告和NASA报告，核工业部收藏有较多的DOE报告，中国科学院文献信息中心收藏的PB报告最全。

12.1.2.2 国内科技报告检索工具

1.《科学技术研究成果公报》

由原国家科委科技成果管理办公室编辑，科技文献出版社出版。1963年创刊，月刊，1966年停刊，1981年5月复刊。它以简介形式报道经国家科委登记公布的国家级重大科技成果，通报各部门、各地方的重大科技成果受奖项目。每

期内容分五大类：农业、林业，工业、交通及环境科学，医药、卫生，基础科学，其他。著录内容包括科技成果名称、登记号、分类号、部门或地方编码、基层编号及密级、完成单位及主要人员、工作起止时间、推荐部门、文摘内容。本工具书分类编排，每年最后一期附有公布项目总索引，以满足人们按分类途径进行回溯性检索的需要。

2.《中国国防科技报告通报与索引》

中国国防科技信息中心编辑，国防科工委情报研究所主办，原名《国防科技资料目录》，月刊。该刊报道该所收藏的中文国防科研、实验、生产和作战训练中产生并经过加工整理的科技报告和有关科技资料。近几年也以数据库的形式对外提供检索服务。

3.《中国机械工业科技成果通报》

由原机械工业部科技信息研究所主办，报道内容包括：基础理论研究成果、科研成果、新产品研制成果、软科学成果、专利成果等，按类编排。

12.1.3　科技报告网络检索方式

12.1.3.1　查找美国四大科技报告的网络检索方式

查找美国四大科技报告主要有以下几种网络检索途径：

1. 国内镜像站点

收藏和报道美国四大科技报告的是 NTIS 数据库。我国已有多家图书馆和文献信息机构购买了 NTIS 网络版文摘数据库。因此，对于国内用户来说，最为常用和方便的是通过国内图书馆的镜像站点来检索美国四大科技报告。目前，用户可以通过设在清华大学的《剑桥科学文摘》（*Cambridge Scientific Abstract*，CSA）中国镜像站点（http：//csa. tsinghua. edu. cn）及美国《工程索引》（Engineering Index，EI）中国镜像站点来进行网站检索。

2. 国家科技图书馆文献中心（http：//www. nstl. gov. cn/index. html）

用户通过登录国家科技图书馆文献中心的网站，可免费检索 NTIS 数据库，不过只能查找题名、作者、报告号、馆藏号、摘要等信息，不能获取原文。

3. NTIS（http：//www. ntis. org）

NTIS 报道的科技报告主要是美国的四大报告，另外包括美国农业部、教育部、环保局、健康与人类服务部、住房与城市部等的科技报告；同时也收录世界其他许多国家，如加拿大、俄罗斯、日本和欧洲各国以及一些国际组织的报告。通过该网站可以检索自 1990 年以来的文献记录，但只提供报告的题目、作者、主题词等信息，无文摘。用户只有缴费后才能看到文摘。

4. NASA 技术报告服务（http：//ntrs. nasa. gov/search. jsp）

NASA 技术报告服务（NASA Technical Report Server，NTRS）由 NASA 科技信息计划资助，可提供上百万篇文献的免费全文检索，不能提供全文的文献可查看摘要、关键词、文档源等信息。提供二次检索。

5. DOE 信息桥（http：//www. osti. gov/bridge）

DOE 信息桥是由美国能源部科技信息办公室提供的一项公共服务，用户可以免费查询 DOE 研究报告文献的全文和引文出处，内容涉及物理、化学、材料、生物、环境科学、能源技术、工程学、计算机与信息科学、可再生能源以及其他和 DOE 职责相关的主题。用户可进行一般检索和高级检索。在高级检索中，用户可根据 DOE 信息桥列出的主题目录进行浏览检索。

12.1.3.2 查找国内科技报告的网络检索方式

1. 万方数据资源系统的《中国科技成果数据库》

《中国科技成果数据库》是国家科技部指定的新技术、新成果查新数据库。其收录范围包括新技术、新产品、新工艺、新材料、新设计，涉及自然科学各个学科领域。该库已成为我国最具权威的技术成果宝库。并且该数据库能及时更新，目前共收录 49 万余条记录。如果用户所在单位购买了万方数据资源系统的《中国科技成果数据库》的使用权，可通过其所在单位的相关接口来对该库进行检索，如果没有的话，则需要首先支付一定的费用，获得用户名和密码后，登录 http：//www. wanfangdata. com. cn 进行检索。

2. 中国科技查新网（http：//www. chaxin. cn/sjkzy. html）

中国科技查新网中《中国科技成果数据库》始建于 1986 年，是国家科技部指定的新技术、新成果查新数据库。数据主要来源于历年各省、市、部委鉴定后上报国家科技部的科技成果及星火科技成果。2001 年《中国科技成果数据库》收录成果已达 24 万余条，在此基础上，每年新增 2 万条最新成果。其收录范围有新技术、新产品、新工艺、新材料、新设计，涉及化工、生物、医药、机械、电子、农林、能源、轻纺、建筑、交通、矿冶等十几个专业领域。《中国科技成果数据库》数据的准确性、详实性已使其成为国内最具权威性的技术成果数据库。它不仅可以用于成果查新和技术转让，还可以为技术咨询、服务提供信息源，为技术改造、新产品开发以及革新工艺提供重要依据。

3.《国家科技成果数据库》（http：//211. 151. 93. 229/grid20/Navigator. aspx? ID=SNAD）

《国家科技成果数据库》收录了 1978 年以来所有正式登记的中国科技成果，按行业、成果级别、学科领域分类。每条成果信息包含成果概括、立项情况、评

价情况、知识产权状况及成果应用情况、成果完成单位情况、成果完成人情况、单位信息等成果基本信息。成果的内容来源于中国化工信息中心，相关的文献、专利、标准等信息来源于 CNKI 各大数据库。可以通过成果名称、成果完成人、成果完成单位、关键词、课题来源、成果入库时间、成果水平等检索项进行检索。成果按照《中国图书资料分类法》（第四版）进行中图分类和按照按 GB/T13745《学科分类与代码》进行学科分类。该库每周进行数据更新。

12.2 会议文献检索

12.2.1 会议文献概述

会议文献是指在学术会议上宣读的论文、产生的记录及发言、论述、总结等形式的文献，许多学科中的新发现、新进展、新成就以及所提出的新研究课题和新设想，都是首先以会议文献的形式公之于众。它具有传播信息及时、主题集中、内容新颖、专业性强、质量较高等特点。

12.2.1.1 会议文献的类型

按出版时间的先后，会议文献可分为：会前文献、会中文献和会后文献。

1. 会前文献

会前文献是指在会议进行之前预先印发给与会者的文献，如论文、论文摘要或论文目录等。会前文献主要有四种类型：会议论文预印本、会议论文摘要、议程和发言提要、会议近期通讯或预告。

2. 会中文献

会中文献是指开会期间发给与会者的文献，主要包括会议议程、开幕词、闭幕词、讨论记录、大会提案和决议。

3. 会后文献

会后文献主要是指会议结束后正式出版的会议论文集，是会议文献中的主要构成部分。通常以会议录、会议论文集、学术讨论会论文集、会议论文汇编、会议记录、会议报告、会议文集等多种名称出版。

12.2.1.2 会议文献的出版形式

会议文献主要有四种出版形式：

1. 图书

大多数会议文献是以图书形式出版，且多数以其会议名称作为书名，或另加

书名，将会议名称作为副书名。一般按会议届次编号，定期或不定期出版。

2. 期刊

有相当部分的会后文献以期刊形式出版，大都发表在有关学会、协会主办的学术刊物中。

3. 科技报告

有些会议文献以科技报告形式出版，如美国四大报告中常编入会议文献。

4. 视听资料

除了以印刷品形式出版外，有些会议还在开会期间进行录音、录像，会后以视听资料的形式出版。

12.2.2 会议文献检索工具

12.2.2.1 国外会议文献检索工具

1.《科技会议录索引》（*Index to Scientific and Technical Proceedings*，ISTP）

《科技会议录索引》是检索正式出版会议文献最权威的工具，也是世界上三大著名的检索系统之一。它由美国费城科学情报研究所（Institute for Scientific Information，ISI）编辑出版，1978 年创刊，月刊。ISTP 主要收录国际上著名的科技会议文献，内容覆盖所有科技领域的会议文献，包括农业和环境科学，生物化学和分子生物学、生物技术、医学、工程、计算机、化学和物理等学科，每年报道论文近 13 万篇，占重要会议论文的 75%～90%。ISTP 出版月刊版和年刊版，月刊版由索引部分和正文部分组成。索引有 6 种：类目索引（Category Index）、作者/编者索引（Author/Editor Index）、主办单位索引（Sponsor Index）、会议地点索引（Meeting Location Index）、轮排主题索引（Permuterm Subject Index）、团体机构索引（Corporate Index）。正文部分按类目索引和会议登记号顺序排列。

2.《工程索引》（*The Engineering Index*，EI）

EI 创刊于 1884 年，由美国工程协会创办，美国工程信息公司出版，是国际上著名的主要收录工程技术期刊文献和会议文献的大型检索系统。主要用于检索文摘。EI 出版月刊和年刊。月刊由正文文摘、作者索引和主题索引（1987 年增设）三部分构成，并附有机构名称首字母缩写表（Acronyms，Initials and Abbreviation of Organization Name），每期收录 15 000 篇文献。年刊由文摘正文、索引和附表组成。索引有五种，分别是：作者索引（Author Index）、主题索引（Subject Index）、作者单位索引（Author Affiliation Index）（1988 年取消）、文摘号对照索引（Number Translation Index）（1987 年取消）、工程出版物索引

(Publication Index for Engineering)。附表有：机构名称首字母缩写表（Acronyms，Initials and Abbreviation of Organization Name)、缩写和单位表（Abbreviation，Units and Acronyms)。

3.《世界会议》(*World Meetings*，WM)

