SHUZIHUAXINXIZHIYUN
JIANSUOYU
LIYONG

★ ★ ★ ★ ★

数字化信息资源检索与利用

欧兆虎　主编

湖南人民出版社

G252.7
0612

数字化信息资源检索与利用

主　编：欧兆虎
主　审：敬　卿
副主编：许建兰　龚晓林　施燕斌　谢永强
编　委：（按姓氏笔画排序）
王　朗　王　群　许建兰　欧兆虎
施燕斌　谢永强　龚晓林　敬　卿

湖南人民出版社

图书在版编目(CIP)数据

数字化信息资源检索与利用/欧兆虎主编.—长沙:湖南人民出版社,2005

ISBN 7-5438-3854-0

Ⅰ.数... Ⅱ.欧... Ⅲ.数字技术-应用-情报检索 Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(2005)第 011429 号

责任编辑:唐长庚

装帧设计:陈 新

数字化信息资源检索与利用

欧兆虎 主编

*

湖南人民出版社出版、发行

(长沙市营盘东路 3 号 邮编:410005)

长沙海德印务有限公司印刷

2005 年 1 月第 1 版第 1 次印刷

开本:850×1168 1/32 印张:10.5

字数:240,000

ISBN7-5438-3854-0

G·903 定价:21.00元

目 录

此为试读，需要完整PDF请访问：www.ertongbook.com

第一章　信息检索基础知识

第一节　信息与信息检索

1. 信　息

信息可以是人脑中的信息，也可以是自然界或人造系统中的信息。著名控制论专家维纳（Wiener）在他的《信息控制论》中说："信息是人们在适应外部世界并使这种适应反作用于外部世界过程中，同外部世界进行交换的内容的总称。"正是信息构成高级社会的一种资源，是物质、能源以外的第三种资源。它像货币一样，流动起来可以增值，产生财富，推动社会进步。对信息检索的用户而言，更有意义的应是认识论层次上的信息意义，即信息是认识主体所感知或所表述的事物运动的状态格式。这一定义告诉我们，若要获得所需信息，必须具备一定的认识能力，其中包括信息意识、信息检索技能、信息组织加工能力和信息分析评价能力。

2. 信息的特征

2.1　客观性与普遍性

信息既不是物质，也不是能量，是客观事物普遍性的表征，

信息是无处不在，无时不有的普遍社会现象。

2.2 流动性与传递性

信息在事物之间的相互联系必定在信息的流动中发生。信息的传递性表现在人与人之间的消息交换，人与自动机、自动机与自动机之间的信息交换，动物界和植物界的信号交换，同时，人类进化过程中的细胞选择、遗传也被看作是信息的传递与交换。

2.3 多样性与综合性

信息在不同的领域具有多种不同的特性或表现形式，如客观事物中的各种自然属性；人工设备的技术特征；人类社会的各种社会特征；人脑中反映客观事物认识的思想、知识；人类交流信息过程中的声音、文字、图像以及用各种编码形式记录下来的数据、新闻、情报、消息等。各种形式的信息又常常以综合的方式表现事物的特征，所谓“多媒体”正是信息多样性和综合性的集中表现。

2.4 相对性与有效性

从信息作为事物相互联系的反映角度看，信息源不确定的程度或者信息源接受信息量的多少，均与信宿的状态有关。这一特征在人作为信宿接受信息的过程中表现得尤为明显。同一信息对具有不同认知水平的人所产生的作用和有效性也不相同。

2.5 积累性与价值性

信息通过人脑思维或人工技术设备的综合、加工和处理，不断积累丰富，提高其质量和利用价值。信息的质量和价值，实际上是对客观事物属性反映的深度和真实程度的认识。虽然信息是人类的一种重要资源，但信息只有被利用才会产生价值，否则，其价值或随时间的流逝而减少，或成为“信息垃圾”。

