普通高等教育“十一五”国家级规划教材

Information Retrieval

信息检索

——原理与技术

主编　王兰成

高等教育出版社
HIGHER EDUCATION PRESS

普通高等教育“十一五”国家级规划教材

信息检索
——原理与技术

Xinxi Jiansuo
——Yuanli yu Jishu

主编　王兰成
编委　陈　洋　朱建华　李　超
　　李小青　徐　震

高等教育出版社·北京
HIGHER EDUCATION PRESS　BEIJING

内容提要

本书是普通高等教育“十一五”国家级规划教材，是教育部高等学校图书馆学学科教学指导委员会推荐的高等学校图书馆学专业核心课程的系列教材之一，也是情报学专业核心课程的教材。

本书针对现代图书情报专业的教学需要，兼顾高校信息检索原理、方法、技术和应用课程的要求编写而成。全书从理论到方法、从技术到应用，系统阐述了信息检索的基本知识和基本原理、计算机信息检索系统及其实现基础、文本信息和多媒体信息的检索原理、信息数据库检索技术、信息内容检索的自动化处理、网络信息检索的原理与技术以及信息检索的效果评估等内容，涉及管理学科、信息学科和人文学科中的信息检索，突出信息检索技术的知识及其应用。

本书可作为高等学校的教学用书，读者对象为从事信息管理（包括图书馆学、情报学、档案学）和计算机科学技术的高校学生。本书也可作为相关专业研究生的学习参考书以及各类图书馆和信息机构岗位培训的教材或业务参考书。

图书在版编目（CIP）数据

信息检索/王兰成主编．—北京：高等教育出版社，2011.3

ISBN 978-7-04-031029-0

Ⅰ.① 信… Ⅱ.① 王… Ⅲ.① 情报检索-高等学校-教材 Ⅳ.① G252.7

中国版本图书馆 CIP 数据核字（2011）第 007195 号

策划编辑 罗雪群 **责任编辑** 郭福生 **封面设计** 赵 阳 **责任绘图** 尹 莉
版式设计 范晓红 **责任校对** 杨凤玲 **责任印制** 张福涛

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100120

经 销 蓝色畅想图书发行有限公司
印 刷 北京七色印务有限公司

开 本 787×1092 1/16
印 张 30.75
字 数 750 000

购书热线 010-58581118
咨询电话 400-810-0598
网 址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.landraco.com
http://www.landraco.com.cn
畅想教育 http://www.widedu.com

版 次 2011 年 3 月第 1 版
印 次 2011 年 3 月第 1 次印刷
定 价 38.00 元

物料号 31029-00

前　言

德国柏林图书馆门前有这样一句话:“这里是知识的宝库,你若掌握了它的钥匙,这里的全部知识都是属于你的。”这里所说的“钥匙”即是指信息检索的方法。美国普林斯顿大学物理系一个年轻大学生约翰·菲利普,在图书馆里借阅有关公开资料,仅用4个月时间,就画出一张制造原子弹的设计图,致使一些国家争相购买他的设计副本。这些信息都来源于图书馆中那些极为平常的、完全公开的资料。

提及信息检索,就会想到国际上的Google、Yahoo!、微软以及国内的百度、中国搜索、搜狗、天网等搜索引擎,这些都是与日常生活关系最为密切的搜索引擎。从某种程度上讲,它们成了信息检索技术的代名词。早期以Okapi、Smart、查询扩展、相关反馈为代表的内容分析技术,后来以PageRank、HITS为代表的链接分析技术,以及近年来的语言模型等,都曾在信息检索发展过程中推进了研究和应用的热潮。那么如何才能具备信息检索的应用能力呢?我们需要了解各种信息来源,需要掌握检索语言,需要熟练使用检索工具,需要学习对检索效果进行判断和评价。这种信息获取能力是信息检索的核心。如何才能具备信息检索的研究能力呢?信息利用、社会进步的过程就是知识不断从生产到流通到再生产的过程。为了全面、有效地利用现有知识和信息,在学习和研究的过程中,对所得信息进行整理、分析、归纳和总结,根据自己的思考和思路将各种信息进行重组,创造出新的知识和信息,从而达到信息激活和增值的目的。其中,信息检索所占的时间比例越来越高。

