

新编高等院校信息管理与信息系统专业核心教材

信息检索导论

Introduce to Information Retrieval

叶继元 主编



电子工业出版社

PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

<http://www.phei.com.cn>

新编高等院校信息管理与信息系统专业核心教材

信息检索导论

Introduce to Information Retrieval

叶继元 主编

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有,侵权必究。

图书在版编目(CIP)数据

信息检索导论/叶继元主编. —北京:电子工业出版社,2003.5

新编高等院校信息管理与信息系统专业核心教材

ISBN 7-5053-8694-8

I. 信… II. 叶… III. 情报检索—高等学校—教材 IV. G252.7

中国版本图书馆 CIP 数据核字(2003)第 034883 号

责任编辑:刘宪兰 特约编辑:叶 林

印 刷:北京牛山世兴印刷厂

出版发行:电子工业出版社 <http://www.phei.com.cn>

北京市海淀区万寿路 173 信箱 邮编 100036

经 销:各地新华书店

开 本:787×980 1/16 印张:26.75 字数:551.2 千字

版 次:2003 年 5 月第 1 版 2003 年 5 月第 1 次印刷

印 数:5 000 册 定价:33.00 元

凡购买电子工业出版社的图书,如有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系。联系电话:(010)68279077

总序

Z O N G X U

20 世纪 70 年代,当强大的信息化巨潮还蕴藏在大洋深处,我们的陆地只有一阵微风吹来之时,有识之士们就开始推动信息化专业人才的培养计划,为迎接即将到来的信息化巨潮扩军备战。他们一方面推动着信息技术的普及;一方面根据不同领域的需求,从不同的角度创办了不同类型的信息化专业,这就是管理信息系统专业、经济信息管理专业、科技信息管理专业、医学信息管理专业、林业信息管理专业、农业信息管理专业……实际上,这些专业培养目标可以概括为:为各行业、各部门培养以 CIO 为目标的信息化专门人才。从这一点上看,这些专业的课程设置应当具有相当大的共同性。1996 年,出于多种考虑,教育部将这些专业合并为一个——信息管理与信息系统专业。

以 CIO 为目标的信息化专门人才是一类管理人才。但是他们所管理的主要对象是信息。这样的知识需求,将信息管理与信息系统专业定位于管理学科,与信息学、经济学、法学等学科交叉。这样的学科特点,给课程建设和教材建设带来不少困难。近 30 年来,尽管我们与许多的同行已经进行了不懈的努力,把信息管理与信息系统专业的课程建设和教材建设向前推进了一大步,但是仍然不尽人意,许多课程和教材还没有体现信息管理专业的特色和需要。在多次有关的研讨会上,大家一致呼吁编写一套真正体现信息管理与信息系统专业特色的教材。

新编和出版一套专业教材是要冒风险的。而编写和出版一套以瞬息万变的信息和信息技术为管理对象的专业教材就要冒更大的风险。国内信息业界著名的出版商——电子工业出版社,以超人的胆略愿意同我们一道承担这一风险,组织编写出版一套新的信息管理与信息系统专业核心教材。这套教材冠以“新编”二字,是试图在其体系上能比已有教材更体现信息专业的特色,同时在内容上要能反映最新信息技术的进步以及最新信息管理思想和方法。

目前,国内开设信息管理与信息系统专业的高等院校已经超过 200 所。这样一个数字一方面表明信息化已经深入人心,信息化队伍的规模正在急速扩大,信息化队伍的素质正在不断提高;另一方面,也给我们增加了巨大的压力,

使我们深感责任重大。好在国内本领域的三位知名学者——黄梯云、陈禹、马费成，以及其他一批著名专家和后起之秀愿意与我们共担风险，鼓舞了我们挑起这副重担的勇气。同时，我们也把这套教材的不断精化寄希望于广大的同仁，愿我们把这套教材越改越好，永改永新。