3. 信息检索

信息检索是从任何信息集合中识别和获取所需信息的过程及其所采取的一系列方法和策略。从原理上看，它包括存储与检索两个方面，如图 1－1 所示。

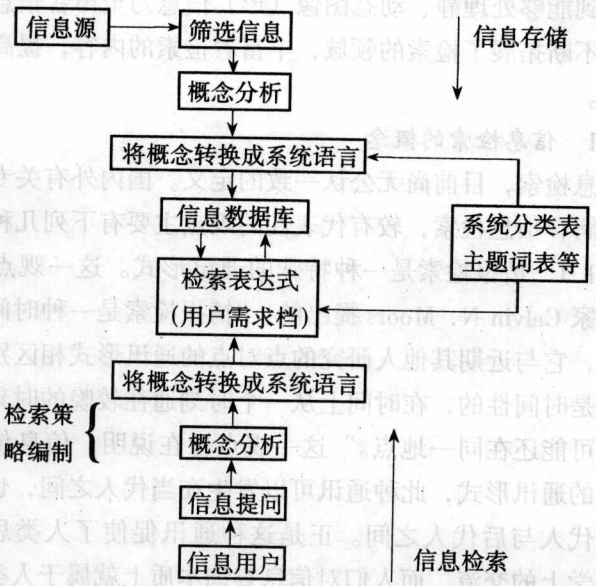

图 1－1　信息存储与检索原理

信息的存储主要包括对在一定专业范围内的信息进行信息特征描述、加工并使其有序化，即建立数据库。检索是借助一定的设备与工具，采用一系列方法与策略从数据库中查找出所需信息。存储是检索的基础，检索是存储的反过程。在现代信息技术条件下，信息检索从本质上讲，是指人们希望从一切信息系统是

高效、准确地查询到自己感兴趣的有用信息，而不管它以何种形式出现，或借助于什么样的媒体。

早期的信息检索，人们主要根据文献的内、外表特征，用手工方式实现。以计算机技术为核心的信息技术，开辟了信息处理与信息检索的新时期。从电脑处理数字信息发展到处理字符信息，又到能够处理静、动态图像（形）信息乃至声音信息等。这一过程不断拓展了检索的领域，丰富着检索的内容，提高了检索的速度。

3.1 信息检索的概念

信息检索，目前尚无公认一致的定义。国内外有关专家从不同角度解释信息检索，较有代表性的观点主要有下列几种：

3.1.1 情报检索是一种特殊的通讯形式。这一观点是由美国数学家 Calvin N. Moors 提出的："情报检索是一种时间性的通讯形式，它与近期其他人研究的点对点的通讯形式相区别"，"此种通讯是时间性的，在时间上从一个时刻通往较晚的时刻，而在空间上可能还在同一地点。"这一观点旨在说明，信息传递是一种特殊的通讯形式，此种通讯可以发生在当代人之间，也可能发生在当代人与后代人之间。正是这种通讯促使了人类思想、文化、科学上的交流。而人们对信息查询本质上就属于人类通讯的范畴。

3.1.2 从信息处理角度定义信息检索。这一观点的提出，旨在超越传统的"文献"范围，把包括动、静态声频、视频信息在内的各种数值信息统统纳入信息检索范围。如果将信息检索作为一门学科，它应该包括矩阵记数法、概率论、最优化理论、模式识别及系统分析技术等各学科领域的内容。

3.1.3 从传统文献检索角度定义信息检索。其代表人物是

前苏联情报检索专家米哈伊诺夫。他指出，情报检索，这是从大量的文献中查寻与情报提问所指定的课题（对象）有关的文献，或者是包括用户所需事实与消息的文献过程，这里谈到的文献，不仅指文献线索，也包括文献的片断，如章、节、段落以及与事实有关的直接情报等。

3.1.4　全息检索。我国王永成教授认为，全息检索就是“可以从任意角度，从存储的多种形式的信息中高速准确地查找，并可以任意要求的信息形式和组织方式输出，出口仅输出人们所需要的一切相关信息的电脑活动”。这里所谓的任意角度，是指要求检索系统从用户可能采用的检索需求作为出发点，并把这些出发点都设计成“检索入口”；所谓多种形式的信息，指的是在现代多媒体技术所能保证存储并输出文本、图像与声音信息的条件下，继续发展直至能输出超声频与超视频信息；所谓任意要求的信息组织形式，是指按用户需求组织对已检索到信息的输出，从而真正实现人－机检索过程中“以人为中心”的服务宗旨；所谓输出一切相关信息，从存储方面看，相关性表现在系统对存储的文本的外部特征、文本的内涵本体特征以及其他辅助性特征等的描述方面存在不可避免的差异，从检索一方看，用户对信息需求的认知能力、表述能力也同样影响检索效果。因此，相关性不仅是传统文献检索也是全息检索的基本特征和评价检索系统的重要参考指标。

