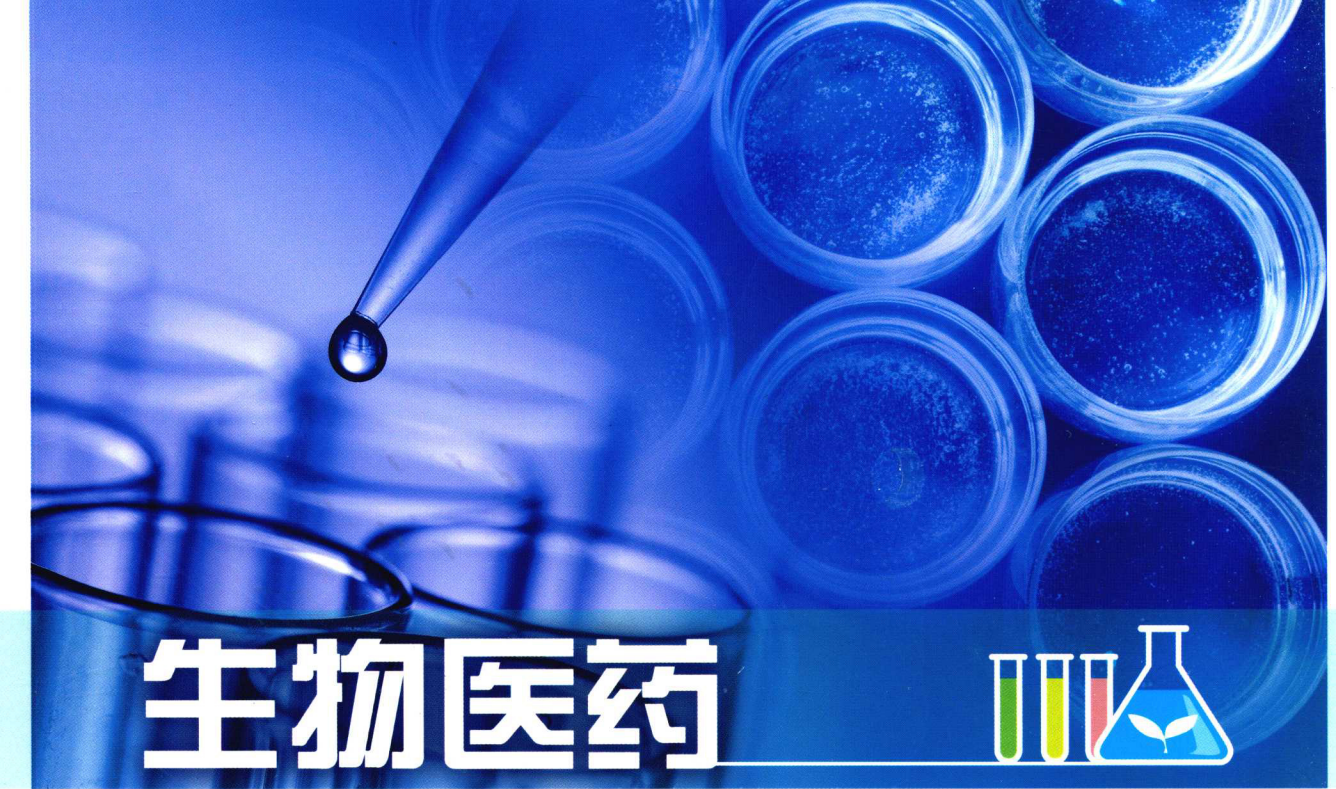




生物医药



信息检索与利用 Shengwu Yiyao

Xinxi Jiansuo Yu Liyong

-----• 主编 薛晓芳 郝继英 陈 锐



军事医学科学出版社

主要内容

生物医药信息检索与利用

主 编 薛晓芳 郝继英 陈 锐
编 者 王 萧 齐凤青 吕 婷 张 玉
何 玮 李芳薇 陈建青 张莉莉
张 萌 陈 锐 张 群 张晓燕
杨志滨 杨晓茹 郝继英 徐 言
殷 亮 胡畔畔 胡德华 程 瑾
薛晓芳 龙旭梅

军事医学科学出版社
· 北 京 ·

内容提要

本书由具有丰富医学信息检索经验的图书馆馆员编写,内容丰富,结构清晰,选材新颖,突出实践,重点介绍了信息检索基础理论方法、各种国内外经典的生物医药文摘数据库和全文数据库、国内外著名的引文数据库、期刊影响因子的查找方法、美国工程索引数据库,并详细介绍了如何查找学位论文、会议、专利、循证医学、药学、生物信息学等方面信息,此外还新增了网络开放存取资源、国内外生物医学专业网站推荐、文献管理软件的使用方法、科研项目申报及查新流程。

本书既可作为生物医药信息检索课程的教材,也可供临床医生、科研人员、图书情报人员、教师等参考使用。掌握了获取信息的途径和方法,能收到事半功倍的效果,并可使终身学习和工作都受益。

图书在版编目(CIP)数据

生物医药信息检索与利用/薛晓芳,郝继英,陈锐主编.

