普通高等教育“十一五”国家级规划教材
普通高等教育信息管理与信息系统专业规划教材

信息存储与检索

INFORMATION STORAGE AND RETRIEVAL

王知津　主编

机械工业出版社
CHINA MACHINE PRESS

普通高等教育“十一五”国家级规划教材
普通高等教育信息管理与信息系统专业规划教材

信息存储与检索

主　编　王知津
副主编　李　培　于晓燕
参　编　（按姓氏笔画排序）
江力波　张收棉　陈芳芳
赵　洪　樊振佳

机械工业出版社

本书为普通高等教育“十一五”国家级规划教材，同时兼顾了存储和检索两个方面，这一点不同于旨在向大学生普及信息检索方法的信息检索与利用类知识。本书内容涉及信息检索的原理、方法、技术、系统以及相关的网络知识等。全书共分9章，包括信息检索基础理论、信息检索模型、文本信息存储与检索、多媒体信息存储与检索、Web信息存储与检索、并行与分布式信息检索、人工智能与自然语言检索、用户界面与可视化、信息检索评价与实验。

本书内容丰富，深入浅出，力图将计算机技术与信息检索紧密结合起来，具有信息检索方面的专业性质，属于侧重“技术”的教材。本书不仅适用于信息管理类学生，还适用于高等院校计算机类专业师生，以及从事信息检索系统、数据库和网站开发、设计的工作者。

图书在版编目（CIP）数据

信息存储与检索/王知津主编 .—北京：机械工业出版社，2009.1
普通高等教育“十一五”国家级规划教材
普通高等教育信息管理与信息系统专业规划教材
ISBN 978-7-111-25874-2

Ⅰ.信… Ⅱ.王… Ⅲ.①信息存贮—高等学校—教材②情报检索—高等学校—教材 Ⅳ.TP333 G252.7

中国版本图书馆CIP数据核字（2008）第205675号

机械工业出版社（北京市百万庄大街22号 邮政编码100037）
策划编辑：易 敏 责任编辑：郭 娟
版式设计：霍永明 责任校对：张玉琴
封面设计：王伟光 责任印制：杨 曦
三河市国英印务有限公司印刷
2009年2月第1版第1次印刷
169mm×239mm·20.25印张·2插页·392千字
标准书号：ISBN 978-7-111-25874-2
定价：29.80元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
销售服务热线电话：（010）68326294
购书热线电话：（010）88379639 88379641 88379643
编辑热线电话：（010）88379721
封面无防伪标均为盗版

前　言

如果在20年前提起“信息检索”，恐怕没有多少人听说过，因为那个时候信息检索还远离广大最终用户，只是作为信息检索专业工作者的专用术语。当然，这并不意味着广大最终用户不需要信息检索，事实恰恰相反，人们在学习、工作和生活的各个领域里，每时每刻都在需求信息和利用信息，只不过绝大多数的检索操作都不是用户亲自进行的，而是由专职人员代替完成的。然而，20年来，随着计算机技术、通信技术和网络技术的飞速发展，特别是Internet的触角延伸到世界的各个角落，成为家喻户晓、人人皆知的大众工具，从而使信息检索也发生了翻天覆地的变化。今天再提起“信息检索”已经不是什么新鲜事了，它已变成了大多数人耳熟能详的常用术语。

信息检索是信息管理领域的核心部分。现代信息检索已经脱离了原来的人工操作方式，而与现代信息技术紧密结合起来，从而进入了一个崭新的历史发展阶段。自20世纪50年代初提出“信息检索”这个概念以来，历经半个世纪的发展和建设，信息检索已成为一门新兴的交叉学科呈现在人们面前。信息检索已经逐渐形成了包括自身的理论、方法、技术和应用领域在内的完整的学科体系，尽管还存在一些没有解决或没有完全解决的课题，但这并不影响它沿着自己的既定方向继续前进。

目前，环顾国内外，关于信息检索的教材数量众多。仅就国内而言，绝大多数此类教材属于“方法”类，主要供在校大学生学习、掌握和运用检索方法，强化利用信息的基本技能和技巧，带有普及性质。还有少数此类教材属于“技术”类，主要供高等学校信息管理类专业的学生使用，旨在使学生深入了解信息检索的原理、技术、系统以及相关的网络知识等，带有专业性质。本书属于后者。

2005年，我们曾翻译出版了《现代信息检索》（机械工业出版社出版）一书。该书主要从计算机专业角度出发，将计算机技术与信息检索紧密结合起来进行介绍，但由于文化和教育背景不同，还不能完全适合我国学生。为此，出版社鼓励我们重新编写一本更加适合我国学生的信息检索教材，这成为我们编写本书的巨大动力。此后，本书被教育部列入普通高等教育“十一五”国家级规划教材，也得到了南开大学教材建设立项资助。

