

■ 高等学校理工科化学化工类规划教材

化学化工信息检索与利用

INFORMATION RETRIEVAL AND USE OF CHEMISTRY AND CHEMICAL ENGINEERING

(第三版)

余向春 主编



大连理工大学出版社
DALIAN UNIVERSITY OF TECHNOLOGY PRESS

■ 高等学校理工科化学化工类规划教材

第3版(911)目錄表封面

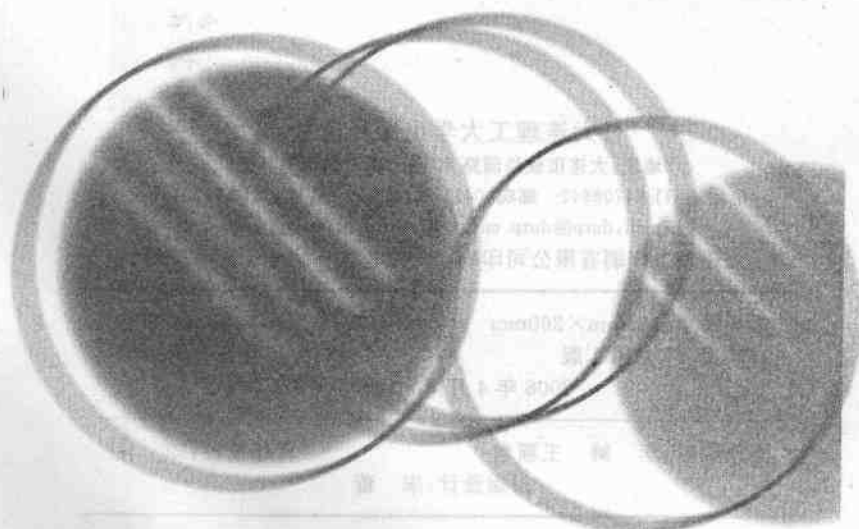
化学化工信息检索与利用

INFORMATION RETRIEVAL AND USE OF CHEMISTRY AND CHEMICAL ENGINEERING

(第三版)

主编 余向春

编著 叶 鹰 黄文林



大连理工大学出版社
DALIAN UNIVERSITY OF TECHNOLOGY PRESS

图书在版编目(CIP)数据

化学化工信息检索与利用/余向春主编.—3版.—大连:
大连理工大学出版社,2008.4
ISBN 978-7-5611-3943-1

I. 化… II. 余… III. ①化学—情报检索②化学工业—
情报检索 IV. G252.7

中国版本图书馆 CIP 数据核字(2008)第 034712 号

大连理工大学出版社出版

地址:大连市软件园路 80 号 邮政编码:116023

电话:0411-84708842 邮购:0411-84703636 传真:0411-84701466

E-mail: dulp@dulp.cn URL: <http://www.dulp.cn>

大连理工印刷有限公司印刷 大连理工大学出版社发行

幅面尺寸:185mm×260mm 印张:15.75 字数:348千字
1990年10月第1版 2008年4月第3版
2008年4月第20次印刷

责任编辑:李 鸽 王颖鑫

责任校对:知 轩

封面设计:宋 蕾

ISBN 978-7-5611-3943-1

定 价:26.00 元

前 言

随着科学技术的不断发展和创新,特别是计算机和网络信息技术的不断进步,目前各种各样的信息资源正在以几何级数的速度逐年递增。作为一名21世纪的大学生、高校教师及科研工作者,如何在浩如烟海的信息资源中,准确、快捷、全面地找出自己真正需要的信息,是必须面对并且需要很好解决的问题之一。本书正是为了很好地解决这一问题,针对化学化工这一专业领域而编写的。

事实上,本书的第1版《化学化工文献检索与利用》作为文献检索与利用课系列教材之一,于1990年10月出版,并于1997年2月进行修订。从20世纪90年代以来,靠手工检索方式已难以全面适应信息社会发展的需要。计算机检索和基于互联网的网络化信息检索,已逐渐成为当今信息检索的主流。为适应这种发展形势,此次对本教材作了全方位信息更新和增补。

本版导论部分主要简述了文献信息发展和检索基本理论知识,以后各章节依次对化学化工信息检索与利用的信息源、检索工具及参考工具,以及如何综合利用这些工具等进行了较全面和系统的梳理,让读者能够多方位地了解海量的网上信息资源,并为读者如何开发和利用信息引路。为了对历史经过和检索知识有较全面的了解,也考虑到部分用户习惯使用印刷型检索工具,本书对部分重要的印刷型检索工具,在原有基础上仍作了不同程度的简明介绍。