3.1.5　概念信息检索。Chank、Kolodner 和 Dejong 认为概念信息检索是基于自然语言处理中对知识在语义层次上的析取，并由此形成知识库，然后根据对用户提问的理解来检索其中的相关信息。它与传统文献检索的不同之处在于，后者是基于关键词（主题词）为核心的标引与检索，而关键词在很多情况下并不适

合用于确切表达文献信息的概念和内容，因此误检与漏检在所难免。而概念信息检索倡导者认为，它可以对输入的原文内容中的概念而不是关键词来进行组织和安排，在对其进行语义层次上的自然语言处理基础上来获取相关的概念和范畴知识，然后，通过记忆机制将它们存储到知识库中以备检索。作为概念信息检索系统，一般由记忆机制、知识库、人机接口等部分组成。目前，国际上开发出一些实用性的概念信息检索系统，如哥伦比亚大学Lebowitz开发的RESEARCHER系统（主要用于阅读和理解用自然语言形式输入的专利文献摘要）；美国通用电气公司人工智能研究室研制的SCISOR－System for Conceptual Information Summarization，Organization，and Retrieval系统（专门用于处理与公司或企业有关的商业信息）。

3.2 信息检索的类型

3.2.1 信息检索按存储和检索的内容划分为文献信息检索、数据信息检索、事实信息检索。（1）文献信息检索。通常指的是检索系统存储的是以二次信息为对象（目录、索引、文摘）的信息，它们是文献信息的外部特征与内容特征的描述集合体。信息用户通过检索获取的是原文的“替代物”。（2）数据信息检索。是指检索系统中存储的是数值型数据，如科学技术常数、各种统计数据、人口数据、气象数据、市场行情数据、企业财政数据等，即事物的绝对值和相对值的数字。检索系统提供一定的运算推导能力，例如外推、内插、填补空缺数据，甚至列出曲线图或进行各种分析等功能。信息用户可用通过检索获得的经过核实、整理的数值信息再作定量分析。（3）事实信息检索。是指检索系统存储的是从原始文献中抽取的关于某一事物（事件、事实）发生的时间、地点和过程（情况）等方面的信息。它是数值信息和

此为试读，需要完整PDF请访问：www.ertongbook.com

系统数据信息的混合。一般先从系统中检索出所需信息后，再加以逻辑推理才能给出结论。例如 MIS 数据库中包含大量公司管理中有关人员、工资、销售统计、预测、产品规模等信息，这类信息主要是用于管理决策的。

3.2.2 信息检索按系统中信息的组织方式划分为全文检索、超文本检索和超媒体检索。(1) 全文检索。是指检索系统中存储的是整篇文章乃至整本书。检索时，用户可以根据自己的需要从中获取有关的章、段、句、节等信息，并且还可以进行各种频率统计和内容分析。随着计算机容量与运算速度的增大和提高，全文检索正迅速由最初的法律、文学领域扩大到更多的学科、专业。(2) 超文本检索。是针对信息在系统中的组织方式不同而言的。从组织结构上看，超文本的基本组成元素是节点（nodes）和节点间的逻辑联接链（links），每个节点中所存储的信息以及信息链被联系在一起，构成相互交叉的信息网络。与传统文本的线性顺序不同，超文本检索强调中心节点之间的语义联接结构，靠系统提供的复杂工具作图示穿行和节点展示，提供浏览式查询。其检索模式是从“哪里”到“什么”。而传统的文本检索系统则强调文本节点的相对自主性，其检索模式是从“什么”到“哪里”。(3) 超媒体检索。是对超文本检索的补充。其存储对象超出了文本范畴，融入了静、动态图像（形）以及声音等多种媒体信息。信息的存储结构从单维发展到多维，存储空间范围在不断扩大。需要说明的是，超文本和超媒体检索，二者的链都是有向的（单、双向并存），均面向浏览式查询。