本书大体上分为4个部分共9章。第一部分是信息检索理论，包括第1、2章，主要介绍信息检索和信息检索系统的基本概念、原理、类型、结构及各种数学模型。第二部分是基本的信息存储与检索，包括第3～5章，重点介绍文本信

息、多媒体信息和Web信息的存储与检索。第三部分是信息存储与检索的提高，包括第6~8章，着重介绍并行与分布式信息检索、智能信息检索、用户界面设计及信息检索可视化。第四部分的第9章是信息检索的评价，侧重介绍信息检索的相关性理论以及评价指标、方法与实验。

本书的编写思路和大纲由王知津提出，经集体反复讨论和修改后确定。各章节的编写者及具体分工如下：王知津（第1章）、赵洪（第2章）、陈芳芳（第3章第1~5、7节）、于晓燕（第4、8章）、江力波（第3章第6节、第5章第1~3节）、张收棉（第5章第4节）、李培（第6、7章）、樊振佳（第9章）。全书书稿的初审由于晓燕和李培完成，王知津负责终审和定稿。

在本书的编写过程中，我们参考和借鉴了大量的中外文书刊资料，由于篇幅所限，未能一一列出，在此对所有参考文献作者表示诚挚的谢意。正是这些参考文献作者的前期工作为本书的完成奠定了基础，并为我们提供了强大的写作动力和丰富的创新素材。本书得以顺利完成，与机械工业出版社易敏编辑所给予的大力支持、鼓励、指导、帮助和建议是分不开的，在此，我们一并表示感谢。

虽然我们尽了自己最大的努力编写好本书，但信息检索毕竟是一个快速发展和不断更新的领域，限于编者的学识、水平和能力，缺点、疏漏和错误在所难免，恳请各位专家、学者和广大读者不吝赐教、指正，以便在本书修订时加以补充、更正和完善。

我们制作了与本书配套的PPT课件，使用本书作教材授课的教师可向出版社编辑索取（yimin9721@163.com）

王知津

目　录

第1章 绪 论

【本章提示】 本章为信息存储与检索提供一个概貌，为后续各章的展开打下基础。本章主要阐述了信息检索的概念、原理和类型等基本理论，介绍了信息检索系统的概念、类型、物理结构和逻辑结构，讨论了信息检索的研究内容、相关学科、产生和发展以及现状与未来趋势。要求重点掌握信息检索基本理论和信息检索系统两大部分，对于信息检索的研究现状与趋势可作一般了解。

1.1 信息检索基本理论

1.1.1 信息检索的概念

“信息检索”（Information Retrieval，IR，我国早期译为“情报检索”）一词最早出现于1952年，由美国学者穆尔斯（C.W.Mooers）提出，从1961年开始在学术界和实践领域中得到广泛的应用。信息检索这一概念首先假设包含相关信息的文献或记录已经按照某种有助于检索的顺序组织起来。信息检索就是对信息项进行表示、存储、组织和存取的全过程。对信息项的表示和组织应该能够为用户提供其感兴趣的信息的方便存取。遗憾的是，对用户信息需求进行全面而准确的描述不是一件轻而易举的事情。例如，在万维网（或者就是Web）环境中考察以下假设的用户信息需求：

找出包含能满足以下两个条件的有关某一学院网球队相关信息的所有网页（即文献）：①该网球队隶属于美国的一所大学；②该网球队参加过美国大学生体育协会（NCAA）举办的网球锦标赛。为了保证查找结果的相关性，检索到的网页必须包括该网球队在过去3年里在全国比赛中的名次及其教练的电子邮箱、地址或电话号码等信息。

显然，在目前的Web搜索引擎界面中，人们不可能直接采用这种对用户信息需求进行完整描述的方式来检索信息，用户必须首先将这些信息需求转换为搜索引擎（或IR系统）能够处理的查询式来查询（Query）。这种转换以其最普遍的形式生成一组关键词（或索引词），而这些关键词能够对用户信息需求的描述进行概括。