-北京:军事医学科学出版社,2015.7

ISBN 978-7-5163-0627-7

I. ①生… II. ①薛… ②郝… ③陈… III. ①生物学-情报检索 ②医药学-情报检索 IV. ①G252.7

中国版本图书馆 CIP 数据核字(2015)第 143632 号

策划编辑:孙 宇

责任编辑:曹继荣

出版:军事医学科学出版社

地址:北京市海淀区太平路 27 号

邮编:100850

联系电话:发行部:(010)66931051,66931049,63827166

编辑部:(010)66931039

传真:(010)63801284

网址:<http://www.mmssp.cn>

印装:中煤涿州制图印刷厂北京分厂

发行:新华书店

开本:787mm×1092mm 1/16

印张:15.75

字数:406 千字

版次:2015 年 8 月第 1 版

印次:2015 年 8 月第 1 次

定价:50.00 元

本社图书凡缺、损、倒、脱页者,本社发行部负责调换

前言

军事医学科学院医学信息检索教研室为研究生开设医学信息检索课程已近 20 余年,虽然是选修课,但是近年来主动选修医学信息检索课程的同学逐年增多,几乎达到研究生总数的 95% 以上。学生进行专业学习,研究生进行选题、开题和中期以及写论文和课题结题等都离不开查找和分析文献信息,而且会在文献调研上花费相当多的时间,因为只有把握了前沿热点信息,掌握了国内外研究动态,才能站在较高的层次做科研。本门课程开设这么多年来,同学们逐渐认识到了学好这门课能大大提高获取信息的能力和效率,起到事半功倍的效果,并为今后的终身学习和科研工作打下坚实的基础。

本书编写成员具备丰富的工作经验,多年来利用书中介绍的各种生物医药信息数据库和平台,为科研人员提供了各种各样的知识服务,如:科研立项、成果鉴定、成果报奖的查新、查收和查引;学科动态追踪;专利技术分析;机构竞争力分析;学科前沿发展态势分析;产业情报;动态监测及预警等。

近年来,生物医药信息产品更新换代较快,本教材也需要及时地更新与修订。本书的编写参考了国内一些优秀教材,吸收了来自学生的需求建议,不仅选取了常用的经典信息数据库产品,还增收一些新颖的信息工具,全部选用数据库当前最新版本讲授,将国内外最新信息产品呈现给学生。本书内容丰富,涵盖信息检索基础理论方法、各种国内外经典的生物医药文摘数据库和全文数据库使用方法、国内外著名的引文数据库利用、期刊影响因子的查找方法、美国工程索引的检索以及各种学位论文、会议、专利、循证医学、药学、生物信息学等方面信息的检索,此外还新增了网络开放存取资源、国内外生物医学专业网站推荐、文献管理软件的使用方法、科研项目查新及申报流程。

通过本教材的学习及实践练习,可以使得学生树立较强信息意识,具备较高信息素养,熟练掌握生物医学信息检索技能,了解国内外常用生物医药信息数据库,充分利用信息工具获得相关生物医药领域前沿信息,最终具备鉴别、评价和综合运用信息的能力。

在使用本教材的过程中,欢迎大家提出使用意见和修改建议,以利于进一步完善和更新。

本书的编写得到军事医学科学院研究生部、解放军医学图书馆、军事医学科学出版社以及相关专家的大力支持和帮助,在此表示衷心的感谢。

编 者

目录

第一章 信息检索基础	(1)
1.1 信息检索概述	(1)
1.1.1 信息与信息资源	(1)
1.1.2 信息检索	(3)
1.2 信息检索技术和工具	(5)
1.2.1 信息检索技术	(5)
1.2.2 信息检索工具	(6)
1.3 信息检索途径与策略	(7)
1.3.1 信息检索途径	(7)
1.3.2 信息检索策略与步骤	(8)
1.4 信息检索语言	(11)
1.4.1 信息检索语言概述	(11)
1.4.2 常用医学信息检索语言	(14)
1.5 信息检索效果评价	(25)
1.5.1 概述	(25)
1.5.2 检索效果评价指标	(25)
第二章 文摘数据库检索	(29)
2.1 中国生物医学文献服务系统	(29)
2.1.1 概述	(29)
2.1.2 检索规则	(29)
2.1.3 检索方法与技巧	(30)
2.1.4 检索结果的处理	(34)
2.2 中文生物医学期刊文献数据库	(35)
2.2.1 概述	(35)
2.2.2 检索方法与技巧	(35)
2.2.3 检索结果的处理	(36)
2.3 PubMed 检索系统	(36)
2.3.1 概述	(36)