本书第1版是在1986年由全国高校文献检索与利用课系列教材编审委员会组织,由浙江大学的余向春、叶鹰、黄文林,北京化工学院的沈松源、郑瑞华,青岛化工学院的许家骥、吴澄、陈志宏、李亚力,武汉化工学院的邹荫生,武汉大学王玉璇,华东化工学院的孙济庆,南京化工学院的杨永厚等七所高校的13位教师协作编写的。第2版修订,由于原班参编人员工作变动等原因,仅以浙江大学的余向春、叶鹰、黄文林,武汉化工学院的邹荫生四位同志参与。本版修订经原余向春、许家骥、邹荫生三位主编互相协商,其中邹荫生同志要求不再担任主编;许家骥先生担任主审。

本版全书共分14章,其中各章分工为:余向春负责修改和撰写第2、6、7、

10、11、13 章；叶鹰负责撰写第 1、5、12 章；黄文林负责修改和撰写第 3、4、8、9、14 章。全书由余向春修改定稿。

数字化信息资源不断涌现，为信息开发和利用提供了更丰富源泉，然而，数字化、网络化信息的分散、无序、庞杂、且动态多变，也给信息检索与利用带来了诸多障碍和困扰。我们在编写过程中，对此也深有体会，并对本书内容的取舍感到难以把握。书中一定会有该舍未舍、该取未取的弊病。希望读者，特别是使用本书作为教材的教师，在使用过程中再去发现和充实。不当之处，敬请指正。

余向春

2008 年 1 月于杭州浙江大学

目 录

第1章 导论 / 1

- 1.1 文献信息基本知识 / 1
- 1.2 信息检索概述 / 4
- 1.3 化学化工文献及其信息检索特点 / 5
- 1.4 信息检索基础理论和技术方法 / 5
- 1.5 信息检索效果评价参数 / 7

主要参考文献 / 8

第2章 化学化工核心期刊及全文电子期刊 / 9

- 2.1 概述 / 9
 - 2.1.1 科技期刊的特点 / 9
 - 2.1.2 科技期刊的类型 / 9
- 2.2 综合性化学期刊 / 10
- 2.3 国内化学化工核心期刊 / 11
 - SCI、SCIE 收录的我国化学化工期刊 / 11
- 2.4 国内电子期刊数据库和网站 / 18
 - 2.4.1 万方数据公司的中国期刊题录数据库 / 18
 - 2.4.2 万方中国期刊全文数据库 / 19
 - 2.4.3 维普中文科技期刊数据库 / 20
 - 2.4.4 中国科学院国家科学数字图书馆网络电子期刊集成目录 / 20
 - 2.4.5 中国科学院文献情报中心 OPAC 查询 / 21
 - 2.4.6 电子期刊服务系统 / 21
 - 2.4.7 国家科技图书文献中心期刊分类目次浏览网站 / 22
 - 2.4.8 CALIS 中西文期刊目次数据库 / 22
 - 2.4.9 清华大学图书馆 INNOPAC 馆藏目录检索系统 / 22

2.4.10 北京大学图书馆 WebCat / 22

2.4.11 浙江大学图书馆基于 HORIZON 馆藏书刊目录查询系统 / 22

2.5 国外电子期刊数据库和网站 / 23

- 2.5.1 美国化学会期刊 / 23
- 2.5.2 Springer 电子期刊数据库 / 23
- 2.5.3 剑桥大学化学期刊 / 24
- 2.5.4 著名出版商 Elsevier Science 的期刊 / 24
- 2.5.5 HighWire 出版社 / 25
- 2.5.6 ND 电子期刊网站 / 25
- 2.5.7 英国皇家化学会(RSC)的电子期刊 / 25
- 2.5.8 UCONN 电子期刊 / 26
- 2.5.9 加州大学 INFOMINE 的电子期刊网站 / 26
- 2.5.10 WILEY InterScience / 26
- 2.5.11 网络世界的资源和服务系列 / 27
- 2.5.12 数字图书馆和档案(DLA)的电子期刊 / 27

2.6 国外化学化工核心期刊及其网站选录 / 27

主要参考文献 / 30

第3章 化学化工开放存取期刊资源 / 31

- 3.1 开放存取出版概述 / 31
 - 3.1.1 开放存取与传统学术出版的比较 / 32
 - 3.1.2 开放存取出版的优势 / 32
 - 3.1.3 开放存取出版的模式 / 33
- 3.2 常见开放存取站点 / 34

- 3.2.1 开放存取期刊 / 34
- 3.2.2 预印本文献 / 39
- 3.3 国内开放存取站点 / 40
 - 3.3.1 Socolar / 40
 - 3.3.2 国内开放存取站点 / 44
 - 3.3.3 其他一些与开放存取有关的网站 / 44
- 主要参考文献 / 45