20世纪90年代以前，知道“信息检索”这个术语的人还不多。随着因特网的形成、发展和普及，信息检索才被越来越多的人所知、所用。就信息检索这个

概念而言，不同的使用者对它有着不同的理解和解释，大体上可以分为两类：

第一类是广义的。对于专门从事信息检索及其系统的研究、开发和设计的少数人来说，“信息检索”的完整含义是“信息存储与检索”（Information Storage and Retrieval，ISR）。也就是说，把“信息检索”当做“信息存储与检索”的简称。这里所谓的信息检索，包括存储和检索两个过程。信息存储是指将有用信息按照一定的方式组织和存放起来；信息检索是指当用户需要这些信息时，再把它们从存放的地方查找和提取出来。因此，对于广义的信息检索来说，存储和检索缺一不可。本书采取信息检索的广义用法，这就要求不仅要知道如何检索，也要知道如何存储，因为如何存储决定了如何检索。

第二类是狭义的。对于普通信息用户来说，在大多数情况下，“信息检索”可以用英文 Information Searching 来表达，其准确含义是“信息查询”或“信息搜索”。也就是说，所谓信息检索，是指按照一定的方式从现有的信息集合或数据库中，找出并提取所需要的信息。可见，狭义的信息检索仅指检索这一个过程，而不关心信息是如何存储的。

1.1.2 信息检索的原理

广义的信息检索的基本原理可以用图 1-1 表示。

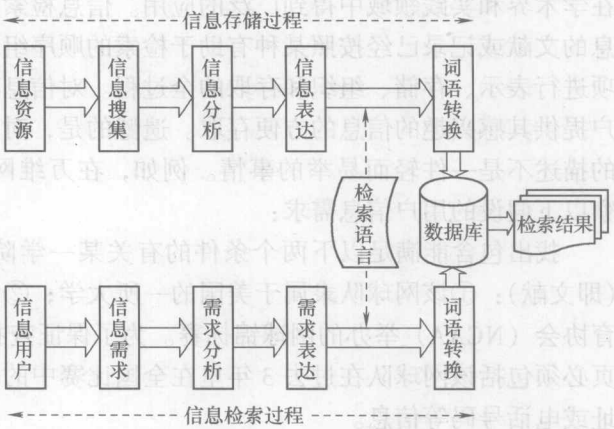

图 1-1 广义的信息检索的基本原理

在存储过程中，专门负责信息检索系统和数据库建立的人从各种各样的信息资源中，搜集有用信息，对有用信息进行主题内容的分析，找出能够全面、准确表达该信息主题内容的概念，借助于检索语言（通常是检索词表）把分析出来的概念转换成检索系统所采用的词语（在自然语言检索系统中，直接使用自然语言而不需要转换），再按照一定的规则和方式将这些有用信息组织成可供检索用的数据库，并存储在一定的介质上。

检索是存储的相似过程。信息用户在工作、学习和生活中产生了信息需求，为了检索并获取自己所需要的信息，用户必须对自己的需求进行主题内容的分析，找出能够全面、准确表达该需求主题内容的概念，也要借助于检索语言（通

常是检索词表）把分析出来的概念转换成检索系统所采用的词语（在自然语言检索系统中，直接使用自然语言而不需要转换），再按照一定的检索规则和方式，制定检索策略，构造检索式，从数据库中查找并获取自己所需要的信息，最后输出检索结果。当然，检索的全过程还应当包括对检索结果进行评价、反馈，或许还要重新制定检索策略，重新构造检索式，反复进行检索，直至检索出满意的结果为止。

从图 1-1 可以看出，信息存储和信息检索有两个交汇处：一个是直接的，即表达信息主题内容的词语与表达需求主题内容的词语之间进行对比的交汇；另一个是间接的，即通过检索语言进行沟通，确保把存储用词和检索用词都统一到同一个检索语言体系中（对于自然语言检索系统来说，不存在存储与检索的间接交汇处）。

由此可见，信息存储和信息检索的直接交汇处是至关重要的，由此形成了信息检索的一致性匹配作用机理，如图 1-2 所示。

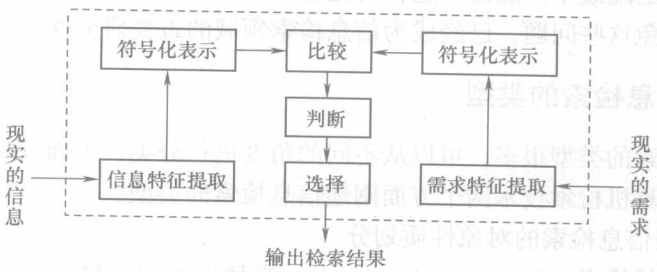

图 1-2 信息检索的一致性匹配作用机理

信息检索的一致性匹配作用机理包括 5 个机理。

（1）提取机理。从现实的信息和现实的需求中提取出能够揭示特定信息和特定需求的语法特征和语义特征。这些特征可以归纳成内容（内部）特征和形式（外部）特征，前者包括特定信息和特定需求的类别（如学科、专业）、主题等；后者包括信息和需求的名称（题名）、作者（责任者）、时间、编号等。

