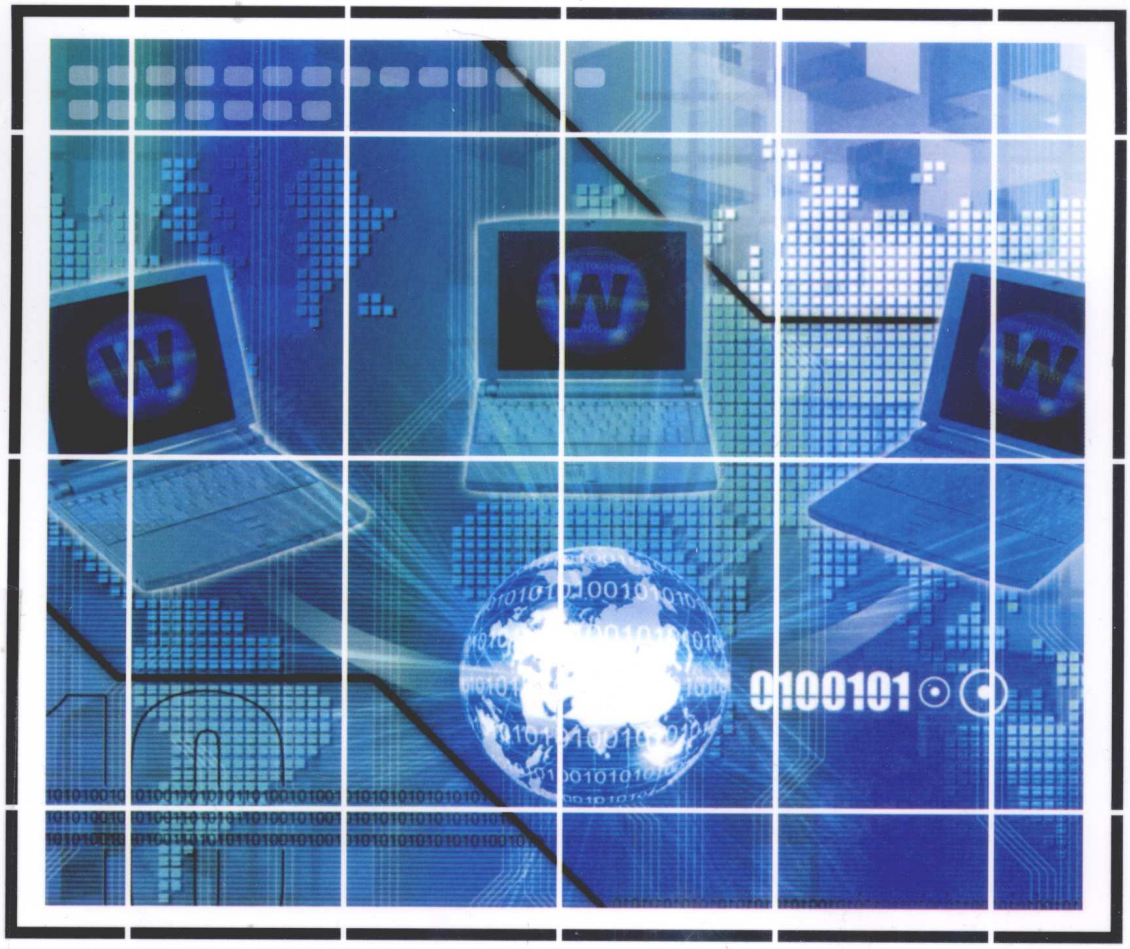
新世纪计算机类本科规划教材
COMPUTER

网络信息检索

董守斌　袁　华　编著



西安电子科技大学出版社
http://www.xduph.com

新世纪计算机类本科规划教材

网络信息检索

董守斌　袁　华　编著

西安电子科技大学出版社

2010

内容简介

本书详细介绍了网络信息检索的原理和技术，内容包括信息检索模型、网络信息的自动获取、网络信息预处理和索引、查询语言和查询优化等。针对网络信息检索的广泛应用，书中对搜索引擎、中文和跨语言信息检索、多媒体检索、并行和分布式信息检索、信息分类和聚类、信息提取与自动问答等重要应用的关键技术也进行了深入的探讨。

本书层次分明，深入浅出；既有原理阐述和理论推导，也有大量的实例分析，阐述力求系统性和科学性。本书可作为高等院校计算机科学与技术、信息管理与信息系统、电子商务等专业的高年级本科生或研究生的教科书和参考书，对广大从事网络信息检索、数字图书馆、信息管理、人工智能、Web 数据挖掘等研究和应用开发的科技人员也有较大的参考价值。

图书在版编目(CIP)数据

网络信息检索/董守斌，袁华编著. —西安：西安电子科技大学出版社，2010.4

新世纪计算机类本科规划教材

ISBN 978-7-5606-2378-8

Ⅰ. 网… Ⅱ. ① 董…② 袁… Ⅲ. 计算机网络—情报检索—高等学校—教材 Ⅳ. G354.4

中国版本图书馆 CIP 数据核字(2010)第 003394 号

策　　划　臧延新

责任编辑　樊新玲

出版发行　西安电子科技大学出版社(西安市太白南路 2 号)

