信息检索概论

（第二版）

祁延莉 赵丹群 /编著

北京大学出版社
PEKING UNIVERSITY PRESS

信息检索概论

（第二版）

祁延莉　赵丹群　编著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

信息检索概论/祁延莉,赵丹群编著. —2 版. —北京:北京大学出版社,2013.6
ISBN 978-7-301-22648-3

Ⅰ. ①信… Ⅱ. ①祁… ②赵… 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(2013)第 129295 号

书　　名:信息检索概论(第二版)
著作责任者:祁延莉　赵丹群　编著
责 任 编 辑:王　华
标 准 书 号:ISBN 978-7-301-22648-3/TP·1292
出 版 发 行:北京大学出版社
地　　址:北京市海淀区成府路 205 号　100871
网　　址:http://www.pup.cn　　新浪官方微博:@北京大学出版社
电　　话:邮购部 62752015　发行部 62750672　编辑部 62765014　出版部 62754962
电 子 信 箱:zpup@pup.pku.edu.cn
印　刷　者:涿州市星河印刷有限公司
经　销　者:新华书店
787 毫米×1092 毫米　16 开本　18.25 印张　360 千字
2006 年 3 月第 1 版
2013 年 6 月第 2 版　2013 年 6 月第 1 次印刷
定　　价:38.00 元

内 容 提 要

本书是为高等院校信息管理本科的信息检索课程编写的教材。全书共分12章，内容包括：信息检索概述、信息源、信息检索系统、文本信息检索技术与方法、多媒体信息检索技术与方法、信息检索系统的用户界面、信息检索的策略与步骤、常用信息检索系统及其使用(一)、常用信息检索系统及其使用(二)、网络信息的组织与检索、常用搜索引擎简介、信息检索评价研究。

本书内容丰富、新颖，注重将计算机检索基础知识与用户检索技能知识相结合，既可以作为高等院校信息管理本科生的教材和教学参考书，也可以作为从事信息检索和信息服务专业人员的学习参考书。

第二版前言

本书自2006年出版后，得到了国内学者与用户的普遍好评。许多高等院校信息管理专业将本书作为信息检索课程的教材、教学参考书以及研究生入学考试指定参考书。近年来，随着信息检索技术的迅速发展和信息服务的不断深化，新的信息检索理论、方法、工具不断涌现。鉴于此，我们对本书进行了修订，以期将信息检索技术和系统的新进展、新功能、新特点纳入本书中，使之更加科学和完善。

此次修订是在初版的基础上完成的，因此初版撰写者对本书的贡献仍然值得肯定。此次修订中，研究生刘蔚参与了第3章、第8章部分内容的修订，研究生李志强参与了第10章、第11章部分内容的编写，祁延莉参与了第3章、第8章、第9章、第10章和第11章内容的编写，并对全书进行了更新、调整和补充。全书最后由祁延莉定稿。

本书在修订过程中参阅和引用了大量的国内外相关文献和网站，在此，我们对作者表示衷心的感谢。

我们还要感谢广大的读者，感谢北京大学出版社的王华编辑，感谢所有关心和帮助过编者的人，没有他们的大力支持和帮助，本书是不可能顺利修订再版的。

敬请读者批评指正！

祁延莉

2012年8月于燕园

第一版前言

随着信息技术的快速发展，尤其是近年来 Internet 的日益普及和网上信息的激增，大大扩展了人们可利用的信息空间。随着计算机研究者的进入，信息存储与检索技术已经成为 IT 产业的核心技术，并取得了一系列的创新性成果。与此同时，信息检索系统无论从技术上还是服务方式上也都向网络化、可视化、便捷化等方向发展，信息检索理论也更加丰富。

本书试图对信息检索领域关注的重要问题进行整合，尽可能全面地介绍信息检索领域最新研究成果、实用的信息检索系统以及信息检索的过程及方法，以适应现代信息管理的需要。本书的主要内容包括信息检索的基本原理、发展历史、信息检索的对象——各种类型的信息源、信息检索系统及其构成、文本信息检索技术、多媒体信息检索技术、信息检索系统的用户界面、信息检索策略与步骤、常用信息检索系统介绍、网络信息的组织与检索、常用搜索引擎介绍、信息检索评价研究等。本书着重介绍信息检索的基本理论、技术和方法，目的是为本科学生今后从事信息检索工作和检索系统的研究与开发奠定基础。