新编高等院校信息管理与信息系统
专业核心教材编委会
2002 年 5 月

前言

Q I A N Y A N

“信息比导弹更重要”，这是人们感叹信息重要性时所作的形象比较。然而，如何才能快而准地获得所需的信息？这正是“信息检索”课程所要解决的问题。

当今社会，信息“爆炸”；信息内容，无所不包；信息载体，多种多样。印刷型信息载体与电子型信息载体各具特色，又相互转化、相互补充、相互交叠。各种检索工具、数据库、搜索引擎层出不穷。如何将海量的各种各样的信息资源有序地组织起来，在需要的时候迅速查找出来，如何编制各种检索工具，如何利用检索工具等，这些都是信息时代和知识时代面临的核心课题。有关这些课题的研究，目前已发展成为一门新兴的边缘学科——信息检索。

信息检索是一门关于信息资源存储、整序和查找理论、方法的学问，其研究内容包括检索理论、检索语言、数据库、信息著录法、标引法和排检法、检索系统、检索策略、检索服务等，它是在传统的文献检索基础上，融合了最新信息检索技术发展起来的。

早在 20 世纪 80 年代，中国教育部在国际上率先确立了“文献检索”课程，该课程如同外语、数学课程一样，是各学科大学生的必修课程。此举曾引起美国同行的羡慕和赞同。但是，随着以计算机为主的新技术的飞跃发展，人们对信息需求的激增，传统的检索方法和检索内容都发生了许多变化。在专业教育方面，原有的五个信息专业合并为一个新专业“信息管理与信息系统”。这些变化迫切需要编写一本融古今中外、手检和机检于一体的，适合于信息管理与信息系统专业大学生的信息检索教材。正是在这样的背景下，本书应运而生。

本书是一本介绍信息检索原理和方法及中外文各学科最常用、最重要的检索工具（包括数据库、搜索引擎等）的教科书。主要内容包括信息检索的原理、方法及意义、检索语言、检索工具的编制与评价、各种印刷型、电子型检索工具，诸如目录、索引、文摘、百科全书、年鉴、传记、词典的特点、作用和使用方法，手工检索、光盘检索、联机检索和网络检索的异同，中外著名检索工具书、数据库的特点、使用范围和出版最新情况等。着重介绍信息资料查找的基本知识和治学入门的必备知识，旨在培养学生的信息意识和信息查找能力，在最短的时间内获得最多的所需信息，以提高信息素养。书后附有可供继续阅读的参考书目和实用性很强的各学科及主要特种文献的常用检索工具、数据库、网址一览表，以有限的篇幅起到“检索工具指南”的作用。

本书与同类教材相比具有如下特色：

1. 全面性。近年来一些教材，论述面较为单一，或仅述网络检索，或侧重传统印刷型文献检索，或主要讲科技文献检索，或强调社科文献检索，或注重机器检索、不讲手工检索，不够全面与系统。本书则整合各种载体、各种文种（中英文为主）、各个学科、各种检索方式、各种检索方法、各种检索工具的知识，将人为分散的检索知识连成一体，满足“宽口径、厚基础、重素质”的教育要求。

2. 新颖性。本书注重介绍近几年新出现的较为可靠的检索方法、检索工具和最新动态，如波普尔世界三理论的新解，都柏林元数据，数据仓库，智能搜索引擎，Z39.50，Elsevier，Kluwer 等外文期刊全文数据库的介绍等。同时在全书结构安排上，也力求有所创新，即改变以往同类教材过多介绍检索工具，忽视或很少讲述信息存储方法的偏颇，适当增加了信息著录法、标引法等内容，使学生不仅会“用”，而且会“编”，从而在更深的层次上理解和掌握信息检索知识。

3. 准确性。本书力求讲述准确，精选和考证有关资料。如在介绍检索工具或网址时，尽可能根据最新版的第一手资料，或亲自上网检索后才下笔。

4. 实用性。本书强调实际查找能力的培养，但不要求死记硬背各种检索工具的具体细节。希望读者掌握各种检索工具的共性和个性，能触类旁通、举一反三。

5. 简明性。本书力求结构清晰，文字简练，具有较强的可读性。

本书的大纲和全书的统稿以及各章节的修改，均由叶继元负责。

各章节的编写具体分工如下：

第 1 章由叶继元编写，第 2 章由徐国利编写，第 3 章由李於洪编写，第 4 章由丁建琴和叶继元编写，第 5 章由杨卫华编写，第 6 章由安艳杰、王同江编写，第 7 章主要由沈艺编写（其中 7.4.4 节为俞德风编写），第 8 章 8.1 节由沈艺编写、8.2 节由徐国利编写、8.3 节由李於洪编写。

