

高等学校教材

化学信息学

第二版

陈明旦 谭凯 编著

HUAXUE
XINXIXUE



化学工业出版社

高等学校教材

化学信息学

第二版

陈明旦 谭凯 编著

JAXUE
XINXIXUE



化学工业出版社

· 北京 ·

本书以化学信息资源检索和化学结构信息可视化为重点, 强调现代数据库检索技术的使用, 全面介绍了化学信息与数据库检索的基础知识、常用文摘数据库、常用全文数据库、常用电子期刊、特种信息资源、常用化学事实数据库及文献数据的管理等内容。另外, 本书重点介绍了化学结构的表示和可视化、分子模型的构建及著名的分子图形软件的使用。为了教学需要, 本书也介绍了重要的印刷类化学文献信息。

本书内容新颖, 实用性强, 适用面广, 可作为高等院校有关专业本科生、研究生的教材, 也可作为科学研究工作者、专业技术人员和图书情报工作者了解掌握信息检索与管理的学习参考。

图书在版编目(CIP)数据

化学信息学 / 陈明旦, 谭凯编著. —2 版. —北京: 化学工业出版社, 2011.2

高等学校教材

ISBN 978-7-122-10412-0

I. 化… II. ①陈… ②谭… III. 计算机应用-化学-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 (2011) 第 009200 号

责任编辑: 宋林青

文字编辑: 孙凤英

责任校对: 宋 玮

装帧设计: 史利平

出版发行: 化学工业出版社 (北京市东城区青年湖南街 13 号 邮政编码 100011)

印 装: 化学工业出版社印刷厂

787mm×1092mm 1/16 印张 17 $\frac{3}{4}$ 字数 471 千字 2011 年 3 月北京第 2 版第 1 次印刷

购书咨询: 010-64518888 (传真: 010-64519686) 售后服务: 010-64518899

网 址: <http://www.cip.com.cn>

凡购买本书, 如有缺损质量问题, 本社销售中心负责调换。

定 价: 32.00 元

版权所有 违者必究

前言

化学信息学是一门化学、化工与信息科学、计算机科学交叉的新兴学科，它采用信息学方法来解决化学问题。由于计算机和网络已经广泛应用到各个化学研究领域，许多化学研究学科都与化学信息密切关联。大量的传统学科与信息科学结合发展成为新兴的边沿学科，共同构成了信息化社会的基础。近年来，借助于计算机和网络技术，在化学实验中使用大量的新仪器和新方法，从而得到了大量的实验数据。信息的急剧膨胀与人们有限的学习时间及吸收能力形成了巨大的矛盾，反过来又要求使用计算机和网络技术对海量信息进行高速和精确的处理和利用，对化学信息处理提出更高和更迫切的要求，以达到产生和处理的相对平衡。化学信息的收集、处理、管理和利用，必不可少地要借助信息学的方法和技术。于是，化学信息学随着计算机在化学中的广泛应用和化学信息网络化的不断发展而引起化学家的广泛关注和高度重视。今天，化学信息学已经成为药物和生命科学研究不可缺少的热门工具。

正由于人们认识到化学信息学对于现代化学研究的重要性，近年来国外部分大学正尝试在化学教育中系统地增加化学信息学课程。化学信息学的发展将推动传统的化学教育模式的改革。2003年德国的Johann Gasteiger 和 Thomas Engel 出版了“Chemoinformatics A Textbook”一书，该教科书系统、全面、深入浅出地介绍了化学信息学的各个研究领域，及其研究现状和今后的发展动向。2004年，梁逸曾等将该书翻译成中文版，即《化学信息学教程》，由化学工业出版社出版。在国内，化学信息学的教学已经引起了高度重视，我国教育部化学类专业教学指导分委员会已将化学信息学列入高等学校化学类专业的化学教学基本内容。目前，已经有多所大学都已开设化学信息学或相关的课程，并出版了数种化学信息学教材。

化学信息学作为一门新的教学课程，其课程的要求、内容、教学方式和教材等是课程建设的一项新任务。国外化学信息学的教学侧重于专业方向教学，交叉性强，涵盖广。在国内，化学信息学作为化学系各专业方向的必修课程之一，由于课时的限制，其教学内容多侧重于化学文献学，其他的相关内容多分别归属于另外的课程，并已经另有专著出版。本书以化学信息资源检索和化学结构信息可视化为重点，强调现代数据库检索技术的使用，全面地介绍了数据库检索的基础知识、常用文摘数据库、常用全文数据库、常用电子期刊、特种信息资源和常用化学事实数据库等内容。另外，本书重点介绍了化学结构的表示和可视化、分子模型的构建及著名的分子图形软件的使用。为了教学的需要，本书也介绍了重要的印刷类化学文献信息。

本书内容新颖，所介绍的应用软件大多是目前的最新版本，包括 Chem3D 12.0、ChemDraw 12.0、HyperChem 7.52、Reference Manager 11、EndNote 9.0、DS ViewerPro 5.0、GaussView 5.0 和 MDL CrossFire Commander 7.0 等功能强大的著名软件。在本书所介绍的内容中，Reference Manager 和 EndNote 是著名的文献管理软件，MDL CrossFire Commander 7.0 是著名的