本书的大纲和全书的框架由祁延莉和赵丹群负责确定，全书的统稿和各章节的修改由祁延莉负责完成。值得一提的是，本书的编者在多年的信息存储与检索课程改造和重组过程中，逐渐形成了信息存储与检索课程的主要框架内容，并在教学中试用，收到了较好的效果。赵丹群老师还编写过简本内部教材在函授生教学中使用。这些有意义的工作对本书大纲的最后形成起了重要的作用。

本书各章节编写具体分工如下：

第 1 章由赵丹群编写，第 2 章由冯兰晓和祁延莉编写，第 3 章（其中 3.3.3 节部分内容由郭强编写）、第 4 章由赵丹群负责编写，第 5 章由赵丹群和朱卫杰编写，第 6 章由祁延莉和周昀编写，第 7 章由祁延莉、赵丹群和王为丰编写，第 8 章和第 9 章由祁延莉编写，第 10 章和第 11 章由贲晓光和祁延莉编写，第 12 章由赵丹群和张秀坤编写。贲晓光还为本书电子版和网络版的制作和出版做了大量的编辑工作。

在本书的编写过程中，我们参考了大量有价值的参考文献。正是这些参考文献作者的前期工作为本书的完成奠定了重要的基础，在此对书中所列的参考文献作者以及没有列出的参考文献作者表示感谢。

本书得以出版，得到了北京大学教材出版委员会提供的资金支持，在此表示感谢，并感谢所有关心、支持和爱护编者的人。

由于编者学识有限，书中难免有不当之处，恳请读者批评指正。

祁延莉

2005 年 8 月于燕园

目　录

第 1 章　信息检索概述

1.1　信息检索的基本概念

人类社会的发展过程中，信息检索(Information Retrieval，IR)的实践活动由来已久，但作为一个比较规范化的学术术语，它最早由美国学者 C. W. Mooers 在 1949 年提出并使用。近年来，随着信息资源的急剧增长，“信息检索”这一学术名词逐渐变得流行起来，并被越来越多的社会成员所认识、了解和使用。那么，信息检索概念的准确含义是怎样的呢？

1.1.1　信息检索的定义

所谓“信息检索”，广义地说是“信息存储与检索”(Information Storage and Retrieval)，它是指将信息按照一定的方式组织和存储起来，并能根据信息用户的需要找出其中相关信息的过程。因此，从本质上讲信息检索是一种有目的和组织化的信息存取活动，其中包括了“存”和“取”两个基本环节。对于“存”来说，主要指面向来自各种渠道的大量或海量信息而进行的高度组织化的存储；对于“取”来说，则要求面向随机出现的各种用户信息需求所进行的高度选择性的查找，并且尤其强调查找的快速与便利。这里，具体的存储载体可以选择卡片、书本、磁带/磁盘、光盘等；存储的内容可以是文献的书目信息、文摘或全文，也可以是图像、音频或视频的数字化信息；而具体的查找途径因存储信息类型的不同而不同，较为常见的有文献的作者、题名、主题或分类号码，图像颜色、物体形状、音乐的节奏或旋律等。

作为一种有目的和组织化的信息存取活动，信息检索中的“存”与“取”之间存在着密不可分的关系。首先，两者是相互依存的：不存储无从检索，不检索存储将失去意义；其次，两者又是互相矛盾和制约的：从存储的角度看，越简单越好，但过于简单的存储，势必影响到检索的质量与效率，即有效的检索需要以增加存储的代价作为前提。信息检索中“存”与“取”之间的这种互动关系在实际检索系统的开发与设计中，需要给予某种合理化的兼顾与平衡。

在通常情况下，大多数人讲到“信息检索”时，一般只涉及“取”，即主要关注如何从存储的信息集合中快速获取各种需要的信息。这时，信息检索也可以称为“信息查询”或“信息查找”(Information Search)。这是对信息检索概念的一种狭义理解。

1.1.2　信息检索的类型

1. 早期分类方法

按照检索对象的不同，早期信息检索一般分为文献检索、事实检索、数据检索三种不同类型。

(1) 文献检索

文献检索是指以文献(包括文摘、题录或全文)为检索对象的一类信息查询活动。典型的文献检索行为多见于以下情形：为了编写教材或撰写综述性论文，某作者需要对论述相关问题的大量文献进行搜集和阅读；为了审查某项专利发明的新颖性和先进性，审查员需要在规定