（2）表示机理。用适当的符号表示信息和需求的各种特征。符号是广义的，可以是文字、数字和符号，也可以是图形、图像、视频和音频。比如，用分类号表示信息和需求的类别，用关键词表示信息和需求的主题。

（3）比较机理。在检索项类型（如题名、作者、分类、关键词）相同的情况下，对代表特定信息的特征符号与代表特定需求的特征符号进行对比。比较的实质是相似性比较或一致性比较，即包括完全一致、部分一致和不一致，也包括等于、不等于、大于、小于。比如，对于两个词或词组来说，它们可以是完全一致、前方一致、后方一致、中间一致；对于两个编号来说，它们可以是相等、大

于、小于。

（4）判断机理。在比较的基础上，对信息是否符合需求以及符合的程度加以判断。两者相符合的信息被检索出来（命中），不相符合的信息被拒绝（不命中）。从符合程度来看，可以是完全符合，也可以是部分符合。在部分符合中，还可以进一步细化。原则上，凡是符合特定检索所规定的比较条件和一致条件的信息，都应该是符合需求的，尽管它们符合的程度有所不同。

（5）选择机理。对于检索出来的结果，按照一定的标准加以选择，带有推荐首选或着重使用的意义。选择的实质就是排序，排序有多种标准和方法，如相关度、权值和（加权检索）、时间（新颖性）、重要作者或单位等。

信息检索的一致性匹配作用机理的实质是简化现实的信息和现实的需求之间的匹配。把内容与形式都非常复杂的信息简化成信息特征的符号化表示，再把内容与形式都非常复杂的需求也简化成需求特征的符号化表示，将这两个非常简单的特征符号化表示进行比较、判断和选择，从而变复杂为简单，化模糊为清晰，大大提高了匹配效率。然而，这种简化也会带来一些弊病，造成误检和漏检。如何解决和避免这些问题，已经成为信息检索领域的重要研究课题。

1.1.3 信息检索的类型

信息检索的类型很多，可以从不同的角度进行分类，下面仅从信息检索的对象性质和计算机检索技术两个方面阐述信息检索的类型。

1. 按照信息检索的对象性质划分

（1）文献检索（Document Retrieval）。文献检索的对象是文献，例如，检索有关“太阳能电池”方面的文献。这里所说的“文献”是指文献单元，即包含一个完整内容的单元，如一篇论文、一本图书、一份报告等，而忽略其物理载体（如纸介质、磁介质、光介质）、出版形式（如图书、期刊、报纸）、加工深度（如一次文献、二次文献、三次文献）等。进一步地说，这里的“文献”可以是完整的原始文献，也可以是原始文献的替代品，如一条目录款目、一条文摘款目或一条索引款目。归根结底，文献检索的目标是检索出原始文献或原始文献的替代品。供文献检索使用的数据库是文献数据库，包括目录、文摘、索引、全文等数据库。

按照文献内容的完整性，文献检索又可以进一步分为书目检索（Bibliographic Retrieval）和全文检索（Full Text Retrieval）。

所谓书目检索，是指检索对象为原始文献的替代品，即文献线索，而不是原始文献本身，要想阅读原始文献，还必须依据文献线索去进一步找到和获取原始文献。书目检索通常借助于文摘数据库、索引数据库、目录数据库来完成。书目检索的首要目标是检索出包含用户所需信息的书目记录，其数据库则由被存储文

献的书目记录构成。

所谓全文检索，是指检索对象为原始文献本身，主要是对全文中的字、词、句、段等进行检索，检索出来的结果就是原始文献，进而可以直接阅读和使用原始文献。全文检索通常借助于全文数据库来完成，通常可以提供报纸、手册、字典、百科全书、统计资料等的文摘或全文，其首要目标是找出能满足用户所需信息的某个实际文本。全文数据库包含文献的实际文本，最终的检索结果也是实际文本。应当指出，全文检索的完整含义不限于检索结果是全文，而是使用全文中的各种元素（如字、词、句、段等）进行检索。因此，如果只使用题名、作者、关键词、摘要等进行检索，而不能使用全文中的各种元素进行检索，即使检索结果同样是全文，也不是严格意义上的全文检索。

无论是书目检索还是全文检索，都假定存在一个有信息需求的目标用户群。当用户提出询问时，系统应能提供包含他们所需信息的书目记录或全文文本。

文献检索是最典型的信息检索，也是信息检索的早期类型。对于学术研究来说，文献检索仍然是目前使用最普遍的信息检索类型。在许多情况下，可以把文献检索直接理解为信息检索的同义语。