第4章 化学化工门户网站及机构库 / 46

- 4.1 化学化工门户网站 / 46
 - 4.1.1 化学信息网 / 46
 - 4.1.2 中国化工信息网 / 49
 - 4.1.3 中国化工门户网 / 51
 - 4.1.4 其他一些化学化工门户网站 / 52
- 4.2 化学化工文献机构库 / 52
 - 4.2.1 奇迹文库 / 53
 - 4.2.2 诺丁汉大学的机构仓储集合 / 55
 - 4.2.3 香港大学的图书馆知识库和香港科技大学 OA 仓储 / 56
 - 4.2.4 其他一些机构库 / 57

主要参考文献 / 59

第5章 化学化工搜索引擎 / 60

- 5.1 搜索引擎技术原理和检索方法 / 60
 - 5.1.1 技术原理 / 60
 - 5.1.2 检索方法 / 62
- 5.2 化学化工专业搜索引擎 / 63
- 5.3 通用万维网搜索引擎和元搜索引擎的应用 / 69

主要参考文献 / 74

第6章 美国《化学文摘》 / 75

- 6.1 概况 / 75
 - 6.1.1 特点 / 75
 - 6.1.2 组织与排列 / 75
- 6.2 文摘著录格式 / 78
- 6.3 索引及其使用 / 78
 - 6.3.1 关键词索引 / 78
 - 6.3.2 作者索引 / 78

- 6.3.3 专利号索引 / 79
- 6.3.4 专利索引 / 79
- 6.3.5 主题索引 / 80
- 6.3.6 普通主题索引 / 81
- 6.3.7 化学物质索引 / 81
- 6.3.8 分子式索引 / 81
- 6.3.9 环系索引 / 83
- 6.3.10 登记号索引 / 84

6.4 索引指南 / 84

- 6.4.1 概况 / 84
- 6.4.2 标题词说明语的类型 / 85
- 6.4.3 附录内容 / 86

6.5 资料来源索引 / 94

第7章 美国化学文摘(CA)网络版 / 96

- 7.1 CA 网络版 SciFinder 概述 / 96
- 7.2 SciFinder 数据库内容 / 97
- 7.3 SciFinder Scholar 使用简介 / 98
 - 数据库用户端检索软件安装 / 98
- 7.4 SciFinder Scholar 数据库检索途径 / 99
 - 7.4.1 研究主题检索 / 99
 - 7.4.2 化学结构检索 / 99
 - 7.4.3 反应式检索 / 99
 - 7.4.4 化学名称和 CAS 登记号检索 / 100
 - 7.4.5 作者名检索 / 100
 - 7.4.6 专利检索 / 100
 - 7.4.7 公司名称和研究机构检索 / 100
 - 7.4.8 期刊目录检索 / 100
 - 7.4.9 部分结构检索 / 100
- 7.5 检索示例 / 101
 - 7.5.1 通过研究主题检索 / 101
 - 7.5.2 通过化学结构检索 / 104
 - 7.5.3 通过标识物质检索 / 107
 - 7.5.4 通过分子式检索 / 108

主要参考文献 / 110

第8章 化学化工相关重要文摘和数据库 / 111

- 8.1 中国化学化工文摘 / 111
 - 8.1.1 中国化学化工文摘 / 111

- 8.1.2 中国化学化工文摘网络版 / 112
- 8.2 俄罗斯《文摘杂志》/ 113
 - 8.2.1 P.Ж. 简介 / 113
 - 8.2.2 P.Ж. 的内容编排、著录和索引 / 114
- 8.3 日本《科学技术文献速报》及相关数据库 / 115
 - 8.3.1 科学技术文献速报 / 115
 - 8.3.2 科学技术振兴机构(JST)数据网站和其他日本网上资源 / 116
- 8.4 其他主要化学化工文摘及网站 / 118
 - 8.4.1 国内其他主要文摘 / 118
 - 8.4.2 国外其他主要文摘及网站 / 120
- 8.5 工程索引及其 Ei Village / 121
 - 8.5.1 使用特点 / 122
 - 8.5.2 检索方法 / 123
 - 8.5.3 检索结果的处理 / 127
- 主要参考文献 / 127

第9章 专利信息及其网络与数据库 / 128

- 9.1 专利文献概述 / 128
 - 9.1.1 知识产权与专利 / 128
 - 9.1.2 专利文献的类型 / 129
 - 9.1.3 专利文献说明书的各国标准缩写 / 129
 - 9.1.4 国际专利分类系统 / 129
 - 9.1.5 专利说明书的结构 / 132
 - 9.1.6 主要的专利文献网站 / 134
- 9.2 中国专利及其检索方法 / 136
 - 9.2.1 中国知识产权局出版物 / 136
 - 9.2.2 中国国家知识产权局中国专利文献数据库 / 138
 - 9.2.3 中国专利信息网 / 138
- 9.3 德温特专利检索 / 138
 - 9.3.1 德温特分类系统 / 138
 - 9.3.2 Derwent 出版物 / 139
 - 9.3.3 德温特专利数据库 / 141
- 9.4 美国专利检索 / 145
 - 9.4.1 美国专利分类号 / 145