的“新颖性调查范围”内查阅有关的专利说明书及其他资料;或者,为了了解某一理论、方法的具体内容或技术细节,研究人员需要查找能提供相关知识的文献等。

(2) 事实检索

事实检索主要针对从文献中提取出来的各种事实(或知识项)所进行的检索活动。例如,微软公司在全球哪些地区设立了分公司,分公司的地址、员工数量、主要负责人等。

(3) 数据检索

数据检索主要以经过选择、整理、鉴定的各种数据信息,例如人口、国民生产总值、建筑材料的各种性能参数等作为检索对象的一类检索操作。

事实上,如果把“事实”看做是非数值型“数据”的话,事实检索可归入数据检索范畴中。文献检索与事实/数据检索有许多共同之处,在信息服务过程中,它们也常常是相互配合、相辅相成的。不过需要强调的是,文献检索与事实/数据检索之间还是存在着本质上的不同,主要表现在以下两个方面:

① 文献检索是一种“相关性检索”,“相关性”的含义是指系统不直接解答用户所提出的问题本身,而只是提供与问题相关的文献供用户参考;

② 事实/数据检索是一种“确定性检索”,“确定性”的含义则是指系统直接提供用户所需要的确切的数据或事实,检索的结果要么是有,要么是无;要么是对,要么是错。

2. 新分类方法

作为一个学术性概念,信息检索的内涵始终处于不断的丰富和发展中。随着信息处理技术的不断发展,尤其是计算机技术与网络技术的不断进步,上述传统的三分方法已经无法适应信息检索发展的现状。与过去相比,现在信息检索的对象已大大丰富。当前,除了文献、事实、数据等这些传统的文本、数值信息外,图形、图像、音频、视频等新型媒体信息急剧增加,异军突起,并逐渐纳入到信息检索的研究视野之内,信息检索的内涵也随之变得丰富起来。当前,信息检索类型出现了一种新的三分方法,即:文本检索、数值检索、音频与视频检索。

(1) 文本检索

文本检索是指以各种自然语言符号系统所表示的信息作为主要检索对象的信息检索活动。文本检索是传统(文献)检索方式的延续,目前在信息检索领域仍占据主要地位并不断获得新的发展,例如:从早期的结构化书目信息检索到当前的无结构或半结构化的自由文本检索;从关键词检索到概念检索甚至语义检索;等等。

(2) 数值检索

数值检索主要是针对数值型数据的查询而发展起来的一类较有特色的信息检索。数值检索不仅能检索出符合特定需求的数据信息,而且还可以在此基础上提供一定的数据运算能力和推导能力。由于数值信息的不断丰富和在某些领域(例如财经、金融、统计等)的广泛应用,自 20 世纪 70 年代起,数值检索逐步获得了独立发展的空间。

(3) 音/视频检索

音/视频检索是主要针对各种数字化音频与视频信息而进行查询的一类新兴的信息检索操作。目前,有关这类信息的检索技术和检索方法正在研究和探索之中,属于信息检索研究的前沿领域。

相对于早期对信息检索概念的细化方法,新的三分法比较全面地反映了信息检索概念的基本内涵和最新发展。

1.2　信息检索的基本原理

随着社会信息化进程的不断深入发展，人类对信息的需求和依赖程度也越来越强，具备一定的信息检索知识与技能也逐渐成为广大社会成员工作、学习、生活的基本需要。那么，信息检索是如何实施信息的组织化存储与快速查询，其基本原理是怎样的？下面就来简单分析一下。

1.2.1　信息检索基本原理描述

在现实生活中，用户的信息需求千差万别，获取信息的方式与途径也各式各样，但如果仔细分析基于不同信息检索设施或工具的检索处理过程，其基本原理却是相同的。在此，我们可以把信息检索的基本原理抽象概括为一句话，即对信息集合与需求集合的匹配与选择，如图 1-1 所示。