在本书编著过程中，张基温教授给予了大力支持与帮助，在此表示衷心的感谢。

由于任务多、时间紧，书中难免有不当之处，恳请读者指正。

叶继元
2003 年 3 月

新编高等院校信息管理与信息系统
专业核心教材顾问
(按姓氏笔画排序)

马费成 陈 禹 黄梯云

新编高等院校信息管理与信息系统
专业核心教材编委会
(按姓氏笔画排序)

马费成	王要武	叶继元
李一军	汪玉凯	陈京民
吴玲达	张维明	张基温
赵国俊	高 阳	戴宗坤

执行主编：张基温

Introduce to I

新编高等院校信息管理与信息系统
专业核心教材

书 目

信息技术原理

计算机系统原理

数据仓库与数据挖掘技术

信息系统安全导论

管理信息系统

信息检索导论

信息系统工程

多媒体技术

信息资源开发与管理

数据库技术与应用

电子政务

电子商务原理

目 录

第 1 章 信息检索概述	(1)
1.1 信息资源的概念与类型	(2)
1.1.1 信息与信息资源的概念	(2)
1.1.2 信息资源的类型	(6)
1.2 信息检索的概念与类型	(11)
1.2.1 信息检索的概念	(11)
1.2.2 信息检索的类型	(12)
1.3 信息检索的研究内容、研究方法与作用	(14)
1.3.1 研究内容	(14)
1.3.2 研究方法	(16)
1.3.3 信息检索的作用	(16)
本章小结	(17)
习题	(17)
第 2 章 信息检索理论与方法	(19)
2.1 信息检索原理	(20)
2.1.1 信息检索策略	(20)
2.1.2 信息检索步骤	(28)
2.1.3 信息检索方法	(30)
2.1.4 信息检索途径	(32)
2.1.5 案例分析	(35)
2.2 信息资源著录法	(36)
2.2.1 传统著录法	(37)
2.2.2 机读目录	(39)
2.2.3 元数据	(52)
2.2.4 都柏林元数据与机读目录之比较	(58)
2.2.5 元数据的应用	(63)
2.3 信息资源标引法	(69)

2.3.1	分类标引	(69)
2.3.2	主题标引	(79)
2.3.3	关键词标引	(89)
2.3.4	名称标引	(91)
2.3.5	其他标引	(91)
2.4	信息资源排检法	(92)
2.4.1	字序法	(92)
2.4.2	类序法	(103)
2.4.3	其他排检法	(107)
	本章小结	(108)
	习题	(110)
第3章	信息检索工具与数据库	(111)
3.1	检索工具概述	(112)
3.1.1	检索工具的界定	(112)
3.1.2	检索工具概况	(113)
3.1.3	检索工具的类型	(114)
3.1.4	检索工具的编制	(123)
3.2	数据库概述	(125)
3.2.1	数据库的界定	(125)
3.2.2	数据库的结构	(127)
3.2.3	数据库的特点与功用	(133)
3.2.4	数据库的类型	(136)
	本章小结	(148)
	习题	(148)
第4章	手工检索	(151)
4.1	手工检索的特点	(152)
4.2	手工检索的策略	(152)
4.3	手工检索的方法	(154)
4.3.1	追溯法	(154)
4.3.2	工具法	(154)
4.3.3	综合法	(155)
4.4	中外著名检索工具书的查检	(156)
4.4.1	目录	(156)