2.3.2	检索方法与技巧	(37)
2.3.3	检索结果的处理	(43)
2.4	Embase 数据库	(45)
2.4.1	概述	(46)
2.4.2	检索运算符与运算规则	(46)
2.4.3	检索方法与技巧	(47)
2.4.4	检索结果的处理	(50)
2.5	BIOSIS Previews 数据库	(51)
2.5.1	概述	(51)
2.5.2	BP 的检索途径	(52)
第三章	全文数据库检索	(54)
3.1	Elsevier ScienceDirect Online 全文数据库	(54)
3.1.1	数据库概况	(54)
3.1.2	数据库操作	(55)
3.2	Wiley Online Library 数据库	(58)
3.2.1	数据库概况	(58)
3.2.2	数据库操作	(58)
3.3	SpringerLink 全文数据库	(60)
3.3.1	数据库概况	(60)
3.3.2	数据库操作	(60)
3.4	ProQuest Health & Medical Complete 医学全文数据库	(62)
3.4.1	数据库概况	(62)
3.4.2	数据库操作	(63)
3.5	中国学术期刊网络出版总库	(65)
3.5.1	数据库概况	(65)
3.5.2	数据库操作	(65)
3.6	其他全文数据库	(68)
3.6.1	Ovid 全文数据库	(68)
3.6.2	万方数据知识服务平台	(69)
3.6.3	维普中文生物医学期刊数据库	(70)
3.7	文献传递服务系统	(71)
3.7.1	中国高等教育文献保障系统	(71)
3.7.2	国家科技图书文献中心	(74)
第四章	引文检索	(76)
4.1	引文检索总述	(76)

4.1.1 引文索引的由来与发展	(76)
4.1.2 基本概念	(76)
4.1.3 引文检索的作用	(77)
4.2 Web of Science™核心合集数据库	(77)
4.2.1 检索运算符与运算规则	(77)
4.2.2 检索方法和技巧	(78)
4.2.3 检索文献的管理与分析功能	(80)
4.3 中文引文数据库检索	(84)
4.3.1 中国科学引文数据库	(84)
4.3.2 中国引文数据库	(86)
4.3.3 中国生物医学期刊引文数据库	(86)
4.3.4 引文在线检索整合系统	(96)
4.3.5 引文数据检索	(100)
第五章 特种文献数据库	(104)
5.1 学位论文检索	(104)
5.1.1 学位论文概述	(104)
5.1.2 国内学位论文数据库	(104)
5.1.3 国外学位论文数据库	(107)
5.1.4 NDLTD 网络博硕论文数字图书馆	(109)
5.2 会议信息检索	(110)
5.2.1 会议信息概述	(110)
5.2.2 会前信息检索	(110)
5.3 专利文献检索	(114)
5.3.1 专利概述	(114)
5.3.2 国内专利检索	(117)
5.3.3 国外专利检索	(121)
5.4 循证医学资源检索	(124)
5.4.1 循证医学概述	(124)
5.4.2 循证医学证据资源概述	(128)
5.4.3 循证医学资源检索	(132)
5.5 药学信息检索	(144)
5.5.1 综合性药学数据库资源	(144)
5.5.2 与药物研发相关的数据库	(148)
5.6 美国工程索引数据库	(158)
5.6.1 概述	(158)

5.6.2	检索方法与技巧	(159)
5.6.3	检索结果的处理	(160)
5.6.4	其他功能	(161)
第六章	生物信息学数据库	(162)
6.1	生物信息学数据来源及数据库的构建	(162)
6.1.1	生物信息学数据来源	(162)
6.1.2	生物信息学数据库构建	(162)
6.1.3	生物信息学数据库的查找	(166)
6.2	生物信息学数据库类型	(167)
6.2.1	序列数据库	(167)
6.2.2	结构数据库	(171)
6.2.3	功能数据库	(173)
6.2.4	其他专业数据库	(174)
6.3	生物信息学数据库的检索	(176)
6.3.1	生物信息学平台	(176)
6.3.2	检索方法和实例	(176)
第七章	网络信息资源检索	(183)
7.1	网络信息资源概述	(183)
7.1.1	Web 及其信息资源	(183)
7.1.2	Web 信息资源的类型	(183)
7.1.3	Web 信息资源的特征	(184)
7.2	搜索引擎	(185)
7.2.1	定义	(185)
7.2.2	类型	(185)
7.2.3	医学研究常用搜索引擎介绍	(186)
7.3	开放存取资源	(196)
7.3.1	开放存取	(196)
7.3.2	开放存取的实现途径	(196)
7.3.3	国外开放存取资源	(197)
7.3.4	国内开放存取资源	(201)
7.4	生物医学专业网站	(202)
7.4.1	美国国立卫生研究院	(202)
7.4.2	美国国立医学图书馆	(205)
7.4.3	美国疾病控制与预防中心	(209)

7.4.4 世界卫生组织	(210)
7.4.5 国家科技图书文献中心	(211)
7.4.6 中国高等教育文献保障系统	(211)
第八章 文献信息管理与利用	(214)
8.1 文献管理	(214)
8.1.1 文献管理的发展	(214)
8.1.2 国内外文献管理软件介绍及功能对比	(214)
8.1.3 EndNote 文献管理软件使用方法	(218)
8.2 科研项目申报与科技查新	(230)
8.2.1 科研项目申报	(230)
8.2.2 科技查新	(236)

1.1 信息与信息资源

1.1.1 信息

(1) 信息的定义: 中国国家标准《情报与文献工作词汇基本术语》(GB/T4304-1985) 将信息定义为: “信息是物质存在的一种方式, 形态或运动状态, 是事物的一种普遍属性, 一般指客观消息中包含的语义, 可以使信息中所描述事件的不确定性减少。”