WM是由美国世界会议信息中心编辑、麦克米兰出版公司（Macmillan Publishing Company）出版的刊物，按季发行。专门预报两年内世界各国将要召开的各种学术会议，报道范围包括自然科学、应用科学、工程技术、社会科学及医学等学科领域。按其报道的地区和内容分为4个分册：《世界会议：美国与加拿大》(*World Meetings*：*United States and Canada*)，1963年创刊。预报近两年内将在美国和加拿大召开的各种医学、自然科学和工程技术会议。《世界会议：美国和加拿大以外的国家和地区》(*World Meetings*：*Outside United States and Canada*)，1968年创刊。预报近两年内将在美国和加拿大以外地区召开的各种医学、自然科学和工程技术会议。《世界会议：社会与行为科学，人类服务与管理》(*World Meetings*：*Social and Behavioral*，*Human Services and Management*)，1971年创刊。预报将在北美和其他国家和地区召开的社会科学方面的会议。《世界会议：医学》(*World Meetings*：*Medicine*)，1978年创刊。专门报道医学方面的会议。WM每分册由正文和索引两部分组成。正文部分按照会议登记号顺序排列。WM提供六种索引途径：关键词索引（Keyword Index)、会议日期索引(Date Index)、会议截稿日期索引（Deadline Index)、会议地址索引（Location Index)、出版物索引（Publication Index)、会议主办单位名录索引（Sponsor Directory and Index)。

4.《会议论文索引》(*Conference Papers Index*，CPI)

CPI由美国数据快报公司（Data Courier Inc.）于1973年创刊，月刊，原名为《近期会议预报》(*Current Program*)，1978年改为现名，1981年改由美国剑桥科学文摘社（Cambridge Science Abstract Inc.）编辑出版。从1987年起改为双月刊，主要报道刚召开或即将召开的各种学术会议上宣读或呈递的论文。属于题录型检索工具，收录学科包括生命科学、化学、物理学、地理科学、工程技术、医学等，年收录会议论文约7.2万篇。有月刊和年累积索引，月刊分正文和索引部分，提供主题和著者索引，年累积索引中增加了会议日期索引、会议地点索引。

12.2.2.2 国内会议文献检索工具

国内会议文献的检索工具主要是《中国学术会议文献通报》和《中国社会科学学术会议通览》。

1.《中国学术会议文献通报》

《中国学术会议文献通报》，1982 年创刊，原名为《国内学术会议文献通报》，季刊，1984 年起改为双月刊，1986 年起又改为月刊，1987 年改为现名。由中国科技信息研究所、中国农业大学主办，科技文献出版社出版。它由文献通报、会议预报和会议动态等三个相互独立的部分组成，内容涉及数理科学和化学、医药卫生、农业科学、工业技术、交通运输、航天航空、环境科学及管理科学。该刊报道方式以题录为主，兼有简介和文摘。每期报道 1 500～2 000条，论文按会议名称集中排列。每期附有《会议名称分类索引》。该刊自 1990 年起将每期的主题索引改为年度索引，在每年的最后一期中报道。用户可通过分类和主题途径检索。

2.《中国社会科学学术会议通览》(1979—1990)

《中国社会科学学术会议通览》由社会科学文献出版社 1992 年出版，书中收入了 1979—1990 年 12 年间在我国召开的国际性以及全国性的社会科学学术会议 5 283 个，并对其中有较大代表性的 818 个会议撰写了“会议述要”。全书内容包括会议召开的时间、地点、主办单位、会议规模、主要议题以及争论焦点，对不同的学术观点也做简要介绍。

12.2.3 会议文献网络检索方式

12.2.3.1 国外会议文献网络检索方式

1. ISI Proceedings（http：//www. isiknowledge. com）

ISI Proceedings 是美国 Thomson Scientific 公司推出的可供检索 ISTP 和 ISSHP（*Index to Social Science & Humanities Proceedings*，《社会科学与人文科学会议录索引》）的检索平台。汇集了世界上最新出版的科技领域会议录资料，包括专著、丛书、预印本以及来源于期刊的会议论文，内容涉及农业、环境科学、生物化学与分子生物学、生物技术、医学、工程、计算机科学、化学、物理学、社会学、公共健康、经济、管理、历史、文学等。它们总共覆盖了从 1990 年至今召开的 60 000 次会议上发表的约 350 万篇论文。每周更新，每年新增超过260 000条记录。国内许多大学的图书馆购买了 ISI Proceedings 的使用权，用户可通过这些图书馆的镜像站点访问 ISI Proceedings 的数据库。

2. EI Compendex Web（http：//www. lib. tsinghua. edu. cn/chinese/EI-village/switch. htm）

美国工程信息公司在继续以印刷版、联机与光盘形式出版 EI 外，也提供了网络版的《工程索引》数据库 EI Compendex Web。Compendex 是目前世界上最

全面的工程领域二次信息数据库。它收录了700多万条数据，这些数据出自5 000多种工程类期刊、会议论文集和技术报告。国内用户可通过清华大学图书馆的镜像站点访问EI Compendex Web。

12.2.3.2 国内会议文献网络检索方式

1.《中国学术会议论文全文数据库》

万方数据资源系统的《中国学术会议论文全文数据库》是国内最具权威性的学术会议论文全文数据库，收录了1998年至今的国家一级学会在国内组织召开的全国性学术会议近7 000个，数据范围覆盖自然科学、工程技术、农林、医学等27个大类，所收论文累计50万篇。用户可登录万方数据资源系统的首页或者已购买该库的学校图书馆中的相关接口进行会议文献的检索。

2.《中国重要会议论文全文数据库》

中国知网中的《中国重要会议论文全文数据库》收录我国2000年以来国家二级以上学会和协会、高等院校、科研院所、学术机构等单位的论文集，年更新约10万篇论文。产品分为十大专辑：理工A、理工B、理工C、农业、医药卫生、文史哲、政治军事与法律、教育与社会科学综合、电子技术与信息科学、经济与管理。十专辑下分为168个专题和近3 600个子栏目。用户可通过直接登录中国知网的首页（http：//www.cnki.net）或已购买该库的学校图书馆的接口进行相关会议文献的检索。

3. 中国学术会议在线（http：//www.meeting.edu.cn）

中国学术会议在线是经教育部批准，由教育部科技发展中心主办，面向广大科技人员的科学研究与学术交流信息服务平台。它是针对当前我国学术会议资源分散、信息封闭、交流面窄的现状，利用现代信息技术手段，为用户提供学术会议信息预报、会议分类搜索、会议在线报名、会议论文征集、会议资料发布、会议视频点播、会议同步直播等服务。

12.3 学位论文检索

12.3.1 学位论文概述

学位论文是指为取得学位而撰写的学术性研究论文，是学位制度的产物。英国习惯称之为Thesis，美国称之为Dissertation。学位论文形式上一般分为学士论文、硕士论文和博士论文。不过由于学士论文的质量较低，人们对其需求少，

也没有专门的检索工具。因此，通常的学位论文检索仅指检索硕士论文和博士论文。

学位论文在研究水平上参差不齐，但一般来说，硕士论文和博士论文多是在导师指导下，经过较长时间完成的，而且要经过相应研究领域的专家审查，其选题具有一定新意，探讨的问题较为专深，阐述也较具体详尽，具有一定的独创性和学术价值，是一种重要的信息资源。

由于所有要取得学位的学生都必须提交学位论文，其数量是非常可观的。学位论文除少数在答辩通过后发表或出版外，多数不公开出版发行，属于非卖品，只在授予学位的院校或研究机构的图书馆和按国家规定接受呈缴本的图书馆保存有副本。因此，学位论文的收集和使用不如公开出版物方便，一般只能通过复制来获取。我国的学位论文一般均由授予学位的学校图书馆自行收藏，要获得我国的学位论文通常可通过馆际互借关系复制取得，另外，中国科技信息研究所和中国国家图书馆是国家法定学位论文收藏单位，集中收藏了大量的科技或社科方面的学位论文。

12.3.2 学位论文检索工具

12.3.2.1 国外学位论文检索工具

1.《国际学位论文文摘》(*Dissertation Abstracts International*，DAI)

DAI创刊于1938年，现为月刊，由美国大学缩微品国际出版公司（UMI）出版，是目前世界上检索学位论文使用最广泛的一种检索工具。目前该刊分为3个分册：A辑《人文与社会科学》(*Humanities & Social Science*)、B辑《自然科学与工程》(*Science & Engineering*)、C辑《世界范围》(*World Wide*)。A辑，月刊，主要报道美国和加拿大的400多所大学及研究机构的人文和社会科学方面的博士论文。B辑，月刊，主要报道美国和加拿大的400多所大学的自然科学与工程技术方面的博士论文。C辑，季刊，原为欧洲学位论文文摘（European Abstracts），报道奥地利、荷兰、比利时、德国等欧洲国家著名大学的博士论文。从1989年起，C辑报道的学位论文扩大到世界范围，名称改为Word Wide，由UMI设在伦敦的分公司出版，报道世界范围各学科领域的博士及博士后学位论文，以西欧国家的学位论文为主，内容包括人文和社会科学领域、科学与工程技术领域。DAI提供关键词索引和著者索引两种检索途径。

2.《国际硕士学位论文文摘》(*Masters Abstracts International*，MAI)

MAI于1962年创刊，双月刊，由UMI编辑出版。主要报道美国和加拿大等国家100所大学的硕士学位论文，内容涉及自然科学、社会科学和应用科学等

各方面。

12.3.2.2 国内学位论文检索工具

1.《中国学位论文通报》

《中国学位论文通报》由中国科技信息研究所编辑，科技文献出版社出版，1985年创刊，原为季刊，1986年以后改为双月刊。它以题录、简介和文摘结合的形式报道我国自然科学领域的学位论文，每期内容包括分类目录、正文和索引三部分，是国内检索自然科学领域学位论文的主要工具。

2.《中国博士学位论文提要》

《中国博士学位论文提要》由国家图书馆编撰，北京图书馆出版社出版，内容覆盖自然科学、人文与社会科学各个领域，是目前检索中国博士学位论文最全的大型工具书。论文提要按顺序编排流水号。每篇论文提要包括论文题目、作者、指导教师、学位授予机构、学位授予年代、页数、提要正文和关键词。正文提要大约在500～800字之间，力求在较短的篇幅内概括出文章的中心思想和主要内容，揭示出文章的精神要义，反映出作者的学术观点和研究成果。书后附有著者索引和关键词索引。

