YIXUE WENXIAN SHUJUKU JIANSUO JISHU YU CAOZUO JIAOCHENG

医学文献数据库

检索技术与操作教程

张桂云 龙莉艳 赵炜 主编

中国科学技术出版社

医学文献数据库检索技术与操作教程

张桂云　龙莉艳　赵　炜　主编

中国科学技术出版社
·北　京·

图书在版编目(CIP)数据

医学文献数据库检索技术与操作教程/张桂云,龙莉艳,赵炜主编. —北京:中国科学技术出版社,2005.8
ISBN 7-5046-4099-9

Ⅰ.医… Ⅱ.①张… ②龙… ③赵… Ⅲ.医学—情报检索—教材 Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(2005)第 072002 号

中国科学技术出版社出版
北京市海淀区中关村南大街 16 号 邮政编码:100081
电话:010-62103210 传真:010-62183872
http://www.kjpbooks.com.cn
科学普及出版社发行部
北京国防印刷厂印刷
*
开本:889 毫米×1194 毫米 1/16 印张:9.5 字数:250 千字
2005 年 8 月第 1 版 2005 年 8 月第 1 次印刷
印数:1—1500 册 定价:26.00 元

《医学文献数据库检索技术与操作教程》
编委会

主　编　张桂云　龙莉艳　赵　炜

副主编　马　良　夏　蕾

编　者　（以姓氏笔画为序）

马　良　龙莉艳　李彭元　张桂云

张　磊　范吉莲　郝俊勤　赵　炜

夏　蕾　梁蜀忠　羡秋盛

责任编辑　许　英

封面设计　陈建生

责任校对　刘红岩

责任印制　王　沛

前　言

随着信息技术的发展和电子产品的普及，计算机医学文献检索已成为当今医生获取信息的一个重要手段。解放军总医院已于1996年将硕士研究生的《医学文献检索与利用》课程改为必修课，并坚持每年为博士研究生开设同样内容的讲座。在授课过程中，我们考虑到这是一门实践性强、内容更新快的科学方法课，因此采取以技术操作为主，以解决学生工作、学习中的实际需求为目的的授课方式，达到以实践加深理论理解的目标，取得很好的效果。近年来我们还发现《医学文献检索与利用》的好教程也在不断增多，从1994年至今，我们已试用三种教程，这些教程不但有信息学基本理论，而且根据计算机技术和网络技术的发展速度及时增加和充实了大量电子信息的介绍、网络检索和信息发布等新技术，内容非常丰富。但尽管这样，授课过程中仍然存在部分学生面对计算机总是在不停地询问下一步怎么办的情况，我们发现仅仅依靠教程，无法帮助学生尽快将检索理论与实践相结合，因此迫切需要一个技术操作教程来帮助他们熟悉各类常用的检索系统，并在操作过程中随时讲述遇到的信息检索理论和信息检索术语，以配合各类以理论为主线的《医学文献检索与利用》教程的使用。基于上述想法，本书依托解放军总医院丰富的临床医学文献资源，根据我们多年的教学经验和电子阅览室读者使用情况分析，挑选了七个常用的中西文临床医学文献数据库，详细介绍其使用流程和检索技巧，并将众多的检索规则从实用的角度进行归纳和总结，以帮助学生尽快掌握信息检索技能。本书可以配合《医学文献检索与利用》教程的使用，也可以单独作为没有机会系统接受医学文献与检索课程的医生的技术操作手册。

本书共分八章。第一章为信息检索基本知识，阐述信息检索的概念和发展趋势、信息检索的原理和分类、信息检索工具和信息检索语言、计算机信息检索技术、信息检索步骤和检索策略的制定、信息检索结果的评价。第二章至第八章主要介绍七个常用的中西文临床医学文献数据库的使用方法，它们分别是MEDLINE数据库、EMBASE（荷兰医学文摘）、CBMDisc（中国生物医学文献数据库）、CMCC（中国生物医学期刊文献数据库）、中国学术期刊全文数据库、LWW医学期刊全文数据库和ProQuest Medical Library全文数据库。

由于信息更新快，可能会发现某些细节与实际不符。同时由于编者水平有限和编写时间紧迫，书中难免错误疏漏，恳请读者批评指正。

编　者

2005年4月30日

目　录

第一章　信息检索基本知识

第一节　信息检索的概念和发展趋势

一、信息

什么是信息(information)？英国科学家波普尔（K. Popper）认为信息分为三大类：第一类是有关客观物质世界的信息，即信息是事物存在的方式及其运动规律、特点的外在表现形式。第二类是有关人类主观精神世界的信息，它反映人类所感受的事物运动状态及其变化方式，处于意识和思维状态的信息。第三类是有关概念世界的信息，它反映人类所表述的事物运动状态及其变化方式，用语言、文字、图像、影视数据等各种载体来表示。信息领域的研究对象是第三类信息。

实际上，信息至今还没有严格统一的定义，关于信息的多种阐述只是不同的领域从不同的专业角度分别给予的解释。广义地来说，信息是指事物运动的状态与方式的反映，是自然界、人类社会和人类思维活动中存在的一切物质的一种属性。任何的消息、知识、数字、文字、程序等都是信息，它负载于不同的载体，表现为不同的形式，能够对人的感官系统形成刺激，进而被人类所利用。Information 在我国有两种比较正规的译名：信息和情报。近 10 年来，我国更倾向于使用信息这个概念，许多以情报命名的单位也因此更名为信息部门。