不同的学科对信息的理解各不相同。信息论的创始人申农(C. Shannon)从通信系统理论的角度考虑, 认为信息是“用来消除随机不确定性的东西”。经济学家把信息看作是物质、能量和开闭的客观世界三大要素之一, 能为管理和决策提供科学依据。哲学家则认为信息是事物的本质、特征和运动状态的反映。

虽然, 信息在不同的学科领域具有不同的含义和属性, 但是当今学术界存在两种较为普遍的观点, 第一种观点认为, 信息是事物的一种存在方式或状态, 即信息不是事物本身, 但是反映了事物的本质和特征, 不同事物有不同特征, 同一事物在不同的条件下也会发生变化, 这种特征和变化就是信息。第二种观点认为, 信息就是一组具有意义的符号或数据。

(2) 信息的特性

①普遍性和客观性: 信息普遍存在于宇宙万物中, 它无时不有也无处不在, 不仅存在于有机界, 也存在于无机界, 既可以反映物质的特征和运动状态, 也可以体现人类大脑思维的结果。同时, 信息是客观存在的, 它可以被感知、获取、处理、存储、传递和利用。

②依附性和转换性: 信息本身是抽象的, 它必须依附于具体的物质载体, 才能被记录、存储和传递。物质载体的多样性决定了信息表现形式的多样性, 而且信息可以从一种载体形态转换到另一种载体形态。不同的信息可以用同一载体承载, 同一信息也可以用不同载体来表现。

③传递性与共享性: 通过信息传递, 人们才能对客观存在的信息产生主观认识, 信息可以按时间或空间传递也可以用空间电磁波传递。信息的可传递性使得信息可以被获取和利用。信息还具有共享性, 信息作为一种资源, 在交换的过程中, 自己不仅得到了新的信息而且原有的信息也没有丢失。

④认知性与创造性: 信息可以被人类感知、感知、认知, 并且, 人们对信息的认识具有主观能动性, 不同的个体对相同信息的理解和感受程度也是具有差异的。人类还可以将获取到的

第一章 信息检索基础

1.1 信息检索概述

1.1.1 信息与信息资源

1.1.1.1 信息

(1)信息的定义:中国国家标准《情报与文献工作词汇基本术语》(GB/T4894-1985)将信息定义为:“信息是物质存在的一种方式、形态或运动状态,是事物的一种普遍属性,一般指数据、消息中包含的意义,可以使信息中所描述事件的不定性减少。”

不同的学科对信息的理解各不相同。信息论的创始人申农(C. Shannon)从通信系统理论的角度考虑,认为信息是“用来消除随机不确定性的东西”。经济学家把信息看作是与物质、能量相并列的客观世界三大要素之一,能为管理和决策提供科学依据。哲学家则认为信息是事物的本质、特征和运动状态的反映。

虽然,信息在不同的学科领域具有不同的含义和特性,但是当今学术界存在两种被大众接受的观点:第一种观点认为,信息是事物的一种存在方式或状态,即信息不是事物本身,但是反映了事物的本质和特征,不同事物有不同的特征,同一事物在不同的条件下也会发生变化,这种特征和变化就是信息。第二种观点认为,信息就是一组具有意义的事实或数据。

(2)信息的特性

①普遍性和客观性:信息普遍存在于宇宙万物中,它无时不有也无处不在,不仅存在于有机界,也存在于无机界;既可以反映物质的特征和运动状态,也可以体现人类大脑思维的结果。同时,信息是客观存在的,它可以被感知、获取、处理、存储、传递和利用。

②依附性和转换性:信息本身是抽象的,它必须依附于具体的物质载体,才能被记录、存储和传播。物质载体的多样性决定了信息表现形式的多样性,而且信息可以从一种载体形态转换成另一种载体形态。不同的信息可以用同一载体承载,同一信息也可以用不同载体来表现。

③传递性与共享性:通过信息传递,人们才能对客观存在的信息产生主观认识,信息可以沿时间线纵向传递也可以沿空间线横向传递。信息的可传递性使得信息可以被获取和利用。信息还具有可共享性,信息作为一种资源,在交换的过程中,自己不仅得到了新的信息而且原有的信息也没有丢失。

④认知性与创造性:信息可以被人类解读、感知、认知,并且,人们对信息的认识具有主观能动性,不同的个体对相同信息的理解和感知程度也是具有差异的。人类还可以将获取到的

初始信息创造成再生信息,还可以利用再生信息创造出新的再生信息。

⑤可塑性与时效性:信息在传递和利用的过程中,人们借助先进的技术可对其进行综合、分析和加工处理,如序化、压缩和提取,也可以把信息从一种形式转换成另一种形式,方便人们选择和利用,这就是信息的可塑性。信息的价值与时间有关,信息的时效性体现在,及时的信息可能价值连城,而一些过时的信息可能分文不值。

⑥可开发性与可增值性:由于客观事物存在复杂性,而且事物之间存在相互关联,各种信息往往交织在一起;又由于在一定的时间范围内人们对信息的认识总是存在一定的局限性。因此,作为反映客观事物的信息是可以开发的,并且是需要开发的,信息在开发的过程中可以引申、推导、创造出更多有价值的信息,从而使信息增值。