12.3.3 学位论文网络检索方式

12.3.3.1 国外学位论文网络检索方式

1. ProQuest Digital Dissertation & Thesis（《博硕士学位论文数据库》，PQDT或称PQDD）

PQDT是美国UMI公司出版的世界著名学位论文数据库，收录欧美1 000余所大学文、理、工、农、医等领域的近200万博士、硕士论文的摘要及索引，是学术研究中十分重要的参考信息源，每年约增加6万篇博士论文信息。是目前世界上最大、使用最广泛的学位论文数据库。与其相对应的书本式的期刊有：*Dissertation Abstracts International*、*American Doctoral Dissertations*、*Comprehensive Dissertation Index*、*Masters Abstracts International*。其中博士论文摘要350字左右，硕士论文摘要为150字左右，1997年以后的博士论文有头24页全文，同时提供网上全文订购服务。PQDT具有收录年限长（从1861年开始）、更新快、提供部分全文等特点。我国有多所高校购买了PQDT数据库的使用权，用户可通过这些高校图书馆的镜像站点来检索PQDT数据库。

从2001年开始，在文摘库的基础上，ProQuest公司开发了电子版的学位论文全文服务方式，由国内高校、科研机构、公共图书馆等单位联合组成的Pro-

Quest 博士论文全文中国集团自 2002 年起开始订购 PQDT 中的部分博硕士论文全文。目前 ProQuest 学位论文全文数据库中收录有 4 万余篇博硕士论文，内容涵盖社会科学、哲学、宗教、环境学、生物学、语言、文学、教育、信息和艺术等多个学科。

2. Dissertation. com（http：//dissertation. com）

Dissertation. com 是一个帮助博士和硕士出版其英文版的电子学位论文的网站，并且可在世界范围内多个大型图书销售商的网站（Dissertation. com、Books. Google. com、Amazon. com）上进行销售。用户可通过科学浏览和关键词两种方式进行学位论文检索，可免费查看论文题目、作者、导师、机构、ISBN 号、页数、摘要、前 25 页论文等信息，付费后才能获取全文。目前，Dissertation. com 收录有人文与社会科学、自然科学与工程技术等方面的学位论文。

12.3.3.2 国内学位论文网络检索方式

1.《中国优秀博硕士学位论文全文数据库》（CDMD）

CDMD 是目前国内相关资源最完备、收录质量高、连续动态更新的博硕士学位论文全文数据库。目前它是中国知网数据库系列的一部分，分成 10 大专辑，168 个专题数据库。它覆盖的学科范围广泛，收录全国 652 家博硕士培养单位的优秀博硕士学位论文 43 万多篇。该数据库比较好地整合了国内的学位论文资源，集题录、文摘、全文文献信息于一体，实现一站式文献信息检索。用户可通过登录中国知网的首页或高校图书馆站点进行检索。

2.《高校学位论文数据库》（http：//opac. calis. edu. cn）

CALIS 中文全称为中国高等教育文献保障系统。CALIS 是由 CALIS 自建的数据库项目之一，由 CALIS 全国工程文献中心负责组织、协调 83 所高校合作建设的文摘索引数据库。它采用统一规范、分散加工、集中建库的运作模式，由工程文献中心制定数据规范，各个参建单位使用统一的录入软件，分散加工数据，并定期通过 FTP 方式向工程文献中心提交数据，工程文献中心对汇总的数据进行质量控制和检测后，通过 CERNET 提供服务。该项目于 1999 年 3 月启动，2000 年 4 月开始向高校用户提供服务。内容涵盖自然科学、社会科学、医学等各个学科领域。该库提供题录与文摘，没有全文，提供基本查询与高级查询两种检索方式。基本查询中可通过论文题名、作者、导师、作者专业、作者单位途径来检索。支持截词检索与逻辑与、逻辑或和年代限制选择。高级查询可进行最多 4 个检索词的复合检索。

3.《中国学位论文数据库》(CDDB)

CDDB收录了自1989年以来全国各高等院校、研究生院及研究所向中国科技信息研究所提交的自然科学领域的硕、博、博士后论文。目前，CDDB是万方数据资源系统的一部分，用户可通过直接登录万方数据库网站或者高校图书馆的镜像站点进行学位论文检索。

12.4 专利文献检索

12.4.1 专利文献概述

专利是指取得专利权的发明创造。专利文献是包含已经申请或被确认为发现、发明、实用新型和工业品外观设计的研究、设计、开发和试验成果的有关资料，以及保护发明人、专利所有人及工业品外观设计和实用新型注册证书持有人权利的有关资料的已出版或未出版的文件（或其摘要）的总称。专利文献有广义和狭义之分。广义的专利文献包括专利申请审批全过程产生的各种文件（如专利申请说明书、专利说明书等），以及专利公报、专利分类表、专利索引等出版物。狭义的专利文献仅指专利申请说明书和专利说明书。在专利文献的各种出版物中，专利说明书的出版量最大，世界上年出版量为100～110万件。

专利文献有如下特点：

(1) 资料新颖。据统计，世界上的技术发明有90%～95%发表在专利文献上。它以最快的速度报道最新的发明创造，有助于了解国内外科学技术的最新进展、水平动态。专利文献已成为申请专利“查新”的依据并可防止侵权行为，避免法律纠纷。

(2) 内容广泛。每年全世界公布的专利数量庞大，内容遍及各学科门类的物品、生产工艺和技术、方法等领域。通过专利文献，可以获取大量经济和技术信息，促进本部门业务开展和新市场开拓。

(3) 实用性强。各种专利说明书报道的内容具体、可靠，有的还包括附图，有助于对有关问题进行分析、借鉴。

(4) 分类逐渐趋向统一，格式标准化。为了便于国际交流与合作，1968年以来，越来越多的国家采用《国际专利分类法》(International Patent Classification，IPC)，或在专利说明书上标有IPC分类号，为专利文献的分类检索提供了方便。专利文献的撰写具有统一的方式和风格。

12.4.2 专利文献检索工具

12.4.2.1 国外专利文献检索工具

1.《世界专利索引》(*World Patent Index*，WPI)

WPI是检索世界各国专利文献的主要工具，由英国德温特公司制作出版。从1970年开始，出版《中心专利索引》(*Central Patents Index*，CPI)，包括药物、农业化学、塑料、化工等专业的专利文摘刊物共12种。1974年开始出版《世界专利索引》，内容涉及综合、机械、电气和化工等4个方面。年报道量有78万件，占世界专利文献总量的70%以上，以周报形式出版。1975年起又出版《世界专利文摘杂志》(*World Patent Abstract Journal*，WPA)，以文摘形式报道专利。加上CPI，与WPI配套使用，形成一套检索世界主要专利国家和各个专业门类专利文献的检索工具体系。目前，德温特公司专利检索体系共报道37个国家和地区、2个专利组织的专利文献，已成为世界上最著名、规模最大的专利文献检索系统。WPI由索引周报和文摘周报组成。索引周报分为四个分册：《综合分册》(P)、《机械分册》(Q)、《电气分册》(S-X)、《化工分册》(A-M)。每个分册由四种索引组成：专利权人索引（Patentee Index)、国际专利分类索引(IPC Index)、登记号索引（Accession Number Index)、专利号索引（Patent Number Index)。文摘周报分《电气专利索引》(*Electrical Patent Index*，EPI)、《化学专利索引》(*Chemical Patent Index*，CPI)、《工程专利索引》(*General & Mechanical Patent Index*，GMPI)。EPI对应于S-X分册，CPI对应于A-M分册，GMPI对应于P、Q分册。

2.《美国专利公报》

《美国专利公报》是美国专利商标局（United States Patent and Trademark Office，USPTO）权威性刊物。创刊于1872年，每周出版一次，主要连续报道USPTO专利事务方面的各种通知、命令，报道各类授权专利的著录项目（专利号、发明名称、发明人、专利权人、申请日期、申请号、国际专利分类号、美国专利分类号)、主权利要求和附图。每期专利公报后有两种索引：专利权人索引和分类索引，用户可通过分类、专利号、专利权人、发明人等途径检索专利文献。

3.《美国专利年度索引》

《美国专利年度索引》是美国专利文献检索的主要工具，由美国专利商标局出版。每年分两册出版：《专利权人索引》、《分类索引》。《专利权人索引》按发明人和专利权人字母顺序混合排列。发明人名称下列出：发明题目、专利号、批

准日期、分类号及受让人名称。专利权人名下列出：专利发明人、专利号及分类号。《分类索引》依大小类号顺序列出专利号，只有掌握了确切的分类号方可使用。《分类索引》分为两部分，前半部是主分类（Original Classification），把一年中标有原始分类号的专利，按分类号顺序排列，后半部是参见类（Cross Reference Classification），按参考类号排列。

12.4.2.2 国内专利文献检索工具

1.《中国专利公报》

中国专利局在1985年9月10日发布了首批《中国专利公报》，按发明类型分为《发明专利公报》、《实用新型专利公报》、《外观设计专利公报》三种。《专利公报》是查找专利说明书的文摘型检索工具。三种公报的编排结构基本一致。每期由目录、专利文摘、专利事务、各种索引及号码对照表四部分组成。报道内容按固定次序编排。

2.《中国专利年度索引》

《中国专利年度索引》是三种公报的年度索引，即分类年度索引和申请人、专利权人索引，分两册出版。前者将本年度三种公报上报道的全部公开、公告和批准的各种专利分别按IPC分类号或外观设计分类号编排。后者按申请人或专利权人的名称或译名的汉语拼音编排。

3.《中国专利分类文摘》

《中国专利分类文摘》是由原中国专利局文献馆在《专利公报》基础上加工编辑的一套检索工具，分两册出版：《中国发明专利分类文摘》、《中国实用新型专利分类文摘》。它报道我国年度公开的全部发明和实用新型专利说明书摘要。

