XIANDAI
XINXI JIANSUO
LILUN YU SHIJIAN

现代信息检索
理论与实践

唐曙南 编著

时代出版传媒股份有限公司
安徽科学技术出版社

现代信息检索理论与实践

唐曙南　著

时代出版传媒股份有限公司
安徽科学技术出版社

图书在版编目(CIP)数据

现代信息检索理论与实践/唐曙南编著. —合肥:安徽科学技术出版社,2011.5
ISBN 978-7-5337-5250-7

Ⅰ.①现… Ⅱ.①唐… Ⅲ.①情报检索-高等学校-教材 Ⅳ.①G252.7

中国版本图书馆 CIP 数据核字(2011)第 113917 号

现代信息检索理论与实践 唐曙南 编著

出 版 人:黄和平 选题策划:王 霄 责任编辑:王 霄
责任校对:盛 东 责任印制:梁东兵 封面设计:朱 婧
出版发行:时代出版传媒股份有限公司 http://www.press-mart.com
安徽科学技术出版社 http://www.ahstp.net
(合肥市政务文化新区翡翠路 1118 号出版传媒广场,邮编:230071)
电话:(0551)3533330
印 制:合肥市星光印务有限责任公司 电话:(0551)4235059
(如发现印装质量问题,影响阅读,请与印刷厂商联系调换)

开本:889×1194 1/32 印张:7.25 字数:210 千
版次:2011 年 5 月第 1 版 2011 年 5 月第 1 次印刷

ISBN 978-7-5337-5250-7 定价:15.00 元

前　言

随着科学技术的突飞猛进，社会信息化程度不断提高，知识经济迅速兴起。在这个信息时代，要成为现代社会的新型人才，必须具备信息利用的能力，才能适应科学技术、社会经济飞速发展的局面，才能利用信息为学习、工作和研究提供更多的方便。随着中国高等教育改革的逐渐深入和社会信息化环境的逐步形成，培养大学生自学能力、独立研究能力和创新能力的信息素质教育，已成为提高当代大学生综合素质的重要途径。掌握一定的检索基础理论与技术，可以帮助学生有效地改进学习、工作的方法，提高学生搜集信息、运用信息的能力，尤其对学生毕业论文或者课程设计的实践环节有重要帮助，它能有效提高学生利用和研究信息的能力，最终形成科学严谨的作风。

信息检索是指信息用户为了满足其信息需要所采取的信息获取与处理的各种行动。通过高效率的信息检索，有利于节省科研时间，提高工作效率，提高自学能力，培养创新人才，促进智力资源的开发与利用，推动社会进步与发展。

本书的主要特点是理论与实践相结合。就信息检索来讲，在教材编写以及教学过程中基本有两种方法：第一种，注重理论与实践；第二种，直接讲解或者介绍检索工具和系统。对于后者，编者认为仅仅是“授之以鱼”，此法虽然效果明显、快捷，但是短效。信息检索，应该是一门方法课，它的独特之处在于有强大且内容丰富的理论基础。它包括信息的存储和信息的检索两个方面，只有掌握了信息的存储过程，才可以知其然然后知其所以然，而信息的存储涉及的就是信息检索大量的理论知识。因此，笔者认为，即使不是专业的信息检索人员，在掌握信息技能的过程中，也不应该急功近利，只有在扎实的理论指导下，面对信息技术日新月异的今天，才不至于只知其一，不知其二。本书在理论

的讲解和实际的检索上，力求翔实、全面、系统。数据和实例尽量采用最新的、第一手的资料，并加入了大量的图示，以便于理解。

本书共由七章组成，分为理论与实践两大部分。第一章全面介绍信息、知识、情报、文献的概念以及信息的重要性；第二章介绍信息检索的概念、信息系统的组织以及信息检索语言；第三章介绍信息检索的方法、技术、步骤以及检索效果的分析及评估；第四、第五章分别介绍国内外数据库信息的检索与利用；第六、第七章分别介绍特种文献的检索和Internet网络信息检索。前三章为理论，后四章为实践。