(3)信息的类型

①按照信息的来源分:自然信息和社会信息。自然信息比如风、雨、雷、电等自然现象,表现自然界中事物的状态及特征;而社会信息比如文化、语言等反映人类社会中各种事物和现象的形态及性质。社会信息按照信息描述的内容不同又可以分为经济信息、政治信息、科技信息、军事信息、文化信息等。

②按照信息加工深度分:零次信息、一次信息、二次信息和三次信息。零次信息指未公开于社会的信息如实验数据、书信、手稿等。一次信息又称原始信息。原始信息未经加工,是分散无序的,往往无法被存储、检索、传递与利用,需要进一步加工处理后才能有效利用,如会议记录、统计报告等。二次信息是指对一次信息进行加工处理,形成有序的、有规则的信息,如书目、文摘、索引等。三次信息是指在已有的基础上,经过分析、研究、综合、整理产生的新的信息,如评论、综述等。

③按照信息的载体分:口头信息、文献信息和实物信息。口头信息是指通过交流、讨论、报告等方式传播的信息,它反映了人们的观点和看法,是推动研究的最初起源。口头信息出现早、传播方便,但缺乏完整性和系统性,不易保留和存储,但是却是一种挖掘潜力很大的极为丰富的资源。文献信息是指用文字、图形、符号、音频、视频等方式记录下来的知识和信息资源,是人们筛选、归纳和整理过的信息,其优点是易识别、易传播、易保存。实物信息是指事物本身存储和表现的信息,比如文物、样品、模型等,它直观性强、信息量大。

④按照信息的运动状态分:静态信息和动态信息。静态信息比如历史文献和资料等已成为比较稳定形态;动态信息是指及时地反映现实生活中出现的新事物的新闻和情报等,体现事物的发展、变化状态。

⑤按照信息的价值:有用信息、冗余信息和有害信息。有用信息值对我们的生产、生活和学习有积极作用的信息;冗余信息就是我们暂时不能利用,但也没有害处的信息;有害信息则是给我们的生产生活带来危害的信息。

此外,还可以按照其他标准对信息进行分类,如按照信息被人感知的方式可分为直接信息、间接信息;按照信息来源的稳定性,可分为流动信息、固定信息和偶然信息;按照信息表达的真实程度,又可分为真实信息和虚假信息;按照信息的传播范围,可分为秘密信息、内部信息和公开信息。

1.1.1.2 信息资源

(1)信息资源的涵义:信息资源是指人类经过筛选、组织、加工、整理等信息活动后积累起来的,能够满足人类需求的各类信息要素(信息技术、设施、设备、信息生产者等)的集合。

信息是普遍存在的,但只有满足一定条件的信息才是信息资源。与人类需求相关的信息才是信息资源,它只是信息海洋中的一部分。信息资源是可以被人类利用的信息,是人类依据当前生产力水平和研究水平开发和组织的信息;人类的参与是信息资源形成的必要条件。

(2) 信息资源的类型

1) 按信息加工程度不同,信息资源可以分为零次信息资源、一次信息资源、二次信息资源和三次信息资源。

① 零次信息资源是指未经加工,直接记录在某种载体上的原始信息,如手稿、实验数据、调查材料等。这些尚未融入正式交流渠道的信息,往往能反映研究工作的最新动态,或是遇到的最新问题,或是科研人员针对某些问题的最新想法等,这些是形成创造性思维的最佳素材。

② 一次信息资源是指在零次信息资源的基础上,作者对实验数据、手稿等资源进行组加工形成原始创作,如期刊论文、学位论文、科技报告等。

③ 二次信息资源是对大量无序、分散的原始信息进行收集、整理、著录等深层次的加工,如,对一次信息资源的著者、篇名、主题、分类、出处等信息进行编目,形成新的文献形式供研究人员检索和使用,包括索引、文摘、目录等。

④ 三次信息资源是用一定的方法对一次、二次信息进行高度浓缩、提炼加工、总结分析而成的信息,包括综述、述评、数据手册、辞书、统计年鉴、百科全书等。

2) 按信息记录载体不同,信息资源可分为实物型信息资源、记录型信息资源和智力型信息资源。

① 实物型信息资源包括天然实物信息资源和人工实物信息资源,是事物本身存储和表现的,如各种产品、模型、样机等。它们本身代表着一种技术信息。

② 记录型信息资源是指记录或存储在各种载体上的信息。根据记录的载体不同又可以分为印刷型、机读型、视听型和缩微型。

③ 智力型信息资源是指由人脑存储的知识、经验等信息,有的难以明确表达和记录下来,有待人们进一步开发。

3) 按信息内容不同,信息资源可分为经济信息资源、政治信息资源、军事信息资源、生活信息资源和文化信息资源等。

4) 按信息资源的流通范围不同,信息资源可分为公开信息资源、外部信息资源、内部信息资源、秘密信息资源。

5) 按信息资源的形态不同,可分为动态信息资源和静态信息资源。

6) 根据信息生产领域不同,可划分为政府信息资源、公益机构信息资源、中介团体信息资源、企业公司信息资源等。

1.1.2 信息检索

1.1.2.1 信息检索的涵义

信息检索有广义和狭义之分。广义的信息检索指将信息按照一定的方式组织和存储起来并根据用户的需要查找相关信息的过程,包括信息存储和检索(Information Storage and Retrieval)两个过程。信息存储是指将大量分散的、无序的信息集中起来,根据信息源的内容特征和形式特征,经过收集、整理、分类、标引等处理,使其有序化、系统化,并按一定的技术要求建成一个具有检索功能的检索系统或工具;而信息检索是指利用编制好的检索系统或检索工具,查