4.4.2	索引	(162)
4.4.3	文摘	(168)
4.4.4	百科全书	(175)
4.4.5	类书	(176)
4.4.6	年鉴	(177)
4.4.7	传记工具书	(178)
4.4.8	词典	(179)
本章小结		(181)
习题		(181)
第 5 章	光盘检索	(183)
5.1	光盘概述	(184)
5.1.1	光盘的发展	(184)
5.1.2	光盘的类型	(184)
5.1.3	光盘的彩皮书标准	(187)
5.2	光盘检索特点	(189)
5.2.1	光盘检索与手工检索的比较	(189)
5.2.2	光盘检索与联机检索的比较	(190)
5.2.3	光盘检索的局限	(191)
5.3	光盘检索要求	(192)
5.3.1	光盘检索设备	(192)
5.3.2	光盘存储技术	(196)
5.3.3	光盘检索步骤	(198)
5.3.4	检索表达式	(200)
5.4	中外著名光盘数据库的选介	(203)
5.4.1	二次文献数据库光盘	(203)
5.4.2	一次文献数据库光盘	(212)
5.4.3	其他数据库光盘	(214)
本章小结		(219)
习题		(219)
第 6 章	联机检索	(221)
6.1	联机检索概述	(222)
6.1.1	联机检索系统的构成	(222)
6.1.2	联机检索的基本原理	(223)

6.1.3 联机检索功能	(224)
6.2 联机检索特点	(224)
6.3 联机检索的步骤与方法	(226)
6.4 联机检索的影响因素	(228)
6.4.1 人为因素	(228)
6.4.2 机器因素	(228)
6.5 中外文著名联机检索数据库系统	(230)
6.5.1 中文联机检索数据库系统	(230)
6.5.2 外文联机检索数据库系统	(241)
本章小结	(251)
习题	(252)
检索实例 6-1	(252)
第 7 章 网络检索	(257)
7.1 因特网信息检索的特点	(258)
7.1.1 开放性	(258)
7.1.2 链接性	(259)
7.1.3 简便性	(261)
7.2 传统网络信息检索服务工具	(262)
7.2.1 远程登录	(262)
7.2.2 文件传输服务	(262)
7.2.3 电子邮件	(263)
7.2.4 电子公告牌	(264)
7.2.5 Archie	(265)
7.2.6 广域消息服务	(265)
7.2.7 Gopher	(266)
7.2.8 WWW	(266)
7.2.9 基于 Z39.50 的信息检索服务	(267)
7.2.10 代理服务器和 NAT	(268)
7.3 搜索引擎	(269)
7.3.1 搜索引擎的类型	(270)
7.3.2 搜索引擎的特点	(273)
7.3.3 著名搜索引擎的使用	(274)
7.3.4 搜索引擎的局限	(293)

7.3.5 智能搜索引擎	(295)
7.4 中外著名网络数据库的检索	(300)
7.4.1 数据型数据库	(300)
7.4.2 事实型数据库	(304)
7.4.3 文献型数据库	(307)
7.4.4 全文型数据库	(311)
本章小结	(334)
习题	(334)
第 8 章 信息检索效果评价	(337)
8.1 各种检索方法的比较	(338)
8.1.1 信息检索查全率	(339)
8.1.2 信息检索查准率	(340)
8.1.3 信息检索漏检率	(341)
8.1.4 信息检索误检率	(342)
8.1.5 信息检索响应时间及其他指标	(342)
8.2 提高检索效果的方法	(343)
8.2.1 提高检索人员素质	(343)
8.2.2 优选检索工具和数据库	(344)
8.2.3 优化检索策略与步骤	(347)
个案分析 8.1: Go 搜索引擎的优化策略	(348)
个案分析 8.2: Google 搜索引擎的优化策略	(348)
8.2.4 精选检索词	(348)
8.2.5 巧构检索提问式	(350)
8.2.6 熟悉检索代码与符号	(352)
8.2.7 鉴别检索结果	(353)
8.3 学术信息检索实例评价	(355)
8.3.1 人文科学学术资源检索实例评价	(355)
8.3.2 社会科学学术资源检索实例评价	(358)
8.3.3 自然科学学术资源检索实例评价	(360)
8.3.4 工程技术学术资源检索实例评价	(363)
本章小结	(368)
习题	(368)
附录 A 各学科及主要特种文献常用检索工具、数据库、网址一览表	(369)

