

化学信息学导论

繆强 编

高等教育出版社

北京 2001.6

内容提要

本书是教育部“ 高等教育面向 21 世纪教学内容和课程体系改革计划 ”的研究成果,是面向 21 世纪课程教材。全书共分三部分。第一部分是化学、化工信息资源的检索与利用;第二部分介绍化学数据信息的处理方法;第三部分是化学知识的计算机推演和计算机实验方法。

本书可作为高等院校化学化工专业的“ 化学信息学 ”课程的入门教材。书中附有一些软件的应用说明、示例、重要的科学与技术信息源和化学、化工网站地址,因而也可以作为相关领域科研工作人员的参考书和应用工具书。

图书在版编目(CIP)数据

化学信息学导论/ 缪强编 .—北京:高等教育出版社,2001.6

ISBN 7 - 04 - 009331 - 6

.化... . 缪... .化学 - 情报检索 - 高等学校 - 教材 . G252 .7

中国版本图书馆 CIP 数据核字(2001)第 25837 号

化学信息学导论

缪强

出版发行	高等教育出版社				
社 址	北京市东城区沙滩后街 55 号		邮政编码	100009	
电 话	010 - 64054588		传 真	010 - 64014048	
网 址	http:// www hep .edu .cn http:// www hep .com .cn				
经 销	新华书店北京发行所				
印 刷					
开 本	787×960 1/ 16		版 次	年 月第	版
印 张	14		印 次	年 月第	次印刷
字 数	250 000		定 价	12.30 元	

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换。

版权所有 侵权必究

责任编辑	周传红
封面设计	张楠
责任绘图	陈钧元
版式设计	周顺银
责任校对	朱惠芳
责任印制	

前 言

化学是一门古老的科学,在漫长的发展进程中,积累了大量的文献。随着以计算机与网络为标志的现代信息技术的迅速发展,化学信息的数量日趋增加,形式也呈现多样化。Internet 是当今世界上最大的信息传播媒介,它把世界的各个大学、机构和研究室联系在一起,把学习者、教育者、研究人员联系在一起,共享各种信息资源。World Wide Web 极大地影响着化学信息学的形成和以后的发展。在线文献检索、在线数据库检索、在线交流、网络远程教学和远程计算都已经成为现实。

化学信息学(Chemical Informatics)是以化学、化工、计算机与信息技术为基础的一门新兴边缘学科,包括:化学、化工文献学;化学知识体系的计算机表示、管理与网络传输;化学图形学;化学信息的解析与数据挖掘处理;化学知识的计算机推演;化学教育与教学的现代技术与远程信息资源。近年来,化学信息科学在许多方面都取得了很大的进展:(1) 化学结构的计算机编码和图形检索方法的发展使传统的文献信息产品实现电子化、网络化;(2) 化学计量学中的多元统计分析、人工神经网络、遗传算法、小波分析等先进的数据挖掘技术在化学的许多领域都有成功的应用;(3) 根据已有知识和定理通过计算机推演、大规模计算预测与模拟可以得到物质的物理和化学性质,化工仿真也取得了很大成功;(4) 可视化和虚拟技术影响着化学、化工信息产品的发展,其中有些软件在Internet上可以免费得到。

信息技术是现代社会生活的重要支柱。培养大学生的信息意识和获取、利用信息的能力,在世界各国的高等教育中都占有重要的地位。美国化学会化学信息分会和图书馆协会在 Indiana 大学成立了化学信息教学资料交换中心 CCI-IM(the Clearinghouse for Chemical Information Instructional Materials),负责收集、发布和分发化学信息源。美国、欧洲、澳大利亚的很多大学开设了化学信息学类的课程,美国 Indiana 大学已经有超过 40 人获得了化学信息学硕士学位。我国教育部理科化学教学指导委员会也把“化学信息学”列入了 1998 年 10 月制定的化学专业和应用化学专业化学教学基本内容中。近年来,我国不少高校陆续开设了文献检索与利用、计算机化学,受到学生们的普遍欢迎,取得了良好的效果。但是,在 Internet“信息高速公路”成为最大的信息媒体的今天,这些课程对化学信息学整体的介绍,尤其是在网络应用方面的介绍还远远不够。为了从整体上提高我国大学生从计算机和网络上获取和利用化学信息的能力,我们编