信息检索是指信息用户为处理解决各种问题而查找、识别、获取相关的事实、数据、文献的活动及过程。信息检索作为当代大学生和广大科技工作者的一项基本功,其训练和培养对适应现代社会和作品发表、科研创新都极其重要。善于从信息系统中获取信息的大学生、科研人员,较不具备这一能力的人必然有更多的成功机会。美国一家期刊将交互网络检索专家列为未来十大热门职业之一。这些情况说明以计算机为代表的信息检索方法与技术越来越重要。随着用户信息需求的增长和信息技术的发展,国内外有关信息检索的研究不断深入。从CNKI数据库关于“信息检索”文献的最新检索结果分析,不论是文献篇数还是基金论文数,发展趋势都是总体向上,信息检索的研究内容越来越深入、越来越广泛,信息检索研究成果的数量越来越多,质量也越来越高。

适应网络化、智能化及个性化的需要是目前信息检索技术发展的新趋势。信息检索已发展到网络化和智能化的阶段,其对象从封闭、稳定、简单数据库系统集中管理的信息内容扩展到开放、动态、分布异构及管理松散的超文本内容、多媒体资源,其用户也由原来的图书情报专业人员扩展到包括管理人员、教师、学生和各专业人士在内的最终用户,人们对信息检索的方式与结果提出了更高、更多样化的要求。因此,信息检索(搜索引擎)因Internet的普及而日益变成一个热门学科,各种相关学科的技术都用于信息检索,而信息检索也用于各个领域;信息检索又是一门

复杂的学科，涉及信息编码、数据结构、实现算法、自然语言处理及知识的表达和应用、管理学、信息学、数学，甚至哲学，它几乎涉及信息科学和技术的所有方面。

目前，关于信息检索方面的书已有很多。信息检索学科涉及自然科学、社会科学、人文科学等许多学科和领域。教育部曾多次发文强调在高校开设、改进和发展文献检索课程的意见。涉及信息检索的广大学生和读者，一方面希望有一套科学地进行知识更新和积累的治学方法，以充分挖掘图书情报信息机构中的资源；另一方面希望提高自己的计算机应用能力，通过加大运用信息技术的力度，节省查找文献的时间，适应信息时代和信息社会的实用需求。本书不仅详细介绍了检索系统和检索工具的使用方法介绍，而且更注重内容的新颖性和系统性，加强技术性，尽量反映最新的相关技术。

本书从信息检索的基本理论研究和进展、方法技术原理和进展、项目与实验成果和进展以及用户信息检索需求等多个角度出发，向广大读者全面展现图书情报领域和计算机领域关于信息检索的新知识、新研究和新进展。内容既有宏观方法面的介绍，又不乏具体技术面的知识，使本书的适应对象面更宽。本书力求从适合信息管理（包括图书馆学、情报学、档案学）学科各类文、理科读者学习目标的角度来论述信息检索的新方法和新技术，以体现内容的先进性；以融合信息管理学科与计算机学科的新研究成果来提升当前信息检索学习的技术含量，从而体现交叉学科的优势；以系统介绍信息检索的理论、方法、技术和实现等知识，来体现教学内容的完整性和新颖性。