本书编著者从事信息检索的教学与研究工作已有15个年头，本书是作者长期从事教学研究工作之成果。在本书的编写过程中，参阅和引用了国内外有关专家学者的论著，主要的参考文献已集中列于文后，因篇幅所限，可能未一一列出，在此特向这些参考文献的作者表示衷心的感谢！限于作者的学识和水平，本书错误及缺漏在所难免，望广大读者及同行不吝赐教。

唐曙南

目　录

第1章　信息概论

1.1　信息引论

人类要在世界上生存与发展，就必须获取物质、能量、信息这三大要素，由这三个基本要素分别形成材料科学、能源科学和信息科学，这被人们认为是现代科学的三大支柱。

曾经有这样一个测试，某海军陆战队在原始森林进行为时一个月左右的生存实验，具体要求如下：第一，每个队员除了身上穿的衣服外，随身只能带三件物品，每件物品不能超过2千克；第二，队员都是由飞机空降到半径为1000千米的原始森林的中心地带，要求在一个月时间内从森林里走出来。问题是队员带哪三件物品最合适？最理想的答案是队员必须携带钢刀、火石、指南针。因为钢刀能获取食物，火石可以取火，指南针可以指明方向，是获取信息的一种工具。有了它，陆战队员们就可以获取方向信息；没有它，即使物质和能量资源都很丰富，可能一辈子都走不出原始森林。

1.1.1　信息的概念

信息的概念十分广泛。世间万物的运动，人间万象的更迭，都属于信息的范畴。树木上的年轮、蜜蜂的飞舞、小鸟的歌唱，这是自然界的动植物在传递信息。听收音机、看电视、阅读报纸杂志、发送手机短信、上网漫游，这是人类在接受、查找、利用和交流信息。特别是在现代的信息社会中，信息更是无处不在、无时不有，信息的生产与利用已经成为现代人类生活的基本组成部分。从购物到旅游，从升学到择业，直到跨国公司的兼并与收购，都是对信息的收集、分析和在此基础上的行

动。在万象纷呈的现代信息海洋中，人们不断地创造出新的信息财富，推动着社会政治、经济、文化的飞速发展。

汉语中很早就有“信息”这个词。唐代诗人李中有诗句“梦断美人沉信息，目穿长路倚楼台”(《暮春怀古人》)。这里“信息”指的是音信、消息。南宋诗人陈亮也有诗句“欲传春信息，不怕雪埋藏”(《梅花》)。这里“信息”则指的是一种意境。

如今，“信息”已经成为使用频率最高的词汇之一，但对于“信息”这个具有多重属性的重要概念，至今还没有一个统一的定义，不同领域的研究者站在各自的角度，对信息的内涵有着不同的描述。

1928年，奈奎斯特(Nyquist)和哈特莱(R. V. Hartley)在一篇题为《信息传输》的论文中，把“信息”理解为选择通信符号的方式。他指出不管符号所代表的意义是什么，只要从符号表中选择的数目一定，所发出的信息的数量也就一定。他们的思想和研究成果为后来信息论的创立奠定了基础。

1948年，美国贝尔实验室的科学家、信息论的创始人克劳德·申农(Claude E. Shannon)从通信系统的角度，给“信息”下了定义。在他著名的论文《通信的数学理论》中，他将“信息”解释为：两次不定性之差，说明通信的意义在于消除某种“不确定性”。在这里，信息是指有新内容、新知识的消息。他认为，信息的多少意味着消除了的“不确定性”的大小。人们看电视、阅读报刊，所得到的消息的内容可能是已经知道的，也可能是还不知道的。事先已经知道的不是信息，因为它没有从中消除不确定性，事先不知道的才是信息，因为它消除了一些不确定性。这是从信息在通信过程中所起到的作用的角度提出来的，该论文也成为信息论诞生的标志。