12.4.3 专利文献网络检索方式

12.4.3.1 国外专利文献网络检索方式

1.《德温特创新索引》（Derwent Innovation Index，DII，http：//isiknowledge.com）

DII将《世界专利索引》（WPI）和《专利引文索引》（*Patents Citation Index*，PCI）的内容整合在一起，通过学术论文和技术专利之间的相互引证的关系，建立了专利与文献之间的链接。数据库每周更新，并且可以回溯到1963年，是检索全球专利的最权威的数据库。DII采用Web of Science的界面，与美国《科学引文索引》（SCI）、《期刊引文索引》（JCR）、《科技会议录索引》（ISTP）、《生物技术数据库》（BIOSIS Previews）四种数据库检索界面相同，可从五种数

据库直接进入检索，或五种数据库的跨库检索。DII 提供了快速检索、表格检索、引用检索、高级检索四种检索方式。一般检索方式有主题、专利权人、发明人、专利号。表格检索提供了通过主题、专利权人、发明人、专利号、国际专利分类号、德温特分类代码、德温特手册代码或德温特专利人藏号等检索入口。引用检索提供被引专利号、被引专利权人、被引专利发明人、被引专利的德温特人藏号进行检索的功能。高级检索即直接输入检索式进行检索，熟练的专家可以组织出精确复杂的检索策略，以查找所需要的专利信息。此外，DII 还支持逻辑运算符和通配符检索。国内用户可通过 ISI 在中国的镜像站点来检索 DII 数据库。

2.《美国专利商标局专利库》(http：//patents. uspto. gov/patft/index. html)

《美国专利商标局专利库》是由 USPTO 开发的数据库，收录了 1790 年以来的美国专利文献，数据每周更新，用户可免费进行检索，并能获取全文。《美国专利商标局专利库》分为两部分：《授权专利数据库》（Issued Patents）和《公开专利申请数据库》(Published Applications)。《授权专利数据库》提供 1976 年以来公布的美国专利全文和 1790 年以来公布的专利图像，提供快速检索、高级检索和专利号检索三种方式。《公开专利申请数据库》自 2001 年 3 月开始提供服务，数据库中的内容包括美国专利申请的题录、文摘、公开的美国专利申请说明书的全文，提供快速检索、高级检索和出版号检索。

3. 欧洲专利局专利信息网（esp@cenet）(http：//ep. esp@cenet. com)

esp@cenet 是欧洲专利局（European Patent Organization，EPO）从 1998 年开始向 Internet 用户提供的免费检索专利信息的系统。它共包括三个数据库：EP-esp@cenet、Worldwide、WIPO-esp@cenet。EP-esp@cenet 可用于检索近两年来由欧洲专利局出版的专利数据，可获取专利文摘和全文，该数据库每三周更新一次。Worldwide 提供了世界上 63 个国家和地区近 30 年来的专利文献数据，20 个国家 1920 年来的专利扫描图像以及 10 个专利机构的英文文摘和全文。WIPO-esp@cenet 可用于检索最近两年由世界知识产权组织出版的 PCT（Patent Cooperation Treaty）专利数据，该数据库每周更新。

12.4.3.2 国内专利文献网络检索方式

1.《中国专利全文数据库》

该数据库是中国知网的重要组成部分，收录了 1985 年 9 月以来的 230 余万条专利，包含《发明专利》、《实用新型专利》、《外观设计专利》三个子库，进一步根据国际专利分类（IPC 分类）和国际外观设计分类法分类，准确地反映中国最新的专利发明。专利的内容来源于国家知识产权局知识产权出版社，相关的文献、成果等信息来源于中国知网的各大数据库。可以通过申请号、申请日、公开

号、公开日、专利名称、摘要、分类号、申请人、发明人、地址、专利代理机构、代理人、优先权等检索项进行检索，并下载专利说明书全文。

2.《中国专利技术数据库》

《中国专利技术数据库》是万方数据资源系统的一部分，收录从1985年至今的发明专利、实用新型专利、外观设计专利数据信息，包含专利公开（公告）日、公开（公告）号、主分类号、分类号、申请（专利）号、申请日、优先权等数据项。目前，该数据库已收录290万余条专利数据。

3. 中国专利信息网（http：//www. patent. com. cn）

中国专利信息网始建于1998年5月。于2002年1月推出了改版后的新网站，集专利检索、专利知识、专利法律法规、项目推广、高技术传播、广告服务等功能为一体。用户在该网站注册后，即可获得有限的免费检索服务，可以免费检索近期相关专利的题名、摘要等信息。

4. 中华人民共和国国家知识产权局网（http：//www. sipo. gov. cn/sipo/zljs/default. htm）

中华人民共和国国家知识产权局网由国家知识产权局主办，收录了我国1985年来的全部发明专利、实用新型专利和外观设计专利，记录内容包括专利的完整题录信息、文摘、申请公开说明书和审定授权说明书。用户无需注册即可进行免费检索，检索途径有专利号、名称、摘要、申请日、公开日、公告日、IPC分类号等。

5. 中国知识产权网（http：//www. cnipr. com）

中国知识产权网是国家知识产权局知识产权出版社在国家的支持下于1999年6月创建的知识产权综合性服务网站。用户在该网站注册后，可进行国内外专利信息的检索。

12.5 标准文献检索

12.5.1 标准文献概述

标准是对重复性事物和概念所做的统一规定，它以科学、技术和实践经验的综合成果为基础，经有关方面协商一致，由主管机构批准，以特定形式发布，作为共同遵守的准则和依据。标准文献是指记录标准的一切物质载体，包括标准（Standard）、规范（Specification）、规则（Rules，Instruction）、工艺

（Practice）等。它具有明确的适用范围和针对性、可靠性和时效性强、有法律约束力、编排格式统一规范等特点。标准文献是标准化工作的成果，也是进一步推动科研、生产标准化进程的动力，它有助于了解各国的经济政策、生产水平、资源情况和标准化水平，对开发新产品、改进老产品有着重要的参考作用。

12.5.1.1 标准文献的类型

1. 按适用范围分

（1）国家标准。根据我国《国家标准管理办法》规定，强制性国家标准用“GB”为代号，推荐性国家标准用“GB/T”为代号。

（2）部（行业、专业）标准。根据我国《行业标准管理办法》规定，强制性行业标准的代号，用行业名称的两个汉语拼音字母表示，推荐性行业标准的代号，则在该拼音字母后加“/T”表示。

（3）指导性技术文件。用部（行业、专业）标准代号为分子，以“Z”为分母表示。

（4）企业标准。根据我国《企业标准管理办法》规定，企业标准的代号，用“Q”加斜线“/”加企业的数字代号表示。

（5）地方标准。自从我国《地方标准管理办法》颁布后，强制性地方标准的代号用“DB”加省、直辖市、自治区代码前两位数加斜线“/”表示，推荐性地方标准的代号在斜线后再加上“T”表示。

2. 按研究的对象和性质分

（1）技术标准。是对需要协调统一的技术事项所制定的标准。包括基本标准，产品标准，方法标准，工艺标准，设备标准，原材料、半成品和外购件标准，安全、卫生、环保标准等。

（2）管理标准。是对需要协调统一的管理事项所制定的标准。包括管理目标、管理项目、管理程序、管理方法和管理组织方面的标准。

（3）工作标准。是按工作岗位制定的有关工作质量的标准。具体内容有岗位目标、工作程序和工作方法、业务分工与业务联系（信息传递）方式、职责与权限、质量要求与定额、对岗位人员的基本技能要求和检查与考核办法。

3. 按成熟程度分

（1）强制标准。是指国家以法律条文、行政手段或国际组织之间以缔结条约的形式而颁布的标准，必须予以强制执行。

（2）推荐标准。是指行业协会或国际组织为适应某种发展趋势而推荐适用的标准。

12.5.1.2 我国标准文献的编号

1. 国家标准编号

国家标准的编号由国家标准代号、国家标准发布的顺序号和国家标准发布的年代号构成，即“国家标准代号—顺序号—年代号”。以“GB”表示强制性国家标准，以“GB/T”表示推荐性国家标准。

2. 行业标准编号

行业标准的编号由行业标准代号、行业标准发布的顺序号和行业标准发布的年代号构成，即“行业标准代号—顺序号—年代号”。强制行业标准的代号用行业名称的两个汉语拼音大写字母表示，如农业行业标准用“NY”表示，化工行业标准用“HG”表示，轻工行业标准用“QG”表示等。推荐行业标准代号是在两个汉语拼音大写字母后加斜线“/”和“T”表示。

3. 企业标准编号

企业标准的编号由企业标准代号、标准顺序号和发布标准年代号组成，即“企业标准代号—顺序号—年代号”。企业标准代号由汉语拼音字母“Q”加斜线“/”和企业代号组成。

4. 地方标准编号

地方标准的编号由地方标准代号、标准顺序号和发布标准年代号组成，即“地方标准代号—顺序号—年代号”。地方标准的代号，由汉语拼音字母“DB”，加上省、自治区、直辖市行政区代码前两位数再加斜线，组成强制地方标准代号；再加“T”，组成推荐地方标准代号。如北京市强制地方标准代号为“DB11/”，北京市推荐地方标准代号为“DB11/T”；天津市强制地方标准代号“DB12/”，天津市推荐地方标准代号“DB12/T”。

12.5.1.3 我国标准文献的分类

我国标准文献的分类主要采用《中国标准文献分类法》。《中国标准文献分类法》是在原国家标准组织编制的《中国标准文献分类法（试行）》基础上进行修订而成的，是目前国内用于标准文献管理的一部工具书。该分类法由24个一级大类目组成，用英文字母表示，每个一级类目下分100个二级类目，二级类目用两位数字表示。类目的设置以专业划分为主，适当结合科学分类。《中国标准文献分类法》的一级类目如下：

A 综合	N 仪器、仪表
B 农业、林业	P 工程建设
C 医药、卫生、劳动保护	Q 建材

D 矿业	R 公路、水路运输
E 石油	S 铁路
F 能源、核技术	T 车辆
G 化工	U 船舶
H 冶金	V 航空、航天
J 机械	W 纺织
K 电工	X 食品
L 电子元器件与信息技术	Y 轻工、文化与生活用品
M 通信、广播	Z 环境