4. 常用的检索方法

查找（Searching）就是实施检索策略、搜寻所需文献信息的

过程。如何查找，并没有一定之规可供遵循。同一个问题不同的检索者可能就有不同的查找方法，这是因为检索者在主观上受到他们的实际经验、知识结构、对检索工具了解的广度和深度、认识问题的方法、心理品质等因素的影响；在客观上又受制于检索工具、检索时间和物理环境等因素的影响。以下几种方法，无论对于计算机检索还是手工检索，都是常用的方法。

4.1 广度优先法

就使用网址而言，在不了解查询某一专题信息的 URL 地址时，可从提供信息总目的 Web 页面开始浏览，沿着专题链接层层查找，直至找到有关的内容为止。然后用“书签”保存这个页面的 URL，转向另一个分支。这种方法可以迅速获得较多的相关地址，然后进行筛选。就使用搜索引擎而言，国内外专家也建议先用链接页面多、响应时间快的引擎。

4.2 引文法（跟踪法）

文献之间的引证和被引证关系揭示了文献之间存在的某种内在联系，引文法（也有称为跟踪法）就是利用文献后所附的参考文献、相关书目、推荐文章和引文注释查找相关文献的方法。这些材料不仅指明了与读者需求最密切的文献线索，而且往往包含了相似的观点、思路、方法，具有启发意义。循着这些线索去查找，不仅利用了前人的劳动成果，省却了很多时间和精力，而且可能在原来的基础上有新的发现。利用引文法高效率地查找文献的最有用工具是利用引文索引。

引文法又可分为两种，一种是由远及近地搜寻，即找到一篇有价值的论文后进一步查找该论文被哪些其他文献引用过，以便了解后人对该论文的评论，是否有人对此作过进一步的研究、实验结果如何、最新的进展怎样等等。由远及近地追寻，越查资料

越新，研究也就越深入，但这种查法主要依靠专门的引文索引，如《科学引文索引》(Science Citation Index)、《社会科学引文索引》(Social Science Citation Index)。另一种较为普遍的查法是由近及远地追溯，这样由一变十，由十变百地获取更多相关文献，直到满足要求为止。这种方法适合于历史研究或对背景资料的查询，其缺点是越查材料越旧，追溯得到的文献与现在的研究专题越来越疏远。因此，最好是选择综述、评论和质量较高的专著作为起点，它们所附的参考文献筛选严格，有时还附有评论。

4.3　常规法

引文法的一个主要缺点是作者个人收集文献数量有限，不可能列出有关专题的全部文献，这一不足可用常规法来弥补。所谓常规法就是利用常规检索工具查找有关文献的方法，是信息时代应掌握的最基本的信息查找方法。现在对文献的书目控制手段已日趋完善，各种印刷版、缩微版、光盘版和网络版的检索工具层出不穷，有很大的挑选余地。用户应根据自己的检索知识和条件选用一种或几种检索工具。

常规法可分为顺查法、逆查法和抽查法。顺查法是以课题研究的起始年代为出发点，利用选定的检索工具如书目、索引、文摘由远及近地逐年查找。逆查法则相反，是由近及远地查找，起点是从最近发表的文献开始，直到设定终止的年代或查到所需资料为止。由于这两种方法都是利用检索工具，又是逐年逐卷地查找，遗漏重要文献的可能性就减少了，查全率比引文法高。但逐年查找的缺点是费时费力、检索工作量大，因此可以利用抽查法。抽查法是基于这样一个规律来查文献的，即任何一门学科的专题研究大体都像波浪起伏般发展，时而高潮，时而低潮。由于兴旺时期发表的文献量大，各种学术观点较为集中，如果针对课

题研究处于兴旺时期的若干年查找，则付出较少的时间可获得较为满意的检索结果。这是一种效率较高的查法，但必须熟悉学科或研究专题发展的历史。

4.4 交替法

交替法就是把引文法和常规法结合起来查找文献的方法，即先利用常规检索工具找出一批有用文献，然后利用这些文献所附的引文进行追溯查找，有此获得更多文献。这一方法是针对单纯用引文法所获得的情报价值越来越小的缺点提出来的。按照引文规律，有价值的文献在发表后最初几年（例如五年）内被引用的次数较多，但以后趋于减少。因此，追溯的年期应予限制。跳过追溯的那几年再用常规检索工具查出具有新价值的文献，然后再根据所附的参考文献追溯，并依此进行第三次或多次循环，直到获得的文献符合要求为止。