电　　话　(029)88242885　88201467　　邮　　编　710071

网　　址　www.xduph.com　　电子邮箱　xdupfxb001@163.com

经　　销　新华书店

印刷单位　陕西华沐印刷科技有限责任公司

版　　次　2010 年 4 月第 1 版　2010 年 4 月第 1 次印刷

开　　本　787 毫米×1092 毫米　1/16　印　张　22.375

字　　数　529 千字

印　　数　1～3000 册

定　　价　32.00 元

ISBN 978-7-5606-2378-8/G・0035

XDUP 2670001-1

＊＊＊如有印装问题可调换＊＊＊

序

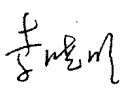
随着互联网上的信息越来越丰富，人们一方面越来越相信所需要的信息能够在网上找到，另一方面也常常要为花不少时间才能找到所需的信息而烦恼。于是，搜索引擎在我们工作和生活中扮演的角色越来越活跃，关心和研究如何从网络上有效获取信息的人也越来越多起来。“网络信息检索”一方面是亿万人每天都要进行的实践，另一方面也成为一个生机勃勃的研究领域。这从“全国搜索引擎与网络信息挖掘学术研讨会”近年投稿红火的情况可见一斑。同时，在教学方面，我国一些大学纷纷开设了相关的课程，多数在研究生层次。据我所知，华南理工大学是最早针对本科生开设这类课程的，本书作者即为其主讲教师，本书是她们几年来教学和科研实践的结晶。

读者会发现这本书是有用的。其有用性，在于相比起我国先前出版的几本与网络信息检索相关题材的书籍而言，内容是最丰富的。其内容在时空上的跨度之大令人兴奋，使得这本书不仅可以作为教材，还可以作为打算进入这个领域的研发人员的入门参考。例如第一章绪论，从网络信息的特点、信息检索的概念开始，对网络信息检索的基本含义进行了一个概要介绍，同时也概览了其发展的历史，列举了本领域知识与技术在多方面的应用。从中读者可以感受到网络信息检索既是由来已久，也是方兴未艾的一个重要领域。

读者会发现这本书是有特色的。其最大的特色，就是和我国已经出版的几本类似的书相比，这本书定位在教材，而且很好地体现了这种定位。本书的作者在网络信息检索领域工作多年，对内容的选取和篇章结构的安排颇有讲究。在介绍技术性内容的章节，除了后面有思考题、练习题之外，其中还包含有大量举例，对于教材来说，这是很有意义的。同时，我们还可以看到，作者不仅掌握了大量文献资料，而且在具体写作中融入了自己工作的体会，从而使得本书具有较强的感染力。例如，将信息检索的要义概括为“两个表示，一个比较”，就很有教益，值得读者仔细体会。

读者还会发现这本书是先进的。其先进性，体现在它用了将近一半的篇幅(从第 8 章开始)介绍前沿技术。它没有局限于尽管很重要但毕竟比较“单纯”的搜索引擎，而是在更大的空间上介绍网络信息检索技术，包括 P2P 网络信息检索、跨语言信息检索、多媒体信息检索、问答系统等，为读者展示了一个宽广的领域前沿。而第 7 章对于搜索引擎的专门介绍，篇幅虽不多，但以一种很实用的方式把握了搜索引擎系统的精髓。

网络信息的无比丰富，给信息检索领域带来了新的生命力和发展空间，也为计算机专业教学提供了新的内容。读过这本书，看到作者能够在多年科研实践和几轮教学实践之后及时将网络信息检索这样一个内容广泛的题材组织成一本教材，令人感到欣喜，我向她们表示祝贺。当然，网络信息检索领域依然处于动态性很强的发展阶段，面对日新月异(内容和表现形式)的海量网络信息，新的应用不断提出，它们带动新的技术不断发展，相信我们的作者能够与时俱进，不断推陈出新。



2010 年 1 月，于北京大学

前　言

信息检索是信息处理领域的重要基础。随着网络的快速发展和成熟，传统的信息检索技术在网络信息处理上得到了广泛的应用，并逐步形成了新兴的网络信息检索(Network Information Retrieval)技术。网络信息检索的典型应用如搜索引擎，已经成为互联网的重要基础应用。本书系统阐述了网络信息检索的基本原理、技术和应用。

本书共分两大部分，第一大部分由第1～7章组成，主要介绍网络信息检索技术的基本原理和搜索引擎的应用；第二大部分由第8～12章组成，主要介绍网络信息检索的核心技术和应用。具体来讲，各章的主要内容如下：