与申农同时代的美国科学家、控制论专家维纳(N. Wiener)在《时间序列的内插、外推和平滑化》一文和《控制论》一书中提出，信息就是我们适应外部世界，并把这种适应反作用于外部世界，同外部世界相互联系、相互作用、相互交换的一种内容，它既不是物质，也不是能量。维纳的表述为我们提供了一条深入揭示信息本质的正确途径，从此，“信

息”的概念才被广泛应用。

我们可以这样认为，信息是用语言、文字、数据或信号等形式，通过一定的传递和处理，表现各种相互联系的客观事物在运动变化中所具有的特征内容的总称。它是事物存在的方式、形态和运动规律的表征，是事物具有的一种普遍属性，与事物同在，存在于整个自然界和人类社会中。

信息是无处不在，无时不有，无人不用的。在古代战争中，士兵用烽火台上的狼烟传递敌人来犯的信息，官员通过微服私访、民间采风等活动了解民情，收集百姓信息；现代生活中，人们通过比较商品的价格、产地、原料、保质期等信息选择商品，通过看电视、电影等进行娱乐活动。冰消雪融、草木发芽，这是自然界带给我们季节变换的信息；蜜蜂的舞蹈和小鸟的歌唱，是动物之间在传递信息；新闻报道、商品广告，这是社会带给我们的信息；闹钟的铃声把我们从睡梦中叫醒，这是生活中的信息。听课、看书、读报、看电视、听广播、上网，是我们接受信息；点头、摆手、跺脚、摸鼻子、说、唱等，一举一动都在发出或传递信息。

可见信息对我们的生活是多么重要，学会获取信息、存储信息、处理信息和传递信息，已经成为现代人必备的基本技能之一，而从事信息的管理和服务，则是当今最为热门和尖端的活动。

1.1.2　信息的特征

信息的特征是指信息区别于其他事物的本质属性。

1. 信息具有普遍性和客观性

信息是现实世界中各种事物运动与状态的反映，世间一切事物都在运动着，都有一定的运动状态和状态方式的改变，因而一切事物随时都在产生信息，即信息的产生源于事物，是事物的普遍属性。同时，信息又是客观存在的，不是虚无缥缈可以随意想象和创造的，它的存在是不以人的意志为条件的，是可以被感知、处理、存储、传递和利用的。

2. 信息具有特殊性和相对性

世间一切不同的事物都具有不同的运动状态和方式，会以不同的特征展现出来，因而不同的事物给人们带来不同的信息。同时，由于用户感受能力、理解能力、目的性的不同，对于同一对象，其所获得的信息内容和信息量因人而异。例如，对于考古发掘出来的一个瓦片，在普通人的眼中，它只不过是一个破瓦片而已，毫无价值，但在考古工作者眼中，它却代表了特定历史年代的政治、经济、文化水平，从它身上可以看出当时的宗教、文化和风俗习惯等很多情况。

3. 信息具有实质性和依存性

事物在运动过程中和形态改变上所展现出的表征，是其属性的再现，被人们认知后，就构成了信息的实质内容。信息必须依靠语言、文字、图像、符号等记录手段来表述，依靠纸张、声波、磁性材料、化学材料等存储和传递，不可能脱离物质载体而单独存在，并且其内容不会因为记录手段或物质载体的改变而发生变化。

4. 信息具有中介性和共享性

信息源于事物，但不是事物本身，是人们用来认识事物的媒介。信息能够共享是区别信息不同于物质和能量的最主要特征，即同一内容的信息在同一时间或不同时间、同一地域可以被两个以上的用户使用，其分享的信息量不会受分享用户的多少的影响，原有的信息量也不会因之而损失或减少，并且能被用户多次反复使用。