Beilstein/Gmelin 数据库检索平台, Current Contents 是 ISI 的低价位产品, 而 DS ViewerPro 和 HyperChem 均是优秀的三维分子图形软件。对它们的使用, 目前鲜有详细和全面的介绍。另外, 本书详细介绍了 SciFinder 和 Web of Science 的使用。多数高等院校师生可方便地使用本书所介绍的大部分数据库、电子期刊和分子图形软件。本书具有较强的实用性和普及性, 可作为高等院校化学、化工和材料及相关专业本科生、研究生的教材, 也可作为科学研究工作者、专业技术人员和图书情报工作者了解掌握信息检索与管理的学习参考。

本书的修订工作由厦门大学化学系周朝晖、李耀群、谭凯完成, 得到了厦门大学网络中心郭嘉老师的指正, 也得到化学工业出版社的热心支持和大力帮忙, 化学系本科生杨静、陈善慈等同学帮助进行了部分素材的制作, 在此表示衷心的感谢。由于时间短促, 书中难免存在遗漏和不妥之处, 敬请读者批评和指正。

编著者
2010 年 11 月

目 录

第一章 化学信息概论	1
第一节 化学信息与化学信息学	1
一、化学信息学的产生和发展	1
二、化学信息学的研究领域	2
三、信息资源检索的意义与作用	3
四、化学信息学的课程内容	4
第二节 电子信息和数据库	5
一、电子信息	5
二、数据库	6
三、在线化学数据库	8
第三节 信息检索的方法和步骤	13
一、信息检索方法	13
二、信息检索效果的评价	14
三、信息检索技术	15
四、信息检索步骤	17
第二章 印刷类化学文献信息	22
第一节 图书分类法简介	22
一、图书馆的重要地位和作用	22
二、图书馆图书的分类系统	22
第二节 重要工具书、专著和大全	24
一、工具书	25
二、专著和大全	27
第三节 重要化学期刊简介	29
一、综合性自然科学期刊	30
二、化学学科综合性期刊	30
三、专业性化学期刊	31
四、综论类期刊	34
第四节 化学文摘 (Chemical Abstracts)	35
一、CA 类目	36
二、文摘编排顺序和格式	37
三、文摘著录内容和格式	37
四、CA 中的索引	40
五、CA 检索的注意事项	42
第三章 常用文摘数据库	44
第一节 ISI 数据库	44
一、ISI 数据库简介	44
二、科学引文索引 (SCI)	45
三、Web of Knowledge	46
第二节 Current Contents 数据库	52
一、数据库简介	52
二、浏览文献	52
三、检索文献	55
四、检索结果及处理	58
第三节 Web of Science 数据库	61
一、数据库简介	61
二、数据库检索	62
三、检索结果及处理	66
第四节 INSPEC 数据库	69
一、数据库简介	69
二、数据库检索	70
第五节 BIOSIS Previews 数据库	72
一、数据库简介	72
二、数据库检索	72
第六节 Chemical Abstract 数据库	74
一、化学文摘社 (CAS)	74
二、CA 网络版 (SciFinder Scholar) 数据库	76
第七节 EI Village 2 数据库	85
一、数据库简介	85
二、数据库检索	85
三、检索结果及处理	88

第四章 常用全文数据库	91
第一节 ProQuest 数据库	91
一、数据库简介	91
二、数据库检索	92
三、检索结果及处理	93
第二节 EBSCO 数据库	94
一、数据库简介	94
二、数据库检索	95
三、检索结果及处理	97
第三节 Science Direct 数据库	98
一、数据库简介	98
二、期刊/书籍浏览	99
三、数据库检索	99
四、检索结果及处理	101
第四节 中国期刊全文数据库	102
一、数据库简介	102
二、数据库专栏目录	102
三、数据库检索	103
四、检索结果及处理	104
第五节 万方数据化期刊	106
一、数据库简介	106
二、数据库检索	106
第五章 常用电子期刊	109
第一节 SpringerLink	109
一、概述	109
二、浏览	109
三、检索方法	110
四、检索结果及处理	110
第二节 Wiley InterScience	111
一、概述	111
二、浏览	112
三、检索方法	113
四、检索结果及处理	113
第三节 Kluwer Online	116
一、概述	116
二、浏览	116
三、检索方法	117
四、检索结果及处理	118
第四节 American Chemical Society	118
一、概述	118
二、浏览	119
三、检索方法	119
四、检索结果及处理	120
第五节 Royal Society of Chemistry	121
一、概述	121
二、浏览	121
三、检索方法	122
四、检索结果及处理	122
第六节 American Institute of Physics	122
一、概述	122
二、浏览	123
三、检索方法	123
四、检索结果及处理	124
第六章 特种信息资源	126
第一节 专利数据库	126
一、专利基础	126
二、中国专利数据库	128
三、美国专利数据库	129
四、欧洲专利数据库	134
五、专利搜索引擎——Google Patent (谷歌专利)	136
第二节 会议文献检索 (Proceedings)	138
一、会议文献概述	138
二、会议录引文索引 (Conference Proceedings Citation Index)	138
三、中国重要会议论文全文数据库	139
第三节 学位论文检索	141
一、PQDD 博硕士学位论文数据库	141
二、CALIS 学位论文库	142
三、中国博士学位论文和优秀硕士学位论文全文数据库	143
第四节 论文评价与引文索引检索	144
一、论文评价	144
二、引文索引	144
三、期刊引文报告 (Journal Citation Reports)	145