第1章介绍网络信息检索的发展历史，并对网络信息检索的应用进行了综述；第2章主要介绍信息检索的经典模型及多个扩展模型，大部分检索模型都附有实例，以加深读者的理解；第3章对网络信息搜集进行了详细的阐述，首先概述网络信息的载体和分布特点，在此基础上阐述网络信息搜集的基本流程，并详细讨论了搜集策略和性能等问题；第4章主要讲述网络信息处理和索引的关键技术，其中重点是网页去噪和倒排索引等；第5章介绍Web查询语言和查询方式，并重点论述了查询扩展和相关反馈等查询处理技术；第6章讲述评价检索性能的各种量化指标，以及主流的检索评价标准和方法等；第7章介绍搜索引擎的发展历史和工作原理，重点讲述经典链接分析算法，以及大规模搜索引擎中的体系结构和数据结构设计，并以开源系统为例，对搜索引擎的核心排序算法进行了剖析。

针对海量网络信息的处理问题，第8章主要介绍并行和分布式信息检索的原理以及应用；第9章讲述中文信息检索的关键技术，重点介绍了中文分词技术，以及中文信息检索模型和跨语言信息检索模型；第10章讲述基于内容的多媒体信息检索技术，重点介绍了基于内容的图像信息检索方法；第11章主要讲述自动分类和自动聚类这两种重要并且非常实用的技术，内容包括类的基本知识、特征提取、分类和聚类算法以及方法评测等；第12章介绍Web信息抽取技术以及自动问答系统的相关知识。

本书适合初学网络信息检索的读者，可作为专科生、本科生、研究生的网络信息检索及相关课程的教材。如果以本书作为本科生教材，相关内容的学时(讲授学时为32，实验学时为16)建议如下，教师可根据不同专业的要求进行调整。

第1章	网络信息检索概述	2学时
第2章	布尔模型、向量空间模型、概率模型	4学时
第2章	扩展模型	4学时
第3章	网络信息分布特点、网络信息搜集、搜索性能问题	4学时
第4章	文本特性、网页去噪、倒排索引	4学时
第5章	查询扩展、相关反馈	4学时
第6章	检索性能评价指标、评测标准和方法	4学时

第 7 章	搜索引擎工作原理、链接分析、排序算法	6 学时
第 8～12 章	高级应用课题	建议本科生自学
实验	设计一个搜索引擎原型系统	约 16 学时

本书也可供研究生教学使用。在用于研究生教学时，可适当减少讲授前 7 章的内容，增加后面 5 章的内容。

本书的编撰获得了广东省计算机网络重点实验室的龙卫江老师、方卫东老师、何克晶老师、张晶老师和李粤老师的大力支持，在此表示衷心的感谢；同时要感谢实验室的研究生曹鸿、张元丰、陈晓志、许洋洋、刘鹏飞、何剑飞、张倩、蔡智、胡俊刚、李嘉林、陈晓峰、农双、温泽逢、陈车前、张丽平、叶力洪等同学，感谢他们所做的资料搜集和整理等琐碎而辛苦的工作。

本书初稿成文于 2006 年 1 月，至今已数易其稿，并已在华南理工大学计算机学院本科和研究生教学中多次试用。感谢华南理工大学计算机学院修读“网络信息检索”以及“信息检索与 Web 挖掘”课程的同学们，他们在使用本书初稿作为辅助教材和讲义时提出了许多很好的建议和意见。

在编写本书的过程中，参考了大量的论文和网络资料，在此向这些参考文献的作者深表感谢。在表述中，本书尽量引用较为经典、规范的表述，并给出了详细的标注。如有缺失，敬请谅解。

网络信息检索的特点是系统性强，涉及面广，技术新并且发展非常快，而我们的学识和水平有限，因此书中难免出现疏漏和不足之处，敬请读者指正。

作者 2010 年 1 月于
广州华南理工大学北校区

目 录

第1章 绪 论

互联网(Internet)的出现和普及是人类文明发展历史上的一个重要的进步，它为人们存储、加工、传递和利用信息提供了一个有效的载体，并且突破了传统的信息载体在时间和空间上的限制，使得人类在传递和共享信息的效率上得到了前所未有的提高。

互联网在全球范围内的迅速发展与成熟，促使社会各领域信息飞速膨胀，为人们提供了丰富的信息源，但要在浩如烟海的信息中准确定位所需的信息却是一个极大的挑战。网络信息检索就是着力解决网络信息的组织和检索问题的一门新兴学科。

1.1 网络信息检索概述

1.1.1 网络信息

网络信息是指通过互联网可以利用的各种信息资源的总和。随着互联网的迅速发展，网络信息作为一种新型的信息资源，发挥着越来越重要的作用。与传统的非网络信息资源相比，网络环境下的信息资源具有以下几个方面的特点：

(1) 网络信息内容丰富。互联网已经成为全球最大的信息资源基地，同时其信息资源的增长十分迅速。在互联网上几乎可以获得任何领域的信息，其内容涉及政治、经济、文化、科学和娱乐等各个方面，涵盖社会科学、自然科学、人文科学和工程技术等各个领域。

(2) 网络信息变化频繁。在互联网上，信息地址、信息链接和信息内容经常处于变动之中，信息资源的更换和消亡更是无法预测。因而，网络信息时时刻刻处在变化和发展之中。

(3) 网络信息结构复杂。互联网对网络信息资源本身的组织管理尚未形成完全统一的标准和规范，网络信息呈全球化分布结构，信息资源物理地存储在世界不同地区各种不同类型的服务器上。因此，在信息的组织和检索方面比较复杂。