5. 信息具有动态性和可加工性

客观事物本身都在不停地运动变化，运动状态和方式会随时间的推移而改变。信息可以被分析或综合，扩充或浓缩，也可以被转换形式。

6. 信息具有可传递性和受干扰性

信息可以通过不同的载体形式和手段进行跨越时空的传递。在传递过程中，由于受到各种因素的干扰，会产生信息失真现象，对用户的利用产生影响。

1.1.3　信息的功能

在人类步入信息社会以后，信息同物质、能量构成人类社会的三大资源。物质提供材料，能量提供动力，信息提供知识与智慧。因此，信息已成为促进科技、经济和社会发展的新型资源，它不仅有助于人们不断地揭示客观世界，深化人们对客观世界的科学认识，消除人们在认识上的某种不确定性，还源源不断地向人类提供生产知识的原料。

1. 信息是感知世界的中介

信息是介于物质世界和精神世界之间过渡状态的东西，是人们用来认识事物、感知世界的不可缺少的中间环节，且贯穿认知活动的始终。人类认识世界和改造世界的过程，就是一个不断地从客观世界获得信息，并对信息进行加工处理，形成新的知识，然后通过实践活动反作用于客观世界的过程。人们看柳树发芽，就知道春天来了；看到天上乌云滚滚，就知道要下雨了，这都是通过自然界发出的信息来感知世界。

2. 信息是管理决策的依据

管理决策的决定性因素取决于对客观实际的了解，对未来形势发展及后果的正确判断，而这些都必须依赖于全面、及时、准确的信息分析研究。信息活动贯穿科学决策的全过程，并渗透到决策过程的每一个环节。

3. 信息是科学研究的条件

人类知识的继承性和共享性，使得任何一项科学研究都必须借鉴前人的研究成果和依靠同时代同行的帮助，所以需要时间和空间上的信息传递。另外，学科的交叉与融合也需要信息的连接与交汇。

4. 信息是社会发展的资源

信息资源是人类借以对其他资源进行有效管理的工具，它对推动社会经济发展和社会进步起着日益重要的作用。物质作为材料，能源作为动力，信息作为智慧，相当于人的身体、体力和智力，必须协调发展。只有三者健全发展的人，才是一个真正健康的人。

世界上任何组织和个人都需要信息来指导其活动。国家决策部门需要利用各种国内外信息使国家机器正常运转；管理机构能否成功，取决于是否有效地组织了各分支部门的信息交流；商品购买者需要了解、比较不同商品的价格、质量等信息；足球教练需要了解球员状况、对手等信息；科学研究则更为重视信息的作用，只有掌握最新的学科前沿信息，才能使科研具有创新性，避免重复无谓的劳动，跟上时代的步伐。

1.1.4 信息资源

1. 信息资源的含义与特点

信息资源是由信息与资源两个概念整合而衍生出的新概念。如前所述，信息是事物的一种普遍属性；资源，就其一般而言，是自然界及人类社会中一切对人类有用的材料。结合资源概念来考察信息资源，我们可以这样来描述信息资源。信息资源是信息世界中对人类有价值的那一部分信息，是附加了人类劳动的、可供人类利用的信息，并非所有信息都能成为资源，只有经人类开发与重新组织后的信息才能成为信息资源。因此，构成信息资源的基本要素是：信息、人、符号、载体。信息是组成信息资源的原料，人是信息资源的生产者和利用者，符号是生产信息资源的媒介和手段，载体是存储和利用信息资源的物质形式。信息资源与其他资源相比，具有可再生性和可共享性的特点。可再生性是指它不同于一次性消耗资源，它可以反复利用而不失去其价值，对它的开发利用愈深入，它不仅不会枯竭，反而还会更加丰富和充实。可共享性是指它能为全人类所分享而不失去其信息量。

2. 信息资源的类型

信息资源的类型，可根据多种依据来划分。

(1)按开发程度划分，可分为潜在信息和现实信息。潜在信息资源是指人在认识和思维创造的过程中，存储在大脑中的信息，只能为本人所利用，无法为他人直接利用，是一种有限再生的信息资源。现实的信息资源是指潜在人脑中的信息通过特定的符号和载体表述后，可以在特定的社会条件下广泛地传递并连续往复地为人类所利用，是一种无