第五节 Reference Manager 的使用.....	148	一、EndNote 简介.....	155
一、RM 简介.....	148	二、EndNote 文献库的操作.....	156
二、文献数据的导入.....	149	三、EndNote 文献数据的导入.....	158
三、RM 的检索.....	150	四、EndNote 的在线检索.....	159
四、RM 与 Word 的结合使用.....	153	五、EndNote 与 Word 的结合使用.....	161
第六节 EndNote 的使用.....	155		
第七章 常用化学事实数据库.....	164		
第一节 CrossFire Beilstein/Gmelin 的使用.....	164	一、概述.....	179
一、数据库简介.....	164	二、检索方法.....	180
二、检索程序界面.....	165	三、检索结果.....	181
三、数据库检索.....	166	第四节 科学搜索引擎 (Scirus)	181
四、检索结果及处理.....	170	一、概述.....	181
五、应用实例.....	171	二、检索方法.....	182
第二节 NIST Chemistry WebBook.....	176	三、检索结果及处理.....	184
一、数据库简介.....	176	第五节 维基百科全书.....	185
二、数据库检索.....	177	一、概述.....	185
三、检索结果及处理.....	178	二、检索方法.....	186
第三节 谷歌学术搜索 (Google Scholar)	179	三、检索结果.....	186
第八章 化学结构的可视化.....	187		
第一节 化学结构的表示和可视化.....	187	三、分子结构图的绘制.....	199
一、化学结构的表示.....	187	四、电子转移和原子特征的表示.....	200
二、二维化学结构的表示和可视化.....	188	五、化学反应式的绘制.....	201
三、三维化学结构的表示和可视化.....	189	六、其他功能.....	201
四、大分子化学结构的表示和可视化.....	191	第三节 ChemDraw 的使用.....	202
五、分子表面的显示.....	192	一、简介.....	202
六、化学结构的网络表示.....	193	二、操作界面.....	203
第二节 ISIS/Draw 的使用.....	197	三、二维分子结构的绘制.....	206
一、简介.....	197	四、其他功能.....	210
二、操作界面.....	197		
第九章 三维分子模型设计.....	213		
第一节 分子力学和分子建模.....	213	第三节 Chem3D 的使用.....	227
一、分子力学基本原理.....	213	一、简介.....	227
二、分子力场参数化.....	216	二、操作界面.....	228
三、三维分子模型的构建.....	217	三、分子模型的操作.....	231
第二节 DS ViewerPro 的使用.....	218	四、分子模型的构建.....	233
一、简介.....	218	五、分子结构数据的测定.....	236
二、操作界面.....	219	六、分子的计算.....	236
三、分子模型的绘制.....	221	七、分子表面和分子轨道图形的显示.....	237
四、分子模型的显示.....	222	第四节 HyperChem 的使用.....	238
五、其他功能.....	226	一、简介.....	238

二、工具栏的使用	239	三、分子模型的操作	265
三、菜单命令	240	四、分子模型的构建	267
四、使用实例	259	五、分子结构数据的测定	270
第五节 GaussView 的使用	261	六、分子的计算	271
一、简介	261	七、计算结果的显示	271
二、操作界面	262		
参考文献			274

第一章 化学信息概论

第一节 化学信息与化学信息学

一、化学信息学的产生和发展

信息是事物的存在方式和运动状态的记录，它精确地描述物体或事件，并且可借助于一定的物质载体进行存储和传播。计算机和网络通信中的二进制编码，是现代信息技术的基础。随着科学技术的进步，计算机和网络技术逐步应用于各个科学研究领域，以其高速、精确的计算手段，处理巨大数量的研究信息，进一步推动了科学技术的发展。如今，与社会经济发展密切相关的信息科学、生命科学、材料科学、能源科学和环境科学等，已发展成为当今的热门学科。大量的传统学科与信息科学结合发展成为新兴的交叉学科，共同构成了信息化社会的基础。

化学是一门古老的实验性学科，它是由五千多年前青铜器时代的冶炼术逐步发展而来的。在长期的发展过程中，积累了大量的实验事实、数据与文献。化学家们通过分析实验结果，结合数理知识，发展化学模型，演绎与归纳出化学理论基础。目前，物质的分子结构、能量和某些分子性质，虽然可以用量子化学的薛定谔（Schrödinger）方程描述，但是受到计算机硬件和软件发展的限制，还有许多计算需要使用一系列的近似方法。例如：针对化学反应的计算，在考虑反应条件（如溶液、温度和催化剂）的影响时，精确的计算常是无能为力的。

尽管目前的理论化学研究无法处理复杂的化学问题，然而化学家仍然进行了许多卓有成效的研究工作，合成了大量具有特定性质的化合物，并将实验室进行的小型反应转化为化学工业的大型生产。化学家进行一系列的实验，分析实验结果，发现其共同点和不同点，并构建可能重现实验的模型，进行合理的推论，以新实验进行模型的检验、证实、否认和修改。经过实验、数据分析、建模、再实验、再验证的不断完善的过程，最后将感性的化学实验信息提升为理性的化学知识，这就是化学家成功的秘诀。以上所有的工作成功完成，都要充分利用大量的化学信息和化学知识。