(4) 网络信息格式多样。网络信息的媒体形式多种多样，包括文本、图形、图像、声音和视频等，各种类型的媒体信息都有多种不同的信息描述格式，例如文字信息的格式有HTML、TXT、PDF、DOC等格式；图像信息的格式有BMP、GIF、JPG等格式，因此网络信息格式呈现多样化。

(5) 网络信息价值差异。由于网络信息的发布具有很大的自由度和随意性，且缺乏必要的质量控制和管理机制，因而，网络信息资源的价值差异较大，既有较大参考价值的有用信息，也有毫无用处的垃圾信息，甚至还有不少有害的信息，可谓良莠不齐。因此，如何评价、选择和过滤信息成为网络信息组织和检索的重要任务。

1.1.2 信息检索

信息检索(Information Retrieval，IR)泛指用户从包含各种信息的文档集合中查找所需要的信息或知识的过程。信息检索将信息按一定的方式组织和存储起来，再根据用户的需求查找所需信息，并返给用户。信息检索包括信息的存储、组织、表现、查询、存取等各个方面，一般而言，主要包括以下三个环节：

(1) 处理搜集：对信息内容进行分析与编码，产生信息记录及检索标识；

(2) 组织存储：将全部记录按文件、数据库等形式组成有序的信息集合；

(3) 检索服务：对用户提问进行处理和输出相应的检索结果。

信息检索的关键部分是信息提问与信息集合的匹配和选择，即对给定提问与集合中的记录进行相似性比较，根据一定的匹配标准选出有关信息。

信息检索最初应用于图书馆和科技信息机构，后来逐渐扩大到其他领域，与信息检索有关的理论、技术和服务构成了一个相对独立的知识领域，是信息学和计算机科学的交叉学科，这里引用1997年Kowalski对信息检索系统的定义[1]："信息检索系统是对信息的存储、检索和维护，信息可以是文本、图像、音频、视频或其他多媒体对象"。

信息检索系统一般由信息收集、处理、索引、存储、检索等部分组成，信息检索结构可以用图1-1表示。从图1-1中可以总结出"两个表示，一个比较"来概括信息检索的精髓，所谓"两个表示"就是通过预处理和特征提取，把信息和查询分别表示为一定的数学形式，如向量；"一个比较"是把这两个数学表示进行相似性比较，以判定某信息是否可以作为该查询的结果进行输出。

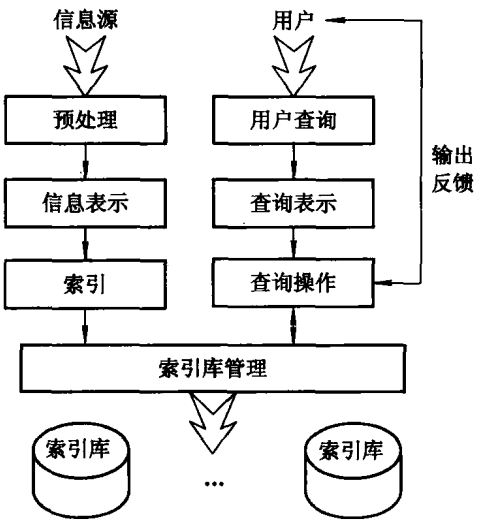


图1-1 信息检索系统结构示意图

这里需要区分两个概念：信息检索和数据检索。数据(data)和信息(information)是两个完全不同的概念，数据是对客观事物的数量、属性、位置及其相互关系的抽象表示，以适合于用人工或自然的方式进行保存、传递和处理。而信息是指经过系统组织、整理和分析的数据。例如表1-1中的"80亿+"是一个数据，"被Google索引的页面为80多亿"则是信息。

表 1-1 与 Google 相关的数据和信息

数 据	信 息
80 亿+	被 Google 索引的页面为 80 多亿
10 亿+	被 Google 索引的图片为 10 多亿
10 亿+	被 Google 索引的 Usenet 信息为 10 多亿
100 多种	Google 界面的可用语言为 100 多种
35 种	Google 搜索结果所采用的语言约 35 种

数据可以很容易地被组织和存储，对数据的检索相对容易，也容易做到准确地检索；数据检索如一般的数据库检索，处理的是结构化数据；数据检索的条件一般具有清晰的定义，要求取得满足特定条件的所有对象，因此它的准确率可以达到百分之百；数据检索效率的评价标准一般是响应时间或存储空间等方面的开销。而信息检索一般是从非结构化或半结构化的文档集中找出与用户需求相关的信息，包括新闻、科技论文等文本数据，HTML 和XML 等网页，图像、图形、视频和音频等多媒体数据。信息检索的条件描述本身就是一个难题，一般很难做到完全准确，而用户的需求描述也可能是不准确的。造成这一点的主要原因是，信息检索通常是对自然语言进行处理，而自然语言本身没有很好的结构，语义上也存在模糊性。因此，信息检索的评价也更难，一般使用检索精度(Precision)和召回率(Recall)等评价标准来衡量信息检索的效果。