4.5 排除、限定和合取法

这实际上是将信息加工的方法融入检索中去。思维中使用排除这一概念，是指对查找对象的产生和存在的状态在时间和空间上加以外在否定。把这一方法移植到检索中就是在时间和空间上极大地收缩检索范围。如要查中国网络资源建设的文章，确定1994年以前 Internet 未进入中国，则可排除 1994 年以前的报刊资料，这就是排除法。限定法是相对于排除法而言的，指对查找对象在时间和空间上加以内在的肯定。排除的结果必然是限定，反之亦然。令人满意的答案往往不是完整地记录在某一篇文献中的。如果把不同资料中设计所需信息的记载都裁取下来，汇集在一起，再经过去粗取精、去伪存真的加工，构成一个完整的答案，这就是合取法。采用这一方法，不仅要对各类工具书触类旁通，灵活运用，还要学会分析来自各方面的庞杂的材料。

值得指出的是，不同的科学研究领域对文献信息的需求会有不同，对信息载体和检索手段的利用也会有差别，尤其是在社会科学领域与科学技术领域。科技文献具有较强的知识积累性，文献老化周期大大短于社科文献，新发表的论文大体上能将在它之前已有的相关知识吸收进去。因此科技信息的查询更注重期刊论文、研究报告、会议录和专利这类文献，在检索手段上更强调利用时效性强的现代化检索工具。而社会科学的研究除了要吸收新知识，也要注意以往的研究成果，几十年前的研究成果同样具有权威性和参考价值。它对图书这种包含比较成熟、定型的知识的文献往往给予更多的关注。社科文献的检索除了利用较新的载体和手段，也不能偏废传统的载体和手工检索方式，特别是多年积累的文献还没有条件电子化，这是读者在考虑检索方法时要注意的。

第二节　信息的类型与存储载体

1. 信息的类型

信息与人类智能活动有关的知识、技术、科学、文化、社会等密切联系在一起，其涉及范围如此之广，以至于很难用统一的标准进行分类。

1.1　按信息表现形式划分

1.1.1　文字信息。文字是人们为了实现信息交流、通信联系所创造的一种约定的形象符号。广义的文字还包括各种编码，如 ASCII 码、汉字双字节代码、国际电报与单元代码以及计算机中的二进制数字编码等都是一些符号的约定。这些文字、符

号、代码均是信息的表述形式，其内容再现于它们的结构属性之中。如基本笔画的不同组合，二进制码“0”和“1”的不同排列等，分别代表不同的信息内容。

1.1.2 图像信息。图像（形）是一种视角信息，它比文本信息直接，易于理解。人工创造的图像（形），如一张纸、一幅画、一部电影，大自然的客观景象等都是抽象或间接的图像信息。随着多媒体技术的发展，各类图像信息库将会极大地丰富人类生活。

1.1.3 数值数据信息。数值数据是“信息的数字形式”或“数字化的信息形式”。狭义的“数据”是指有一定数值特性的信息，如统计数据、气象数据、测量数据以及计算机中区别于程序的计算数据。广义的数据是指在计算机网络中存储、处理、传输的二进制数字符编码，文字信息、图像信息、语言信息以及从自然界直接采集的各种自然信息等均可转换为二进制数码，网络中的数据通信、数据处理和数据库等就是广义的数值数据信息。

1.1.4 语音信息。人讲话实际上是大脑的某种编码形式的信息转换成的语音信息的输出，是一种最普遍的信息表现形式。音乐也是一种信息形式，是一种特殊的声音信息，它是通过演奏方式表达丰富多彩的信息内容的。

1.2 按信息的出版类型划分

1.2.1 图书。包括专著、教科书、各种科普读物及各种专业参考工具书等。图书经编著者精心选择，反复斟酌后写成，其内容系统、成熟、定型，信息经筛选，可靠性强，是人们从事学习、研究不可缺少的信息来源。不过，传统印刷业图书出版周期较长，体积大，更新速度慢，电子版图书的出现将弥补这一缺陷。

此为试读，需要完整PDF请访问：www.ertongbook.com