12.5.2 标准文献检索工具

12.5.2.1 国外标准检索工具

1.《国际标准目录》（*ISO Catalogue*）

《国际标准目录》是由国际标准化组织（International Organization for Standardization，ISO）编辑出版，年刊，以英法两种文字出版，报道上一年度的全部现行标准，包括新近批准生效的标准和作废的标准。它主要由五部分组成：（1）分类目录，ISO 标准的分类按制定标准的技术委员会（Technical Committee，TC）的名称设立类目。分类号由字母加数字组成，如 TC55。1993 年以后，使用《国际标准分类表》（International Classification for Standards，ICS）。（2）主题索引，该索引采用文中关键词排检。（3）标准序号目录，包括标准号、TC 号。（4）技术委员会序号索引，包括技术委员会 TC 号、标准号和标准在分类目录中的页码。（5）废弃目录，在目录下列出已作废标准的标准号，同时对照现行标准的标准号，内容根据作废标准的标准号顺序排列。

2.《ISO 技术规则》

《ISO 技术规则》由国际标准化组织编辑出版，年刊，报道可视为国际标准的文件和已达到委员会草案阶段和国际标准草案阶段的全部文件。

3.《国际电工委员会出版物目录》（*Catalogue of IEC Publications*）

《国际电工委员会出版物目录》是由国际电工委员会（International Electrotechnical Commission，IEC，主要负责电气和电子领域中标准化组织和协调工作，制定电子、电力、电信和原子能等的国际标准）编辑出版，年刊，以英法两种语言对照出版。由两大部分组成：标准序号目录（Numerical List of IEC Publications）和主题索引（Subject Index）。标准序号目录按标准序号顺序排列，主题索引按主题词字顺排列，其后有相应标准号。用户可检索到标准号、标准制定

年份、标准名称、页数、价格、版次、简介等信息。

4.《美国国家标准目录》(*ANSI Catalogue*)

《美国国家标准目录》是由美国国家标准协会（American National Standards Institute，简称 ANSI）编辑出版，年刊。该目录由三部分组成：主题索引、分类索引、标准序号索引。主题索引是目录的正文部分，按产品名称字母字顺排列，可检索到主题词、标准名称、标准号、价格等信息。分类索引采用 ANSI 自己制定的标准分类体系进行编排，可按分类号和标准号检索。标准序号索引按各专业标准的序号排列。

12.5.2.2　国内标准文献检索工具

1.《中华人民共和国国家标准目录总汇》

《中华人民共和国国家标准目录总汇》由国家质量技术监督局编，中国标准出版社出版，年刊。由分类目次、目录正文和辅助索引三部分组成。自 1999 年起，每年上半年出版新版，载入截止到上一年度批准发布的全部现行国家标准信息，同时补充载入国家标准清理整顿、复审、补充、修改和更正等相关信息。

2.《中国标准化年鉴》

《中国标准化年鉴》由国家标准局编辑，中国标准出版社出版，1985 年创刊。该年鉴主要分说明和目录两大部分及标准号索引。说明部分包括中国标准化发展概况、标准化工作、标准化学术活动和有关的统计数字等，用中英文对照排印，每年增加新的内容。目录部分即国家标准分类目录，按专业分类，每一类又按标准号排列。

3.《中国国家标准汇编》

《中国国家标准汇编》自 1983 年起陆续出版，由中国标准出版社出版，收集了我国正式发布的全部现行国家标准。自 1995 年起，新增出版在上一年度被修订的国家标准的汇编本。

4.《中国标准导报》

《中国标准导报》由中国标准出版社主办，1992 年 6 月 1 日创刊，双月刊。它是集政策、学术、技术、信息于一体的标准化综合性刊物。主要宣传国家标准化工作的方针和政策，报道标准化的发展和动态，介绍国内外标准化领域的最新研究成果，提供最新标准发布、出版、废止及代替信息。

5.《国家标准代替、废止目录》

《国家标准代替、废止目录》由中国标准出版社出版，每年出版一次，提供国家标准的最新代替、废止和转化信息。所涉及的领域包括：包装、船舶、城建、电力、地质矿产、核工业、纺织、供销、化工、电子、冶金、机械、林业、

农业、医疗器械等 28 个行业。该目录由四个部分组成：被代替国家标准目录、国家标准转化行业标准目录、国家标准废止目录、索引。被代替国家标准目录包括：被代替标准编号（指最近一次被代替的标准编号）、历次修订编号、现行标准编号、现行标准名称。索引包括“现行标准编号与被代替的标准编号对照表”和“历次修订情况中非同号被代替标准编号与现行标准编号对照表”。

12.5.3 标准文献网络检索方式

12.5.3.1 国外标准文献网络检索方式

1. ISO 网站（http：//www.iso.org）

ISO 已经制定了 17 000 余条国际标准，并且每年会出版 1 100 条新的标准。所有的标准文献都可通过登录 ISO 网站进行检索。ISO 提供的产品有手册（Handbooks）、包（Packages）、目录清单（Check-lists）、数据库（Databases）和杂志（Magazines）五种形式。用户可以通过主题词进行检索，可免费查看摘要、标准名称、页数、价格、版本、TC 和 ICS 号等信息，付费后才能获取全文。用户还可通过 ICS 和 TC 号进行浏览检索，也可以通过标准名称的字顺索引进行浏览检索。此外，用户在该网站上还可了解到关于 ISO 组织、发展、最近消息、教育和培训等相关信息。

2. IEC 网站（http：//www.iec.ch）

IEC 网站是国际电工委员会提供的标准文献的网络检索方式，用户可通过快速检索（IEC 编号）、文本检索（主题）和参考文献检索方式来进行标准文献的检索。可免费获得标准文献名称、摘要、IEC 标准号等信息，付费后才能获取 PDF 格式的全文，有的全文提供双语版本。

3. PERINORM 标准数据库（http：//www.cssinfo.com）

该数据库收录了世界上 45 万余条工业技术标准文献及规范，包括 ISO、IEC、IEEE 等组织制定的标准，其中大约有 5 000 条标准的 PDF 格式的标准全文，可直接下载。

4. 其他网络检索方式

世界上其他著名的标准文献检索方式有：世界标准服务网（http：//www.wssn.net）、IEEE 标准（http：//standards.ieee.org）、美国国家标准系统网络（http：//www.nssn.org）、英国标准研究所（http：//www.bsi.org.uk）、加拿大标准委员会（http：//www.scc.ca）、德国标准协会（http：//www.din.de）、法国标准协会（http：//www.afnor.fr）、新西兰标准组织（http：//standards.co.nz）等。

12.5.3.2　国内标准文献网络检索方式

1. 中国标准咨询网（http：//www. chinastandard. com. cn）

由中国技术监督情报协会、北京中工技术开发公司与北京世纪超星信息技术发展有限责任公司合作创建，为我国各行各业及科研单位面向世界走向国际市场提供技术监督法规信息、国内外标准信息、产品抽检信息和质量认证信息等全方位的网上咨询服务。中国标准咨询网是国内首家标准全文网站，有 ISO 标准、IEC 标准、ASME 标准、ASTM 标准、BS 标准、DIN 标准、JIS 标准、GB 标准、HB 标准、GBJ 标准、IEEE 标准、ANSI 标准等数十万标准。提供简单检索和高级检索，可以从中文标准名称、发布日期、发布单位、实施日期、英文标准名称等多个途径进行检索。

2. 中国标准服务网（http：//www. cssn. net. cn）

中国标准服务网是中国标准化研究院开发的国家级标准信息服务门户，是世界标准服务网的中国站点，其标准信息主要来自国家标准化管理委员会、中国标准化研究院标准馆及科研部门、地方标准化研究院（所）以及国内外相关标准化机构。目前，中文数据库包含的标准种类有：国家标准和国内所有行业标准，以及 ISO、IEC、ANSI、BS、DIN、NF、JIS、ASME、ASTM、IEEE、UL 等国际和国外标准。网站设有“标准化新闻”、“标准化动态”、“标准检索”、“政策法规”、“参考资料”等栏目。提供标准号、中文标题、英文标题、中文关键词、英文关键词、被代替标准、中标分类号、国际分类号等检索途径。

3. 中国标准网（http：//www. zgbzw. com）

中国标准网由北京北标科技发展有限公司和北京浩瀚角雅典书屋共同提供，设有“图书目录”、“光盘/录像带”、“标准知识”、“每日新闻”等栏目。提供国家及行业标准查询、图书查询以及部分国际标准查询（ISO、IEC、UL）。检索途径有标准号、中文标准名称、英文标准名称、分类号等。

4.《中外标准数据库》

《中外标准数据库》是万方数据资源系统的一部分。该库收录了国内外的大量标准，包括中国国家发布的全部标准、某些行业的行业标准以及电气和电子工程师技术标准；收录了国际标准数据库、美、英、德等的国家标准，以及国际电工标准；还收录了某些国家的行业标准，如《美国保险商实验所数据库》、《美国专业协会标准数据库》、《美国材料实验协会数据库》、《日本工业标准数据库》等。

5. 其他标准文献网络检索方式

其他可以检索标准文献的网站有：中国标准化研究院（http：//

www. cnis. gov. cn）、中国标准咨询服务网（http：//www. chinagb. org）、中国标准出版社网（http：//www. bzcbs. com）、国家标准化管理委员会网（http：//www. sac. gov. cn）。

12.6 档案文献检索

12.6.1 档案文献概述

档案文献是人们在实践活动中形成的历史记录。它的主要特点是历史性、真实性、确定性和知识性等。档案文献具有凭证价值和参考价值，因此在社会的方方面面发挥着广泛而重要的作用。档案文献的作用主要有：（1）行政作用。档案是国家以及各级政府机构制定政策法规、处理社会问题的依据，也是提高政府机构决策和管理科学性的必要条件。（2）业务作用。档案文献记载了业务活动的开展、结束、所取得的成果以及经验教训等信息，可以为后续活动的开展提供信息支撑和保障，具有重要的参考和业务指导作用。（3）文化作用。档案是人类历史文化的积累，反映了一个国家或民族的文化特征，是文化传播和发展的手段，也是文化创新的基础。（4）法律作用。档案是当事人在当时、当地所形成的原始记录，真实性、可靠性强，因此可以在解决争端、处理案件等活动中发挥证据作用。（5）教育作用。档案文献因真实性而具有很强的说服力和感染力，因而是一种重要的教育资源。在我国，档案在对青少年的精神文明教育、爱国主义教育和历史文化教育方面一直都发挥着不可替代的重要作用。