（2）数值检索（Numeric Retrieval）。数值检索有时也叫数据检索（Data Retrieval）。数值检索的对象是以数字形式表示的具体数值，如生产指标、统计数据、物价、股票及理化特性等，主要应用于科学研究、工程设计和经济统计等领域。数值的范围不限于数字本身，还包括图形、图表、数学公式、化学分子式及结构式等非数字型的数值。数值检索的目标是检索出能满足给定条件的、能够直接使用的数值，如钢铁产量、GDP、CPI、汽车的价格、黄金的密度、聚氯乙烯的分子结构、尼罗河的全长、喜马拉雅山的高度等。供数值检索使用的数据库是数值数据库，例如，物理数据库可以提供有关物质的密度、比热、沸点、熔点、拉力和压力等参数；热力学数据库可以提供有关物质的热力学特性和计算公式；建筑数据库可以提供有关建筑材料的型号、强度、刚度及其他理化特性，还可以提供有关建材产品的型号、规格和价格等。

数值检索是新型的信息检索，其发展速度已超过了文献检索。由于数值检索的结果可以直接使用，数值数据库必须具有高度的准确性和浓缩性，所以，数值的收集、加工和输入必须非常仔细，不能有半点马虎。此外，数值的鉴定也是一项非常重要而又十分复杂的工作。

（3）事实检索（Fact Retrieval）。事实检索的对象是某一特定的客观事实，反映事物或事件发生的时间、地点和过程等实际情况，例如，“长江哪一年汛期的水位最高”、“克隆羊最早是由谁研制成功的”、“世界上最大的空难是哪一次”，等等。回答这类问题，事先必须有详细记载。与文献检索和数值检索不同，事实检索一般不能通过简单检索直接提供问题的答案，而必须进行比较复杂的对比、

分析、推理后才能得出最终结果，从而满足给定条件。事实检索是在数值检索的基础上发展起来的。

在国外，有时对数值检索和事实检索并不加以区分，而是把两者都概括在数据检索或事实检索之下，这样的检索系统应能查找出某项具体的事实或数据。例如，有一个办公信息数据库，包含职工姓名、职位、工资等，还有一个超市信息的数据库，包含商品名称、价格、数量等，数据或事实检索应能检索出某位经理的工资和某种香水的价格。

文献检索是一种相关性检索，主要是确定某一文献集合中的哪些文献包含了用户查询中的关键词，然而，只有这些关键词通常是不能满足用户的信息需求的。文献检索获得的结果具有不确定性和概率性，也就是说，检索结果出来后，还不能确定它们是否满足要求、在多大程度上满足要求，而这些只有在阅读或浏览了原始文献之后才能确定。因此，只能说检索出来的结果与检索课题是相关的。这就是说，文献检索检出的结果可以是不准确的，并且可能有觉察不出来的错误。

相比之下，数值检索和事实检索是确定性检索，检索出来的结果要么有、要么无，要么是、要么否，要么对、要么错，直接回答用户的具体问题，毫不含糊。例如，在检出的1000个结果中，如果只有一个是错误的，就意味着本次检索在整体上是失败的。产生这种区别的主要原因是，一方面，文献检索所处理的通常是自然语言文本，而人们总是不能使自然语言文本很好地结构化，并且自然语言文本可能会有语义上的歧义。另一方面，数值检索和事实检索所处理的通常是事先已经定义好的结构和语义的数据。此外，如上所述，事实检索是3种检索类型中最复杂的。

文献检索是信息检索的核心和主体，数值检索和事实检索是由文献检索派生出来的，但很有发展前途。与数值检索和事实检索相比，文献检索的内容更丰富、方法更灵活，是信息用户最经常使用的。

2. 按照计算机检索技术划分

(1) 脱机检索 (Off-line Retrieval)。脱机检索是计算机检索的最早技术。脱机检索的存储介质是磁带，输入介质是穿孔卡片或穿孔纸带，不使用通信和终端设备，采用成批处理方式，用户不直接使用计算机，检索作业由专职的检索人员完成。作为计算机检索技术的一个发展阶段，脱机检索在计算机检索技术的历史上占有一席之地，但现在已经很少使用了。目前使用较多的计算机检索技术包括联机检索、光盘检索和网络检索。

(2) 联机检索 (On-line Retrieval)。联机检索以联机检索提供商为中心，联机提供商开发自己的检索软件，建立自己的联机检索系统，数据库则是从数据库生产商那里购买的。用户利用联机检索终端，通过专用的或公用的电话线路等数