在信息时代，借助于计算机和网络技术，化学研究发展了许多新的实验方法，从而得到了大量的实验数据。目前已知的化合物超过 4 千多万，不同的化合物可使用不同方法合成，它们有不同的光谱图，表现出不同的物理、化学和生物学性质，由此产生和积累了极丰富的化学信息。另外，Internet 上具有大量基于 Web 的化学信息网站和化学信息数据库，网络逐步成为各种信息资源传递的重要载体。信息的急剧膨胀与人们有限的学习时间及吸收能力之间形成了巨大的矛盾，反过来又要求使用计算机和网络对海量信息进行高速和精确的处理和利用，以达到产生和处理的相对平衡。化学信息的收集、处理、管理和利用，必不可少地要借助具有高效和共享性的计算机和网络。由此，化学、化工与信息科学、计算机科学交叉的新兴学科——化学信息学应运而生。1973 年在荷兰举办了“Computer Representation and Manipulation of Chemical Information”会议。1975 年美国化学会将其期刊“Journal of Chemical Documentation”改名为“Journal of Chemical Information and Computer Sciences”。化学信息学使用计算机和信息技术产生数据，将数据转换为

信息,再将信息提升为知识,已达到解决化学问题的最终目的。化学信息学给新世纪的化学研究带来全新的面貌,它随着计算机在化学中的广泛应用和化学信息网络化的不断发展而引起化学家的高度重视。

化学信息学最通用的英文名称是: Chemoinformatics 或 Cheminformatics。由于化学信息学是新兴的学科,不同的学者对化学信息学进行了不同的定义。迄今为止,化学信息学还没有一个统一的被广泛接受的定义。1987年,诺贝尔化学奖获得者法国化学家 J. M. Lehn 在获奖报告中首次提出了化学信息学概念,尽管他本人以后未进一步深入研究这些概念的本质。K. Brown 于 1998 年把化学信息学定义为:“化学信息学是将数据转化为信息和信息提升为知识的总称,其主要目的是在药物先导化合物的辨认及优化领域做出快速决定”。G. Paris 提出了一个更具一般性的定义:“化学信息学是一个一般的术语,它包括化学信息的设计、建立、组织、管理、检索、分析、判别、可视化及使用”。缪强认为:“化学信息学涉及化学、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”。邵学广等人对化学信息学的定义为:“化学信息学是近几年发展起来的一个新的化学分支,它利用计算机和计算机网络技术,对化学信息进行表示、管理、分析、模拟和传播,以实现化学信息的提取、转化与共享,揭示化学信息的内在实质与内在联系,促进化学学科的知识创新”。李梦龙等人对化学信息学的定义为:“化学信息学是利用计算机和计算机网络技术,对化学信息进行表示、管理、分析、模拟和传播,实现化学信息的提取、转化与共享,揭示化学信息的内在实质与内在联系的学科”。英国 Sheffield 大学对化学信息学的表述是:“化学信息学是创建、检索、组织、分发和处理数据库中的化学信息的一个令人激动的新学科”。J. Gasteiger 给出了一个通用的定义:“化学信息学是应用信息学方法解决化学问题的学科”。

化学信息学作为一个化学与信息科学、计算机科学的交叉新学科,其定义及其基本内容均需要进一步的探讨。由于计算机和网络已经广泛地应用到各个化学研究领域,许多化学研究学科多多少少都与化学信息有关联。正是由于化学信息的这种广义性,对化学信息学的定义和内涵,目前国内外的学者有很大的争论。不少学者认为,自己所从事的与计算机和信息有关系的化学研究是化学信息学的一部分。因此,在解释化学信息学的名词和内涵时,都更多地强调自己原来的研究领域。但是有些研究学科在传统上都已经自成体系,并已经自有很多专著和教科书出版。正由于许多化学研究学科都广义地与化学信息有关联,对化学信息学进行确切的定义也不容易。因此,在化学学科的科研领域,对化学信息学的定义和内涵并非一定需要严格的界定。

二、化学信息学的研究领域

化学信息学应用现代信息技术构建信息处理系统,用于处理长期积累的大量化学信息资源,帮助化学家组织、分析和理解已知的科学数据,正确地预测化学物质的性质,开发新化合物、材料和方法。另外,化学信息学的研究,涉及分子模拟、化学结构的编码、检索、化学数据的可视化等研究领域。化学信息学是信息技术应用在化学上的表现。

对于化学信息学的研究内容的包涵范围,不同的学者也有不同阐述。缪强认为包括以下六个方面:“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”。徐筱杰认为化学信息学内容包括:“化学信息的组织、管理、检索和使用;分子结构的编码、描述、三维结构的构建;化学信息的加工、处理及深化;计算组合化学;化学体系中信息的交换及传递;分子的物理化学性质预测”。邵学广等人认为化学信息学的研究内容主要包括:“利用计算机技术和计算机网络技术对化学信息进行表示和管理;利用计算机网络技术对化学信息进行收集、传播和共享;化学体系的计算机模拟或建模;利用计算机技术对复杂的化学信息进行解析,以快捷



方便的方式最大限度地提取和利用有用信息”。

化学是研究化学物、化合物性质和化合物转换的学科。化合物和化学反应是化学的两个主要研究对象,对化合物的研究是静态,而对化学反应的研究是动态。进行从反应物到产物的化学反应设计并合成具有理想性质的化学物是化学家的主要工作之一。化合物的性质与化学结构是密切相关的,对化学反应的反应产物进行化学结构表征,研究结构与性质、结构与活性关系(SPR、SAR)是化学的基本研究工作之一,更进一步的研究是定量结构与性质、定量结构与活性关系(QSPR、QSAR)。由于这些化学的基本研究工作的复杂性,目前无法使用量子化学方法解决。基于理论化学研究的局限性,化学研究更多表现出的是推论学习过程。使用推论学习将这些数据提升为知识,加深对化学的理解。