- 9.4.2 美国专利文摘光盘数据库 / 146
- 9.4.3 美国专利商标局网站专利数据库 / 146
- 9.5 ESP@CENET 专利数据库 / 150
 - 9.5.1 检索方法 / 151
 - 9.5.2 欧洲专利局世界专利出版物的覆盖范围 / 152
- 9.6 日本特许厅网站专利数据库 / 154
- 主要参考文献 / 160

第10章 特种文献及其网络检索 / 161

- 10.1 美国科技报告 / 161
 - 10.1.1 概况 / 161
 - 10.1.2 美国政府的科技报告 / 161
 - 10.1.3 科技报告的印刷版检索工具 / 162
- 10.2 美国政府报告数据库及有关网站 / 166
 - 10.2.1 NTIS 数据库 / 166
 - 10.2.2 NASA 技术报告服务器 / 166
 - 10.2.3 DOE 信息桥 / 167
 - 10.2.4 GrayLIT Network 科技报告全文数据库 / 167
 - 10.2.5 美国四大科技报告的其他检索网址 / 167
- 10.3 其他国家科技报告 / 168
 - 10.3.1 中国科技报告的检索 / 168
 - 10.3.2 欧洲各国科技报告 / 169
 - 10.3.3 科技报告原文的索取 / 169
- 10.4 会议文献及其网络资源 / 169
 - 10.4.1 概况 / 169
 - 10.4.2 印刷版会议文献检索工具 / 170
 - 10.4.3 化学化工会议文献网络资源 / 170
- 10.5 学位论文及其网络资源 / 172
 - 10.5.1 概况 / 172
 - 10.5.2 印刷版学位论文检索工具 / 172
 - 10.5.3 学位论文网络资源 / 173
- 10.6 标准文献及其网络资源 / 174
 - 10.6.1 国际标准 / 174
 - 10.6.2 国内标准 / 177

10.6.3 其他常用标准 / 179

10.6.4 标准文献网络及其检索 / 183

主要参考文献 / 185

第11章 参考工具书及其网络资源 / 186

11.1 概述 / 186

11.1.1 参考工具书的编排结构 / 186

11.1.2 参考工具书内容的补充和更新 / 187

11.1.3 参考工具书的查找 / 187

11.2 字典、词典 / 188

11.2.1 字典、词典的类型 / 188

11.2.2 常用化学化工名词、术语词典 / 189

11.2.3 网上检索词典 / 191

11.3 常用化学化工数据手册 / 191

11.4 百科全书 / 194

11.4.1 概况 / 194

11.4.2 贝尔斯坦有机化学大全
(印刷版) / 195

11.4.3 盖墨林无机化学大全
(印刷版) / 200

11.4.4 贝尔斯坦/盖墨林化学大全
网络版 / 200

主要参考文献 / 203

第12章 DIALOG 及联机检索 / 204

12.1 DIALOG 检索技术 / 204

12.2 DIALOG 检索步骤 / 212

12.3 STN 系统 / 216

主要参考文献 / 219

第13章 《科学引文索引》及其派生出版物 / 220

13.1 概述 / 220

13.2 《科学引文索引》印刷版中的索引
及其检索 / 222

13.2.1 作者引证索引 / 222

13.2.2 专利引证索引 / 222

13.2.3 轮排主题索引 / 223

13.2.4 机构引证索引 / 223

13.2.5 来源索引 / 224

13.3 《科学引文索引》扩展版及其检索 / 225

13.4 《ISI 期刊引用报告》介绍 / 227

13.5 中国科学引文数据库及其派生
出版物 / 228

主要参考文献 / 228

第14章 化学综合利用 / 229

14.1 网络与网上化学化工文献
综合利用 / 229

14.1.1 网络会议 / 229

14.1.2 化学类电子论坛 / 230

14.1.3 博客 / 231

14.1.4 化学软件 / 231

14.2 查找信息的策略和技巧 / 232

14.2.1 查找信息的策略 / 232

14.2.2 信息查找的非常规技巧 / 233

14.3 信息的鉴别与整理 / 234

14.3.1 信息资料的鉴别 / 234

14.3.2 信息资料的加工整理 / 235

14.3.3 信息资料的阅读 / 236

14.3.4 信息研究 / 237

14.4 信息的综合归纳与撰写 / 237

14.4.1 信息的综合归纳 / 237

14.4.2 综述的编写 / 238

14.4.3 科技论文摘要的编写 / 239

14.4.4 科技论文的撰写 / 240

主要参考文献 / 241

导论

文献信息,浩如烟海,据不完全统计,现在全世界每年出版学术期刊 10 多万种、图书 60 多万种、专利约 100 万件、技术标准约 20 万件、科技报告约 90 万件、各类论文数百万篇,并且不断增长。面对如此众多的信息资源,任何人都不可能通阅。如何从浩瀚的文献信息“海洋”中找出自己真正需要的“情报”,就是文献检索或信息检索的任务。