写了《化学信息学导论》。书中内容已在近几届本科生和研究生中进行了试讲,取得了较好效果。

本书是大学化学信息学课程的入门教材,强调实践并侧重于计算机与网络的应用,许多高级课题并不包含在内,例如化学反应的计算机表示、化学数据库、化学结构的匹配与检索技术、化学图形学、人工智能与化学化工专家系统等等。介绍这些专门课题需要更多的背景知识,不宜作为基础课教材内容。本书主要介绍了化学、化工及其相关领域中获取网络信息的方法,学科信息资源,化学信息产品与利用,数据信息处理,以及通过计算机推演获取化学化工信息的方法。书中选择介绍了部分常见和易于普及的专业软件如 Origin、ChemWindow、WebLab viewer、HyperChem、PrSim 等。有些软件,如 ChemOffice,虽然非常优秀且在国外也用得很多,但由于价格偏高,在国内暂不易普及,故书中没有详细介绍。

全书大致可分为三个部分,共十章。书中各章相对独立自成系统。第一部分从第一至第四章为化学、化工信息资源的检索与利用,其中除了介绍美国《化学文摘》(CA)及其电子版(CA on CD)、科学引文索引(SCI)网络版和在线专利检索以外,还有化学化工网站、网上数据库和部分网上化学资源等等内容。第二部分讨论化学数据信息的处理方法,在第五、第六章中介绍了常用的统计分析和谱图解析的方法和应用软件,第七章是化学中的多元分析简介。第三部分主要是化学知识的计算机推演和计算机实验方法,是非传统实验室化学信息的获取方法,内容有化学结构的计算机表示与图形显示、计算化学方法、构效关系分析与结构信息的利用、分子与材料设计简介、化工计算与仿真等。

本书具有普及性和实用性。适用于我国高等院校化学、化工及相关专业理工科本科的“化学信息学”入门课程的教材。通过本书的学习,可以掌握从网上获取化学、化工信息及其有关信息的处理方法。书中附有一些软件的应用说明、示例、重要的科学与技术信息源和化学、化工网站地址,因而也可以作为相关领域科研工作人员的参考书和应用工具书。

在本书的编写过程中,得到了忻新泉教授的关心和指导,得到了教育部理科化学教学指导委员会的热情鼓励,得到了高等教育出版社和南京大学的大力支持,本书的完成还得益于导师陈念贻教授在作者步入这个领域初期时所给予的诸多教诲,得益于多年教学实践中与教研室同事、历届同学和研究生的讨论,得益于全国第一次化学信息学研讨会上与师长和同行们的探讨。历年来张海燕高级工程师、屠庆云高级工程师和白志平副教授给予了作者很多帮助。对曾经帮助过作者的各位师长、先生和女士们致以由衷的感谢。对本书利用参考的专著、论文、资料的作者们致以深深的谢意。