化学信息主要来自于由实验测量或理论计算的数据的组合。这些数据包括化合物的物理、化学性质,成分,结构信息和光谱图形等。化学信息经过收集、浓缩、提取和推论才能提升成为化学知识。广义上说,使用信息学方法处理上述化学基本问题都是化学信息学研究领域。换言之,化学信息学所涉及的研究领域为:应用计算机科学方法或信息学解决化学问题,对化学信息进行有效的存储、操作和处理,使化学信息合理地提升为化学知识。

生物信息学的主要研究集中于基因和蛋白质。基因和蛋白质作为化合物,也是化学家很感兴趣的研究对象。化学家在核酸和蛋白质的结构和功能的说明方面做出了实质性贡献,因此,化学信息学与生物信息学没有严格的区分。在传统上,化学信息学以小分子或中等分子(含数百或上千个原子)为研究对象,而生物信息学的研究对象是更大的分子。如果将化学信息学和生物信息学结合使用,将会更大地促进基因和蛋白质的结构、性质和功能的研究。例如在药物设计中,生物信息学可为新药开发提供识别蛋白质目标,而药物是小分子化合物,化学信息学方法可用于开发出新的先导结构并优化可能的药物。

三、信息资源检索的意义与作用

通过化学实验可获取大量的化学信息,通过理论化学计算也可获得化学信息。这些化学信息通过书籍、期刊、专利、数据库以及音像资料等媒体记录和传播,使后人得以享用。借助计算机技术和互联网通信技术,快速、高效地检索媒体形式的化学信息,是另一个获得化学信息的途径。

科学技术的发明创造需要依靠经验和知识的积累。任何一项创造发明或一个新的理论的建立,都是在人类已经取得成就的基础上,进行创新性的探索,这是事物发展的客观规律。信息资源既是过去经验的总结,又是未来工作的向导。古今中外一切有成就的科学家,都是在广泛吸收前人和同代人知识的基础上,受到启发而取得成功的。牛顿曾说过一句名言:“如果我比笛卡尔看得远些,那是因为我站在巨人们的肩上的缘故”。

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。化学信息的检索、管理和利用具有以下的重要意义与作用。

1. 启迪创造

科研工作具有继承和创新两重性,科学研究的兩重性要求科研人员在探索未知进行创新之前,应该尽可能地继承和利用与之相关的信息。科研人员的工作投入会有相当多的时间是与信息密切相关的,信息检索是科学研究的重要前提,它在整个科研过程中占有重要的地位。科研人员在开始从事一项新科研项目时,都应该对该研究项目所涉及的研究范围进行全面的检索,系统地了解该研究项目的国内外研究现状,充分了解前人已取得的研究成果,认识已有研究工作的不足,对其发展动向作出判断,从信息的海洋获取新鲜营养,启发新颖的构思,周密地完成选题、立项、可行性论证等工作,避免重复他人的劳动,做到有所发现、有所创新、有所前进。

难度大的科研项目可能要经历漫长的研究过程。除了在选题、立项阶段进行全面的资源



检索外,在整个研究过程中还要密切关注该研究项目的发展动态。对于影响项目进展的困难问题,需要再进行必要的信息检索,探讨解决困难的其他途径,修改已制定的研究方案。课题立项、具体实施、项目结题、成果鉴定等各阶段都需要信息检索。由于信息检索贯穿了科研工作的始终,因此它是科研工作的重要组成部分。

2. 拓宽视野

21世纪是科学技术高速发展的时代。随着计算机、网络技术的普遍应用,信息与知识的积累日新月异,出现了大量的边缘科学和交叉科学。面对知识频繁更新的世界,狭窄的专业化教育已不适应飞速发展的新形势。科研人员只有接受终身教育,在不断的教育中更新知识、拓宽视野、适应环境的变化,才有可能不落伍。信息资源的检索、管理和利用是快速获取最新知识的最重要途径之一。

3. 培养能力

现代教育的职能不单纯是知识的传授,还要大力进行各种能力的培养,其中包括自学能力、思维能力、研究能力、表达能力、创造知识能力、终生教育能力、组织管理能力和收集处理信息的能力(检索各种信息资源的能力、分析与综合利用信息资源的能力)。学习和掌握信息的检索、管理和利用,是提高这方面能力的重要途径之一,它引导学生发挥自己的创造力,为学生打开自己动手获取补充知识的大门。这将会对学生的事业发展产生极大的效益和深远的影响。

4. 培养素质

素质教育是当前教育改革的热门话题。信息素质是信息化社会人才素质的重要组成部分,具备了优良的信息素质的人才具有比较强的判断力、决策力。信息素质主要是指信息能力和信息意识。信息能力是指信息搜集、检索、判断、选择、管理和利用能力。信息意识是指对信息的性质、地位、价值和功能的认识和反映,它直接影响对信息的捕捉、判断和利用的自觉程度。学生可利用教育信息化的环境,通过对信息的检索、收集、管理、利用,实现对知识的探索 and 发现,这对培养创新人才具有重要的意义。面向信息社会的跨世纪人才要掌握信息,具备信息意识,会使用信息系统获取自己所需的信息。信息素质是信息社会中每个人赖以生存、用于学习的基本能力,它是进入信息社会的通行证。

