



普通高等教育“十二五”规划教材



化学信息学 (第三版)

邵学广 蔡文生 编著



科学出版社

013027829

06-05
05-3

普通高等教育“十二五”规划教材
高等学校化学类专业规划教材·名校名师系列

化学信息学

(第三版)

邵学广 蔡文生 编著



科学出版社

北京



北航

C1637096

06-05
05-3

内 容 简 介

本书介绍了 Internet 上化学资源的使用方法,并对化学信息学方法及其在化学、生物化学、药物化学等领域中的应用进行了详细论述。本书共 10 章,包括联机文献检索、网络图书与网络期刊、数据库资源、化学信息资源查询、化学信息的计算机表示与建模、计算机辅助结构解析与合成设计、分子模拟、进化计算与优化算法、小波分析、多元校正与因子分析。

本书可供高等学校化学、化工、生物化学、药物化学以及相关专业的师生和广大科技工作者参考阅读。

图书在版编目(CIP)数据

化学信息学/邵学广,蔡文生编著. —3 版. —北京:科学出版社,2013. 3

普通高等教育“十二五”规划教材 高等学校化学类专业规划教材·名校名师系列

ISBN 978-7-03-036904-8

I. ①化… II. ①邵… ②蔡… III. ①计算机应用-化学-情报检索-高等学校-教材 IV. ①G252. 7

中国版本图书馆 CIP 数据核字(2013)第 041494 号

责任编辑:丁 里 / 责任校对:李 影

责任印制:阎 磊 / 封面设计:迷底书装

科学出版社 出版

北京东黄城根北街 16 号

邮政编码:100717

<http://www.sciencep.com>

天津新科印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2001 年 6 月第 一 版 开本:787×1092 1/16

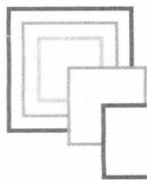
2005 年 4 月第 二 版 印张:17 1/2

2013 年 3 月第 三 版 字数:445 000

2013 年 3 月第五次印刷

定价:49.00 元

(如有印装质量问题,我社负责调换)



第三版前言

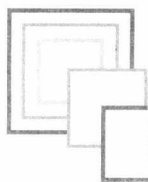
《化学信息学》(第一版)自 2001 年夏出版以来,深受广大读者的欢迎,同时,也收到读者对本书的一些宝贵意见。因此,于 2005 年初编写了《化学信息学》(第二版),删减了计算机和计算机网络部分的内容,但丰富了化学计量学方法的相关内容。化学信息学在过去的几年中得到了长足进步和发展,Internet 上的化学资源迅速增加并日趋完善,新的化学信息学方法不断被提出并在化学领域中得到了新的应用。因此,我们在第二版的基础上编写了《化学信息学》(第三版)。

第三版在内容和章节安排上进行了调整。首先,删除了第二版的第 1 章,即与计算机和 Internet 基本知识的相关内容。第二版的第 2~5 章没有进行大的调整,作为第三版的第 1~4 章,分别对联机文献检索、网络图书与网络期刊、数据库资源和化学信息资源查询进行介绍。但是,随着 Internet 服务方式和服务内容的改变,对其中的具体内容进行了更新。第 5~10 章主要介绍化学信息学方法与应用。第二版的第 6 章分别变为第三版的第 5 章和第 6 章,分别介绍化学信息的计算机表示与建模和化学信息在结构解析和合成设计中的应用,并增加了分子描述符和定量构效关系的相关内容。对第二版的第 7 章在内容上进行了扩充,系统介绍了量子力学、分子力学和分子动力学模拟方法与应用,形成了第三版的第 7 章。为了增加系统性,将第二版的第 9、10、12 章合并形成了第三版的第 8 章,并增加了新的内容。第二版第 8 章和第 11 章分别调整为第三版的第 9 章和第 10 章,并适当增加了多元校正方法与应用的部分内容。另外,在第三版中还增加了一些应用实例,目的是指导读者直接参考本书即可开展实际工作。

由于作者水平所限,书中难免存在疏漏和不妥之处,敬请读者批评指正。

作 者

2012 年 12 月于南开大学



第二版前言

《化学信息学(第一版)》自 2001 年夏出版以来,深受广大读者的欢迎,已经进行了两次印刷。化学信息学是近几年发展起来的新兴学科,发展非常迅速,特别是 Internet 服务和 Internet 资源方面的内容,不仅日趋完善而且变化很快。同时,也收到读者对本书的一些宝贵意见。因此,我们在第一版书稿的基础上编写了《化学信息学(第二版)》。