化学信息学是一门新兴的并在快速发展中的学科,对学科的定义和内涵还

有待讨论与统一。由于编者的学识所限,难免有疏漏与不当之处,为此我们欢迎各位化学信息学的研究者、老师和同学们提出宝贵意见。

作 者

2001 年 2 月

于南京大学

目 录

第一章 化学信息学与 Internet	(1)
第一节 化学信息的构成及化学信息学	(1)
第二节 因特网(Internet)概述	(4)
第二章 化学化工网站与网上数据库	(5)
第一节 通用信息检索引擎	(6)
第二节 重要的化学化工综合网站.....	(10)
第三节 网上化学结构数据库、反应数据库和物性数据库	(16)
第四节 网上化学化工分类信息与信息资源.....	(20)
附 1. IUPAC 专业分会与组织	(29)
附 2. 美国化学会(http:// www .acs .org)	(31)
附 3. 网上化学信息资源	(32)
第三章 化学化工文献检索	(40)
第一节 文献概述.....	(40)
第二节 美国《化学文摘》(印刷出版物).....	(41)
第三节 CA 索引和手工检索方法	(45)
第四节 CAS 的电子出版物	(49)
第五节 CA on CD 的使用方法	(50)
第六节 美国科学信息研究所 ISI 和《科学引文索引》SCI 简介	(55)
第七节 其他文献数据库.....	(62)
附 1. CA 主题分类	(63)
第四章 网上专利检索	(67)
第一节 中国专利文摘数据库.....	(67)
第二节 北京市经济信息中心的中国专利页.....	(68)
第三节 美国专利数据库.....	(71)
第四节 其他专利数据库与知识产权保护组织.....	(74)
第五章 化学实验数据统计与谱分析	(77)
第一节 实验数据的统计分布.....	(77)
第二节 分布的数字特征计算.....	(80)
第三节 化学分析的采样.....	(82)
第四节 比较实验.....	(84)
第五节 回归分析.....	(87)
第六节 数据的最小二乘平滑处理.....	(93)

第七节	曲线拟合和谱峰分辨.....	(95)
第八节	谱的微分与积分.....	(96)
第六章	科学数据处理与绘图工具软件 Origin	(98)
第一节	Origin 的窗口组成	(98)
第二节	数据的输入、编辑和保存	(99)
第三节	数据的统计处理	(102)
第四节	数据关系的图形表示	(106)
第五节	谱与曲线的处理	(111)
第七章	化学中的多元分析方法	(115)
第一节	化学与多元分析	(115)
第二节	相关分析与变量分组	(121)
第三节	降维、信息压缩与显示技术.....	(125)
第四节	分类与判别	(128)
第五节	聚类分析	(132)
第六节	因子分析	(133)
第七节	其他的信息挖掘技术	(137)
第八章	化学结构的表示与可视化.....	(142)
第一节	分子拓扑结构的矩阵表示与矩阵指数	(143)
第二节	分子拓扑结构的线性标记方法	(145)
第三节	有机分子结构式绘图软件 ChemWindow	(151)
第四节	分子几何结构的表示	(154)
第五节	三维分子结构显示软件 WebLab ViewerPro	(158)
第九章	分子结构信息与分子设计	(164)
第一节	分子及其聚集体的计算与模拟方法评述	(164)
第二节	定量构效关系方法	(171)
第三节	分子设计中的重要参数	(177)
第四节	分子计算与模拟软件举例	(181)
第五节	分子建模与分析软件 HyperChem 功能简介	(182)
第六节	分子与材料设计简介	(186)
第十章	化工过程计算与仿真	(193)
第一节	物性估算	(194)
第二节	化工仿真软件 PrSim 简介	(205)

第一章 化学信息学与 Internet

第一节 化学信息的构成及化学信息学

1. 什么是化学信息

什么是信息？信息是“含义”，是一种“代码”，它被用来精确地描述物体或事件(以消除不确定性)。信息是独立于物质和能量以外的自然的第三属性。例如，生物基因包含了物种遗传所需的生物信息，语言是人们用来表达思想、情感和传递知识的人类信息。计算机和网络通讯中的二进制编码是现代信息技术的基础。当今世界是信息社会，信息已经成为与材料、能源同等甚至更为重要的资源，成为促进和发展科学、教育、政治、经济的基础。

化学信息包括化学化工及其相关领域的科学研究、化学工业生产技术、投资与商务活动、教育与教学过程中所涉及到的化学结构的信息、化学与化工反应信息、图谱与物理化学性质等数值信息、化学文献信息、专利、化学化工商务信息、远程教学、远程会议与在线交流、网络上的其他信息等等。更为广义的看法认为化学信息应该包括化学过程中分子间的相互适应、相互识别，例如在超分子的化学形成过程和分子自组装过程中所表达的分子信息。化学信息学是化学、化工与信息科学、计算科学的交叉学科，它涉及化学化工信息的获取、管理、处理与控制、计算推演与模拟和图形表示的技术与方法。学习化学信息学的目的就是要培养学生的信息意识和提高获取信息的能力。