12.6.2 档案文献检索工具

伴随着我国政务公开和档案开放的进程，越来越多的机关工作人员、科技人员以至普通公民开始了解和利用档案文献这种具有特殊价值的信息资源。我国档案数量庞大，内容丰富，时间跨度大，分布面广。据国家档案局 1998 年底的统计，全国各级各类档案馆及地、师级以上的机关档案室共保存档案约 2.7 亿卷，排架长度约 6 500 公里，内容涉及党务、政务、经济、科技、军事、文化等社会生活的方方面面。要想从这浩如烟海的档案中获得特定的档案文献或信息，需要借助于合适的档案文献检索工具。

档案文献检索工具种类较多，根据以下标准可对档案文献检索工具进行划分：

1. 按编制方式分

(1) 目录：由揭示档案特征的条目汇集而成并按照一定次序编排的档案检索工具。如分类目录、主题目录、专题目录、案卷目录、卷内文件目录等。

(2) 索引：将档案的某些特征按一定次序编排并注明出处的档案检索工具。如文号索引、人名索引、地名索引等。

(3) 指南：以文章叙述的方式，综合介绍档案情况的一种检索工具。如档案馆指南、全宗指南、专题指南等。

2. 按载体形式分

(1) 卡片式检索工具：将条目著录于卡片上，将卡片按一定顺序排列而成的检索工具。

(2) 书本式检索工具：将著录条目顺序排列并装订成册的检索工具。

(3) 缩微式检索工具：用缩微摄影方式制作的以胶片为载体的检索工具。

(4) 机读式检索工具：将档案的内容和形式特征存储在计算机存储介质上，利用计算机进行检索的工具。

3. 按内容范围分

(1) 综合性检索工具：以一个或若干个档案馆的全部档案或以一全宗的档案为检索和介绍对象的检索工具，如全宗文件目录、分类目录、全宗指南等。

(2) 专题性检索工具：以某一专题的档案为对象的检索工具，如专题目录、专题指南等。

4. 按功能分

(1) 馆藏性检索工具：反映档案实体整理体系及其相互关系的检索工具，如全宗目录、案卷目录等。

(2) 查检性检索工具：从档案的某一内容或形式特征提供检索途径的检索工具，如分类目录、主题目录、专题目录、人名索引、地名索引、文号索引等。

(3) 介绍性检索工具：介绍和报道档案内容及其有关情况的检索工具，如专题指南、全宗目录、档案馆指南等。

12.6.3 档案文献网络检索方式

随着高新技术的飞速发展，以计算机、网络通信技术为代表的互联网的兴起，传统的档案检索也开始发生变化，以网络为依托，数据库检索成为主要形式。

1.《美国国家档案馆档案数据库》(Access to Archival Databases, AAD, http://aad.archives.gov/aad)

AAD是美国国家档案馆（National Archives and Records Administration,

NARA）在“电子文件档案馆”项目（Electronic Records Archives，ERA）的支持下发展起来的第一个公开的可利用的应用系统，主要用于检索 NARA 电子文件。该系统可以在线检索超过 20 个美国联邦机构所产生的近 5 000 000 份涵盖各个主题范围的电子文件，用户能够检索所需要的含有特殊信息的文件。它拥有重要的背景信息，帮助用户更好地理解文件，包括代码列表、说明性注释，以及一些档案的相关文献。用户可按关键词进行检索，也可根据分类目录进行浏览检索（可按家族/个人历史、个人部分、战争/国际关系、时间间隔、地区、政府开支、索引项等分类），还可根据主题分类进行浏览检索。此外，AAD 还列出了最新的电子档案文件和最受用户欢迎的电子档案文件。对于每一份电子文件，用户可查看到文件标题、创建者、档案载体形式、描述程度、其他标题、地点、日期范围、功能与用途、内容、访问限制、普通注释、索引词汇等信息。

2. 北京市档案信息网（http：//www. bjma. gov. cn）

北京市档案馆现有馆藏 171 万卷（册），排架长度一万多米，包括纸质、录音、录像、影片、照片等各种载体，内容十分丰富。档案馆通过网上利用、档案阅览、举办展览、史料出版等多种途径为社会提供档案利用服务。已建立了 6 个档案目录数据库，即《民国时期档案目录数据库》、《中华人民共和国时期档案目录数据库》、《档案资料目录数据库》、《北京市劳动模范档案目录数据库》、《诉讼档案目录数据库》、《工商税务档案目录数据库》，共有 796 686 条目录数据。同时，用户可对数字化档案进行在线阅览。由于该系统还处在试运行阶段，目前只提供了北平市政府、北平市社会局、北平市民政局、北平市教育局、北平市卫生局等 7 个专题近 180 余万页的数字化档案。提供关键词检索途径。

3. 上海档案信息网（http：//www. archives. sh. cn）

上海档案信息网分为公共服务、珍档荟萃、馆藏指南、网上展览、档案文库、兰台纵横、专题精粹、档案博览、申城变迁、沪上机构、海上人物、上海掌故等板块，用户可直接查看里面的全文。除了文字形式的文件外，还有大量的图片。该网站提供了全宗指南和专题指南，也可直接查看全文。

【案例】

华为进军美国遭遇思科阻击[①]

华为是中国高科技企业的一面旗帜。截至 2005 年 6 月，华为已经申请各

① 李启章、吴辉、张璇、裴宏、曾旭辉、徐进：《对部分知识产权典型案例的分析报告之三》，见 http：//www. sipo. gov. cn/sipo2008/albd/2005/200804/t20080402 _ 367658. html，2008-05-30。

种专利6 500多项，累计授权1 300多项，且其中多为发明专利。但就是这样一家企业，由于没有提前做好专利布局，在美国市场的拓展就碰上了专利的高墙。

与华为发生争执的是全球最大的网络通信设备厂商思科。自2000年开始，华为就把路由器作为自己的主打产品，推出了Quidway中低端路由器，还发布了Netengine高端路由器。华为产品不单在中国市场阻击思科，而且还积极向东南亚、俄罗斯、埃及和南美等地渗透，甚至将价格竞争战术推广到西欧和北美的发达国家。

2003年1月22日，思科在美国起诉，指控华为技术有限公司及其在美国的两家全资子公司侵犯了思科拥有的知识产权。

华为一方面尽快停止了涉嫌侵权的路由器在美国市场的销售，以期减少可能在未来发生的天价侵权赔偿额，另一方面加快了与美国3COM公司的合作，该公司作为美国土生土长的专利战悍将，有900多项美国专利。2003年6月11日，3COM也正式介入诉讼，成为第三方，要求法院判决3COM与华为合资生产的产品没有侵权，以保证其与华为的新合资公司产品的顺利销售。

2004年7月，诉讼三方达成和解协议，华为同意停止销售诉讼中所提及的产品，并且在全球范围内只销售经过修改后的新产品。

思科、华为一案说明，专利在企业发展和开拓市场的过程中具有重要的作用。中国高科技企业开拓国外市场，需提前做好专利布局，尽量避免并准备好应对竞争对手的专利战。中国企业在向国外出口产品前，要进行有关的知识产权调查和有关专利文献的检索，如果发现存在侵权的可能，应及时对产品进行修改。

关键术语

科技报告	会议文献	学位论文	专利文献
标准文献	档案文献	检索工具	网络检索

思考题

1. 科技报告有哪些特点?
2. 简述美国的四大科技报告及其检索工具。
3. 简述ISTP和EI。
4. 国内会议文献的网络检索方式主要有哪些?

5. 简述国内学位论文的主要数据库。
6. 简述国内专利文献的检索工具及主要网络检索方式。
7. 简述标准文献的类型及编号方式。
8. 简述 ISO 和 IEC。
9. 国内标准文献主要检索工具有哪些？
10. 简述档案文献的价值和作用。
11. 档案文献检索工具有哪些？
12. 简述 AAD。

REFERENCE

主要参考文献

图书

1. 赵丹群．现代信息检索原理、技术和方法．北京：北京大学出版社，2008
2. 赵泉．信息检索．北京：机械工业出版社，2008
3. 马张华，黄智生．网络信息资源组织．北京：北京大学出版社，2007
4. 朱俊波．实用信息检索．成都：西南交通大学出版社，2007
5. 白冰．中文工具书实用．上海：上海古籍出版社，2007
6. 刘英华，赵哨军，汪琼．信息资源检索与利用．北京：化学工业出版社，2007
7. 彭奇志．信息检索与利用教程．北京：北京大学出版社，2006
8. 李朝云，傅正．现代信息检索与利用．合肥：安徽大学出版社，2006
9. 段明莲，沈正华．数字时代的图书馆信息资源组织．北京：北京图书馆出版社，2006
10. 陈雅芝．信息检索．北京：清华大学出版社，2006
11. 刘阿多．科技网络信息资源检索与利用．南京：东南大学出版社，2005
12. 王云娣．数字信息资源的开发与利用研究．武汉：武汉大学出版社，2005
13. 张燕飞．信息组织的主题语言．武汉：武汉大学出版社，2005
14. 冯惠玲．档案文献检索．北京：高等教育出版社，2004
15. 李晓明，闫宏飞，王继民．搜索引擎——原理、技术与系统．北京：科学出版社，2004
16. 戴维民．信息组织．北京：高等教育出版社，2004
17. 马文峰．人文社会科学信息检索．北京：北京图书馆出版社，2004
18. 孙建军．信息检索技术．北京：科学出版社，2004
19. 张琪玉，侯汉清．情报检索语言实用教程．武汉：武汉大学出版社，2004
20. 焦玉英．信息检索进展．北京：科学出版社，2003
21. 董小英，马张华等．互联网信息资源的检索利用与服务．北京：北京大学出版社，2003
22. 卢小宾，李景峰．信息检索．北京：科学出版社，2003