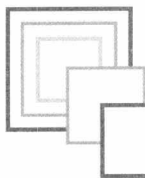
第二版在内容以及章节安排上都进行了较大的调整。首先,随着计算机知识和使用技术的普及,将第一版第一章的内容进行了删减,并与第二章合并组成了第二版的第 1 章。由于 Internet 中的化学资源发展较快,其内容不断丰富,因此,在第二版中将第一版第三章的内容分解为第二版第 2~5 章,分别对联机文献检索、网络图书与网络杂志、数据库资源和化学信息资源查询进行了介绍。这样,当本课程的学时数较少时,只对本书的第 1~5 章进行教学即可,而后面的内容则可以安排在研究生阶段学习。

第一版的第四、五两章基本没有改动,分别变为第二版的第 6、7 两章。但对近几年发展较快的化学计量学方面的内容(第一版第六章)进行了较大变动,分解为第二版第 8~12 章,分别对多元校正与因子分析、人工神经网络、遗传算法和模拟退火、小波分析和免疫算法进行了介绍。在第二版中,每一部分都增加了新的内容或新的知识,在某些新的方法中还增加了程序设计的有关内容,目的是指导读者直接参考本书即可开展实际工作。

由于作者水平所限,书中难免存在缺点和错误,敬请读者批评指正。

作 者

2005 年 1 月于中国科学技术大学



第一版前言

自从化学学科出现以来,信息的记载、组织与交流对化学学科的发展起了重要的作用,同时也成为化学学科的一个重要组成部分。这是因为化学实验的记录资料具有长远的实践意义。在化学学科中,化学家根据百年以前的记录资料从事科学实验的例子并不稀奇。另外,化学物质结构的记录与检索需要建立独特的记录系统。随着计算机技术的发展,化学家必须建立自己的信息表示、记录与管理以适应时代的要求。

化学信息可分为两大组成部分,即化学物质的化学信息和媒体形式的化学信息。前者是利用科学的原理和方法通过测量得到的化学成分的相关信息,如物质的物理、化学性质,物质中各成分的定性、定量以及结构信息,分子间的相互作用信息(包括化学反应信息)等。后者是化学信息的记录形式,如图书、期刊、专利、数据库以及音像资料等。化学信息的传播使化学家们共享测量的原理、方法及测量结果。

化学信息学(chemoinformatics)是近几年发展起来的一个新的化学分支,它利用计算机技术和计算机网络技术,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,揭示化学信息的实质与内在联系,促进化学学科的知识创新。化学信息学的研究内容主要包括:(1)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计算机管理;(2)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;(3)利用计算机对化学信息和化学体系进行模拟;(4)收集、传播和共享化学信息。

在本书的编写过程中,参考了大量的 Internet 资源及有关参考书,在此谨对为本书提供帮助的网站和作者表示衷心的感谢。由于作者水平所限,书中难免存在缺点和错误,敬请读者批评指正。

作 者

2001 年 2 月于中国科学技术大学



第三版前言

第二版前言

第一版前言

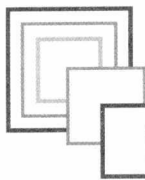
第1章 联机文献检索	1
1.1 美国化学文摘	1
1.1.1 美国化学文摘和 SciFinder 简介	1
1.1.2 SciFinder 的著录内容与格式	2
1.1.3 SciFinder 的功能与使用方法	6
1.2 Web of Science	11
1.2.1 Web of Science 简介	11
1.2.2 Web of Science 的著录内容与格式	12
1.2.3 Web of Science 的功能与使用方法	13
1.2.4 ISI Web of Knowledge 的其他功能	18
1.3 Ei Village	19
1.3.1 Ei Village 简介	19
1.3.2 Ei Village 的著录内容与格式	20
1.3.3 Ei Village 的功能与使用方法	21
1.4 期刊全文数据库	24
1.4.1 ScienceDirect	24
1.4.2 ACS Publications	27
1.4.3 RSC Publishing	29
1.4.4 SpringerLink	30
1.4.5 Wiley Online Library	31
1.5 国内学术期刊数据库	32
1.5.1 CNKI	32
1.5.2 中文科技期刊数据库	34
1.5.3 万方数据知识服务平台	35
1.6 其他文献检索系统简介	36
1.7 专利文献查询	39
1.7.1 专利文献数据库	39
1.7.2 专利文献检索	40
上机训练与习题	45
第2章 网络图书与网络期刊	46
2.1 网上图书馆	46