1.1.3 网络信息检索

网络信息检索是指能够通过网络接受用户的查询指令，并向用户提供符合其查询要求的网络信息资源的过程。可以把网络信息检索理解为检索对象为网络信息的信息检索。网络信息检索系统的结构示意图如图 1-2 所示。

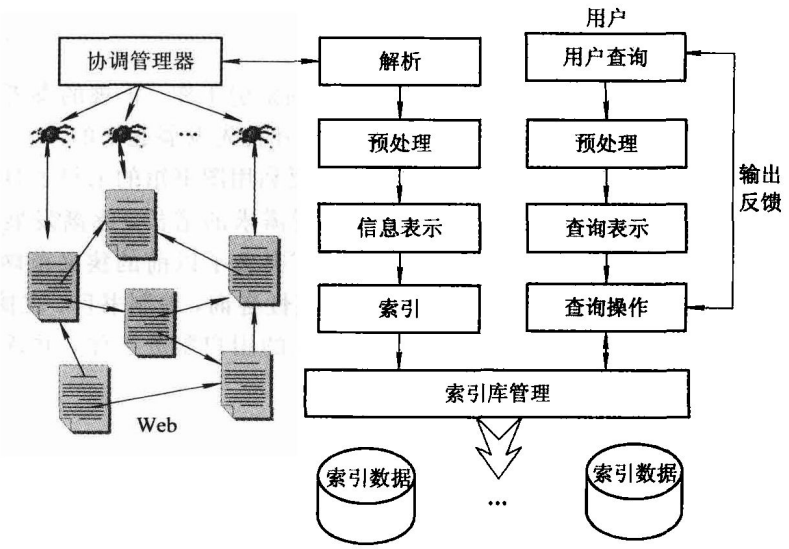


图 1-2 网络信息检索系统结构示意图

可见，网络信息检索系统与传统意义的信息检索系统在总体结构上大致相同，所不同的只是信息的来源不一样。传统信息检索系统的来源一般是图书、事先录入的信息等，而网络信息检索的信息来源于互联网，大都是 Web 页面、文件、图像和音视频媒体等。

1.2 信息检索的发展

信息检索起源于图书馆的参考咨询和文摘索引，从 19 世纪下半叶首先开始发展。当时，信息存储和传播主要以纸质为载体，信息检索活动也围绕着文献的获取和控制展开。至 20 世纪 40 年代，索引和检索已成为图书馆独立的工具和用户服务项目，“文献检索”(Document Retrieval)一度成为信息检索的同义词。随着 1946 年世界上第一台电子计算机的问世，计算机技术逐步走进信息检索领域，并与信息检索理论紧密结合起来；人们开始使用“情报检索”这个概念，脱机批量情报检索系统、联机实时情报检索系统相继研制成功并商业化，当时的信息检索，是更接近于数据库检索的一种形式。20 世纪 60 年代到 80 年代，在信息处理技术、通信技术、计算机和数据库技术发展的推动下，随着信息载体类型的多元化以及传播手段的改进，情报检索和文献检索逐渐归于信息检索这一具有兼容性的概念，研究范围也日趋扩展，信息检索在教育、军事和商业等各领域高速发展，并得到了广泛的应用。

目前，信息检索已经发展到网络化和智能化的阶段。信息检索的对象从相对封闭、稳定一致、由独立数据库集中管理的数据信息扩展到开放、动态、更新更快、分布广泛、管理松散的网络信息；信息检索的用户也由原来的情报专业人员扩展到包括商务人员、管理人员、教师学生、各专业人士等在内的普通大众，他们对信息检索从方式到结果提出了更高、更多样化的要求。适应网络化、智能化以及个性化的需要是目前信息检索技术发展的新趋势。

具体来说，信息检索经历了从手工检索、计算机检索到网络信息检索的发展过程。

1.2.1 手工检索

信息检索直接发源于图书馆的参考咨询工作和文摘索引工作。正规的参考咨询工作是由美国的公共图书馆和大专院校图书馆于 19 世纪下半叶首先发展起来的。

20 世纪初，多数图书馆成立了参考咨询部门，主要利用图书馆的书目工具来帮助读者查找图书、期刊或现成的答案。随着文献的激增和读者需求的增长，逐渐发展到从多种文献源中查找、分析、评价和重新组织情报资料，“索引”突破了以前的狭义范畴，成为独立的检索工具。到 20 世纪 40 年代又进一步包括回答事实性咨询，编制书目、文摘，进行专题文献检索，提供文献代译等。“检索”从此成为一种独立的用户服务工作，并逐渐从单纯的经验工作向科学化方向发展。

1.2.2 脱机批处理检索

1946 年世界上第一台电子计算机问世之后，就有人开始研究计算机在信息检索领域的应用。20 世纪 50 年代中期至 60 年代后期是信息检索的脱机批处理阶段。当时计算机还没有连接成网络，也没有远程终端装置，不能提供实时检索，只能进行现刊文献的定题检