找出与用户需求相匹配的特定信息。

狭义的信息检索仅指对信息的查找,即指根据用户的需求,利用检索系统或检索工具,在大量信息中迅速准确地查找出符合用户特定需要的信息的过程。

1.1.2.2 信息检索的基本原理

信息检索包括信息的存储和检索两个既相互独立又密切联系的互逆过程。

信息存储过程是指,从各种各样的分散的信息资源中,搜集有用的信息,对搜集到的信息进行主题内容的分析,总结出能够准确表达该信息主题内容的概念,并借助检索语言(如检索词表)把分析出来的概念转换成检索系统能识别的词语(一些自然语言检索系统不需要转换),再按照一定的规则将收集整理过的信息组织成可供检索利用的数据库,并存储在一定的载体上。

信息检索与信息存储过程相反。信息用户在生活和学习中产生了信息需求,对信息用户的信息需求进行主题内容分析,找出能准确表达该主题内容的概念,并借助检索语言(如检索词表)把分析出来的概念转换成检索系统能识别的语言,再按照一定的规则制定检索策略,形成检索提问式,提交到检索系统中查找并获取用户所需要的信息,然后输出检索结果。

1.1.2.3 信息检索的类型

(1)根据检索手段分:手工检索、计算机检索。

①手工检索:利用印刷型检索工具如文摘、目录、索引、题录等检索工具达到查找和获取文献信息的目的。

②计算机检索:可分为联机检索、光盘检索和网络检索,用户借助计算机和通讯网络,达到查找和利用数据库中的信息资源的目的。与手工检索相比,计算机检索具有检索范围广、检索速度快、检索效率高等突出优点。

(2)根据数据格式和检索技术的层次分:文本信息检索、多媒体信息检索、超媒体信息检索。

①文本信息检索:针对文本的信息检索,主要包括对文摘数据库检索和全文数据库的检索,前者仅能提供文献的线索(即题录)和文摘,而后者则可以直接根据文献资料的内容进行检索,并获取全文,发展较快也较成熟,文本信息检索仍然是大部分搜索引擎的基础。

②多媒体信息检索:是基于内容的信息检索技术以多媒体信息(如文字、声音、图像、动画等信息)为检索对象的信息检索。

③超媒体信息检索:是基于超媒体系统进行的检索,超媒体是一种非线性的网状结构,用户不需要按顺序阅读而是可以通过交叉链接的方式选择感兴趣的内容阅读,超媒体可以容纳包括文字信息、声音、图像、动画等各种静态或动态信息。

(3)根据存储和检索的内容分:文献检索、数据检索和事实检索。

①文献检索:是指检索用户所需文献的线索或者文献原文的检索过程。文献检索结果一般是有关某课题或满足用户特定需求的一组文献。例如检索有关“肝癌介入治疗”的文献。

②数据检索:是指以数值型信息为检索对象,查找用户所需要的数据信息。其检索对象包括化学分子式、调查数据、统计数据、公式等。例如查找中国某一年各省的GDP或某一种传染病的发病率等。

③事实检索:是指以特定的事实信息作为检索对象,利用事实数值型数据库或特定的工具书查找用户所需要的描述性事实。其检索对象包括企业、机构、人或事物的基本情况。例如查

找某个机构的机构名称或某一个专业术语或某人的生平事迹等。

1.2 信息检索技术和工具

1.2.1 信息检索技术

1.2.1.1 布尔逻辑检索(Boolean Logical Searching)

利用布尔逻辑算符进行的检索。常用逻辑运算符:逻辑与(AND),表示用 AND 连接的两个检索词 A 和 B 必须同时出现,文献才能被检索到,常用来缩小检索范围,提高查准率;逻辑或(OR),表示用 OR 连接的两个检索词同时出现或者只要有其中一个词出现在文献中该文献就能被检索到,常用来扩大检索范围,提高查全率;逻辑非(NOT),比如 A NOT B,表示被检出的文献中只包含 A 不包含 B,常用来缩小检索范围,提高查准率。

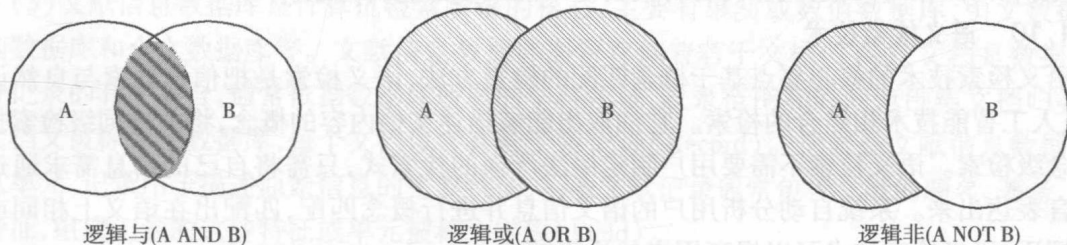


图 1-1 布尔逻辑算符

1.2.1.2 截词检索(truncation searching)