信息、知识和情报是三个既有联系又有区别的概念。信息是物质的一种普遍属性，是物质存在的方式和运动的规律与特点，它普遍存在于自然界、社会和人类思维活动中。知识是人们在改造世界的实践中所获得认识和经验的总和，是系统化、理论化的信息。情报是被传递、被利用的知识，是知识中的精华，是产生新知识的催化剂。人类不断接受客观事物发出的信息，经过思维加工，获得对事物本质及其运动规律的认识，将信息转化成知识。人类为解决问题去搜寻所需要的知识，这部分对人类有用的、进入人类社会交流的运动着的知识转化为情报。随着人类认识世界能力的加强，信息不断转化为知识，然后又不断被转化为情报用于改造世界，人类获取知识和利用情报后再去创造新的信息，如此反复循环，信息、知识和情报都越来越丰富。

现在我们处在高速发展的信息社会，它的特点是信息量呈指数增长，信息检索技术发展迅速，信息产品渗透到各个领域，信息用户扩展到各个行业。因此加强信息建设应该包括两方面的内容，一是要求信息工作者加强对信息检索领域的研究，能够及时、有效地加工信息并提供优秀的信息产品；二是要求加强对信息利用者的信息素质的培养，要求最终利用者能够敏锐地捕捉信息，并具备信息查询、分析和再利用的能力，从而摆脱信息工作者这个信息中介所带来的信息滞后、信息歧义等缺点。

信息素质的概念最早是由美国信息产业协会主席保罗·车可斯基于 1974 年提出来的。他把信息素质定义为“利用大量的信息工具及其主要信息源使问题得到解答的技术和技能”，后来又将其解释为“人们在解答问题时利用信息的技术和技能”。现在通常将信息素质定义为人所具有的对信息进行识别、加工、利用、创新、管理的知识、能力与情意等各方面基本品质的总和。它包括信息意识、信息能力和信息道德三方面。其中信息意识主要表现为人们对信息重要性认识的自觉程度，捕捉信息的敏感程度，能从信息的角度出发来感受、理解和评论自然界、社会中的各种现象等。

信息能力是即人们获取信息，处理信息，利用信息和创造信息的能力。它是构成信息素质的核心。信息道德指在整个信息活动中调节信息创造者、信息服务者、信息利用者之间相互关系的行为规范的总和。只有信息利用者提高信息意识，丰富信息检索知识，强化信息检索技能，加强信息道德建设，才能把握科学研究的主动权，才能推动科学进步和社会发展。

二、信息检索的概念、作用

信息检索是指从任何信息集合中查找并获取特定的相关信息的过程和活动。广义的信息检索包括信息存储和检索两个过程（Information Storage and Retrieval）。信息存储是指工作人员将大量无序的信息集中起来，根据信息源的外表特征和内容特征，经过整理、分类、浓缩、标引等处理，使其系统化、有序化，并按一定的技术要求建成一个具有检索功能的工具或检索系统，供人们检索和利用。而检索是指运用编制好的检索工具或检索系统，查找出满足用户要求的特定信息。狭义的信息检索（Information Retrieval）是针对信息用户的，仅指信息的查找和获取过程。

信息检索的意义在于使信息有序化，利于大家快速查找和获取自己所需的信息，从而达到以下几个目的：①借鉴别人的研究成果，节约自己的研究时间；②继承和利用别人的研究成果，避免重复劳动；③客观地评价自己的科研项目和研究成果的水平。

三、信息检索的发展趋势

专业化的信息检索起源于图书馆的参考咨询和文摘索引工作，从19世纪下半叶开始发展，至20世纪40年代，索引和检索已成为图书馆独立的工具和用户服务项目。随着1946年世界上第一台电子计算机问世，计算机技术逐步走进信息检索领域，先后经历了脱机检索、联机检索、光盘检索和网络检索四个阶段，著名的美国国家医学图书馆的MEDLARS系统、DIALOG系统、BIBLIOFILE、WWW分别是这四个阶段的代表。目前信息检索领域呈现联机检索、光盘检索和网络检索三种检索形式并存及一个数据库多种载体形式并存的局面。计算机技术应用到信息检索领域中，同时也促进了信息检索理论的发展，逐步形成了布尔检索模型、向量检索模型、模糊检索模型和概率检索模型等成熟的检索模型。

目前，信息检索已经发展到网络化和智能化的阶段。信息检索的对象从相对封闭、稳定一致、由独立数据库集中管理的信息内容扩展到开放、动态、更新快、管理松散的Web内容；信息存储和检索的方法也相应地由线性组织向网状结构过渡；信息检索的结果从提供目录、文摘等相关的二次信息发展到可以直接获得电子版的全文；信息检索的用户也由原来的情报专业人员扩展到包括商务人员、管理人员、教师学生、各专业人士等在内的普通大众，他们对信息检索从方式到结果都提出了更高、更多样化的要求。因此，适应网络化、智能化以及个性化的需要是目前信息检索发展的新趋势。

1. 智能检索

智能检索是将人工智能技术与信息技术结合，发挥人工智能的作用，它涉及信息检索与知识检索的结合、数据库技术与知识库技术的结合、数据处理与知识处理的结合等。智能检索改变了传统的全文检索技术基于关键词匹配进行检索带来的查不全、查不准的现象，因为智能检索充分利用分词词典、同义词典改善检索效果，进一步还可在知识层面或者说概念层面上辅助查询，通过主题词典、上下位词典、相关同级词典，形成一个知识体系或概念网络，给予用户智能知识提示，最终帮助用户获得最佳的检索效果。另外，智能检索还包括歧义信息的检索处理，通过歧义知识描述库、全文索引、用户检索上下文分析以及用户相关性反馈等技术结合处理，高效、准确地反馈给用户最需要的信息。