索(Selective Dissemination of Information)和回溯性检索(Retrospective Search)，同时利用计算机编辑出版检索性刊物。1954 年，美国海军机械试验中心(Naval Ordnance Test Station，NOTS)使用 IBM 701 型机，初步建成了计算机情报检索系统，这标志着以计算机检索系统为代表的信息检索自动化时代的到来[2]。

在这个时期，信息检索系统面向小型的科学文摘数据库、法律和商业文档，检索模型为基本的布尔模型和向量空间模型，提出向量空间模型[3]并付诸实践的康奈尔大学(Cornell University)的 Salton 教授和他的学生成为这个领域的先驱。

1.2.3 联机检索

1967 年，美国系统发展公司(System Development Company，SDC)研制成功 ORBIT(On-line Retrieval of Bibliographic Information Time shared)联机情报检索软件，开始了联机情报检索阶段[4]；与此同时，美国洛克希德公司成功研制了国际联机情报检索系统 Dialog(http：//www. dialogweb. com)。20 世纪 70 年代卫星通信技术、微机计算机技术以及数据库技术的同步发展，使得用户得以冲破时间和空间的障碍，实现了国际联机检索。远程实时检索多种数据库是联机检索的主要特点。计算机检索技术从脱机阶段进入联机信息检索时期。联机检索是计算机技术、信息处理技术和现代通信技术三者的有机结合。

图 1－3 所示的是美国国家医学图书馆的 MEDLINE 系统(http：//www. ncbi. nlm. nih. gov/PubMed/)，Dialog 系统作为这一时期的信息检索领域的代表，至今仍是世界上最著名的信息检索系统之一。

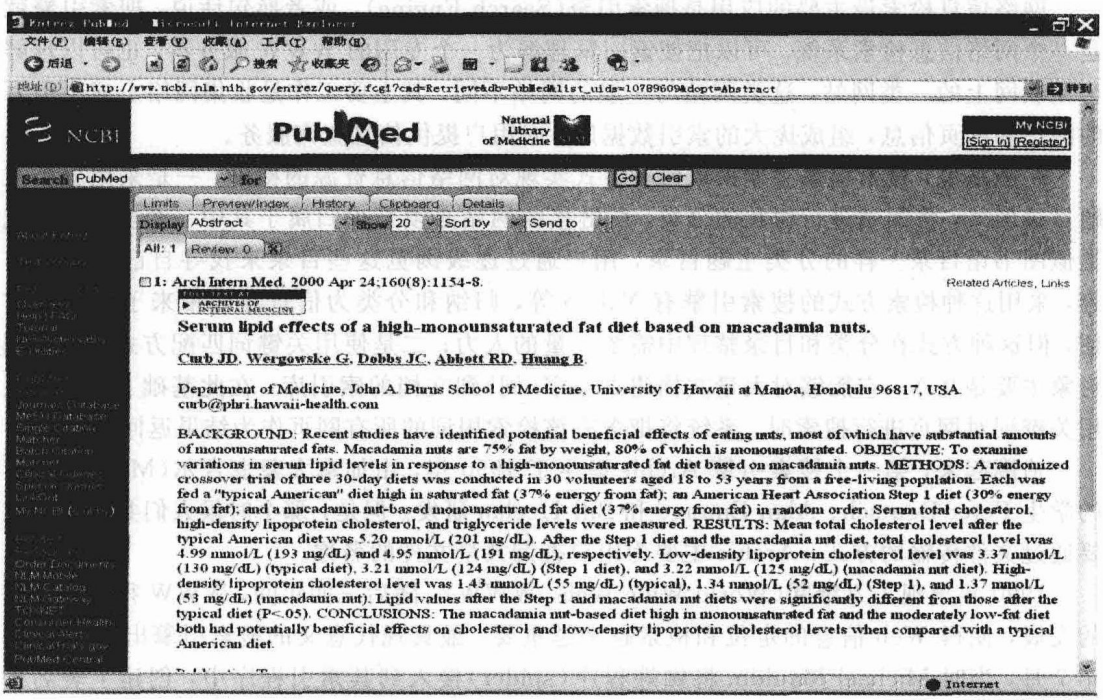


图 1－3 医学信息检索系统 MEDLINE

1.2.4 网络信息检索

互联网在二十世纪六七十年代初见雏形，八十年代末九十年代初迅速流行。此时，单纯的手工检索和机械检索都显现出各自或多或少的缺点，因此，极有必要发展一种新型的信息检索方式，网络信息检索应运而生。网络信息检索系统几乎包括了计算机在信息检索领域表现出来的全部优点，它是联机检索的高级阶段，使人们可以在很短的时间里查找到分布在全球各个角落的信息。网络信息环境的出现，使得信息检索研究的对象和范围不断扩大，研究队伍也突破了原有的以图书情报领域的专家学者为主的模式，众多的科研机构以及商业公司加入到研究信息检索技术的行列。可以说，网络使计算机信息检索技术进入了一个崭新的发展阶段，而网络信息检索又使网络信息的利用率提高，信息的组织更加有序和高效。