23. 叶继元．信息检索导论．北京：电子工业出版社，2003
24. 张帆．信息存储与检索．北京：高等教育出版社，2003
25. 刘景会，杨建民，范明祥．文献信息检索通论．北京：中国科学技术出版社，2003
26. 李国辉，汤大权，武德峰．信息组织与检索．北京：科学出版社，2003
27. 史田华．信息组织与存储．南京：东南大学出版社，2003
28. 马张华．信息组织．北京：清华大学出版社，2003
29. 谢新洲．数字出版技术．北京：北京大学出版社，2002
30. 肖明．信息资源管理．北京：电子工业出版社，2002
31. 高润芝．现代信息资源检索与利用．北京：经济管理出版社，2002
32. 沈固朝．信息检索（多媒体）教程．北京：高等教育出版社，2002
33. 郭太敏．信息资源检索与利用．北京：中国矿业大学出版社，2002
34. 华薇娜．网络学术信息资源检索与利用．北京：国防工业出版社，2002
35. 王云庆，苗壮．现代档案管理学．青岛：青岛出版社，2002
36. 刘俊熙．信息检索．北京：北京图书馆出版社，2002
37. 毕强，杨文祥．网络信息资源开发与利用．北京：科学出版社，2002
38. 钟守真．信息资源管理概论．天津：南开大学出版社，2001
39. 焦玉英，符绍宏，何绍华．信息检索．武汉：武汉大学出版社，2001
40. 信息检索利用技术编写组．信息检索利用技术．成都：四川大学出版社，2001
41. 谈大军，李志义．文献与网络信息检索．广州：华南理工大学出版社，2001
42. 陆建平．信息检索 从手工到联机、光盘、因特网．上海：华东师范大学出版社，2001
43. 王子荣．Internet 基础．北京：电子工业出版社，2001
44. 王梦丽等．信息检索与网络应用．北京：北京航空航天大学出版社，2001
45. 冯惠玲，张辑哲．档案学概论．北京：中国人民大学出版社，2001
46. 马张华．信息组织．北京：清华大学出版社，2001
47. 俞君立，陈树年．文献分类法．武汉：武汉大学出版社，2001
48. 张清华．信息检索基础教程．北京：中国林业大学出版社，2000
49. 董源．信息检索学．北京：中国林业出版社，2000
50. 方正，蔡梅，杨琼，谢立虹．社会科学信息检索与利用．长沙：中南大学出版社，2000
51. 刘俊熙，赵伯兴，蒋时雨．网络环境下的社会科学信息检索．上海：上海大学出版社，2000
52. 李进，祝闽．Internet 学习教程．北京：北京大学出版社，2000
53. 符绍宏等．因特网信息资源检索与利用．北京：清华大学出版社，2000
54. 清源计算机工作室．初探世界——教你利用网络资源．北京：机械工业出版社，2000
55. 曹树金，罗春荣．信息组织的分类法与主题法．北京：北京图书馆出版社，2000
56. 冯惠玲．档案检索．北京：高等教育出版社，1999
57. 郭庆光．传播学教程．北京：中国人民大学出版社，1999

58. 郑章飞．现代信息检索．武汉：华中理工大学出版社，1999
59. 马文峰．社科文献检索．北京：中国人民大学出版社，1999
60. 马文峰．社会科学文献信息检索概论．北京：中国人民大学出版社，1999
61. 任胜国，周敬治．文献信息检索教程．北京：北京图书馆出版社，1999
62. 张力治．情报学进展．北京：航空工业出版社，1999
63. 储荷婷，张晓林，王芳．Internet 网络信息检索．北京：清华大学出版社，1999
64. 孟广均，沈英等．信息资源管理导论．北京：科学出版社，1998
65. 朱天俊，李国新．中文工具书基础．北京：北京图书馆出版社，1998
66. 袁正平原著，郑红修订．中文工具书实用教程．成都：四川大学出版社，1998
67. 李琳，秦洪晶．网海拾贝——Internet 信息查询．北京：人民邮电出版社，1998
68. 洪漪．档案信息组织与检索．武汉：武汉大学出版社，1998
69. 曾蕾．联机环境中的情报检索语言．北京：书目文献出版社，1996
70. 钟义信．信息处理科学原理．北京：北京邮电大学出版社，1996
71. 詹德优．中文工具书导论．武汉：湖北教育出版社，1994
72. 刘昭东．信息与信息化社会．北京：科技文献出版社，1994
73. 张惠惠．情报联机检索．上海：上海交通大学出版社，1993
74. 张琪玉．档案检索．北京：书目文献出版社，1993
75. 王秀兰．英文工具书．武汉：武汉大学出版社，1991
76. 臧志芬．参考工作与参考工具书．北京：书目文献出版社，1988
77. 赵国璋，朱天俊，潘树广．社会科学文献检索．北京：北京大学出版社，1987
78. 彭斐章，乔好勤，陈传夫．目录学．武汉：武汉大学出版社，1986
79. 刘湘生．主题法理论与标引．北京：书目文献出版社，1985
80. G. 隆多著；刘钢，刘健译．术语学概论．北京：科学出版社，1985
81. 兰开斯特著；王知津，陈光祚译．情报检索系统．北京：书目文献出版社，1984
82. 张琪玉．情报检索语言．武汉：武汉大学出版社，1983

论文

1. 甘晨，易法令，王圆妹．基于颜色和纹理特征的图像检索技术研究．中国科技信息，2008（8）
2. 汪建中，韩维柱．Ebsco 数据库现状调查与统计分析．情报科学，2008（4）
3. 徐庆，杨维维，陈生潭．基于内容的图像检索技术．计算机技术与发展，2008（1）
4. 张利平．基于综合特征的图像检索技术的研究．图书馆学研究，2007（12）
5. 奉国和．自动文本分类技术研究．情报杂志，2007（12）
6. 章成志．自动标引研究的回顾与展望．现代图书情报技术，2007（11）
7. 康艳，张虹，侯汉清．情报检索语言不是“明日黄花”．图书情报工作，2007（10）
8. 汪维华，汪维清．基于内容的多媒体检索技术．计算机工程与设计，2007（10）