四、化学信息学的课程内容

由于人们逐渐认识到化学信息对于现代化学的重要性,因此,近年来国内外某些大学正尝试在化学教育中系统地增加化学信息学课程。化学信息学的发展,对改革传统的化学教育模式的巨大推动作用也是可以预期的。在国内的高等院校,化学信息学的教学已经引起高度的重视,化学教学指导委员会已将化学信息学列入化学教学的基本内容。许多大学都开设了化学信息学或相关课程,并出版了数种化学信息学教材。

英国 Sheffield 大学的化学系开设三年学士学位和四年硕士学位化学信息学专业,选择化学信息学专业的学生,除了需要学习和掌握坚实的各学科化学基础知识外,还必须学习相关的化学信息学必修课程,如:基础化学信息学、脚本语言、高级化学信息学、数据库设计和化学信息项目等。英国 UMIST 大学设有硕士化学信息学专业,其中必修课包括:研究方法和可行性研究、化学信息源、分子模拟和设计 I、光谱学和晶体学中的化学信息、数据库设计和编程、分子模拟和设计 II、化学信息学的应用和生物信息学基础等。美国印第安纳大学的信息学学院(School of Informatics)设有化学信息学专业,该专业本科生的必修化学信息学课程包括:化学信息学 I、化学信息学 II(分子模拟),选修课有:化学信息资源和检索、化学信息的计算机资源。该专业硕士生的学习课程包括:信息学导论、信息管理、化学信息技术、计算化学和分子模拟。在 J. Gasteiger 和 T. Engel 出版的“Chemoinformatics A Textbook”(化学信息学教程)中,其内容包



括化合物的表示法、化学反应的表达、数据、化学数据库和数据资源、化学结构检索、物理和化学数据的计算、结构描述的计算、数据分析方法和应用。该书对物理和化学数据的计算、结构描述的计算和数据分析方法和应用部分只作简要的介绍,因为分子力学、量子化学、自由能关系、化学计量学、神经网络、模糊逻辑、遗传算法和专家系统等详细描述本身就可编写成专著,而且已经另有专著出版。

化学信息学作为一门新的课程,其课程的要求、内容、教学方式和教材等是课程建设的一项新任务。通过分析和比较国内外化学信息学的课程,发现国外更多强调的是,将化学信息学视为一个专业方向。学生除了学习专业化学知识外,还必须学习数据库设计、程序的编写、分子模拟等方面内容,加强其计算机和理论化学的知识水平,当然也必须掌握有关化学信息的检索等基础化学信息学知识。其涉及的教学内容需要大量的课时,且大多属于计算机科学、化学计算和数据分析等领域,目前这类专门的课程已有大量教材出版。

在国内,化学信息学是化学系各专业方向的必修课之一。由于课时的限制,在教学内容和教科书的编写方面,较为狭义的化学信息学定义,应是更可接受的。化学信息学除了涉及化学信息的获取、管理、处理、推演和模拟外,近年来化学结构的可视化(以图形的方式对化学信息进行描述)是化学信息学最值得注意的成就之一。作为大学本科生和研究生的学习教材,本书的内容主要侧重于介绍高校师生可方便使用的信息资源,强调以数据库为基础的化学信息的存储、检索、管理和利用,作为化学信息检索的基本要求也介绍了印刷品化学信息的检索,还包括化学结构信息的可视化(图形)表示和分子结构图形的计算机模拟等内容。

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。当前使用化学信息学最多的是医药工业。学生通过本课程的学习,提高了化学信息的获取、管理和使用能力。具有坚实的化学信息学知识和熟练计算机技能的人才,可在制药、农用化学、生物工艺等领域开展开拓性研究工作,具有良好的从业条件。

第二节 电子信息和数据库

一、电子信息

传统的信息以纸介质作为载体。这种介质使用了很长时间,而且仍然会沿用下去,虽然它存在着很多固有的缺点,但仍然无法被完全取代,但其重要性已日益降低,而正发挥着越来越大作用的电子信息将占据主要地位。

电子信息资源起始于 20 世纪 60 年代初,它以电子数据的形式,将文字、声音、图像等形式的信息存储在光、磁等介质上。电子信息以磁性介质和光介质作为信息存储载体,以现代信息技术作为记录手段,信息以数字化的形式存在,既可在计算机内高速处理,又可借助于通信网络进行远距离传送,使得全球资源共享得以实现。电子信息表现形式为文本、超文本、多媒体和超媒体。电子信息使得信息的组织方式发生了质的变化。信息单元由线形排列转变为可按照自身的逻辑关系组成网状结构,具有直接、相互联系、非线性特点,便于进行计算机的检索与利用,提高了信息资源的利用价值。

电子信息资源主要是指通过计算机等设备以数字信号传递的数字信息资源组成的数据库。从信息检索的角度,数据库信息资源可分为下列几类。

1. 联机信息资源

联机信息资源始于 20 世纪 70 年代,用户使用检索终端设备,通过通信设施,与远程计算机

相连,检索远程数据库中的信息资源,实现了电子信息资源的远距离传输,使信息资源在全球范围内得到广泛的传播和高度的共享。使用国际联机数据库可实时、快速地检索到全面和实用价值较高的电子信息,但是专用数据库的联机费用昂贵,一般用户难以承受,多为单位用户使用。著名联机信息服务系统 Dialog、STN 和 OCLC 数据库等,它们具有学科覆盖面广、信息资源丰富、信息追溯年代长和查准率高的特点。用户可实时、快速地进行人机对话式检索,可联机或脱机浏览、传送所检索结果,并可根据需求联机修改检索策略。