本书共分为11章。

第1章是全书的导论，介绍数据、信息、知识、信息检索、传统的信息检索工具和计算机信息检索等知识以及信息检索技术的研究与发展。

第2章论述信息检索的基本原理，包括信息检索及其发展、信息检索模型和检索语言、信息检索的一般方法和技术以及跨学科的信息检索研究。

第3章介绍信息检索技术的基础或环境，包括计算机信息处理平台、数据通信网络处理平台、结构化数据与结构化查询、信息的存储与检索。

第4章论述文本信息检索的原理和技术，是第6章和第9章的基础，介绍基本信息检索、扩展信息检索、全文检索、超文本检索和知识层的文本信息检索技术。

第5章论述信息数据库检索及其技术，包括文献信息数据库及其建库技术、规范化设计和数据库查询技术，介绍国内外光盘数据库及其检索。

第6章介绍信息内容检索的自动化处理，包括信息自动标引、自动分类、自动文摘和自动翻译的技术以及信息自动处理中的其他先进技术。

第7章介绍计算机信息检索系统，包括计算机信息检索系统及其开发方法、类型和检索子系统设计，检索系统的发展技术，国内外典型的全文数据库系统。

第8章介绍多媒体信息检索原理与技术，包括多媒体信息检索的原理、处理和检索方法以及多媒体检索研究面临的挑战和发展。

第9章论述网络信息检索的原理及技术，包括网络信息检索的基本原理，网络信息的采集、组织、整合和搜索技术，介绍网络信息检索研究热点及其进展。

第10章介绍网络信息资源的检索与利用，主要包括国内外典型的网络信息检索工具、典型搜索引擎的使用分析以及其他网络数据库资源及其检索。

第11章介绍信息检索的效果评估与展望，包括信息检索效果的评价指标体系和国内外的研究内容，并总结信息检索进一步研究和发展的方向。

本书每章均有本章要点和思考与练习，各章后附有参考文献。

在本书的编写过程中，得到了来自多方的帮助和支持。感谢教育部高等学校图书馆学学科教学指导委员会对我们的信任和推荐，感谢南京政治学院上海分院及军事信息管理系为本书的编写所提供的帮助及优越环境，感谢复旦大学计算机学院副院长汪卫教授对本书的推荐，感谢高等教育出版社的大力支持及相关工作人员为本书的出版所付出的辛勤工作。特别感谢本书参阅和引用的参考文献的作者，他们的研究成果给了我们许多启迪。我的两位同事，陈洋和朱建华博士，在多媒体信息检索和信息检索评价指标的研究方面付出了辛勤的劳动；李超、李小青博士和在读博士生徐震，与我一起反复修改和调整本书中信息检索系统、网络信息检索等内容，逐步形成了本书的主体内容。

本书由王兰成任主编，负责全书的策划、章节内容框架设计、统稿和定稿工作，并撰写了前言和第1、2、6章，参与编写了其他各章；徐震编写了第3章和第5章，李小青编写了第4章和第7章，陈洋编写了第8章，李超编写了第9章和第10章，朱建华编写了第11章。陈洋、朱建华和徐震参与了全书的校稿工作。

尽管我们十分努力，但本书仍然可能存在缺点甚至错误，恳请各位专家和广大读者不吝批评、指正，以期再版时逐一改正。另外，本书撰写过程中广泛参考了相关文献，包括书籍、论文及网页资料等，已尽可能列在参考文献中，但其中难免有所遗漏；从网上下载并引用的一些资料，因历时较长、几易书稿和网页内容变更等原因，无法详细列清它们的来源，谨向这些作者表示深深的歉意。我们的联系方式：wanglancheng@163.com。

王兰成　谨识

2010年10月于上海

作者简介

主　编

王兰成　1962年生，博士，上海市人，南京政治学院上海分院教授，博士生导师；主持国家社科基金项目两项、军队和省部级课题多项；著作有《数字图书馆技术》、《知识集成方法与技术》、《Oracle数据库管理员基础教程》等；主要论文有《数字信息群的知识集成研究进展与关键问题》、《国外知识组织技术研究的现状、实践与热点》、《基于本体的知识检索模型及呈现技术研究》、《基于EMM中文抽词算法的XMARC主题信息挖掘》、《全文数据库建库原理和应用技术》、《互联网军事网络舆情分析系统研究》和*Knowledge Indexing Based on Concept Lexicon and Segmentation Algorithm*等。

编　委

陈洋　1975年生，博士，安徽芜湖人，南京政治学院上海分院信息管理系副教授；完成和参与研究国家社科基金项目、省部级以上立项课题多项，著作两部，译著一部，发表学术论文10余篇；主要论文有《云计算与数字化图书馆技术发展》、《面向领域的政工指挥信息系统分析设计方法》、《国内外档案网站建站技术和功能调查分析》和《军事信息文本的自动分类方法及其应用研究》等。

朱建华　1974年生，博士，江苏启东人，南京政治学院上海分院信息管理系讲师；完成和参与研究国家社科基金项目、省部级以上立项课题多项，参编教材4部，在专业期刊发表学术论文10余篇；主要论文、著作有《基于Nutch的军事网络信息资源搜索引擎设计研究》、《知识集成服务中的知识组织和知识网格技术》、《军队政治工作信息化应用技术》等。

李超　1981年生，博士，湖南衡山人，南京政治学院上海分院博士研究生毕业，研究方向为数据库与信息处理；承担或参研多项省部级以上立项课题，参编教材两部，在专业期刊发表学术论文10余篇；主要论文有《基于主题语义关联的扩展参考检索》、《应用领域本体的Web信息知识集成研究》、《一种基于主题和分众分类的信息检索优化方法》和《本体论方法在档案信息检索系统中的应用研究》等。

李小青　1982年生，博士，陕西泾阳人，南京政治学院上海分院博士研究生毕业，研究方向为数据库与信息处理；承担或参研多项省部级以上立项课题，参编教材两部，在专业期刊发表学术论文10余篇；主要论文有《网格环境下的数字图书馆资源管理研究》、《基于网格的数字图书馆资源发现和获取机制研究》、《论基于普遍心理分层理论的Web用户体验模型》和《基于用户体验的Web信息构建模型研究》等。