用截词符号“?”“*”或“\$”加在检索词的前后或中间,以检索同一词根或概念相关的词。截词检索分为有限截词和无限截词两种方式,如“*”用在词尾,代表任意个字符,“?”表示零个或一个字符,n个“?”并列使用表示可添加的字符数小于等于n。截词检索可以扩大检索范围,提高查全率。

1.2.1.3 词干检索(autostemming)

系统自动将相同词干的词检出,如输入 work 后,系统会将 working、worked、works、worker 等检出。有的检索系统用功能选项的方式,有的则用运算符“\$”。适用于西文数据库检索,但对中文网站的优化也有借鉴意义。

1.2.1.4 邻近检索(proximity searching)

邻近检索(proximity searching)是用来控制检索词之间位置关系的检索技术,通过用位置算符来实现,比如常用的位置算符 near、same。“near”表示两个检索词位置邻近,形如“冠心病 near 高血压”,两词出现顺序没有要求。“same”要求两个检索词出现在同一个句子中。

1.2.1.5 字段检索(field searching)

又称字段限定检索,利用检索词出现在文章的不同字段进行的字段限定检索。常见形式:ti=糖尿病,要求检索出在题名字段中出现“糖尿病”的文献。

1.2.1.6 精确短语匹配

精确短语匹配要求所检短语以完整形式出现,在许多检索系统中用双引号限定。如“肝

癌的介入治疗”。

1.2.1.7 相关信息反馈检索

相关信息反馈检索:一般由检索系统自动进行反馈,将与检出结果存在某种程度相关的信息检索出来的检索技术,如 PubMed 检出结果的每条记录下都有“Related Articles”。

1.2.1.8 可视化信息检索技术

可视化信息检索技术是指将用户提问、文献信息、情报检索模型以及各种信息检索的过程中不可见的内部语义关系等信息以可视化的形式在有限的电脑屏幕前系统高效地呈现出来的一种技术。方便快捷地帮助用户理解检索结果之间的关联,更好地把握检索方向,以提高信息检索的效率与性能。目前,信息可视化信息主要有文档的可视化和 Web 的可视化。

1.2.1.9 多媒体信息检索技术

多媒体信息检索技术是基于内容特征的检索,如图像的形状、颜色,声音的音色、音调等。它结合模式识别、图像处理、计算机视觉等多种基础技术,直接对视频、图像、音频的内容进行分析,抽取特征和语义,建立以内容特征为基础的索引并进行检索。

1.2.1.10 语义检索技术

语义检索技术的核心要点基于概念匹配的检索方法,语义检索是把信息检索与自然语言技术、人工智能技术相结合的检索。它抽取能够描述文档内容的概念,将关键词级检索提升到概念级检索。语义检索不需要用户制定语法严谨的检索式,只需将自己的信息需求通过自然语言表达出来。系统自动分析用户的语义信息并进行概念匹配,匹配出在语义上相同或相近的词语,语义检索技术可以提高用户的检索效率。

1.2.2 信息检索工具

信息检索工具是指报道、存储信息并提供检索途径的工具。它包括两个方面的职能:一方面是著录相关信息的特征,形成信息线索;另一方面提供检索途径,使用户按照一定的方法可以查找到相关信息。

按照加工处理信息的手段不同,通常将信息检索工具分为印刷型检索工具和计算机检索工具。

1.2.2.1 印刷型检索工具

印刷型检索工具是指对文献信息进行汇集、加工形成以印刷型载体形式出现的目录、文摘、索引等检索工具,是 20 世纪 80 年代以前应用最为广泛的一种检索工具。印刷型检索工具具有检索方便、灵活、判断准确的优点。但是检索速度相对较慢,不能随时根据需求修改检索策略,难以实现多元概念的检索。目前,许多印刷型检索工具都已开发出电子版本,成为计算机检索工具,如国内的《中国药学文摘》《中文科技资料目录》;国外的《科学引文索引》(Science Citation Index, SCI)、《医学索引》(Index Medicus, IM)、美国《化学文摘》(Chemical Abstract, CA)等。

1.2.2.2 计算机检索工具

计算机检索工具又称计算机检索系统,是用电子技术、计算机技术、网络通讯技术等构成的用于存储和查找信息的检索系统。具有更新速度快、检索途径多、检索效率高、检索结果输出灵活等特点。

(1) 计算机检索系统的逻辑构成主要指它所包含的子系统,通常包括:①信息源选择与采

集子系统:它的任务是连续不断地采集各种信息,以满足服务对象的需求;②内容分析与标引子系统:根据一定的规则和程序对信息进行分析,提炼信息的外部特征和内容特征,使之成为存储和检索的依据;③建库子系统:建立并维护可以用于检索的数据库;④提问处理子系统:识别用户的各种检索操作;⑤用户接口子系统:在用户与系统之间建立通讯接口;⑥词表管理系统:管理和维护词表,支持用户的各种查询操作。