2. 知识挖掘

知识挖掘是按照既定的目标对大量的数据进行探索，揭示隐含其中的规律并进一步将之模型化的先进、有效的方法。知识挖掘的目的是将大量非结构化的多媒体信息融合成有序的、分层次的、易于理解的信息，并进一步转换成可用于干预预测和决策的知识。知识挖掘其实是一个智能化、自动化的过程。知识挖掘主要包括摘要、分类（聚类）和相似性检索等方面内容。

知识挖掘是信息处理新技术，又是一门受到多学科领域的研究者共同关注的边缘学科，因此“知识挖掘”还有“数据发现”、“数据开采”、“知识抽取”、“信息发现”、“知识发现”、“智能数据分析”、“探索式数据分析”、“信息收获”和“数据考古”等提法和不同的定义。目前研究最多的是文本挖掘技术，它的目的是帮助人们更好地发现、组织、表示信息，提取知识，满足信息检索的高层次需要。

3. 异构信息整合检索

在信息检索分布化和网络化的趋势下，信息检索系统的开放性和集成性要求越来越高，需要能够检索和整合不同来源和结构的信息，这是异构信息检索技术发展的基点，包括支持各种格式化文件的处理和检索；支持多语种信息的检索；支持结构化数据、半结构化数据及非结构化数据的统一处理；和关系数据库检索的无缝集成以及其他开放检索接口的集成等。从目前实践来讲，该技术的完美实现还有待基于自然语言理解的人机交互以及多媒体信息检索整合技术的进一步突破。美国 OVID 技术公司（Ovid Technologies）的 OVID 检索平台就成功地整合了 LWW、MEDLINE 和循证医学（Evidence - Based Medicine Reviews）等不同数据格式的数据库，为用户提供了统一友好的检索界面。

第二节　信息检索的基本原理和分类

一、信息检索的基本原理

信息检索实际上是指两个互逆的过程：一是存储过程，即把大量分散无序的信息进行科学的分析、标引，然后有序地组织成一个集合；二是检索过程，根据检索系统的组织规则，制定检索提问策略，将满足用户需求的信息提取出来。在存储与检索这两个过程中通过一定的方法和手段使所采用的特征标识达到一致，以便有效地获得和利用信息源。其中，存储是检索的基础，检索是存储的目的，只有有序地存储才能有效地检索。信息存储和检索的全过程可用图 1 - 1 表示。

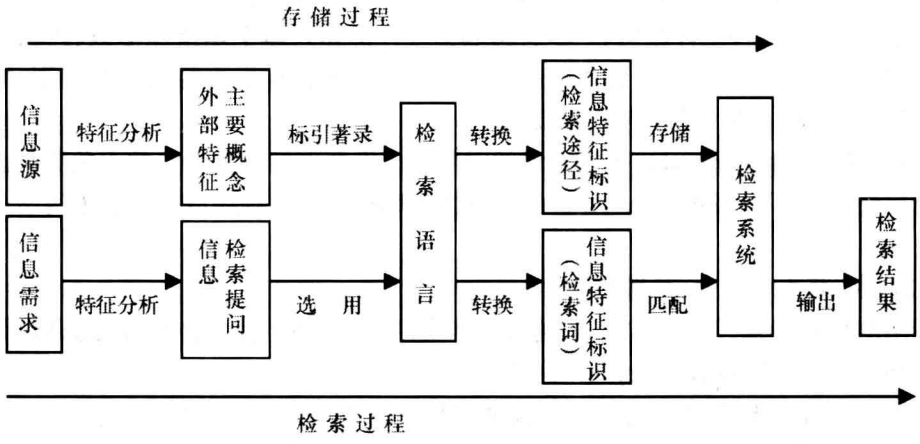

图 1 - 1　信息检索基本原理示意图

存储过程，主要对信息源进行分析、标引，将其外表和内容的特征（如信息源的标题、作者、来源和主题等）用特定的检索语言转化为一定的信息特征标识（如主题词、分类号和类目名称等），再将这些标识按一定的顺序编排后输入检索系统，从而为检索提供有规可循的途径。该过程主要包括对原始信息的概念分析、标引著录和形成索引等几部分，主要由信息工作者来完成。

检索过程，主要对信息需求进行分析，形成检索提问，然后根据检索语言转化成检索提问标识，提交检索系统，找出相匹配信息。这部分可以由信息工作者和专业人员完成。

从信息检索的基本原理来看，为了保证信息能存得进、取得出，就必须使信息存储与信息检索所依据的规则尽量做到一致。也就是说，为了检索过程的顺利进行和达到较高的检索效率，除了在存储和检索过程的各个环节必须依据一定的方法和规则外，还必须有统一的检索语言和名称规范作为存储人员和检索人员的共同依据。

二、信息检索分类

1. 信息检索分类

按照不同的分类标准，信息检索有多种分类方法。如按照检索方式分为手工检索和计算机检索；按照检索结果类型分为直接检索和间接检索。目前计算机信息检索占据主导地位，以信息线索为检索结果的间接检索仍是比较成熟、比较常用的检索方式。