2.1.1 图书馆主页	46
2.1.2 WWW 虚拟图书馆	47
2.2 网上书店	49
2.3 网络期刊	50
2.4 数字化图书	52
2.5 出版商主页	53
上机训练与习题	54
第3章 数据库资源	55
3.1 晶体结构数据库	55
3.1.1 剑桥结构数据库	55
3.1.2 蛋白质数据库	57
3.1.3 核酸数据库	58
3.2 波谱数据库	61
3.2.1 NIST Chemistry WebBook	61
3.2.2 化合物谱图数据库	64
3.3 网上化学手册	67
3.3.1 化学元素周期表	67
3.3.2 化合物基本性质数据库	69
3.3.3 物理化学常数	70
上机训练与习题	71
第4章 化学信息资源查询	72
4.1 查询	72
4.2 导航	77
4.3 ChIN 简介	78
4.4 商业信息	81
4.5 化学机构信息	83
上机训练与习题	84
第5章 化学信息的计算机表示与建模	85
5.1 化合物结构编码	85
5.1.1 线型编码	85
5.1.2 ZXLN 线型编码	86
5.1.3 内部表示	87
5.2 分子图形学	88
5.3 分子描述符	93
5.4 定量构效关系	95
5.4.1 基本方法	95
5.4.2 应用举例	96
上机训练与习题	98
第6章 计算机辅助结构解析与合成设计	100
6.1 化学信息数据库	100

6.1.1 数据库系统的构成	100
6.1.2 数据模型和数据库的类型	101
6.1.3 数据库的存取过程	101
6.1.4 化学信息数据库	102
6.2 化学人工智能与专家系统	105
6.2.1 人工智能	105
6.2.2 专家系统	105
6.2.3 知识库	106
6.2.4 推理方法	109
6.3 计算机辅助结构解析	111
6.3.1 结构解析系统的构成	111
6.3.2 结构解析的工作过程	112
6.3.3 结构解析中的有关技术	112
6.4 计算机辅助合成设计	114
6.4.1 计算机辅助有机合成	114
6.4.2 计算机辅助分子设计	121
上机训练与习题	124
第7章 分子模拟	126
7.1 量子力学	126
7.1.1 基本理论	126
7.1.2 Gaussian 软件	129
7.1.3 应用举例	133
7.2 分子力学	136
7.2.1 分子力场	136
7.2.2 应用举例	142
7.3 分子动力学	144
7.3.1 基本方法	144
7.3.2 软件	145
7.3.3 应用举例	148
7.4 波谱模拟	153
7.4.1 ^1H -NMR 谱的量子化学模拟	153
7.4.2 ^{13}C -NMR 谱的计算机模拟	156
上机训练与习题	161
第8章 进化计算与优化算法	162
8.1 人工神经网络	162
8.1.1 模型	162
8.1.2 学习算法	164
8.1.3 应用举例	167
8.2 遗传算法	168
8.2.1 自然进化与遗传算法	169

8.2.2 基本过程	170
8.2.3 遗传算法的发展	172
8.2.4 应用举例	175
8.3 模拟退火算法	178
8.3.1 固体退火与模拟退火算法	178
8.3.2 基本过程	180
8.3.3 模拟退火算法的发展	182
8.3.4 退火演化算法	183
8.3.5 应用举例	187
8.4 免疫算法	191
8.4.1 基本原理	191
8.4.2 免疫优化算法	192
8.4.3 一种用于重叠信号解析的免疫算法	195
8.5 其他优化算法简介	199
8.5.1 蚁群算法	199
8.5.2 粒子群优化算法	201
上机训练与习题	203
第9章 小波分析	205
9.1 小波及小波分析	205
9.1.1 小波的定义	205
9.1.2 傅里叶变换	206
9.1.3 小波变换	208
9.2 小波分析的基本算法	209
9.2.1 多尺度分析	209
9.2.2 多尺度信号分解(MRSD)算法	210
9.2.3 MRSD 算法的改进	212
9.2.4 小波包变换	213
9.3 小波分析的程序设计	214
9.3.1 Matlab 工具箱	214
9.3.2 WaveLab 简介	215
9.3.3 MRSD 算法的程序设计	216
9.3.4 连续小波变换的程序设计	219
9.3.5 小波包分析的程序设计	221
9.4 小波分析的应用	222
9.4.1 数据压缩	222
9.4.2 平滑和滤噪	224
9.4.3 背景扣除与基线矫正	227
9.4.4 近似导数的计算	228
9.4.5 重叠信号解析	230
9.4.6 谱图分辨率的改善	231