从学理上定义,信息检索(Information Retrieval)是从信息集合中找出所需情报的过程,其本质是情报需求与储存信息之间的选择和匹配,在技术上也可以说是对信息项进行表示、存储、组织和获取。现代意义的信息检索作为一个独立领域,是 1946 年计算机出现后在国际上逐步得以确立的。自从被确立为一个独立研究领域以来,信息检索一直是情报学中最活跃的研究分支之一,并成为现代专业研究人员尤其是科技人员必须掌握的技能,也是当今信息素养(Information Literacy)教育的核心内容之一。

20 世纪 90 年代后,随着计算机和网络信息技术的发展,靠“手翻、眼看、大脑判断”的手工检索方式已难以全面适应信息社会发展的需要,计算机信息检索和基于互联网(Internet)的网络化信息检索迅速发展并逐渐成为当今信息检索的主流,进而促使本教材不断修订再版。

本书结构框架如下:导论简要阐述文献信息发展状况及其与化学化工研究的关系,从介绍信息检索的理论知识开始依次对化学化工信息检索与利用涉及的信息源(包括专业期刊、全文期刊数据库、网上开放存取资源、门户网站等)、检索工具(包括搜索引擎、CA 等文摘及其数据库、DIALOG 系统等)和参考工具书等进行系统梳理,最后落实到文献信息综合利用上,希望能有效引导读者从事化学化工信息检索与利用。

1.1 文献信息基本知识

信息、材料和能源一道并称为现代社会三大支柱,因而可将信息视为一种客观存在。载体化的有一定逻辑完整性的信息通常称为文献,文献作为人类思想的记录,是人类理性与智慧的结晶,是精神财富的实体形态。已有多种文献的定义,如:文献是记录有知识的一切载体(GB3792.1-83);文献是典籍与宿贤(辞海);文献是记录在载体上的有意义的信息等。科技文献中图书、期刊、专利、技术标准、科技报告、会议文献、学位论文、政府出版物、产品样本和档案被称为十大科技文献或十大科技信息源,其中图书、期刊又称普通文献,专利、技术标准、科技报告、会议文献、学位论文、政府出版物、产品样本和档案又叫

特种文献。所有文献均可数字化,全文数据库即由数字化全文文献构成,数字化文献也可以包含所有文献类型。

按信息加工深度,文献可分为一次文献(原始文献)、二次文献(检索文献)和三次文献(集成文献)。一次文献即信息源,二次文献是检索工具,三次文献的主体是工具书。由于网络的发展和普及,一、二、三次文献均以数字化形态在网上展现,形成丰富多彩的网络化文献信息世界。

文献具有下列性质:

(1)载体性

文献离不开载体,没有载体就没有文献,文献与载体不可分割,这是文献的第一特性。

(2)逻辑性

记录在载体上的信息,其内容必须具有逻辑性,否则不能称为文献。

(3)独立性

同一文献信息内容不因载体、语种和出版形式的不同而改变,甚至不因作者的存亡而改变,这就是文献的独立性,这一性质使翻译或转载等得以成立。

(4)累积性

从文献整体上看,文献信息内容具有累积性,这是人类知识与文化不断丰富的基础。

(5)自序性

文献具有自然成序的性质,类似耗散结构的自组织状态,这一性质使文献常自相配套和互补。

在以上性质基础上,呈现文献增长与老化、集中与分散、浓缩与冗余等文献分布规律,这就是著名的 Bradford 定律、Lotka 定律等文献计量规律的来历。

在文献信息处理的方法论方面,已经形成一些专门的方法,主要是分类法与编目法、主题法与索引法等。

(1)分类法与编目法(Classification & Cataloging):

十大文献信息源的分类方法各有特点,因而文献分类自然形成图书分类、期刊分类、专利分类等各种分类法,其中尤以图书分类法较为成熟和发达。现在我国广泛应用的图书分类法主要是大多数公共图书馆和学校图书馆采用的《中图法》(中国图书馆图书分类法的简称)和科学院系统图书馆采用的《科图法》(中国科学院图书馆图书分类法的简称);国外应用较多的是《UDC》(国际十进分类法的简称)、《DDC》(杜威十进分类法的简称)和《LC》(美国国会图书馆图书分类法)。期刊分类的一种常见方式是采用图书分类法的简表来操作;专利分类法则基本上趋向于统一采用《IPC》(国际专利分类法);其他类型的文献尚无通用的分类法。