化学是一门古老的科学。在很长的发展历史中，积累了大量的实验事实、数据与文献。在中世纪，炼丹术的秘方是由师傅通过口授方式传递给他的徒弟的；15 世纪以后，化学信息开始以教科书的方式传递；从 19 世纪开始，科学技术的飞速发展引起信息的大爆炸；近 50 年来，新的实验方法、实验技术所产生的信息量极其浩繁，化学文献更是以每年数倍的速度迅速增加。近代科学的显著特点之一，是学科之间的相互渗透。化学家为了进行科学研究，所需要的信息范围甚广，所涉及的学科范围很宽。随着以计算机与网络为标志的现代信息技术的迅速发展，化学信息的数量日趋增加，形式也呈现多样化。

今天，被称为信息高速公路的 Internet(因特网)上存在着数以百万计的计

算机主机, Internet 上的信息就分布在这些世界各地的主机上, 信息资源极其丰富, 约占全部信息资源的 20% 以上, 随着图书情报界、图书出版界进入网络化、电子化和图书馆虚拟化的时代, 这个百分数正在迅速增加。Internet 是当今世界上最大的信息传播媒介, 它把世界的各个大学、机构和研究室联系在一起, 把学习者、教育者、研究人员联系在一起, 共享各种信息资源。World Wide Web 极大地影响着化学信息学的形成和以后的发展。在线文献检索、在线数据库检索、在线交流、远程学习和远程计算都已经成为现实。因此, 在信息时代, 对化学文献、化学数据仅仅以传统的方式处理和传递就不会适应实际的需要了, 必须学习运用计算机管理、处理和表达化学信息, 通过互联网获取与传递化学信息。

获取化学信息的三个基本途径是: 检索、阅读文献; 实验测量、查阅数据或图谱、数据处理; 根据已有化学知识和物理定理通过计算机推演、大规模计算预测与模拟或再现化学结构、化学性质和化学过程。这三个基本途径效率的提高大都得益于计算机技术和互联网通讯技术的高速发展。

2. 化学信息学领域范围

化学信息学曾经被狭义地定义为化学文献学, 主要是指化学、化工及其相关领域文献的检索与利用, 而计算机与信息技术的发展和日益普及大大地拓展了它的内涵。现在, 化学信息学 (Chemical Informatics) 是以化学、化工、计算机与信息技术为基础的边缘学科, 并涉及生物、材料、环境、能源、地球与空间资源、冶金、器件 (传感器、分子器件与纳米机器) 等相关领域, 它包括以下六个方面:

- 化学、化工文献学: 传统方式的和电子与网络时代的文献信息检索与个人资料管理;
- 化学知识体系的计算机表示、管理与网络传输: 化学结构、化学反应的计算机表示, 化学数据库技术, 化学信息的网际通讯语言;
- 化学图形学: 化学信息可视化和虚拟技术, 化工制图;
- 化学信息的解析与处理: 化学实验设计, 实验数据处理, 图谱的分辨与解析, 生物分子的信息解析, 多元分析与数据挖掘技术;
- 化学知识的计算机推演: 结构与性质关系, 分子及其聚集体系的计算与模拟, 分子与材料设计, 化学反应的分析与设计, 化工过程计算与仿真, 专家系统;
- 化学教育与教学的现代技术与远程信息资源。

3. 化学信息学的成就与发展

化学信息科学在近年来最值得注意的成就之一是化学结构的计算机编码和图形检索方法的发展。传统的文献信息产品经过发展与转化后无一不可实现电子化、网络化, 例如, 美国《化学文摘》(CA) 的光盘版 CA on CD、《科学引文索引》