9.4.7 小波分析的其他应用	232
上机训练与习题	233
第 10 章 多元校正与因子分析	235
10.1 引言	235
10.1.1 化学计量学	235
10.1.2 多元校正与分辨	236
10.1.3 因子分析	236
10.2 数据矩阵的构成	237
10.2.1 二维色谱数据	237
10.2.2 三维荧光光谱数据	238
10.2.3 多组分光度分析数据	238
10.2.4 配合物体系的研究	239
10.3 间接校正方法	239
10.3.1 K-矩阵法	240
10.3.2 P-矩阵法	242
10.4 主成分分析	243
10.4.1 原理	243
10.4.2 应用举例	245
10.5 主成分回归	246
10.6 偏最小二乘回归	249
10.6.1 原理	249
10.6.2 算法	250
10.6.3 应用举例	251
10.7 化学因子分析	252
10.7.1 基本步骤	252
10.7.2 目标因子分析	253
10.7.3 秩消因子分析	256
10.7.4 渐进因子分析	256
10.7.5 窗口因子分析	259
10.7.6 启发渐进式特征投影	261
10.8 高阶校正方法简介	264
上机训练与习题	267
主要参考文献	268



第 1 章 联机文献检索

通过 Internet 进行文献联机检索是指通过联机方式,根据用户提供的信息(关键词、作者等)给出相关的文献信息,如论文题目、期刊名称、卷、页、摘要甚至全文。目前已有许多文献检索系统,如 Web of Science、美国化学文摘(CA)以及一些专业性更强的专业文献系统。

访问这些文献系统一般有两种方式,即远程登录和 WWW 浏览,目前主要以 WWW 方式为主。对于免费系统,可从 Internet 或通过其他途径得到有关信息,自由地使用;但对于收费系统,只有在交费后得到有关的账号和密码才能获得使用权限。目前一般是由所在单位统一订阅供本单位的用户使用。

1.1 美国化学文摘

1.1.1 美国化学文摘和 SciFinder 简介

美国化学文摘(*Chemical Abstract*, 简称 CA)是由美国化学会(American Chemical Society)化学文摘服务社(*Chemical Abstract Service*, 简称 CAS)于 1907 年创办的文摘系统。开始以印刷版出版发行,1967 年以前,每半月一期,每年一卷,后来改为每周一期,每年两卷。随着计算机和 Internet 的发展,逐步形成了光盘版(CA on CD)和网络版。

CA 的主要特点是它的文摘详细、客观地报道化学化工文献,文摘质量高,不加任何评论,报道的内容包括文献的研究目的和范围,新化学反应、化合物、材料、工艺、程序、工具和资源,新知识的应用,以及观察的结果和作者的解释与结论;收录的范围广泛,系统、全面地收录世界上化学化工方面 98% 的文献,其中 70% 的文献来自美国以外的国家和地区,共收录 150 多个国家或国际组织的 56 种文字出版的 14 000 余种出版物,包括期刊文献、会议文献、专利文献、学位论文、图书文献和科技报告等;另外,CA 的使用非常方便,在印刷版中有各种卷索引、五年累积索引、十年累积索引、八十年累积索引等。CA 的全部文献分为 5 部分 80 类,其中生物化学部分为 1~20 类,有机化学部分为 21~34 类,大分子部分为 35~46 类,应用化学与化学工程部分为 47~64 类,物理化学、无机化学和分析化学部分为 65~80 类。所收录的文摘按类别编排,每一个文摘拥有一个文摘号。

CA 的电子出版物始于 1969 年,当时的计算机发展水平还比较落后,采用磁带作为记录载体,但实现了化学文献的自动、高效检索,推动了文献检索的发展。随着计算机水平的提高,采用了光盘作为记录材料,形成了 CA 的光盘版,即 CA on CD。CA 的网络版则是随着 Internet 的发展而建立起来的新的文献检索方式,先后开发了基于客户端软件的 SciFinder Scholar 和基于网络访问的 SciFinder Web 版。2012 年底,SciFinder Scholar 的客户端服务已经停止,目前主要是通过 SciFinder Web 版进行服务。光盘版和网络版的使用给化学文献的检索提供了更为方便的方式,不仅检索方式灵活多样,更重要的是可以

系统、全面地进行检索。

SciFinder Web 版整合了 Medline 医学数据库、欧洲和美国等 50 多家专利机构的全文专利资料以及化学文摘 1907 年至今的所有内容。它涵盖的学科包括应用化学、化学工程、普通化学、物理、生物学、生命科学、医学、聚合体学、材料学、地质学、食品科学和农学等诸多领域。它可以透过网络直接查看 CA 1907 年以来的所有期刊文献和专利摘要以及七千多万的化学物质记录和 CAS 注册号。