每种分类法的核心结构都是类号和类目,其差异也体现在类号和类目上,如《中图法》大类为

A 马列主义、毛泽东思想	B 哲学	C 社会科学总论
D 政治	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	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
Z 综合性图书		

《LC》大类为

A 总类、丛书	B 哲学、宗教	C 历史
D 历史与地志	E~F 美洲	G 地理学、人类学
H 社会科学	J 政治学	K 法律
L 教育	M 音乐	N 艺术
P 语言学及文学	Q 科学总论	R 医学
S 农业与矿业	T 工艺学	U 军事学
V 海事学	Z 目录学与图书馆学	

《UDC》大类为

000 总论	100 哲学	200 宗教
300 社会科学	400 (留用)	500 纯粹科学
600 技术科学	700 美术	800 语言学、文学
900 历史		

《IPC》八大部为

A 人类生活必需	B 作业、运输	C 化学、冶金
D 纺织、造纸	E 固定建筑物	F 机械工程
G 物理	H 电学	

完整的 IPC 类号由部(字母)、大类(数字)、小类(字母)、大组/小组几部分构成,如 C09K11/72。

编目(Cataloguing)是文献信息的专门处理技术,通常以图书编目为中心。编目在技术操作上体现为按照特定格式的著录,较通用的国际标准是基于美国 MARCII 格式制定的 ISO2709-1973(E)。为使各国的机读书目信息能相互交换和共享,IFLA 参照 ISO2709 制定了 UNIMARC,中国也以 UNIMARC 为依据制定了 CNMARC。

UNIMARC 和 CNMARC 都由头标区、目次区和数据区三部分构成,数据区采用 3 位数字划分的“字段”,字段分配为

0XX: 标识	1XX: 代码信息	2XX: 著录信息
3XX: 注释	4XX: 款目连接	5XX: 相关题名
6XX: 主题分析	7XX: 责任者说明	8XX: 国际交流
9XX: 国内使用		

具体、详细的用法可查阅相关手册。

在数字化网络化时代,编目在很大程度上正被元数据(Metadata)取代,技术标准则以

XML-RDF 为主流。

(2) 主题法与索引法 (Subject & Indexing):

主题法与分类法关系密切,依据是主题词表。通常,分类表按类号排列,而主题词表则按字顺排列。学术界存在两类观点:一是主张分类、主题一体化,二是主张分类、主题二元化。

索引法即标引方法,其成果是各种索引。从技术的角度看,标引可分为利用规范词表(如主题词表、叙词表等)进行的前控标引和利用非规范词表(非关键词词表、非单元词词表)进行的后控标引,生成的索引相应就是前控索引(如各种主题索引)和后控索引(如各种关键词索引)。目前,标引和索引制作的发展趋势是自动标引,与之密切相关的是文摘法,发展方向也是自动文摘。

1.2 信息检索概述

信息检索 (Information Retrieval) 是在 1949 年国际数学会议上由 Galvin W. Mooers 首次提出的,由于英文 information 可以翻译成汉语的“信息”或“情报”,信息检索和情报检索就被作为同义词使用了。如果要区分“信息”和“情报”,则可以把情报定义为具有主观价值的信息,这样,前面所说信息检索是从信息集合中找所需情报就有了特定涵义。

信息检索历史可以粗略地划分为手检时代和机检时代。20 世纪 60 年代前属于手检时代,20 世纪 70 年代后属于机检时代,其间有交叉。20 世纪 90 年代后,进入网络化机检时代。

在手检时代,信息源是纸质原始文献,检索工具是文摘、索引,手册、辞典等必备参考工具书,信息检索过程是通过使用检索工具和工具书查找并获取存储于图书情报机构中的原始文献。这时,专业期刊和 Chemical Abstracts 是化学化工研究机构的基础资源,The Merck Index、CRC Handbook of Chemistry and Physics 等是化学家的重要案头参考书。

在机检时代,信息源逐步数字化,检索工具发展成各种数据库和机检系统,参考工具书入网,信息检索成为利用各种检索工具查找并通过网络获取全文的过程。这时,全文期刊数据库成为基础资源,机检系统和搜索引擎成了通用检索工具,一台电脑加一条网线成了化学家的案头参考源。

这些变化就发生在近 30 年内,而且就发生在我们身边。手检与机检的差如表 1-1 所示。

表 1-1

手检与机检差别

比较项	手检	机检
总体特征	手翻、眼看、大脑判断	选库、策略、机器匹配
标引及索引特点	检索点较少	检索点较多
检索时间	较慢	较快
检索知识要求	专业知识、外语知识、检索工具知识	专业知识、外语知识、机检系统知识
查全查准率	查准率较高	查全率较高
综合效率	较低	较高