最常用的分类标准是按信息存储和检索的内容分为文献检索、数据检索和事实检索。

（1）文献检索（Document Retrieval）：以文献为检索对象的信息检索称文献检索。关于文献（Document，Literature），我国国家标准《文献著录总则》（GB/T 3792.1—1983）中明确定义为："文献是记载有知识的一切载体。"具体地说，文献是将知识、信息用文字、符号、图像、音频等记录在一定的物质载体上的结合体。文献是一种重要的情报源，医学科研成果大多数是以文献的形式记载并得到学术认可。因此文献检索是信息检索中最重要、最广泛使用的一种，文献检索技术也是信息检索技术中研究最多、最成熟的一种。

早期的文献检索主要以二次文献为检索对象，以帮助用户获得原始文献的线索，如 MEDLINE 数据库是世界上收录范围最广、使用频率最高的医学文献数据库。近年来，全文数据库以其检索方便、结果直接而受到广大用户的欢迎，LWW、UMI 都是很受医生欢迎的外文全文数据库，《中国期刊网全文数据库》等中文全文数据库的发展也非常迅速。

（2）数据检索（Data Retrieval）：数据检索的对象是数值性数据，即具有数量性质并以数值形式表示的量化信息，包括各种统计数据、科学实验数据等。例如人体生理指标参数、药物的各种理化参数等都可以建立数据型数据库。这类数据库为用户提供多种检索途径和运算规则，方便用户查询，检索的最终结果可以供用户参考或直接利用。

（3）事实检索（Fact Retrieval）：事实检索的对象是已经存在的各种事实的有关资料。如美国国立癌症研究院建立的 Physician Data Query（PDQ）数据库提供有关癌症治疗和临床实验最新研究进展的内容。

2. 全文检索、超文本检索、超媒体检索

随着网络时代的到来，信息载体形式越来越丰富，信息传播途径多样化，因此出现了多种新的信息存储和检索方式，并得到广泛的应用和研究。

（1）全文检索：全文检索就是提供按照数据资料的内容而不是外在特征来实现的信息检索手段，并帮助用户获得信息源的全部原始资料。它能提供快捷的数据管理工具和强大的数据查询手段，快速帮助人们进行大量文档资料的整理和管理工作，并使人们能够快速方便地查到他们想要的

任何信息。全文检索由于其包含信息的原始性、信息检索的彻底性、所用检索语言的自然性等特点在近年来发展比较迅速，成为深受人们关注的一种非常有效的信息检索技术，尤其在文献检索中占据重要的位置。

(2)超文本检索：超文本以结点为单位组织信息，在结点与结点之间通过表示它们之间关系的链加以连接，构成表达特定内容的信息网络。超文本组织信息的方式与人类的联想记忆方式有相似之处，从而可以更有效地表达和处理信息。它是随着 WWW 的发展而产生的一种新型的信息存储和阅读方式。与传统文本检索相比，超文本是非线性组织，将信息内容按其内在联系划分为不同的层次和不同关系的知识单元，然后将这些知识依照其层次和关系组成一个网状结构，从而提供一种跳跃式扫读文本内容的手段。另外，它不仅显示对象内容，而且显示对象间的关系。

(3)超媒体检索：允许超文本的信息结点存储多媒体信息(图形、图像、音频、视频、动画和程序)，并使用与超文本类似的机制进行组织和管理。但在实际中，管理和组织多媒体信息比单纯的文本信息复杂得多，要将超文本的知识表示方法与文本、图形、图像、音频、视频、动画等多媒体信息的存储和处理技术相结合。超媒体技术广泛应用于与各种信息查询有关的方面，如教学、信息检索、字典和参考资料、商品介绍展示、旅游和购物指南、交互式娱乐等。

第三节　信息检索工具和信息检索语言

一、信息检索工具

1. 信息检索工具的概念

信息检索工具是存储、报道和查找信息线索或原始信息全文的工具，它通常是以压缩的形式出版。它是对信息进行搜集整理、特征分析和组织加工后的产物，同时又是信息检索的主要手段。

人们可以通过直接翻阅获得信息，也可以通过信息检索工具来系统查询信息。

2. 信息检索工具的分类

信息检索工具根据不同的分类标准有不同的分类方法，如按加工、处理信息的手段不同可分为手工检索工具和计算机检索工具；按照存储对象的内容可分为文献型检索工具、事实型检索工具和数据型检索工具；也可按照信息出版形式或载体形式等进行分类。

根据不同性质的信息著录格式不同，将检索工具分为目录型检索工具、题录型检索工具、文摘型检索工具、索引型检索工具、全文检索工具、主题指南和搜索引擎。

(1)目录型检索工具：目录型检索工具是记录具体信息的出版单位、收藏单位及其他外表特征的工具。它以一个完整的出版或收藏单位为著录单元，一般著录信息的名称、著者、出处等。我国西汉时期的目录学家刘歆撰写了我国第一部系统目录《七略》，可以算做我国信息检索行为的雏形。目录的种类很多，如国家书目、联合目录、馆藏目录，是用户查找书籍、期刊、报刊等信息源的重要工具。

(2)题录型检索工具：题录型检索工具是以单篇文献为基本著录单位来描述文献外表特征(如文献题名、著者姓名、文献出处等)，无内容摘要，快速报道文献信息的一类检索工具。它与目录的主要区别是著录的对象不同。目录著录的对象是单位出版物，题录的著录对象是单篇文献。