SciFinder 可检索的数据库包括:

CAPLUS(3600 万条参考书目记录, 包含 185 个国家以及 50 多种语言的参考文献, 始自 1907 年)

CAS REGISTRY(6800 万种独一无二的有机和无机物质, 每天更新约 15 000 条, 每种化学物质有唯一对应的 CAS 注册号, 始自 1800 年)

CASREACT(4710 万条反应记录, 每周更新约 150 000 条, 始自 1840 年)

CHEMCATS(7000 万条商业化学物质记录, 来自 990 多家供应商的 1125 多种目录)

CHEMLIST(29.6 万种管制品化合物的详细清单, 每周更新约 50 条)

CIN(169 万条最新商业新闻的详细记录, 始自 1974 年)

MARPAT(93.9 万种已公开的 Markush 结构, 超过 385 000 条包含 Markush 结构专利的参考文献, 每周更新 60~75 条参考专利及 150~200 个 Markush 结构, 始自 1988 年)

SciFinder 具有多种先进的检索方式, 如化学物质、化学结构式、Markush 结构式和化学反应式检索等。

1.1.2 SciFinder 的著录内容与格式

1. 期刊文献的著录内容与格式

期刊文献的著录内容如图 1-1 所示, 主要包括上部的链接部分、中部左侧的文献信息、中部右侧的期刊和录入信息以及下部的文献引用信息。

顶部: 涉及的物质链接(Get Substances)、涉及的反应链接(Get Reaction)、引用文献链接(Get Cited)、引用本文献的文献链接(Get Citing)、全文链接(Full Text)等。

左侧: 文献标题、作者姓名、摘要内容、主题名称(CA 分类号)、涉及的主要概念(Concepts)、涉及的主要物质(Substances)等。

右侧: 文献来源、出版物类型及出版时间、CODEN 代码、ISSN 编号、DOI 编号、作者所在机构及地址、登录号(Accession Number)、出版商、语种。某些文献的著录内容中还包括 CA 交叉参考(cross-reference)分类号。

下部(图中未显示): 附加术语(Supplementary Terms)和施引文献(Citations)。在施引文献清单中还提供了超级链接, 可以直接查看施引的文献。这个功能对实现文献之间的交叉检索具有重要意义。