因此,如今的信息检索,大体经选择数据库、设计检索策略、上机操作即可完成。但为了对历史经过和检索知识有较全面了解,后面还会介绍一些传统知识。

1.3 化学化工文献及其信息检索特点

化学化工领域是文献信息数量最大、品种最多的科技领域之一,从美国化学文摘(CA)一个世纪的文摘量数据中可见一斑:

年代	CA 文摘量	年代	CA 文摘量
1907	11 847	1967	242 527
1917	15 945	1977	410 137
1927	33 491	1987	476 178
1937	64 735	1997	716 564
1947	39 288	2007	850 000
1957	102 525		

庞大的专业文献量导致化学化工研究对信息检索技能的较高要求。据估计,化学化工科研时间约 50% 花费于查阅文献和信息检索,因而信息检索技能是化学化工研究人员必备的基本功。

化学化工领域也是涉及文献类型最全面的科技领域。CA 收录的文献类型几乎覆盖除产品样本外的所有十大科技信息源,对各种专业文献的识别和判断能力也成为对化学化工研究人员的基本要求。

因此,化学化工文献具有数量多、类型复杂的特点,使化学化工信息检索具有多样性和复杂性。尤其是 20 世纪 90 年代以来,网络化浪潮席卷世界,化学化工文献信息也大量数字化、网络化,进而在数字化、网络化信息检索层面提出更高要求。

为深入学习化学化工信息检索,应先掌握信息检索基础理论知识。

1.4 信息检索基础理论和技术方法

在信息检索理论中,布尔检索模型、向量检索模型、概率检索模型是三个经典的检索理论模型。这三类检索模型的差异可用文献 d_j 和检索提问 q 之间的相似度 $\text{sim}(d_j, q)$ 来说明。

布尔检索模型的相似度为

$$\text{sim}(d_j, q) = \begin{cases} 1, & \text{if } \exists q_c (q_c \in q_d) \wedge (\forall k_i, g(d_j) = g_i(q_c)) \\ 0, & \text{others} \end{cases} \quad (1-1)$$

其中, q_d 表示检索提问 q 的析取范式, q_c 表示 q_d 的任意合取分量。

向量检索模型的相似度为

$$\text{sim}(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} \quad (1-2)$$

概率检索模型的相似度为

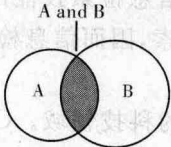
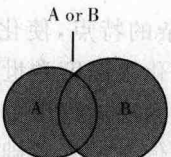
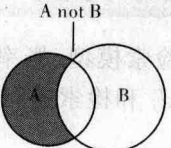
$$\text{sim}(d_j, q) = \frac{P(R|d_j)}{P(\bar{R}|d_j)} = \frac{P(d_j|R) \times P(R)}{P(d_j|\bar{R}) \times P(\bar{R})} \quad (1-3)$$

其中,条件概率 $P(R|d_j)$ 表示文献 d_j 与检索提问 q 相关的概率,条件概率 $P(\bar{R}|d_j)$ 表示文献 d_j 与检索提问 q 不相关的概率;条件概率 $P(d_j|R)$ 表示从相关文献集 R 中选择文献 d_j 的概率,条件概率 $P(d_j|\bar{R})$ 表示从文献补集 $\bar{R}$ 中选择文献 d_j 的概率,概率 $P(R)$ 、 $P(\bar{R})$ 分别表示从整个文献集中随机选出相关文献、不相关文献的概率。

尽管现代信息检索发展了不少新兴理论和模型,布尔检索模型、向量检索模型、概率检索模型仍然是三个基本的检索理论模型。尤其是布尔检索模型,依然是最基本的检索理论模型。

作为信息检索的基本技术,布尔检索模型通过检索词的逻辑组配编制检索策略,即主要通过“与(and,*)”、“或(or,+)”、“非(not,-)”等逻辑运算将检索词联系起来,例如,对 A、B 两词而言,and、or、not 的逻辑含义用维恩(Wienn)图表示(见表 1-2):

表 1-2 A、B 的逻辑含义维恩图

运算符	图解	含 义
AND		逻辑与。A 和 B 都为真时,结果才为真
OR		逻辑或。A 或 B 中只要有一个为真时,结果就为真
NOT		逻辑非。A 为真、B 为假,结果才为真

将检索策略输入检索系统,通过计算机匹配即可获得检索结果而完成检索。因此,基于布尔检索模型的检索涉及的主要问题有:检索策略及其设计、检索系统及其技术、检索操作及其实现等。对检索结果的评价则涉及几个评价检索效果的参数。这里先说明检索策略及其设计和检索效果评价参数,检索系统及其技术和检索操作及其实现将在后面结合 DIALOG 系统介绍。