(SCI)的网络版 Web of Science 和 Chemistry Server 等。

化学计量学中的多元统计分析、人工神经网络、遗传算法、小波分析等先进的数据挖掘技术在化学的许多领域都有成功的应用。

今天化学领域的知识并非都来自于传统实验室的经验测量,其中部分是根据已有知识和定理通过计算机推演、大规模计算预测与模拟得到的。美国化学会在期刊 Journal of Chemical Information and Computer Sciences(化学信息与计算机科学)上增加了副标题“ Includes Chemical Computation and Molecular Modeling ”(包含化学计算和分子建模)。化学信息学正在设法提供一个以用化学结构为框架的通用化学语言来组织化学领域的全部知识。

近年来,可视化和虚拟技术、分子建模和计算化学、分子的物理性质和光谱技术、化工仿真技术取得了巨大成功,影响着化学、化工信息产品的发展,例如 ChemOffice、Cerius²、WebLab Viewer、Gaussian、MOPAC、AspenPlus、Designll、WinSim,等等。有些软件在 Internet 上甚至可以免费得到。典型的化学信息产品有:网络与文献信息产品,分子建模图形与设计软件,化工仿真与工艺设计软件,化学化工专家系统,化学化工数据库(化学与物性数据库,化学反应数据库,晶体结构数据库,图谱数据库,基因、蛋白质等生物大分子数据库),各种化学应用软件,化学教学课件与远程课程等。

化学信息学将在以下领域取得进展:

化学 3D 结构的计算机表示、存储与检索;

合成反应的自动分析;

化学信息的网络传输;

大型化学化工数据库;

与数据库相结合的化学化工数据挖掘技术;

高效率的非局域优化技术;

化学信息的可视化技术与虚拟技术;

分子动力学从头算方法;

生物大分子建模、构象分析与第一原理计算;

基因、蛋白质分子的生物信息解析;

溶液结构;

分子器件、纳米组装的模拟与计算;

材料结构与物性;

相图计算与化工过程仿真技术;

远程教育系统。

我国的现代计算机化学与技术研究起步于 20 世纪 70 年代,南京大学忻新泉教授运用计算机开始在数值处理和非数值运算方面展开工作,并率先在南京

大学开设了计算机化学类课程,推动了我国计算机辅助化学教学的研究活动;中国科技大学张懋森教授开展计算分析和专家系统领域的研究工作,并成为 1984 年我国的第一本计算机化学期刊《计算机与应用化学》的主编;中国科学院上海冶金研究所陈念贻教授将化学模式识别和计算机模拟方法应用于材料研究;中国科学院北京化工冶金研究所许志宏教授提出建立化学化工数据库。国内还有许多学者在计算机化学和计算化学的重要领域中作出了贡献:中国科学院上海有机化学研究所郑崇直教授、湖南大学俞汝勤教授、中国科学院长春应用化学研究所许禄教授、中国科学院药物研究所嵇汝运教授、中国药科大学安登魁教授、清华大学胡鑫尧教授、北京化工大学麻德贤教授、南开大学林少凡教授等。

第二节 因特网(Internet)概述

Internet 是由异型计算机、异型系统、异型网络系统协议所组成的联系世界 150 多个国家和地区各种大小计算机网络的互连体,正像其英文名称所表现的:Internet = Interconnect + Networks。

20 多年前,计算机只能在高规格的实验室看到,Internet 还只是美国很少的几台大型计算机连成的网络,使用者只限于个别国家的小部分研究人员。在那个年代,“个人计算机”这个概念都显得很荒谬,更不用说把无数台计算机连成一个全球网络了。科学发展史表明,任何一种成功的新技术对社会的贡献都在于它可以扩大人类的潜能。

今天的信息高速公路 Internet 具有在瞬间跨越万水千山交流信息的能力,使得人与人之间的距离缩短了,世界变小了。一进入 Internet 即可领略到其信息海洋之浩瀚:在 Internet 上可以轻而易举地到达世界各国的站点,到达许多公司、机构、大学的计算机,浏览或索取他们公开的软件或信息;可以参加各种各样的新闻组或电子论坛;可以发布自己的消息;可以查寻几乎所有的信息;可通过主题、特定的关键词和词组或描述性语言来检索 Internet 站点上提供的数据库信息。