(2)计算机检索系统的物理构成主要包括:①硬件:是系统各种硬件设备的总称,主要包括主机、输入装置、输出装置、外存储器和网络设备等;②软件:指检索程序,它由系统检索软件和维护软件构成,系统检索软件用户通过检索软件进行检索,是用户与系统的接口,通常有菜单式、指令式和智能接口等类型,系统维护软件如数据库管理程序,能保障检索系统的高效运转;③数据库:它是计算机检索系统的重要组成部分,是计算机存储设备上存放的数据的有序集合;④通讯网络:利用通信技术和公共数据传输技术将各个计算机连接起来,网络上的每台计算机都可对网络中的数据库进行访问,实现资源的共享。

(3)文献信息数据库是计算机检索系统的核心,主要有事实或数值数据库、引文数据库、文摘数据库和全文数据库等。文献信息数据库通常存储着若干文档(File),文档是数据库中部分记录的有序集合,通常依据数据库所属的学科领域、收录范围和时间范围给文档归类,因而文档又被称为子数据库,每个文档包含有若干条记录(Record)。记录是文献信息数据库的信息单元,记录用于描述原始信息的主要特征,一条文献记录通常包含文献的题名、著者、出处等特征,组成记录的这些特征或单元被称为字段(Field)。

1.3 信息检索途径与策略

1.3.1 信息检索途径

信息检索途径即查找所需信息的路径和方法,是检索系统的检索入口,检索系统往往会提供搜寻获取资源的各种标识,在计算机检索中一般表现为字段检索。信息检索途径紧紧依赖于信息存储系统,然而信息存储系统又是根据信息特征对信息进行加工和整序。根据不同的信息特征用户可以根据需求选择不同的信息检索途径。

1.3.1.1 自由词检索途径

自由词检索途径是利用自由词(Text Word)作为检索入口查找文献的途径。自由词是检索工具中常出现的任意词汇,主要包括关键词、题名或文摘或全文中直接出现的词等,但一般不包括无实际意义的词汇。自由词检索用词直观、自由、灵活,符合用户习惯,因而是应用最为广泛的检索途径。但自由词不能准确反映概念间的等级从属关系或一一对应关系,因而应用自由词检索途径时应同时考虑自由词的近义词、同义词等相关词进行检索,避免漏检。

1.3.1.2 主题词检索途径

主题词检索途径是利用叙词(Controlled term)作为检索入口查找文献的途径。叙词是经过规范化和优化处理的词汇,由主题词表控制,如《中医药学主题词表》。主题词检索途径可以提高查准率,但维护成本高。提供主题词检索途径的检索系统主要有PubMed、CBMDisc、EV2(工程索引)等。

1.3.1.3 分类检索途径

分类检索途径是把文献信息在分类体系中的位置(分类号或类目名称)作为检索入口查找文献的途径,可方便查找某一学科或某一专业的全部文献信息文。常用的分类法有《中国图书馆分类法》。

1.3.1.4 著者检索途径

著者检索途径是利用文献上署名的编者、作者、专利权人或机构团体名称等作为入口检索文献的途径。查外文数据库时,需注意通常采用姓在前用全称,名在后用首字母缩写的方式进行检索,如:Ulrica Ann Simth 需要用 Simth UA 进行检索。由于存在著者同名等问题,因此,需要用户在检索时还要根据作者的专业领域进一步鉴别。

1.3.1.5 题名检索途径

题名检索途径是利用题名(如文献篇名、书名等)作为入口检索文献的途径。题名是文章内容的高度浓缩,无论是手工检索还是计算机检索,题名检索途径都占有较重要的地位,将提问词限制在题名中可以显著提高查准率。

1.3.1.6 序号检索途径

序号检索途径是利用文献特有序号作为检索入口查找文献的途径。常见序号有 ISBN 号、报告号、专利号、化学物质登记号等。由于序号通常具有唯一性、简短和实用性等特点,因此利用序号来查找相关信息能提高检索速度和检索的准确性。

1.3.1.7 机构检索途径

机构检索途径是以机构名称为检索入口查找信息的途径。有的数据库将机构名称合并到地址字段(Address)。

1.3.1.8 引文检索途径

引文检索途径是指从参考文献入手查找文献的途径。文献之间的相互利用不仅能体现科研人员的相互交流,而且还能表现有关文献在内容上的联系。如美国的《科学引文索引》(SCI)。

1.3.1.9 缺省检索途径

缺省检索途径指系统自动在多个默认检索字段中同时进行检索。例如在 CBM 的缺省字段输入检索词检索则系统同时在篇名、摘要、主题词、关键词及刊名等字段进行检索。

1.3.1.10 其他检索途径

其他检索途径,有些数据库还提供了能满足特殊需求的检索途径,比如一些专利数据库提供专利号(patent number)检索,《化学文摘》提供分子式(formula)检索;《生物学文摘》提供生物体(organism)检索等。

1.3.2 信息检索策略与步骤

1.3.2.1 信息检索策略

检索策略有广义和狭义之分,从广义上说,检索策略是在充分分析课题内容的基础上,选择合适的检索平台,发掘检索点、确定检索词、确定检索工具与检索途径及其相互间的逻辑关系,直到确定最佳实施方案的一系列的科学措施。狭义上则仅仅把检索表达式的构造称作信息检索策略。好的信息检索策略可以显著提高信息检索的效率,逐渐完善检索策略能使检索过程和检索结果达到最优化。