2. 光盘信息资源

近年来快速发展起来的光盘存储技术,具有信息存储密度高、容量大和成本低等优点,用光盘作为信息存储载体的光盘数据库大大地降低了电子信息资源的使用成本,已经得到广泛的应用。目前,大多数著名的数据库公司都有商品化光盘产品,单机版光盘数据库只可进行单机检索,网络版的光盘数据库可进行局域网上检索,还可以与联机检索系统联网进行联机检索。光盘数据库的检索与联机数据库检索相似,但其信息量较少,更新周期较长,但使用费用大大低于联机检索费用。因此,在网速较慢、经费有限时,利用光盘数据库仍是一种获取信息的有效途径。

3. Internet 网络信息资源

Internet 使用 TCP/IP 标准通信协议将全世界的计算机和计算机网络互联,构成一个资源极丰富的信息库,它涵盖了各个领域、学科和种类的信息资源,信息资源更新速度快,而且大部分资源是免费的。无边无际的 Internet 资源已成为人类进行信息交流的最方便的共享空间。Internet 的所有信息的发布、获取以及反馈都可以在其没有国界的网络上完成。

WWW (World Wide Web) 是 Internet 上最热门的资源,它采用客户机/服务器的模式,以超文本的方式链接分散在各地的其他服务器中的信息。用户通过 WWW 浏览器可以方便地获取遍布世界各地的信息资源。WWW 信息服务的特点是可在网上进行多媒体信息的收集、分类、存放、发布和交流,并向网上的用户提供信息检索及其他交互式服务,在页面设计上具有结构合理、可读性强、用户界面友好、及时性强和使用费用低的优点。目前,许多数据库产品都有 Internet 在线产品。

二、数据库

数据库是以特定方式合理地组织相互关联的数据集合。数据库以大容量的计算机为存储介质,它是计算机技术和信息检索技术相结合的产物。数据库是电子信息资源的最主要载体,它是信息检索和管理系统的核心部分。

(一) 数据库结构

1. 记录型文献数据结构

记录是组成文档的基本数据单位。一个记录又包含若干字段。字段是组成记录的基本数据单位。在书目数据库中,一条记录相当于一篇参考文献,常用的字段有:标题、作者、地址、期刊名、学科、文件类型、语种和摘要等。在有些字段中,又包含多个子字段,子字段是字段的下级数据单位。文档是按一定结构组织的相关记录的集合。文档是书目数据库数据组织的基本形式,文档的组织方式与检索系统的硬件和软件功能密切相关。表 1-1 是具有 6 条记录和 6 个字段的文档,其字段分别是:记录号、标题、作者、地址、期刊和年/期。每条记录中信息量的多少和数据库中记录总数的数量是评价数据库的重要指标。不同的数据库,其记录中的字段数量、类型和长度是不同的。



表 1-1 具有 6 条记录和 6 个字段的文档

记录号	标 题	作 者	地 址	期 刊	年/期
0001	化学信息学发展现状	陈泓	四川理工学院化工系	化学研究与应用	2004/04
0002	化学信息学——药学教育的“新基石”	孟宪伟	哈尔滨医科大学分校	当代医学	2004/02
0003	漫谈化学信息学的产生与发展	彭彤	山东大学化学与化工学院	山东化工	2002/06
0004	化学信息学网络考试系统的构建	罗海彬	厦门大学化学化工学院	大学化学	2002/04
0005	化学信息学及其课程建设	邵学广	中国科学技术大学化学系	大学化学	2002/03
0006	化学信息学的涵义及教育	徐筱杰	北京大学化学与分子工程学院	大学化学	2002/01

2. 记录的排序和索引

为了提高记录的检索速度,经常会使用排序和索引技术。在书目数据库中,文档结构主要分为顺排文档和倒排文档。

(1) 顺排文档:将其记录按记录号顺序排列,记录之间的逻辑顺序与物理顺序是一致的,如表 1-1 所示。顺排文档是构成数据库的主体部分,但其他字段排列呈无序状态。进行检索时,必须以完整的记录作为检索单元,从头至尾逐条查询,检索时间长,实用性较差。

(2) 倒排文档:将顺排文档中各个记录按照某一字段的值进行排列(如标题字段、作者字段和地址字段等)。使用标题字段进行倒排的文档,称为标题索引文档。索引是一种加快检索速度的有效手段,它不需要移动数据库中记录而占用大量的存储空间。计算机进行检索时,通过索引检索查找有关信息的对应记录号,然后再进入顺排文档按记录号查找完整的记录。