Internet 已成为一种全球性的文化现象,成为并列于电视、电话的第三大通信系统。Internet 中的各个子网络由不同的国家管理,但又不受国界的限制,它们在网上传递着各自不同的风格和习俗而又协调地交叉在一起。现代信息社会把信息作为最重要的资源和产业,它和材料、能源一起并列成为社会物质和精神活动的三大重要资源。据估算,有效地利用信息、合理地提取其中的有用成分能产生数倍的附加经济效益。目前只是处于 Internet 时代的黎明,它的发展将对社会进步、科技发展、经济增长等起越来越重要的作用。信息技术已成为世界各国在科技战线上争夺的战略制高点。

第二章 化学化工网站与网上数据库

化学信息的需求是多方面的,除了大量的化学专业信息以外,还会涉及物理、数学、天文、地质、环境、计算机技术、工业技术、市场、投资、金融、法律、教育。我们应根据不同需要,掌握一定的检索技巧,灵活运用各种检索策略,学习在 Internet 上快速、准确、全面地获取信息的方法。

Internet 上的 WWW(World Wide Web)环球网是一个可以检索 Internet 上几乎所有信息资源的信息查询系统,它利用超级文本的格式将相关联的信息文件有机地组织在一起。WWW 中的超文本(Hypertext)像 Windows 中的帮助文件,其中每个关键词都被用高亮度的色彩显示,但它的功能又远远超过帮助文件。WWW 的超文本文档形式不仅限于文本,而且是图文并茂、色彩丰富并可带有声音、动画等影视信息的超媒体,具有超出常规描述的多种传播媒体所具有的功能。这些文档往往带有超级链接(超级链接就是用醒目色彩的文字或图片链接着与它们有关的背景材料和参考文章或图片等,简称超链接),使得浏览者只需轻轻地点击它就可以非常方便地进入被链接处,可以从一篇文章转移到相关内容的另一篇文章、从一台主机转到另一台主机浏览下去。

Internet 的发展十分迅速,一方面信息非常广泛而且很多信息在定期或不定期更新,另一方面信息资料的混乱度也在增加,信息的平均质量严重下降,系统化整理成本越来越高。人们不得不花大量宝贵时间去阅读价值不大的资料。因此,信息质量的重要性远胜过数量,信息挖掘的深度胜过广度。Internet 信息检索工具尤其是第二代检索工具可以帮助用户系统地浏览所要检索的目标,筛选和提高信息检索的速度和质量。

Internet 上的化学资源非常丰富,除了有提供通用信息检索的网站以外,还出现了针对化学领域的综合性信息服务和各种化学化工专业的站点。它们有些提供文献和化学化工信息检索,有些提供各种化学数据库,有些是化学论坛,还有些是为化学品、化学仪器、化学软件采购提供帮助。

化学数据库包括化学文献数据库、化学结构数据库和化学物质的物性数据库。数据库不仅提供原始数据,也是进行信息再加工的数据源,在科学研究和工程技术领域有着非常重要的作用。功能强大、内容丰富的大型数据库应用范围广,而小型、实用的专业数据库价格便宜使用方便。随着 Internet 的普及与发展,许多数据库提供网上查询服务。化学文献数据库在其他章节中介绍,本章中的数据库主要是指结构数据库、反应数据库和物性数据库。

第一节 通用信息检索引擎

Internet 检索工具一般具备以下功能:首先有一个简单易用、功能强大的前端界面,使得查询方式容易设定;检索方式比较完备,可以按用户所需的方式进行查询;检索结果具有较好的准确性和可读性,且检索速度较快。网络检索工具大都采用 Server/ Client 结构和交互式的 Windows 前端界面,检索途径多,可使用“书签簿”保留检索历史,在检索窗口中按一定规则输入检索式,发出检索命令后即可得到检索结果。检索结果通常包括题目、内容简介、URL 地址等,由此可进行更进一步的检索,获得更详细的内容。由于检索引擎能够帮助用户在比较少的时间里方便地找到所需要的信息资料,因而成为用户上网的基本辅助工具。