2. 图书文献的著录内容与格式

图书文献的著录内容如图 1-2 所示。内容分左右两部分显示。

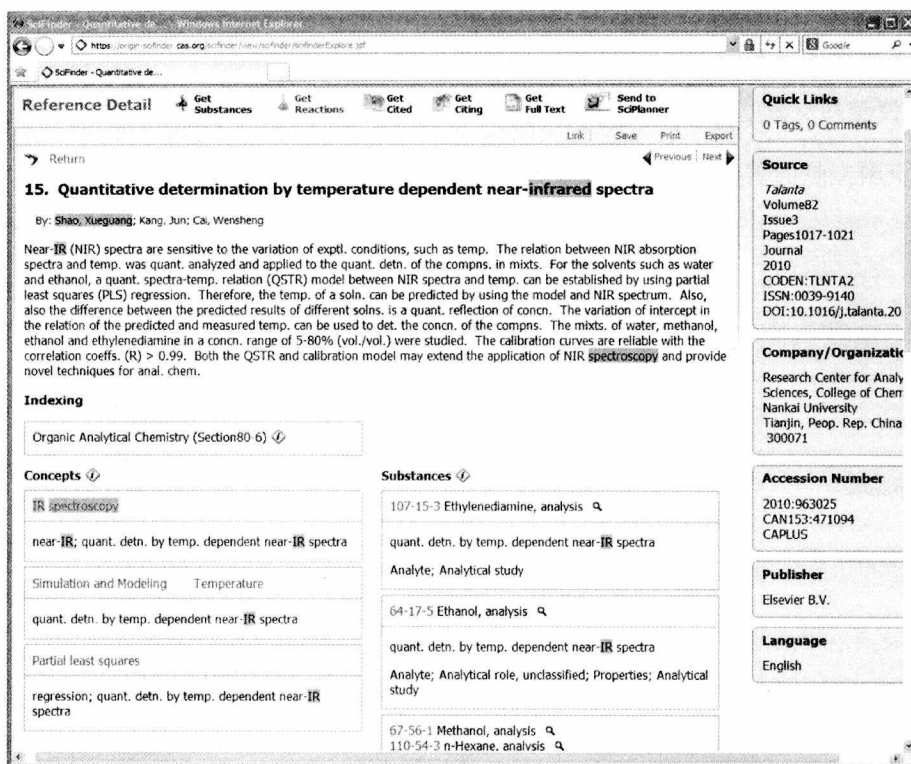


图 1-1 期刊文献的著录结果的内容与格式

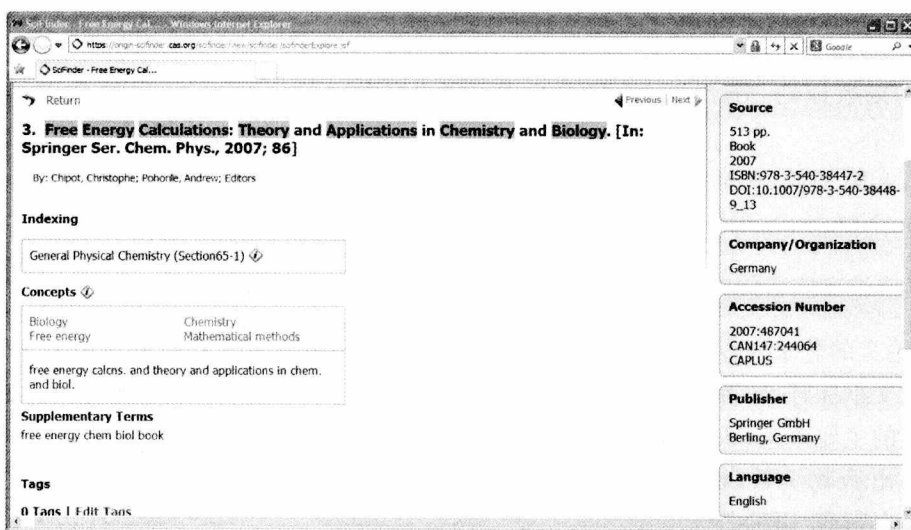


图 1-2 图书文献的著录结果的内容与格式

左侧:文献标题、作者/编辑姓名、主题名称(CA分类号)、涉及的主要概念(Concepts)、附加术语(Supplementary Terms)。

右侧:出版物页数、出版物类型及出版时间、ISBN 编号、DOI 编号、作者所在机构及地址、登录号(Accession Number)、出版商、语种。

在某些图书文献的著录内容中还包括原文语种的书名的拉丁译音、定价等。另外,对于图书光盘和多媒体形式的图书,在著录格式上稍有不同。

3. 会议资料文献的著录内容与格式

会议资料文献的著录内容如图 1-3 所示。内容分左右两部分显示。

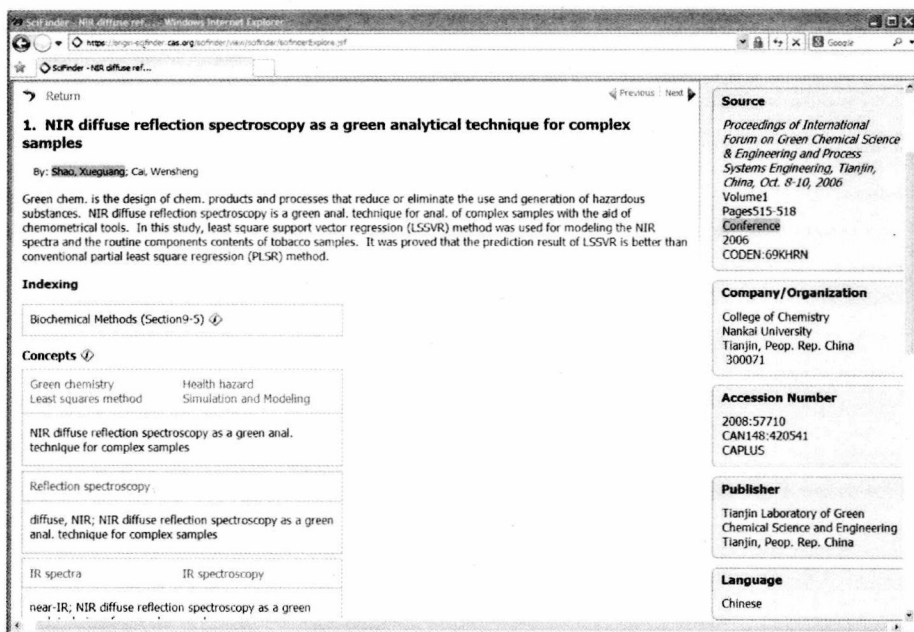


图 1-3 会议资料文献的著录结果的内容与格式

左侧:文献标题、作者姓名、摘要内容、主题名称(CA 分类号)、所涉及的主要概念(Concepts)等。

右侧:文献来源、出版物类型及出版时间、CODEN 代码、作者所在机构及地址、登录号(Accession Number)、出版商、语种。

4. 学位论文的著录内容与格式

学位论文的著录内容如图 1-4 所示。内容分左右两部分显示。

左侧:文献标题、作者姓名、主题名称(CA 分类号)及 CA 交叉参考(cross-reference)分类号、所涉及的主要概念(Concepts)等。

右侧:出版物页数、出版物类型及出版时间、作者所在机构及地址、登录号(Accession Number)、出版商、语种。

5. 专利文献的著录内容与格式

专利文献的著录内容如图 1-5 所示。内容分左右两部分显示。

左侧:文献标题、作者姓名、代理人(Assignee)、摘要内容、专利信息、主题名称(CA 分类号)以及 CA 交叉参考(cross-reference)分类号、所涉及的主要概念(Concepts)、所涉及的主要物质(Substances)等。