9. 龚立群，孙洁丽．OAI、SRW/U 及 OpenURL 的比较及协同使用研究．情报科学，2007（7）

10. 谢琳惠．我国数据库产业的现状、问题及对策．中国图书馆学报，2007（5）

11. 耿骞，赖茂生．自然语言检索的实现及其关键问题．情报科学，2007（5）

12. 颜端武，岑咏华，毛平，成晓．领域知识本体的可视化检索研究．中国图书馆学报，2007（4）

13. 李爱明，刘冰．个性化信息检索系统的用户模型研究．情报杂志，2007（3）

14. 曹红兵．搜索引擎的个性化检索研究．图书情报工作，2007（3）

15. 赵丹群．试论信息检索系统的后控制机制．图书情报工作，2007（1）

16. 季春．音频信息检索技术的发展及应用．现代情报，2007（1）

17. 杨永升．搜索引擎质量评价体系实证研究．合肥工业大学硕士学位论文，2006

18. 刘志芳．网络环境下的个性化检索定制服务．大学图书情报学刊，2006（10）

19. 冯向春．网络工具书资源的评价与利用．现代情报，2006（6）

20. 潘瑞冰．论网络信息的自然语言检索．图书馆学刊，2006（5）

21. 李育嫦．自然语言检索中的词汇控制研究．图书馆学研究，2006（4）

22. 田甜．文档自动分类的方法探讨．情报杂志，2006（2）

23. 秦春秀，赵捧未，窦永香．基于用户兴趣的个性化检索．情报学报，2005（24）

24. 马慧芳，郭治成．文本自动分类在搜索引擎中的应用研究．情报杂志，2005（12）

25. 林彤，江志军．Internet 的搜索引擎．计算机工程与应用，2005（5）

26. 杨先明，但碧霞．网络信息资源的分布特点及其利用对策分析．图书馆论坛，2005（5）

27. 龚芳，耿骞，王洋．关于自然语言检索问题．中国图书馆学报，2005（4）

28. 何灵巧，陆宗城．情报检索语言的发展方向问题——分类主题一体化新论．图书情报知识，2005（2）

29. 曹树金，杨涛．自动分类在搜索引擎性能优化中的应用．情报科学，2004（22）

30. 耿骞，汤艳莉．面向网络信息资源的自然语言检索．情报科学，2004（7）

31. 张琪玉．关于自然语言检索问题．图书馆论坛，2004（6）

32. 黄崑，赖茂生．Web 信息检索技术及研究进展．现代图书情报技术，2004（5）

33. 沈燕，任晓健．基于内容的多媒体检索技术在数字档案馆中的应用．情报杂志，2004（4）

34. 袁勇智．基于 Web Service 架构的 Z39.50．情报杂志，2004（1）

35. 熊回香．网络信息检索及其发展趋势研究．华中师范大学硕士学位论文，2003

36. 朱琳，杨梁彬．网络信息资源自动标引——面向中文网络文本信息的研究．北京大学校长基金论文集（2003）

37. 马费成，裴雷．网络信息资源的分布规律．情报科学，2003（11）

38. 雷怀光．因特网信息检索及其发展趋势探析．现代情报，2003（8）

39. 李玮，李利．Web 搜索引擎与全文检索技术．情报科学，2003（3）

40. 王苏海，颜惠，袁丽芬．Internet 上免费实用参考工具书网站导览．现代情报，2003（3）
41. 郑腾锐，范智军．网络搜索引擎的现状与使用．现代情报，2003（2）
42. 张颖．Internet 信息检索方法与技巧．辽宁商务职业学院学报，2003（1）
43. 李梅，王庆林．中文全文检索技术的研究及实现．情报学报，2003（1）
44. 乔东梅．搜索引擎现状与发展研究．郑州大学硕士学位论文，2002
45. 阁智．国外网络电子资源概况．现代情报，2002（9）
46. 仇恢．信息时代信息检索工具发展的新走向．情报科学，2002（8）
47. 王国红，孙平．网上中文参考工具的资源现状与分析．情报科学，2002（7）
48. 周宁，文燕平．检索结果的可视化研究．中国图书馆学报，2002（6）
49. 柏鹏英．传统信息组织方法在网络环境下的运用．情报资料工作，2002（5）
50. 刘银红，袁琳．网络信息检索标准化探析．情报理论与实践，2002（5）
51. 黄如花，张春蕾．网络信息检索的发展趋势．图书情报知识，2002（4）
52. 李冠南．网络信息检索工具及发展．图书馆工作与研究，2002 年增刊
53. 陈晋．网络数据库及图书馆数字化资源建设．福建商业高等专科学校学报，2002（4）
54. 张颖，贺亚锋．网络信息检索展望．现代图书情报技术，2002（3）
55. 殷占兵．论工具书（电子版）的新突破．高校图书馆工作，2002（2）
56. 刘霞．Internet 上参考工具书查询技巧．情报探索，2002（2）
57. 陈定权．Web 信息检索技术最新进展．现代图书情报技术，2002（2）
58. 何小清．数据库服务方式的发展趋势．情报学报，2002（2）
59. 黄小强．Internet 上电子科技信息的检索和获取——商业性网络数据库介绍．电子科技，2002（1）
60. 胡岷．传统联机检索系统与搜索引擎的比较．江西图书馆学刊，2002（1）
61. 孙延蘅．漫谈因特网中文搜索引擎．泰安师专学报，2002（1）
62. 何慎怡．使用中文工具书应注意的问题．深圳教育学院学报，2002（1）
63. 王凤英，姚艳芝．Internet 新一代信息检索研究——智能化、个性化、分布处理．情报杂志，2002（1）
64. 夏定元．多媒体网络中的图像搜索引擎技术．电子技术，2001（10）
65. 雷鸣，王建勇，赵江华，单松巍，陈葆珏．第三代搜索引擎与天网二期．北京大学学报，2001（9）
66. 杜文芝．网络搜索引擎的现状和发展趋势．现代情报，2001（5）
67. 张蓬花．工具书信息的传播、开发与利用．河南图书馆学刊，2001（5）
68. 刘白秋．国内网络数据库与高校图书馆网络资源利用．集美大学学报（哲学社会科学版），2001（4）
69. 张国强．关于工具书数字化发展趋势的几点思考．辞书研究，2001（3）
70. 庄云勇．试论联机检索的命运．情报探索，2001（3）
71. 李晓玲．Internet 信息检索探讨．重庆交通学院学报，2001（3）

72. 马明霞. Internet 网上最先进的信息检索工具 WWW. 现代情报，2001（3）
73. 赵荷晴，刘华英. 谈工具书的编排和使用问题. 高校图书馆工作，2001（2）
74. 夏立新. 商业联机检索系统与因特网信息资源之比较. 图书情报知识，2001（2）
75. 宋玲，马军. Internet 信息检索分析与研究. 信息检索技术，2001（1）
76. 沈固朝. 个性化的信息检索网上自导式教学. 大学图书馆学报，2001（1）
77. 王建仑，王淑凤. 网络数据库资源的收集与利用. 农业图书情报学刊，2000 年增刊
78. 王德英，周蓉. 网络数据库信息检索探讨. 现代图书情报技术，2000 年年刊
79. 曾福兴. 因特网信息资源搜索工具. 情报科学，2000（11）
80. 许磊. 论分类主题一体化的情报检索语言. 聊城师范学院学报（哲学社会科学版），2000（5）
81. 何华连. 工具书的鉴别与评价. 浙江师大学报（社会科学版），2000（4）
82. 董晓芬. 因特网是否将取代光盘. 图书情报工作，2000（4）
83. 叶新明. 对期刊论文全文数据库光盘的理性思考. 中国图书馆学报，2000（3）
84. 谢志耘. 光盘网络信息检索系统的发展趋势. 现代图书情报技术，2000（3）
85. 张永和. 关于联机检索和 Internet 网络检索的思考——由 DIALOG 新政策所想到的. 四川图书馆学报，2000（2）
86. 肖珑. 国外网络数据库的引进与使用. 现代图书情报技术，2000（2）
87. 冷红中，张联民. 中国学术期刊（光盘版）及其评价. 铁道师院学报，2000（1）
88. 张晓娟. 网络信息资源：概念、类型、特点. 图书情报工作，1999（2）
89. 安玉斌. 工具书浅议. 乌鲁木齐成人教育学院学报（综合版），1998（4）
90. 闫淑侠. 简谈工具书中附录的功能及特点. 图书馆建设，1998（4）
91. 都云程，卢献华. 中文搜索以年轻现状展望. 中文信息学报，1998（3）
92. Kingoff A. Comparing Internet Search Engines. Computer，1997，30（4）
93. 何华连. 我国中文工具书编纂出版分期概观. 浙江师大学报（社会科学版），1995（1）
94. 文图. 学知识的钥匙——工具书. 江西图书馆学刊，1994（4）

网址

1. http：//162. 105. 138. 230/
2. http：//211. 151. 93. 229/grid20/Navigator. aspx? ID=SNAD
3. http：//aad. archives. gov/aad
4. http：//adam. ac. uk
5. http：//baike. baidu. com/view/1154. htm
6. http：//cnis. gov. cn
7. http：//compass. net. edu. cn：8010
8. http：//course. cug. edu. cn/cugFirst/info _ structure/t3-1. html>

9. http：//csa. tsinghua. edu. cn
10. http：//dissertation. com
11. http：//dmoz. org
12. http：//e. pku. edu. cn
13. http：//ep. esp@cenet. com
14. http：//library. usts. edu. cn
15. http：//ntrs. nasa. gov/search. jsp
16. http：//opac. calis. edu. cn
17. http：//patents. uspto. gov/patft/index. html
18. http：//pctgazette. wipo. int
19. http：//search. sina. com. cn
20. http：//standards. co. nz
21. http：//standards. ieee. org
22. http：//wanfang. calis. edu. cn/kjxx/kjwx2. html
23. http：//webcrawler. com
24. http：//www. 9wh. net/Article/chuantong/ruxue/200512170317262582. html
25. http：//www. afnor. fr
26. http：//www. allexperts. com
27. http：//www. alltheweb. com
28. http：//www. altavista. com
29. http：//www. aol. com
30. http：//www. archives. sh. cn/default. htm
31. http：//www. ask. com
32. http：//www. baidu. com
33. http：//www. bjma. gov. cn/Default. ycs
34. http：//www. bjpopss. gov. cn/bjpssweb/n3181c48. aspx
35. http：//www. bsi. org. uk
36. http：//www. bzcbs. com
37. http：//www. bzcn. net/abc/abc/7/200432536. htm
38. http：//www. cei. gov. cn/
39. http：//www. chaxin. cn/sjkzy. html
40. http：//www. chinagb. org/homepage. htm
41. http：//www. chinajournal. net. cn/index. htm
42. http：//www. chinastandard. com. cn
43. http：//www. clearinghouse. net
44. http：//www. cnipr. com
45. http：//www. cnki. net

46. http：//www. cssinfo. com
47. http：//www. cssn. net. cn
48. http：//www. digiway. com/digisearch
49. http：//www. din. de
50. http：//www. dogpile. com
51. http：//www. european-patent-office. org/
52. http：//www. excite. com
53. http：//www. exin. net/patent/
54. http：//www. ez2www. com
55. http：//www. galaxy. com
56. http：//www. google. com
57. http：//www. gov. cn
58. http：//www. goyoyo. com
59. http：//www. gpoaccess. gov/index. html
60. http：//www. grokker. com
61. http：//www. highway61. com
62. http：//www. hotbot. com
63. http：//www. huicong. com
64. http：//www. iboogie. com
65. http：//www. iec. ch
66. http：//www. infoseek. com
67. http：//www. inktom. com
68. http：//www. isiknowledge. com
69. http：//www. iso. org/iso/home. htm
70. http：//www. ixquick. com
71. http：//www. jsipp. cn/pub/jsip/jypx/xgzs/zl/200707/t1075. htm
72. http：//www. kenjin. com
73. http：//www. lib. ruc. edu. cn
74. http：//www. lib. tsinghua. edu. cn/chinese/EI-village/switch. htm
75. http：//www. libnet. sh. cn/sztsg/fulltext/reports/1999/dc. htm>
76. http：//www. linefan. com/search
77. http：//www. looksmart. com
78. http：//www. lycos. com
79. http：//www. mamma. com
80. http：//www. mckinley. com
81. http：//www. meeting. edu. cn
82. http：//www. metacrawler. com

83. http：//www. nssn. org
84. http：//www. nstl. gov. cn/index. html
85. http：//www. ntis. gov
86. http：//www. oingo. com
87. http：//www. osti. gov/bridge
88. http：//www. paper. edu. cn/xxzy _ hyzy _ hywxjqjs. php
89. http：//www. patent. com. cn
90. http：//www. patents. ibm. com
91. http：//www. profusion. com
92. http：//www. sac. gov. cn/home. asp
93. http：//www. savvysearch. com
94. http：//www. scc. ca
95. http：//www. sdau. edu. cn/support/search
96. http：//www. searchenginewatch. com
97. http：//www. se-express. com
98. http：//www. sipo. gov. cn/sipo/zljs/default. htm
99. http：//www. sogou. com/dir
100. http：//www. sohu. com
101. http：//www. sosig. ac. uk
102. http：//www. sourceoecd. org
103. http：//www. spc. net. cn/daobao/daobao. asp
104. http：//www. stats. gov. cn
105. http：//www. std. cetin. net. cn
106. http：//www. stdcn. com
107. http：//www. un. org/Depts/dhl/
108. http：//www. unsystem. org
109. http：//www. uspto. gov
110. http：//www. vivisimo. com
111. http：//www. vlib. org
112. http：//www. wanfangdata. com. cn
113. http：//www. widewaysearch. com
114. http：//www. wssn. net
115. http：//www. yahoo. com
116. http：//www. yam. com
117. http：//www. zgbzw. com
118. http：//www. zhongsou. com
119. http：//wwwlib. global. umi. com/dissertations