检索策略是对检索的全面策划,在操作上主要指检索式的编制。

由于检索策略在机检中具有决定检出结果的重要地位,故设计检索策略意义极其重大,其一般程序是:①确定检索词和截、位关系;②组配成式并优化;③实检和反馈调节。

例如,对“微藻生长因子”这一课题来说,基本检索策略可编制为

(algae+algal+microalgae) * growth()factor

又如,对“图像理解专家系统”这一课题,考虑到“专家系统”与“人工智能”同类,为预防漏检,基本检索策略可编制为

image()understand * (expert()system+artificial()intelligence)

由于检索策略在机检中具有直接决定检出结果的重要性,故编制和调整检索策略意义极其重大,其一般程序是:①确定基本检索词(实词);②组配成式并优化(分配律、结合律、De Morgan 律对逻辑运算成立,故可作“提取公因式”等运算);③实检并反馈调节。

1.5 信息检索效果评价参数

检索效果评价参数主要有:查全率(Recall Factor)、查准率(也称适中率, Pertinency Factor)、漏检率(Omission Factor)、误检率(Noise Factor,即检索噪音)。

设 n 为检索系统中文献总量, m 为检索输出的文献量, a 为 n 中与检索课题有关的文献量, b 为 m 中与检索课题有关的文献量(检准文献量),则 n 、 m 、 a 、 b 之间的关系如图 1-1 所示。

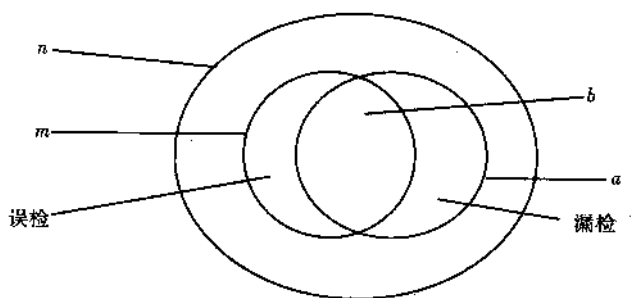


图 1-1 n 、 m 、 a 、 b 之间的关系

令 R 表示查全率、 P 表示查准率、 M 表示漏检率、 N 表示误检率,则 R 、 P 、 M 、 N 定义如下:

$$R = b/a \times 100\%$$

$$P = b/m \times 100\%$$

$$M = (1 - b/a) \times 100\% = 100\% - R$$

$$N = (1 - b/m) \times 100\% = 100\% - P$$

最理想的检索效果是 M 、 N 均为 0,即 R 、 P 均为 100%,但实际上这是不可能的。实验表明: R 和 P 之间存在相反的相互依赖关系,即提高 R 会降低 P ,反之亦然,如图 1-2 所示。

有人认为:在物理、技术科学信息检索范围内, P 提高 1% 将导致 R 降低 3%。实践经验证明:在现代科技信息检索系统中, R 为 60%~70%, P 为 40%~50%。同时,检索特性曲线还与文献专业特点有关,例如在具体概念为主、对象描写明确的化学等专业中,产生误检较少,特性曲线向右移;而在抽象概念多的心理学等领域中,特性曲线则向左移。此外特性曲线还与对查出的文献适用性的评价有关,若评价标准严格,则曲线偏左;反之,曲线偏右。

对于信息检索来说,漏检是影响检索质量的最主要因素,故必须将 M 降低到最低限

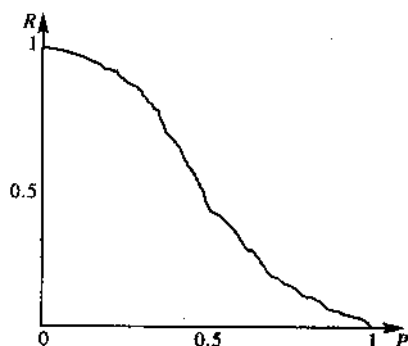


图 1-2 检索特性曲线

度;误检会降低检索的效率,也会影响检索质量。因此,任何检索工具和检索系统必须力争克服漏检(必要条件),同时尽量避免误检(充分条件)。

信息检索不仅是理论方法,而且是操作性很强的技术,具有实用性和技巧性,应通过练习逐步熟悉和掌握。

主要参考文献

- 1 Baeza-Yates, Ricardo et al. Modern Information Retrieval. New York: ACM Press, 1999
- 2 Korfhage, Robert R. Information Storage and Retrieval. New York: Wiley Computer Publishing, 1997
- 3 <http://www.acs.org>
- 4 余向春. 化学文献及查阅方法. 3 版. 北京:科学出版社,2003
- 5 叶鹰. 信息检索:理论与方法. 北京:高等教育出版社,2004