(3)文摘型检索工具：文摘型检索工具是将大量分散的文献，选择重要的部分，以简练的形式做成摘要，并按一定的方法组织排列起来的检索工具。按照文摘的编写人，可分为著者文摘和非著者文摘。著者文摘是指原文著者编写的文摘；而非著者文摘是指由专门的熟悉本专业的文摘人员

编写而成。就其摘要的详简程度，可分为指示性文摘和报道性文摘两种。指示性文摘以最简短的语言写明文献题目、内容范围、研究目的和出处，实际上是题目的补充说明，一般在100字左右；报道性文摘以揭示原文论述的主题实质为宗旨，要做到基本上反映原文内容，讨论的范围和目的，采取的研究手段和方法与所得的结果与结论，同时也包括有关数据、公式，一般500字左右，重要文章可多达千字。

(4)索引型检索工具：索引型检索工具是根据一定的需要，把信息源的有关知识单元或外部特征，如书名、刊名、人名、地名、语词等，按照一定的规则进行标引、编排，并指明出处，为用户提供信息源线索的一种检索工具。它是目前计算机文献检索工具中使用最广泛的一种。索引的类型是多种多样的，检索工具常用的索引类型有：分类索引、主题索引、关键词索引、著者索引、来源索引、引文索引等，因此索引型检索工具的优势在于能够提供多种检索途径，使得检索方便、快捷，缺点是仅获得信息源线索，不能提供原始信息。

(5)全文检索工具：全文检索工具存储的是信息源的原始资料，如整篇文章或整本书等，它是集检索功能与浏览原文功能为一体的检索工具。它的检索原理在于不仅能够满足提供特定关键词或作者、机构等辅助信息等检索途径的需求，而且可提供全文中任意词检索功能。全文检索工具由于其包含信息的原始性、信息检索的彻底性、所用检索语言的自然性等特点在近年来深受人们欢迎。

(6)主题指南和搜索引擎：WWW的出现，推动信息领域出现了新的信息格式和信息存储形式，那就是开放、动态、更新快、管理松散的Web内容和超文本技术。随着Web内容呈指数增长，新的检索工具应运而生。

1)主题指南。又被译为专题指南，或列表查询引擎，是由信息管理专业人员在广泛搜集网络资源及加工整理基础上，按照某种主题分类体系编制的一种可供检索的等级结构式目录。在每个类目及子类下提供相应的网络资源站点地址，并给以简单的描述，使用户能通过浏览该目录，在目录体系的导引下，发现、检索到有关的信息。

主题指南强调的是其浏览功能，优点是人工干预提高了主题指南返回结果的相关性；缺点是很难检索到较专深的信息，难以控制主题等级类目的质量，信息更新速度相对较慢，收录信息数量相对不足。因此主题指南主要适合于综合性、概括性的主题概念或对检索准确度比较高的查询。

2)搜索引擎。是基于Web内容的一种关键词检索工具。它的工作原理和过程是：使用自动索引软件来发现、收集并标引网页，建立数据库，以Web形式提供给用户一个检索界面，供用户输入检索关键词、词组或短语等检索项，代替用户在数据库中找出与提问匹配的记录，返回结果，按一定的相关度排序输出。

它分为独立搜索引擎和元搜索引擎(MetaSearch Engine，又称MegaSearch Engine，Multiple Search Enginge)。每个独立搜索引擎都有自己独有的搜索系统和一个包容因特网资源站点的独有数据库。其数据库由称为“Robots”(或“Spiders、Crawler”)的自动检索程序建立，不需人工干预。较为典型的搜索引擎有AltaVista、HotBot、Excite、Infoseek、Lycos等。元搜索引擎是为弥补独立搜索引擎费事费力之不足而出现的网上辅助检索工具。一般的独立搜索引擎检索范围仅限于其本身的数据库，而元搜索引擎则将用户的检索提问同时送达多个独立搜索引擎的不同数据库中进行检索，在很短时间内就能从这些数据库中检出相关记录的集合。目前，功能较强的元搜索引擎有：Meta Find、Inference Find、DogPile、Metacrawler等。搜索引擎具有检索面广、信息量大、信息更新速度快等优点，非常适用于特定主题词的检索。但其检索噪音较大，为检索带来负面影响。

二、信息检索语言

1. 检索语言的概念

检索语言是应信息的加工、存储和检索的共同需要而编制的人工语言，是表达一系列概括信息内容和检索课题内容的概念及其相互关系的一种概念标识系统。简言之，检索语言是用来描述信息源特征和进行检索提问的人工语言。它可以是从自然语言中或专业文献中精选出来并予以规范化的一套词汇，也可以是代表某种分类体系的一套分类号码，也可以是代表某一类事物的某一方面特征的一套代码等。

2. 检索语言的作用

检索语言在信息检索中起着极其重要的作用，它是沟通信息存储与信息检索两个过程的桥梁。在信息存储过程中，用它来描述信息的内容和外部特征，从而形成检索标识；在检索过程中，用它来描述检索提问，从而形成提问标识；当提问标识与检索标识完全匹配或部分匹配时，结果即为命中文献。检索语言的主要作用如下：