1.3 网络信息检索的应用

随着网络技术和信息检索技术的发展，网络信息检索得到了广泛的应用。除了人们最常用的搜索引擎外，目前流行的还有多媒体信息检索、跨语言信息检索、主题识别和跟踪、信息过滤、问题回答和 Web 数据挖掘等。

1.3.1 搜索引擎

网络信息检索最主要的应用是搜索引擎(Search Engine)，或者换句话说，搜索引擎就是一个网络信息检索系统。可以把搜索引擎理解为一个专用的 WWW 服务器，也可以理解为互联网上的一类网站，这类网站与一般的网站不同，其主要工作是收集网络上成千上万的网站和网页信息，组成庞大的索引数据库，向用户提供信息查询服务。

一般来说，搜索引擎主要采取两种方式实现对网络信息资源的检索，一是采用分类主题目录形式，将网站进行树状的分类，所链接的网站必须至少归属于其中一个类别，形成类似图书馆目录一样的分类主题目录，用户通过逐级浏览这些目录来找寻自己需要的内容，采用这种检索方式的搜索引擎有 Yahoo 等，归纳和分类为信息导航带来了极大的方便，但这种方式在分类和目录整理中需要大量的人力；二是使用关键词匹配方式，其处理对象主要是文本，它能够对大量文档建立由字(词)到文档的索引库，在此基础上，用户使用关键词对网页进行搜索时，系统将把含有该检索用词的所有网页作为结果返回给用户。

追溯起来，第一个搜索引擎 Archie 诞生于 1990 年，由加拿大蒙麦吉尔(McGill)大学的学生 Peter Deutsch、Alan Emtage 和 Bill Heelan 研发。但是，那个时候人们共享数据主要通过文件传输的方式，Archie 主要为用户查询共享文件的名称。

1990 年出现了万维网(World Wide Web，WWW)，随后三四年间，WWW 得到了飞速的发展，使得 Web 信息的定位和检索越来越重要。最具现代意义的搜索引擎出现于 1994 年 7 月，当时 Michael Mauldin 将蜘蛛程序(spider)接入到其索引程序中，创建了著名的 Lycos(www.lycos.com)。Lycos 第一次面向公众开放的时候拥有 5.4 万个文档，主要提供排序的相关检索，受到了用户的广泛认可。到 1995 年 1 月，Lycos 索引的文档数达到 150 万个，1996 年达 6000 万个，比当时其他任何搜索引擎能够提供检索的文档都多。

1994年还发布了很多著名的搜索引擎，如1994年4月，斯坦福(Stanford)大学的两名博士生David Filo和美籍华人杨致远共同创办了超级目录索引Yahoo(www.yahoo.com)，并成功地使搜索引擎的概念深入人心，从此搜索引擎进入了高速发展时期。Infoseek(www.infoseek.com)和AltaVista(www.altavista.com)也诞生于1994年。之后还陆续出现了Looksmart(www.looksmart.com)、Inkotomi(www.inkotomi.com)、AskJeeves(www.askjeeves.com)等著名搜索引擎。

1998年，最具影响力的搜索引擎Google(www.google.com)发布，Google是由斯坦福大学两位博士生瑟盖·布尔(Sergey Brin)和拉里·佩奇(Larry Page)研发的。Google的名字从英文"googol"演变而来，表示10^{100}，代表海量的信息。Google在PageRank技术、动态摘要、网页快照、多文档格式支持、图像搜索、多语言支持、用户界面等方面进行了创新，可支持多种语言，索引页面多，检索面广，搜索信息准确。同年发布的还有微软的MSN(www.msn.com)。1999年北大校友李彦宏和徐勇创办中文搜索引擎百度(www.baidu.com)，专注于中文搜索，收录了大部分的中文网页，更新速度快，有中文搜索的自动纠错和自动提示功能，更符合中国人的使用习惯。

表1-2显示了截至2005年1月世界最大搜索引擎的比较数据[5]，当时的全部网页估计在115亿，可索引网页为94亿。

表1-2 著名搜索引擎之间的一些比较数据

搜索引擎名称	各自报告的网页数量/亿	估计搜集到的网页数量/亿	在索引网页中的覆盖比例/(%)	在整个网页中的覆盖率/(%)
Google	81	80	85.1	69.6
Yahoo	42(估计)	66	70.2	57.4
AskJeeves	25	53	56.4	46.1
MSN	50	51	54.3	44.3