A1	各学科常用检索工具、数据库、网址一览表 (按《中图法》大类排)	(369)
A1.1	马克思主义、列宁主义、毛泽东思想、邓小平理论	(369)
A1.2	哲学、宗教	(370)
A1.3	社会科学总论	(371)
A1.4	政治、法律	(373)
A1.5	军事	(375)
A1.6	经济	(375)
A1.7	文化、科学、教育、体育	(377)
A1.8	语言、文字	(379)
A1.9	文学	(380)
A1.10	艺术	(381)
A1.11	历史、地理	(382)
A1.12	自然科学总论	(383)
A1.13	数理科学和化学	(384)
A1.14	天文、地球科学	(385)
A1.15	生物科学	(387)
A1.16	医药、卫生	(388)
A1.17	农业科学	(389)
A1.18	工业技术	(390)
A1.19	交通运输	(392)
A1.20	航空航天	(393)
A1.21	环境科学	(394)
A2	主要特种文献常用检索工具、数据库、网址一览表	(395)
A2.1	学位论文	(395)
A2.2	专利文献	(396)
A2.3	标准与规程	(398)
A2.4	会议录	(400)
A2.5	研究报告、科技报告	(402)
参考文献		(405)

信息检索概述

第 1 章

在当今信息社会中，信息与空气、水一样重要，它与物质、能源并列构成世界的三大要素。及时获得必要、准确的信息是个人、社会存在与发展的前提条件。信息检索是关于获得所需信息的知识，它不仅是一种技能，而且已发展成为一门学科。理解有关信息与信息资源、信息检索的基础知识，对于以下各章的学习，具有重要意义。

通过本章学习，可以了解（或掌握）：

- 信息与信息资源的定义与特点；
- 信息资源的种类，包括口语信息资源、体态信息资源、实物信息资源和文献信息资源；
- 信息检索的概念、类型与作用；
- 信息检索的研究内容和研究方法。

1.1 信息资源的概念与类型

信息资源 (Information Resource, or Information Resources) 这一术语自 20 世纪 90 年代以来, 在国内外文献中被广泛使用, 然而有关信息资源的定义, 国内外有多种观点, 目前尚无统一公认的定义, 而有关信息资源的类型划分则较为一致。

1.1.1 信息与信息资源的概念

信息与信息资源, 二者既有密切联系, 又有不同之处。因此, 在讲述信息资源概念之前, 有必要先简要介绍一下信息的概念。

1. 信息的概念

信息无时无处不在, 与人类发展历史须臾不可离, 从结绳记事到人上月球, 从日常生活到科学研究, 人们都在自觉不自觉地利用信息。然而, 有关信息的确切定义, 目前不下数十种。各个学科的学者从自身学科的角度, 对信息作过不同的定义, 尽管这些定义说法各异, 但对理解信息概念均有参考价值。

汉语中的“信息”一词, “信”与“息”两字的意思相近, 前者侧重于消息、征兆, 后者强调情况、音讯。据考证, 两字连成一词使用, 最早见于《三国志》中“正数欲来, 信息甚大”的记载。唐代诗人李中《暮春怀故人》中的“梦断美人沉信息”, 宋代诗人王庭《题辰州壁》中的“每望长安信息希”诗句里均有“信息”一词, 其意为消息, 这与当代信息的含义并不完全相同。

英文“Information”一词, 有情报、资料、消息、报道之意, 中国大陆曾译为情报或信息, 后经国家有关部门核准, 建议该词一律译成信息。在中国台湾、香港地区, 该词被译为“资讯”。

从认识论角度上看, 信息既是客观存在的, 又是人的主观认识的产物, 它不同于客观世界 (世界 1), 也不同于精神世界 (世界 2), 而是物质与精神的中介 (世界 3)。作为科学的概念, 信息论创始人美国数学家 C·E·香农把信息定义为“用来清除随机事件的形式的不定性的东西”, 信息就是不定性减少的量, 是两次不定性之差。“负熵”可用来描述信息属性及其运动规律。信息量的大小可用被其消除的不定性的多少来衡量, 即等于信宿消除的不定性的数量。信宿收到信息后, 不定性消除得越多, 获得的信息量就越大。而事物不定性的大小, 又可用概率分布来描述, 即概率越小, 所提供的准确信息能在较大的程度上消除对事物认识的不确定性, 那么, 这一准确信息的信息量就越大。反之, 概率越大, 所消除的对事物认识的不确定性较小, 其信息量越小。概率为 1 或 0 的