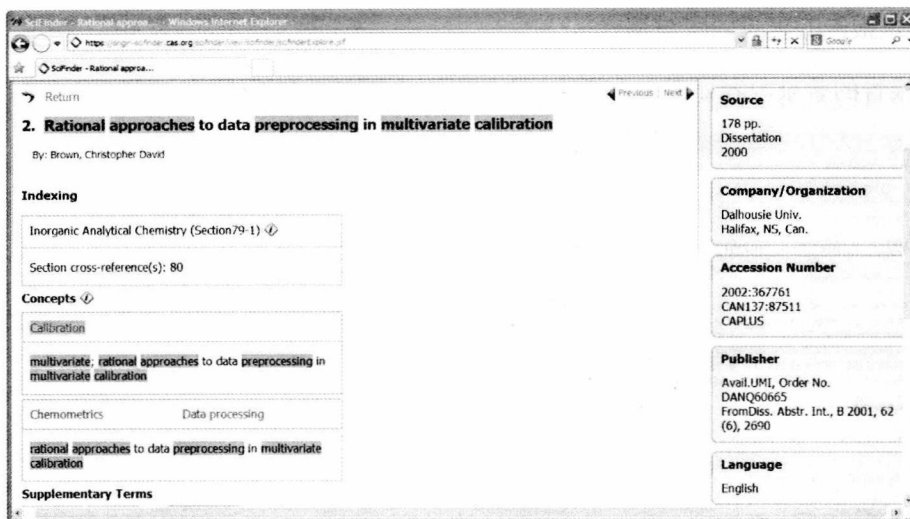


图 1-4 学位论文的著录结果的内容与格式

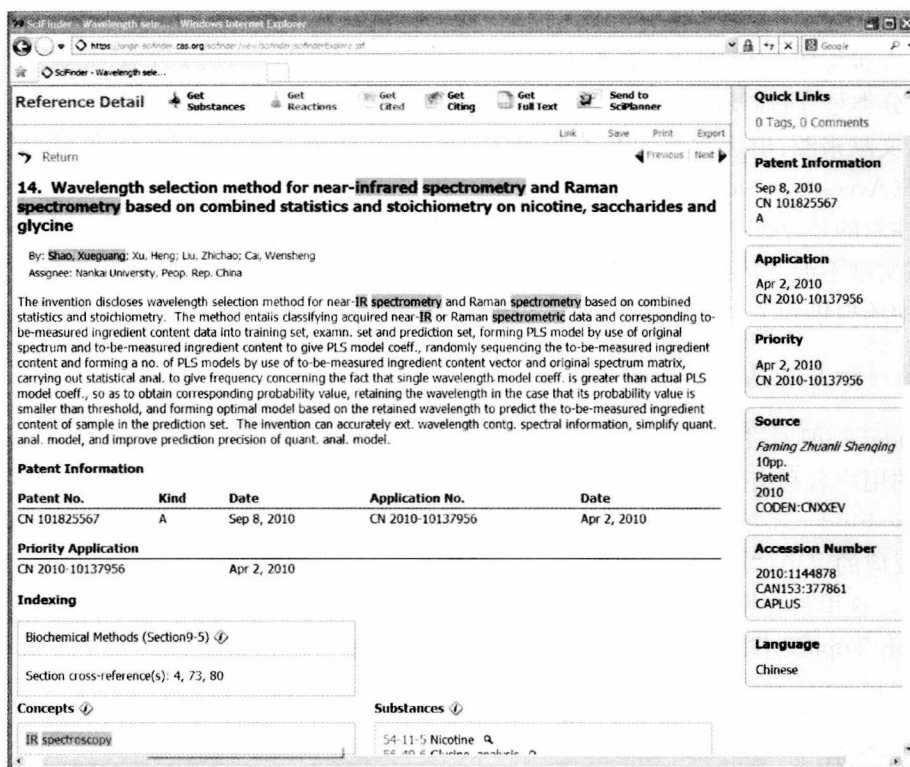


图 1-5 专利文献的著录结果的内容与格式

右侧: 专利公开日期、编号和种类、申请日期以及编号、优先权申请编号以及日期、出版物来源、出版物页数、出版物类型及出版时间、CODEN 代码、登录号 (Accession Number)、语种。

6. 技术报告的著录内容与格式

技术报告的著录内容如图 1-6 所示。内容分左右两部分显示。

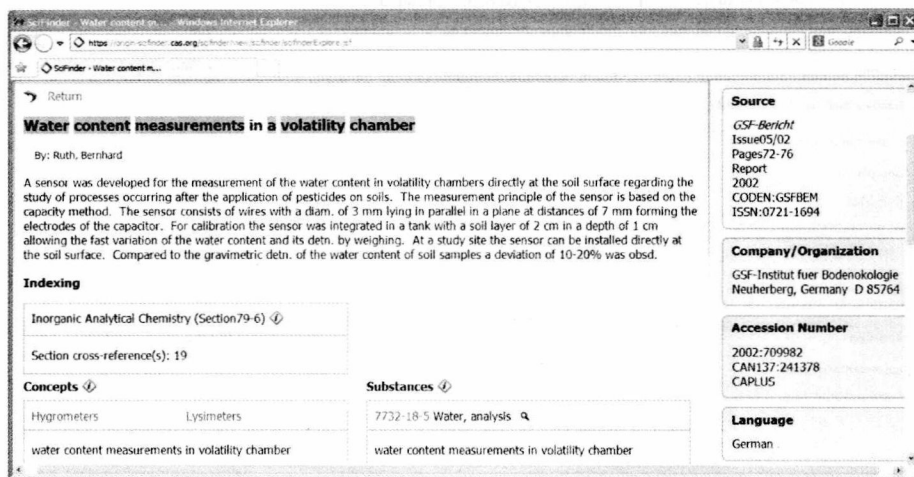


图 1-6 技术报告的著录结果的内容与格式

左侧:文献标题、作者姓名、摘要内容、主题名称(CA 分类号)以及 CA 交叉参考(cross-reference)分类号、所涉及的主要概念(Concepts)、所涉及的主要物质(Substances)等。

右侧:文献来源、出版物类型及出版时间、CODEN 代码、ISSN 编号、作者所在机构及地址、登录号(Accession Number);语种。

值得注意的是,SciFinder 的著录内容并不是固定不变的,只显示文献中涉及的内容。例如,如果该文献不涉及特定的化合物,则不显示“涉及的主要物质(Substances)”栏目。在某些文摘内容中还提供相关文献的超级链接,以便直接访问。

1.1.3 SciFinder 的功能与使用方法

SciFinder 一般通过订阅单位提供的网址进入。在使用之前必须先注册个人账号,然后以个人账号的用户名和密码登录。首次登录时需要同意该系统的一些使用规范。

进入 SciFinder 系统以后,显示如图 1-7 所示的页面。该系统提供了检索文献、化学物质以及化学反应的渠道(“Explore”),即“Explore References”、“Explore Substances”和“Explore Reactions”。这里只介绍文献检索(“Explore References”)部分,共有 6 种检索方式,即主题检索(Research Topic)、作者检索(Author Name)、公司检索(Company Name)、文献识别码检索(Document Identifier)、期刊检索(Journal)和专利检索(Patent)。

1. 主题检索

SciFinder 的默认页面是主题检索(Research Topic)界面。本界面有六个部分。其中,“Research Topic”后的文本框用于填写所要检索的主题,“Publication Year(s)”文本框用于填写文献发表的年份,“Document Type(s)”多选框用于选择文献的类型,“Language(s)”多选框用于选择文献的语言种类,“Author Name”文本框用于填写作者的姓、名和中间名,“Company Name”后的文本框用于填写公司名称。当任意一个或多个检索条件确定后,即可点击“Search”按钮进行相关文献的检索。