徐震　1983年生，湖北黄陂人，南京政治学院上海分院博士研究生，研究方向为数据库与信息处理；承担或参研多项省部级以上立项课题，在专业期刊发表学术论文10余篇；主要论文有《主题检索系统的优化技术研究》、《基于SOA的数字图书馆研究》、《基于业务规则的图书馆信息系统研究》和《数字图书馆流媒体服务平台的设计与实现》等。

郑 重 声 明

目　录

第1章 导 论

本章要点

在介绍数据、信息和知识的基础上，对信息检索的内容和发展方向做了概括。主要包括：

- 传统的信息检索工具。
- 计算机信息检索——文本信息、多媒体信息和网络信息的检索以及信息检索系统。
- 信息检索的发展方向——信息检索的理论、方法、技术和应用。

1.1 数据、信息和知识

1.1.1 从数据到信息

21 世纪是一个高度信息化的社会，信息就是资源、信息就是机会，人人都渴望及时获得有用的信息。有人提出：财富 = 信息 + 技术，可见在激烈的社会竞争中，谁首先获得了最新的信息，谁便获得了发展的主动权，并且拥有了成功和未来。如果说信息搜集是人类赖以生存、发展的一种本能，信息检索则是当代大学生和广大科研人员所必须具备的一种基本技能。因此，学习和具备信息检索的能力，特别是以计算机为代表的现代信息检索能力，就必须掌握信息检索的方法和技术。

按照美国系统科学家拉塞尔·阿克夫(Russell L. Ackoff)的观点，人类思想的内容可分为以下 5 类。

- 数据(data)：符号(symbol)。
- 信息(information)：正在处理的有用的数据，提供资料，解答“谁(who)”、“什么(what)”、“何处(where)”、“何时(when)”的问题。
- 知识(knowledge)：数据和信息的应用，回答“如何(how)”的问题。
- 理解(understanding)：知识的升值，回答“为什么(why)”的问题。
- 智慧(wisdom)：对理解的评估(evaluated understanding)。

阿克夫从区分数据、信息和知识的角度对知识进行了定义。他认为：数据是未经处理的符号；信息是经过处理的、赋予了意义的、有用的数据；而知识是数据和信息的运用。阿克夫表明，前 4 个类别涉及过去，它们指已经有什么或已知什么的处理，只有第 5 类智慧，它与未来有关，因为它包括视觉和设计。有了智慧，人们可以创造未来，而不是只把握现在和过去，但实现智慧是不容易的，人们必须通过实现其他类别才能实现智慧。如图 1 - 1 所示，数据指未经组织的数字、

词语、声音和图像等，信息指以有意义的形式加以排列和处理的数据，知识指用于生产的信息（有意义的信息），信息经过加工处理应用于生产才能转变成知识，智慧是应用知识和创新的能力。本书所介绍的信息检索主要定位在数据检索、信息检索和知识检索这 3 个层面上。

图 1－1　信息、知识和智慧①

国内外许多学者和文献在不同场合和各自领域对数据和信息做了一系列定义。其中数据被定义为：未被解释的符号、简单观察、一组分散的事实、没有回答特定问题的文本、事实和消息等；而信息则被定义为有意义的数据、有目的的相关数据、试图改变接受者认识的消息、回答“何人、何时、何地、何事”问题的文本。数据为未经处理的数字、词语、声音和图像等，信息则是指经过格式化、过滤、已经综合处理的有条件的数据；数据是指各种各样未经组织的数字、语词声音和图像的信号，信息是指按照一定意义排列起来的数据，而知识是指可以被人们认识、掌握和运用的有价值的信息；数据被定义为有关事实的集合，记录和事物有关的原始信息等。

由上述这些定义，可以将数据的一般特征归纳为关于事件和关于世界的一组独立的事实；信息则是已经排列成有意义的形式的数据，是有组织或结构化的数据，是被赋予了关联和因果关系的数据，是放在上下文中并被赋予特定含义的数据，是捕捉了来龙去脉的内容并把它们提炼成经验和想法以后的产出物，是经过一定处理并且有一定意义的数据。

计算机与信息技术经历了半个多世纪的发展，给人类社会带来了巨大的变化与影响。在支配人类社会的能源、材料和信息三大要素中，信息愈来愈显示出其重要性和支配力，它将人类社会由工业化时代推向信息化时代。随着人类活动范围的扩展、信息技术的进步以及网络基础设施的发展，人们能以更快速、更方便、更廉价和更科学的方式来搜集、获取、组织、存储、加工、检索、传送、分析、利用数据和信息，使数据和信息量呈指数级增长。据统计，20 世纪 80 年代，全球信息量每隔 20 个月就增加一倍；进入 20 世纪 90 年代，各类机构所有数据库的数据量增长更快。美国政府部门中一个典型的大数据库每天要接收约 5 TB 的数据量，在 15 s ~ 1 min 的时间里，要维护的数据量达到 300 TB，存档数据达 15 ~ 100 PB。在科研方面，以美国宇航局的数据库为例，每天从卫星下载的数据量就达 3 ~ 4 TB 之多，而为了研究的需要，这些数据要保存 7 年之久。20 世纪 90 年代 Internet 的出现和发展，以及随之而来的企业内部网（intranet）、企业外部网（extranet）及虚拟专用网（Virtual Private Network，VPN）的产生和应用，将整个世界连成一个小小的地