WWW 是 Internet 上最先进的网络信息检索系统,它把超文本技术、网络技术和多媒体技术融为一体,并把 Internet 上的信息按一定的规则组织起来,以提供进一步的查询。从 1994 年 Yahoo 问世以来,已经开发出许多 WWW 检索引擎,有记录可查的检索引擎数量就达到 2 500 个左右。其中既有大型综合型的通用信息检索引擎,也有在特定领域里发挥作用的专业检索引擎。检索引擎的功能也有不少发展,第一代检索引擎是根据关键词在主页中的位置和出现频率来计分和排序,而第二代检索引擎以用户访问该网站时所花费的时间来确定它的人气指数,或者根据该网站被其他网站链接数量来确定它的等级,并向用户返送这些信息。

1. Yahoo([http:// www .yahoo .com](http://www.yahoo.com))

Yahoo 主页见图 2 - 1。Yahoo 提供按主题词分类的索引,并且有功能完备的检索方式,非常有效,但也异常拥挤,每天有上千万人次访问该节点。它已经开发了很长时间,由很多编辑人员维护,所以质量非常高。Yahoo 是目前最常用的检索工具,当不知去哪里时,在 Yahoo 上检索是找到好站点的最好方法,初学者也可以从那里找到较专门的站点。

使用 Yahoo 检索非常简单,输入主题词后点击“ Search ”按钮,或者点击“ Advanced Search ”(高级检索)即可。它将返回三种信息:

- (1) 满足检索条件的 Yahoo 的目录(用户可以利用这些目录进行交叉引用);
- (2) 满足检索条件的 WWW 网站;
- (3) 含有网页索引的 Yahoo 目录。

Yahoo 提供了全面的分类体系,它将 WWW 服务的节点按主题词建立分类索引。我们还可以按这个分类来检索,检索的方法是先选出分类的超链接,在

图 2 - 1 Netscape 浏览器窗口图(显示内容为 Yahoo 主页)

Yahoo 查询结果页面中具有下划线的均为 Yahoo 分类的超链接,如“Business & Economy”(商务与经济)、“Regional”(地区)、“Education”(教育)、“Science”(科学)等等。在选择分类的超链接后,从 Yahoo 就会得到一个查询界面,此时可以按其查询要求进行进一步查询。

Yahoo 还提供一些简单的高级查询,不用通过分类索引直接检索关键词。利用这种方法用户可以省去很多分类检索的步骤,但是注意使用的关键词的含义要相对窄一些。例如,对“chemistry”的检索,会使你面对数万条检索结果,失去了检索的意义。

对输入的关键词, Yahoo 提供 4 种检索方式和 2 种检索范围。它们是:“Intelligent default”(默认的智能方式)、“An exact phrase match”(精确词组匹配方式)、“Matches on all words”(AND, 单词之间是“与”关系)、“Matches on any word”(OR, 单词之间是“或”关系);“Yahoo Categories”(在 Yahoo 的分类网页范围内检索)、“Web Sites”(在 Yahoo 提供链接的 WWW 站点上检索)。Yahoo 还提供在 Usenet 上检索。其中“Intelligent default”是将输入的词作为一个子串,例如输入“meth”,那么 Yahoo 将检索 **meth**yl(甲基)、**meth**amphetamine(脱氧麻黄碱)和 **meth**od(方法)等。点击“Advanced Search”超链接,可以进行检索方式和检索范围的选择设置。

例 2 - 1 检索关于 Cu 与吡啶(pyridine)的反应的信息。

- (1) 点击“ Advanced Search ”;
- (2) 输入关键词“ Cu pyridine ”或“ Copper pyridine ”;
- (3) 将检索方式设为“ Matches on all words(AND) ”;
- (4) 点击“ Search ”,可以检索到数千条有关信息。

在本例中,如果选择关键词“ Copper pyridine reaction ”,缩小检索范围,检索结果的词条还可以少一些。对检索一些不能分开的词组如“ Organic Chemistry ”,Yahoo 规定该词组用双引号“ ”标记,即输入“ Organic Chemistry ”。