搜索引擎把传统的信息检索技术应用到网络信息检索，是典型的网络信息检索系统。目前，搜索引擎已成为人们找寻网络信息的一条主要渠道。据中国互联网络中心(CNNIC)的互联网统计报告[6]，通过搜索引擎获取相关信息的用户占58.2%，直接访问已知网站的占35.7%，其他还有随意浏览、广告、相关链接等方式，共约占6.1%。可见，搜索引擎已经成为信息查询和获取的主要手段。2010年中国互联网络中心(CNNIC)第25次互联网统计报告[7]称，目前中国3.84亿网民中使用搜索引擎的比例是73.3%，即已有近3亿人从搜索引擎获益。与其他国家相比，由于中国互联网仍旧是娱乐功能占主体，总体网民的搜索引擎使用率偏低。在美国，搜索引擎使用率已经超过90%。搜索引擎应用人群的特点决定了它在互联网领域的高商业价值。在中国这样一个网民快速增长和以年轻网民为主的国家，搜索引擎用户将会继续增长。

然而，随着搜索引擎数量的迅速增加，如何准确选择搜索引擎，有效地利用多个搜索引擎的集成资源与检索能力成为重要问题。元搜索引擎(Meta Search Engine)就是一种集成化的检索系统，通过多个成员搜索引擎提供的服务向用户提供统一的检索服务。元搜索引擎的主要目的是综合各种搜索引擎的长处，尽量减少用户的检索过程，提高检索效率。

由于元搜索引擎的结果集通常十分庞大，方便用户快速地找到需要的信息就成为一个十分关键的问题。虽然通过改进页面排序算法，可以尽量使“重要”的页面出现在返回结果的前面，但由于用户职业、兴趣、年龄等各方面的差异，很难让所有的用户都接受系统给出的重要性顺序。另外，统计显示，用户一般不会在结果集中向后翻超过五页。所以，将查询结果以一定的类别层次进行组织，让用户能方便地选择查看类别，可以很好地缩小结果集，从而使用户能更快地找到有用的信息。

图 1－4 所示的 Clusty(www. clusty. com)是美国 Vivisimo 公司开发的一个具有对搜索结果自动进行聚类的元搜索引擎，它能非常快速地将不同类型的网站进行聚类整理并按类别呈现结果。Clusty 在搜索结果页面左侧增加了一个搜索分类目录栏。该目录的作用就是对右侧窗口中的所有搜索结果进行聚类，同时也显示此次搜索结果的总数目。分类目录无需预先定义，是由搜索结果决定的。

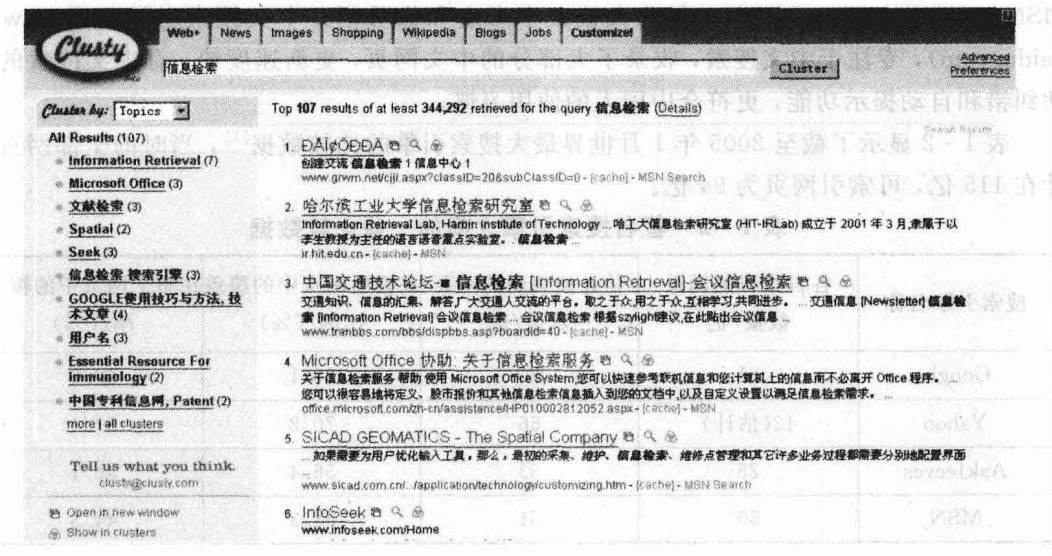


图 1－4　Clusty 的用户界面

1.3.2　多媒体信息检索

传统信息检索技术主要是面向文本的，今天广泛使用的 Google、Yahoo 和百度等搜索引擎主要采用文本检索技术，通常是利用一组关键字或词组成的查询项来搜索定位文本数据库中的相关文本文档，如果某个文档中包含较多查询项，那么就认为此文档比其他包含较少查询项的文档更相关，搜索系统将按照这种相关程度对查询结果进行排序，并依次展现给用户，以便用户浏览和进一步查找。

对图像和视频等多媒体信息集来说，目前，绝大多数检索系统仍采用文本搜索技术，例如 Google 的图像和视频检索功能仍是基于文本关键词的，如图 1－5 所示，这些关键词可能来源于图片周围的文本、文件名等，也可能来源于人工或自动标注(annotation)。然而，对于图像和视频等多媒体信息，一般难以用自然语言进行有效、精确的描述，无法表达其实质内容和语义关系，所以这种依据文本信息检索图片和视频的解决方案很难完全满足人们的查询需要。