（1）使文献信息的存储集中化、系统化、组织化，便于检索者按照一定的排列次序进行有序化检索。

（2）对内容相同及相关的文献信息加以集中或揭示其相关性。

（3）标引文献信息内容及其外表特征，保证不同标引人员表征文献的一致性。

（4）便于将标引用语和检索用语进行相符性比较，保证不同检索人员表述相同文献内容的一致性，以及检索人员与标引人员对相同文献内容表述的一致性。

3. 检索语言的类型

检索语言是各种检索作用的语言的总称，每种语言都同时具有几种属性，按照不同的属性特点作为划分标准，检索语言有多种分类方法。如根据规范程度分为规范检索语言（亦称“规范语言”）和自然语言两大类；根据受控方式分为前控检索语言和后控检索语言；根据检索对象类型不同形成特定的检索语言，如：文献检索语言、网络检索语言（搜索引擎 Search Engine）、图书检索语言、档案检索语言等。

（1）根据检索对象的描述特征分为描述检索对象外部特征的检索语言和内容特征的检索语言。该种分类方法在文献检索系统中使用最多。

（2）描述文献外部特征的检索语言。根据文献的外部特征，如文献题名、著者、文献序号等不能改变的自然属性作为文献存储的表示和文献检索提问的出发点而设计的检索语言。由此而产生的索引系统主要有：文献题名索引系统、著者索引系统、文献序号索引系统、引文索引系统等，为文献查询提供了多种检索途径。

描述文献内容特征的检索语言，按其构成原理，可分为分类检索语言、主题检索语言和代码检索语言。

4. 分类语言

分类语言是用分类号和相应分类款目来表达各种主题概念的，它以学科体系为基础，将各种概念按学科性质和逻辑层次结构进行分类和系统排序。

分类语言一般以数字、字母或字母与数字结合作为基本字符，采用字符直接连接并以圆点（或其他符号）作为分隔符的书写法，以基本类目作为基本词汇，以类目的从属关系来表达复杂概念的一类检索语言。目前采用的分类语言主要有两大类：体系分类法和组配分类法。其中体系分类语言是最常用的一种，它按照学科体系从综合到一般、从复杂到简单、从高级到低级的逻辑次序逐级

展开，如《国际十进分类法》、《美国国会图书馆图书分类法》、《杜威十进分类法》、《中国图书馆图书分类法》等。

5. 主题语言

主题语言是指直接以代表信息主题概念的一组名词术语作为检索标识的一类检索语言。它具有直观、专指性强、使用灵活、适合计算机检索等优点，是信息检索中使用最为频繁的一种信息检索语言。

(1)主题语言具有如下特点：

1)主题词的概念性：主题词是主题词表中表达一定意义的最小的词汇单元。它不仅反映了一定事物的概念，而且它作为事物概念的表达形式而存在。

2)主题词的规范性：主要是指对主题词概念进行控制，使每一个主题词只能表达一个概念，这是主题词区别于自然语言的重要特点。自然语言中存在着各种同义词、多义词、同型异义词等，因此对从自然语言中抽取的词要进行同义规范、词义规范和词类规范等规范化处理。

3)主题词的组配性：由于主题词是建立在概念的基础上，因此可以选定多个主题词来表达和描述文献主题，以正确表达检索过程中复杂的概念。

4)主题词的语义性：可以通过参照系统来正确表达主题词之间的同义、属分和相关等含义上的相互关系，从而形成主题词之间科学的逻辑关系。

5)主题词的动态性：随着科学的不断发展，主题词表需要不断地增删和修改。

(2)主题语言又可分为标题词、元词、叙词、关键词。目前使用最多的是关键词法和叙词法。

1)标题词法：标题词是指从自然语言中选取并经过规范化处理，表示事物概念的词、词组或短语。标题词法是主题语言系统中最早的一种类型，它以严格规范的先组式的标题词作为信息的主题表识，只能选用“定型”标题词进行标引和检索，因此存在反映文献主题概念受到限制、使用不灵活等弊端，不适应时代发展的需要，目前已较少使用。

2)元词法：元词又称单元词，是指能够用以描述信息所论及主题的最小、最基本的词汇单位，在概念上不能再分解。元词法是通过若干单元词的组配来表达复杂的主题概念的方法。元词语言多用于机械检索，适于用简单的标识和检索手段(如穿孔卡片等)来标识信息。但由于单元词法是字面组配，所以时常出现组配错误的情况。组配作为单元词法最突出的特征已被叙词法继承和发挥。

3)叙词法：叙词法以叙词作为信息单元主题标识的主题词法。叙词(Descriptor)是指以概念为基础、经过规范化和优选处理的、具有组配功能并能显示词间语义关系的动态性的词或词组。一般来讲，选做的叙词具有概念性、描述性、组配性。经过规范化处理后，还具有语义的关联性、动态性、直观性。叙词法综合了多种信息检索语言的原理和方法，具有多种优越性，适用于计算机和手工检索系统，是目前应用较广的一种语言。我们常用的美国的《医学主题词表》(Medical Subject headings)、《中医药学主题词表》都是叙词表。

4)关键词法：关键词法是以关键词作为信息单元主题表识的主题词法。关键词(Keyword)是指出现在文献标题、文摘、正文中，能够表征文献主题内容并具有实质意义的语词。它的特点是使用灵活，适合非信息专业人员使用。它主要用于计算机信息加工抽词编制索引，因而称这种索引为关键词索引。中文医学文献检索中使用频率较高的《CMCC》数据库就是采用关键词索引方法建立的。