Yahoo 在高级查询上有些缺陷,如没提供“ NEAR ”等。但是 Yahoo 在其主页的末尾提供了到其他检索引擎如 Altavista、Lycos 等的超链接,如果用户在 Yahoo 上查不到所需资料,可以通过链接到这些地方去检索,从而起到一定的补充作用。

2. 其他检索引擎

第二代检索引擎:网站的质量反映在它内容的丰富性、准确性以及满足用户需求的程度上。第二代检索引擎以被其他网站链接和以大多数用户访问的情况认定一个网站的质量,它们的代表有

<http://www.google.com>;

<http://www.links2go.com>(Link To Go);

<http://www.DirectHit.com>。

精选网站检索引擎:它们提供高质量的网站,对所挑选网站的质量以星的多少来表示。一个网站的星越多,则其内容越好。

<http://www.About.com>(About 让各行各业的专家推荐他们所在行业的最佳网站);

<http://www.eBLAST.com> (Britannica 不列颠百科全书精选网址);

<http://www.RefDesk.com/index.html> (Reference Desk 参考文献检索引擎)。

大型检索引擎 **AllTheWeb** (<http://www.AllTheWeb.com>, <http://www.ussc.alltheweb.com>): AllTheWeb 是目前 Internet 上收录网页最多的检索引擎(约有 2 亿个),他们计划收录 10 亿个网页。同时他们认为 AllTheWeb 的检索速度最快,敢于标明用户每次检索所用的时间。

特色检索引擎

<http://www.Meta-List.net>(Meta List 收录有 20 多万个邮件列表);

<http://www.FindWhat.com>(提供产品与服务信息);

<http://www.InfoJump.com>(收录近 4 000 种报刊的 500 万篇文章)。

检索引擎的检索引擎: Internet 信息检索工具的特点在于它可以让用户通过关键词来查询网上的信息,由于不同的检索工具有不同的组织与分类,因此检索结果的表达方式也各不相同。现在一般用户已经无法知道目前有多少检索引擎,为了方便使用,一些检索引擎目录应运而生。它们专门收集世界各国的各种检索引擎。另外,一种新的称为“网页检索器”的服务正越来越引起人们的兴趣,用户只要把自己感兴趣的网页地址在“网页检索器”中登记,并提供自己的电子函件地址,“网页检索器”就会经常地为该用户检索相关的网页,一旦发现网页内容有新变化,就用电子函件通知该用户。用户登记的网页地址的类型可以包括“http”、“FTP”、“Gopher”等。网页检索器可称为一位忠实的侦察员,它可以及时通知你感兴趣的站点的内容变化,以便于得到最新的信息,这对定期跟踪某些站点的科技信息资源尤其有用,结合检索工具,合理地加以利用,可以收到很好的效果。

<http://www.searchenginecolossus.com> (Search Engine Colossus, 以国家字母顺序排列);

<http://www.searchengineguide.com> (Search Engine Guide, 检索引擎指南);

<http://www.searchpower.com> (Search Power, 分类排列数千个检索引擎);

<http://www.easysearcher.com> (Easy Searcher 2, 专业检索引擎目录);

<http://www.netmind.com> (Net Mind, 网页检索器)。

Internet 上还有许多其他优秀的信息服务商,如:

Infoseek(<http://www.go.com>)包含有许多 Yahoo 中没有的化学专业信息;

Excite(<http://www.excite.com/query.html>)也是一个很好的检索引擎;

AltaVista(<http://www.altavista.com>);

新浪(Sina)(<http://www.sina.com.cn>);

中文 Yahoo(<http://cn.yahoo.com>)中文检索引擎;

OpenText(<http://www.opentext.com>);

WebCrawler(<http://www.webcrawler.com>);

Lycos(<http://www.lycos.com>);

美国在线(AOL)(<http://www.aol.com>);

B to B(<http://www.b2bscene.com>)电子商务网站;

Nightn(<http://www.nightn.com>)。