① Ackoff R L. From Data to Wisdom. Journal of Applied Systems Analysis[J], 1989(16):3 - 9.

球村，人们可以跨越时空在网上交换信息，协同工作。这样，展现在人们面前的已不是局限于本部门、本单位和本行业的庞大数据库，而是浩瀚无垠的信息海洋。根据 Internet 流量监测机构 comScore 最新发布的统计数据，2009 年全球在线搜索市场规模扩大 46%，发达国家和新兴市场的强劲增长使得 2009 年 12 月份在线搜索量超过 1 310 亿次。2009 年，Google 继续保持压倒性领先地位，其网站的搜索量为 878.1 亿次，约占 2009 年 12 月份全球在线搜索总量（1 313.5 亿次）的 2/3，较 2008 年的搜索量增长 58%；Yahoo 排名第二，2009 年 12 月份处理的搜索请求为 94.4 亿次，较 2008 年同期增长 13%；排名第四的 Microsoft 在前 5 家公司中增幅最大，2009 年 12 月份搜索量增长 70%，达到 40.9 亿次，主要是受其推出的必应搜索引擎的带动；百度在线网络技术公司位列第三，2009 年 12 月份搜索量增长 7%，达 85.3 亿次。美国是最大的在线搜索市场，2009 年 12 月份搜索量为 227.4 亿次，占当年全球在线搜索总量的 17%。中国排名第二，搜索量为 132.7 亿次；日本则以 91.7 亿次位居第三。极度膨胀的数据信息量，使人们感受到了“信息爆炸”、“混沌信息空间”（information chaotic space）和“数据过剩”（data glut）的巨大压力。

1.1.2 从信息到知识

1. 信息的概念

“信息”一词的拉丁词源是 informatio，英语为 information。美国的《韦氏大词典》把信息描述为“在观察中得到的数据、新闻和认识”。信息论的奠基者克劳德·香农（Claude E. Shannon）从通信系统理论的角度把信息定义为“用来消除不确定性的东西”，认为信息是组织程度，它能使系统的有序性增强，减少破坏、混乱和噪声。控制论的创始人美国科学家维纳（Norbert Wiener）认为“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的过程中，同外部世界进行相互交换的内容的名称。”美国《未来学家》杂志称“信息是一切容易获得的和不易获得的、有时可供人们参考的事实和思想的总和。”

在一千多年前，我国的唐朝就有了“信息”一词，当时的含义一般指音信、消息，信息就是生活主体与外部客体之间的有关情况的消息。数据是对客观世界中各种事物的性质、特征和变化进行记录的物理符号。这些物理符号不仅指数字，而且包括字符、文字、图形图像和视频动画等，它们是未经加工的事实或一种描述。信息是经过某种方式加工或以更具意义的形式所呈现的数据。也就是说，信息是经过加工的数据，数据是信息的表现形式和原始基础；信息是数据有意义的表示，是数据的加工提炼。信息是客观世界（包括自然世界、人类世界等）呈现出来的运动状态和运动规律。人们通过实践将之采集、处理而得出来的东西，通过这些状态和规律人们可以认识自然、改造世界。

《辞海》把信息描述为“通信系统传输和处理的对象，泛指消息和信号的具体内容和意义。”我国国家标准《情报与文献工作词汇基本术语》（GB/T 4894—1985）给信息下的定义是：“信息是物质存在的一种方式、形态或运动状态，也是事物的一种普遍属性，一般指数据、信息中所包含的意义，可以使信息中描述事件的不确定性减少。”我国学者钟义信对信息的解释是：“信息是事物运动的状态与方式，是物质的一种属性”。据不完全统计，到目前为止，有关信息的定义有上百种，这些定义都从不同的侧面反映了信息的某些特征。随着科学技术的发展，信息一词的含义越来越广泛，外延越来越宽、越来越细，覆盖面越来越广。反映自然界各种事物运动状态和规律的信息叫自然信息，反映人类社会状态和规律的信息叫社会信息，还有管理信息等。因此，目前