6. 主题语言的局限性

主题语言法能有效地提高查全率和查准率，但使用过程中仍存在如下局限：

(1)只适用于信息专业人员，不适用于最终检索用户。

(2)有一部分概念无法准确、专指地用检索语言来描述，规范化词表处理新词汇也有一定的

难度。

(3)标引人员和检索人员在使用检索语言时有很强的主观性,而且越是规范化程度高的检索语言对使用者技能要求越高。

(4)不同的检索系统使用不同的检索语言,兼容性差。

(5)人工标引速度慢,成本高;机器标引的智能化尚有待提高。

相对于受控语言,自然语言具有词汇专指度高、概念表达不失真、符合人们的检索习惯等优点,使得它越来越多地被用于信息检索领域,尤其是全文检索和网络检索领域。但自然语言检索系统效率也不能完全令人满意,由于自然语言的复杂性和计算机处理的机械性,自然语言检索系统中的许多技术尚待解决,因此现在有关专家建议采用粗泛的标引和自由词相结合的标引和检索模式。

第四节　计算机信息检索技术

信息检索技术是指应用于信息检索过程的原理、方法、策略、设备条件和检索手段等因素的总称。在实践中经常使用的计算机信息检索技术主要有:逻辑运算检索、位置检索、截词检索、字段限制检索、加权检索和聚类检索等。

一、计算机信息检索系统

1. 计算机信息检索系统的组成

根据信息存储与检索的原理,计算机信息检索系统分为信息采集、处理和录入子系统,词表和标引子系统,数据库和用户检索子系统四个部分,也是数据库从形成到使用的四个步骤。首先信息采集、处理和录入子系统用来选择和收集信息,然后对信息进行外部特征和内部特征的分析,再参照词表,运用标引子系统将信息的内部特征和外部特征进行描述,最终按照某种规则有序排列信息,形成数据库。以上步骤均由专家、信息工作者来完成,属于数据库的形成阶段。在数据库使用过程中,上述步骤在不断重复,因为新的信息要不断地添加到数据库中。用户检索子系统为用户提供与数据库的接口,用户按照系统规则描述自己的检索要求,系统识别后对数据库进行各种操作,最终为用户提供符合检索需求的信息内容。该部分内容可以用图 1 -2 来表示。

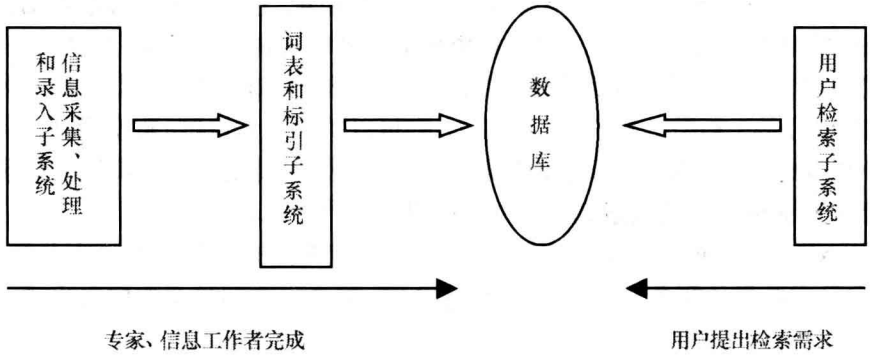

图 1 -2　计算机信息检索系统结构图

与手工检索工具相比,计算机检索系统具有数据更新快、检索途径多的优点。

2. 数据库的概念和结构

数据库(database)是计算机信息检索系统的核心部分,它是各类信息的有序集合,是专家和信

息工作者工作的最终结果，是用户的操作对象。

（1）文档：数据库由多个文档（file）组成。文档是数据库中一部分记录的集合，如 MEDLINE 数据库由于数据多，因此被分为现期文档和若干回溯文档。而且每个数据库都以下列形式来合理组织记录：建立一个顺排文档（serial file）和若干个倒排文档（inverted file），顺排文档和倒排文档之间互相关联，配合完成检索。顺排文档以记录为基本单位，按照信息序号顺序排列。在多数数据库中，顺排文档记载了信息的全部内容，因此对顺排文档进行检索，速度非常慢。倒排文档也称做索引，是从顺排文档中抽取有意义的检索标识，如主题词、作者姓名、出版年、关键词等，并进行排序生成的文档。该文档只包含信息的序号，检索时先在倒排文档中找到序号，然后再在顺排文档中找到该信息的全部内容。使用倒排文档进行检索，提高了检索效率。数据库建立的倒排文档越多，意味着提供的检索途径越多，检索效率也就越高。根据排序的检索标识不同，建立的索引分别被称做主题词索引、作者索引、出版年索引等。不同的数据库结构不同，建立的倒排文档的多少也不同。

（2）记录（record）：记录是数据库的基本单位，记录着每一条信息的全部内容。每条记录在数据库中都有惟一的一个记录序号。

（3）字段（field）：字段是记录的基本单位，描述一条信息的某方面特征。一条信息有若干个内部特征和外部特征，一个字段描述一个特征，用多个字段描述多个特征，就形成一个记录。

以文献数据库为例，一篇文献就是一个记录，文献的内部特征和外部特征分别用题名、作者、文摘、主题词、出版类型、语种等字段来描述，并根据题名、作者、主题词等字段建立题名索引、作者索引、主题词索引等，为用户提供题名检索、作者检索、主题词检索等检索途径。