(二) 数据库的类型

按所提供的化学信息内容,数据库主要可分为文献数据库、事实数据库和结构数据库等类型。

1. 文献数据库

文献数据库是存储参考文献信息的专用数据库,用户在这些数据库中获取信息线索后,还需要进一步查找原文或其他资料。文献数据库包括书目数据库、全文数据库和专利数据库。

(1) 书目数据库只存储参考文献的书目信息,书目信息有:标题、作者、地址、期刊名、学科、文件类型、语种和摘要等。题录数据库和文摘数据库等属于文献数据库,题录数据库和文摘数据库提供的信息内容与印刷型的题录、文摘相似。图书馆的目录数据库,除了具有标题、作者、出版项等书目信息外,还提供用户索取原始信息的馆藏信息。书目数据库提供了参考文献的确定信息来源,为用户指出了获取原始信息的线索。当然通过对文献摘要的浏览,用户可以先对文献有基本的了解和认识。书目数据库漏检率较高,要求用户具有较好的检索技术,还要另行获取原始信息,但其具有检索速度快、成本低的优点,目前仍然是文献数据库的主流。本书的第三章介绍了常用文摘数据库的检索。

(2) 全文数据库是指存储文献全文或主要部分内容的一种数据库。全文数据库将全部文本内容都存储到计算机,构建成由篇、段、句、词和字组成的能被计算机识别和处理的文本。它不像记录型数据库那样由字段存储记录。在全文数据库中进行信息检索称为全文检索。全文检索可以根据需要检索出全文中有关章、节、段、句、词和类等等的信息,也可进行各种统计和分析。早期的计算机文本检索系统,只能对有限的关键词进行搜索,而现在已可对全部文本数据进行快速准确的检索。相对于字段检索,全文检索需要更复杂的处理技术。

① 在检索系统中,词语是表达信息的最小单位。在英语的语句中,词语间有空格作为分隔符,识别词语比较容易。而汉语的词语间无分隔符,词语切分存在二义性。词语切分是中文全文检索的一项特别技术。词语切分要正确地将语句切分为可以识别的字符串(单词、数字、日期和时间等)。为了避免切分的二义性,进行切分操作时,需要利用上下文知识进行处理。

② 短语是词的组合。词的不同组合会具有不同的含义，词的涵盖范围不同也会具有不同的意思。因此，短语的识别有助于提高信息检索的精度。不能正确地识别短语，就可能造成误检或漏检。

③ 在英语中，同一词干的词语代表相同的概念。对词语进行分析，找出各单词的词干，进行基于词干的信息索引，可增加查全率。

全文检索的明显优点是信息全面、漏检率低，但检索速度慢、数据库开发的成本高。本书的第四章介绍通用的全文数据库，第五章介绍的大部分电子期刊的检索也是全文检索，不过只对该电子期刊有效。

(3) 专利数据库是专门存储专利信息数据库。专利是知识产权的重要组成部分，快速准确地掌握专利的最新动态，对科学研究、产品开发等工作是至关重要的，它既可以避免重复劳动、提高研发水平和开发效率，同时对知识产权的确认、保护、合理利用和开发也有很大的促进作用。使用专利数据库进行专利资源检索与传统的通过印刷型检索工具相比，效率要高得多，不受时间、地点限制，也节省了许多费用。目前一些主要的专利数据库都可以通过 Internet 免费检索，相对于文献检索，专利数据库基本是免费的，专利数据库正在成为专利检索的重要途径。专利数据库提供专利的一些主要信息，如专利号、发明者、申请人、关键词或摘要等，从而使用户更容易地筛选所需专利和就近获取专利原文。本书的第六章介绍专利数据库的检索。

2. 事实数据库

事实数据库主要用于存储和检索化合物的原始数据。它是能够直接为用户提供原始资料或具体数据的一种数据库。事实数据库表示的信息以数值为主。与书目数据库比较，数值数据库是对信息进行深加工的产物，可以直接提供所需的数据信息。数值数据库除了一般的检索功能外，还具有准确数据运算功能、数据分析功能、图形处理功能及对检索输出的数据进行排序和重新组织等方面的功能。事实数据库主要有以下几种。

(1) 数值数据库是指专门提供以数值方式表示的一种数据库。它包含化合物的数值数据，例如，物理化学数值和测量值等。此外，它还带有参考文献或提供获取原文的超链接。典型的数值数据库有：Beilstein/Gmelin、DETERM 和 WebBook 等。

(2) 光谱数据库是专门收集和利用光谱的数据库，可以进行光谱数据检索、结构解析和谱图预测等功能。例如，SpecInfo。

(3) 化合物目录数据库是电子版的化合物目录。例如，Chemline、MRCK 等。

(4) 研究计划数据库收集了不同学科领域的研究报告和摘要。例如，UFORDAT、FEDRIP 等。

3. 结构数据库

(1) 化学结构数据库中存储了化合物的化学结构信息，它包括分子中各原子的三维坐标及构成化学键的连接形式。晶体结构具有准确的分子结构信息（晶胞参数和坐标），以及分子间相互作用的信息，使用分子图形软件可将分子结构信息可视化。著名的结构数据库有：ICSD、CSD 和 PDB 等。

(2) 化学反应数据库是提供单步、多步化学反应的反应物、产物和反应条件等信息。内容包括化学反应的类型、反应机理、反应活性、反应的交叉索引及反应的理化参数等。著名的 Beilstein/Gmelin 包含了大量的化学反应信息，这些反应信息也基本上与化学结构有关。

三、在线化学数据库

在线数据库是可以通过 Internet 进行实时检索的数据库。由于数据库的数据变化很快，难以统计完全和最新的数据库信息。以下简单介绍化学和生物化学最重要和最有用的在线数据库。