二、计算机检索技术

1. 布尔逻辑检索

布尔逻辑检索是当今检索理论中最成熟的理论之一，也是构造检索表达式最基本、最简单的匹配模式。它的理论基础是集合论和布尔逻辑。它是通过布尔逻辑运算符来实现的，这些运算符能把一些具有简单概念的检索词（或检索项）组配成为一个具有复杂概念的检索式，用以表达用户的检索需求。布尔逻辑运算符有三种：逻辑与（AND）、逻辑或（OR）、逻辑非（NOT）。

（1）逻辑与（AND）是一种用于交叉概念和限定关系的组配，它可以缩小检索范围，有利于提高查准率。凡是用 AND 的检索式，AND 两侧的检索词必须同时出现在同一条记录中，该记录才算命中。如检索式“肝癌 AND 肺癌”，表明检索文中同时出现肝癌和肺癌的文献。

（2）逻辑或（OR）是一种用于并列关系的组配，它可以扩大检索范围，防止漏检，有利于提高查全率。凡是用 OR 的检索式，OR 两侧的检索词只要有一个在一条记录中出现，该记录就算命中。如检索式“肝癌 OR 肺癌”，表明检索文中出现肝癌或肺癌的文献。

（3）逻辑非（NOT）是一种排斥关系的组配，排斥关系组配用来从原来的检索范围中排除不需要的概念或影响检索结果的概念。如检索式“肝癌 NOT 肺癌”，表明检索文中只出现肝癌，不出现肺癌的文献。

逻辑运算符的优先顺序为 NOT，AND，OR。

布尔逻辑运算符的局限在于它没有反映概念之间的逻辑关系。它把概念与信息源之间的关系简单化，所有的逻辑关系都被简单的匹配代替，因而不能准确地描述文献，造成误检和漏检。比如检索式“肝癌 AND 诊断”，只表明肝癌和诊断两个词同时出现在文献中，但无法确定诊断是肝癌的限定词，即无法肯定命中文献是讲述肝癌的诊断方面的内容。

2. 位置检索

位置检索是控制检索词在原始文献中的相邻位置的一种运算，又称邻近检索、相邻度检索。它弥补逻辑运算带来的局限性，提高查准率。位置运算符隐含了逻辑运算符 AND 的含义，它要求连接的两个检索词必须同时出现，同时还进一步对这两个检索词的位置做了限定。

位置检索可以限定两个检索词同时出现在一个字段中或同一个句子中，或严格要求这两个词之间所允许相隔的单词数或字数。但对于具体检索系统来说，是否有位置运算，运算符检索格式如何，都是有差异的，只是原理相同。如 MEDLINE 数据库中的位置运算符是 NEAR 和 WITH。LWW 医学专集全文数据库中的位置运算符形式则是 ADJ。

3. 截词运算

截词运算符就是使计算机保留检索词中的相同词干部分，允许检索词可有一定范围的变化，这种功能可减少输入步骤，简化检索程序，扩大检索范围，从而节省时间，提高查全率。截词运算符特别适用于英文，一个词可能有多种形式，但它们的词干部分往往相同。中文数据库中，如果数据库索引采用单字索引，那么检索时只要减少检索词的字数，就可起到截词检索的作用。不同的数据库有不同的截词符，MEDLINE 数据库用“?”和“*”，LWW 医学专集全文数据库用“$”。因此使用截词运算符时，要注意的问题是词干部分不能太短。

截词运算分为有限截断和无限截断两种方式。有限截词指截词运算符在检索词中只能代表任意一个未知字符。无限截词指截词运算符在检索词中代表任意一串未知字符。截词运算也可以分为前方一致、后方一致、中间一致和中间屏蔽四种形式。在实际检索过程中，前方一致和中间屏蔽是使用最多的两种情况。前方一致指词干相同，词尾不同。中间屏蔽往往用于英美拼写形式不同的单词，这两种拼写方法在词的中间相差一个字母。

4. 字段限定检索

文献存在数据库中是以记录为存储单元的，每条记录又根据文献的特点分别以字段形式进行描述。同样的词出现在不同的字段含义是不同的，因此一般数据库都提供字段限定检索，以帮助用户准确地描述文献的特征，快速地查找所需内容。不同的数据库字段检索格式不同。如 MEDLINE 数据库中使用的字段限定符为“IN”，字段限定检索格式为“检索词 IN 字段标识”，限定字段检索还可以采用“字段标识 = 检索词”的格式。

5. 加权检索

加权检索是某些检索系统中提供的一种定量检索技术。加权检索的侧重点不在于判定检索词或字符串是不是在数据库中存在、与别的检索词或字符串是什么关系，而是在于判定检索词或字符串在满足检索逻辑后对文献命中与否的影响程度。加权检索的基本方法是：在每个提问词后面给定一个数值表示其重要程度，这个数值称为权，在检索时，先查找这些检索词在数据库记录中是否存在，然后计算存在的检索词的权值总和。权值之和达到或超过预先给定的阈值，该记录即为命中记录。

运用加权检索可以命中核心概念文献，因此它是一种缩小检索范围提高检准率的有效方法。但并不是所有系统都能提供加权检索这种检索技术，而能提供加权检索的系统，对权的定义、加权方式、权值计算和检索结果的判定等方面，又有不同的技术规范。

6. 聚类检索

聚类检索是在对文献进行自动标引的基础上，构造文献的形式化表示——文献向量，然后通过一定的聚类方法，计算出文献与文献之间的相似度，并把相似度较高的文献集中在一起，形成一个个的文献类的检索技术。根据不同的聚类水平的要求，可以形成不同聚类层次的类目体系。在这