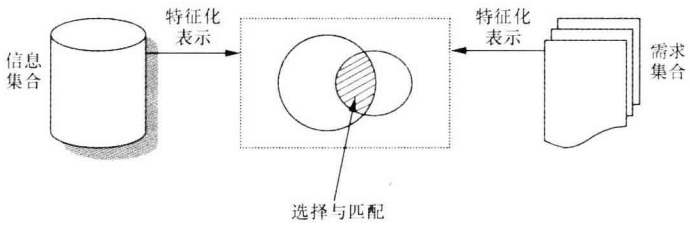

图 1-1　信息检索的基本原理示意图

为了更清楚地阐述图 1-1 所示的信息检索的基本原理，不妨对图中的各个部分进一步地认识和说明。

1. 信息集合

信息集合是指有关某一领域的，经采集、加工的信息集合体。众所周知，现实世界中已产生和积累的信息资源数量是非常庞大的，信息集合中的信息通常需要结合特定的目的和用途，或者面对特定的用户群体，对信息资源进行有选择性地采集，然后加以组织和存储，形成可供用户访问与检索的对象。

在某种意义上说，信息集合是一种公共知识结构，它有可能弥补某个特定用户的知识结构缺陷，即可以向用户提供所需要的知识或信息，或是获取知识的线索，或是提供某种信息去激活人脑中存储的知识。

2. 需求集合

用户的信息需求是在社会实践活动中产生的。当人们在完成某一任务或工作时，经常会觉得缺少某些知识，这就产生了信息需求。由于工作领域和所执行的社会职能的异同，用户信息需求具有较为客观的一面，但由于各自的个性、能力、文化程度、习惯等方面的差异，又使其信息需求带有主观性因素的影响。

众多用户不同形态的信息需求的汇集，就形成了需求集合的存在。信息需求的产生与满足，是实施信息检索行为的前提与基础，也是实施信息检索行为的目的所在。

3. 选择与匹配

面对信息集合与需求集合,如何在两者之间建立起联系与沟通的桥梁,以便能从信息集合中快速获取用户所需要或所缺少的信息与知识呢?这就要求信息检索提供一种"匹配"机制。这种机制的主要功能在于:能够把需求集合与信息集合依据某种相似性标准进行比较与判断,进而选择出符合用户需要的信息。这里,要求匹配机制至少包括以下要素:

① 匹配标准。对于"匹配标准"来说,不同的信息类型,可以选择不同的匹配标准,而且在有些情况下,匹配标准的选择还要依赖于需求的性质及系统的智能水平。例如,对于文本信息而言,最主要、最常用的匹配标准是"内容"(Content)或"主题"(Topic)标准,此外,还有"结构"(Structure)标准等。

② 执行匹配的动因。"执行匹配的动因"主要指匹配动作的执行者或实施者,在通常情况下,它可以是人,也可以是机器,或者是两者同时作用,共同完成匹配操作。

③ 特征化表示。为了保障信息检索的快速与高效,匹配操作还要求在检索匹配之前,分别对信息集合和需求集合进行某种形式化的加工,形成它们的特征化表示。对于信息集合来说,就是要对它们进行分析与标引,使每条信息都获得某种特征化表示,即让原来隐含的、不易识别的特征显性化,并获得相应的标识(例如分类号、主题词等)。正是这些被分析、提取出来的特征及标识,成为组织和查找信息资源的依据和标准。另一方面,对用户提出的信息需求也需要进行类似的加工处理,即分析需求的内容,提取出主题概念或其他属性,并利用与信息集合相同的标识系统(即检索语言)来表示需求中所包含的概念和属性,从而构成用户需求的特征化表示结果——提问(Query)。

这样一来,原先的信息需求与信息集合的匹配就简化为提问与有序的、经过特征化表示的信息集合之间的匹配,即两组有限的词语符号化特征之间的匹配比较。这种简化对于提高匹配和选择的效率来说是非常必要的,但同时它也带来了一些问题,例如检索遗漏、检索错误等。如何减少乃至避免这类问题,也就成了信息检索领域中一个备受关注的研究课题。

1.2.2 手工检索与计算机化检索的对比

手工检索和计算机化检索是当前常见的两种检索方式。虽然从信息检索的上述基本原理上来说,它们并不存在本质上的区别,但由于两者采用了不同的检索设备、选择了不同的匹配动因等,在检索形式上仍呈现出巨大的差别,如表 1-1 所示。由于计算机能存储大量的信息和数据资源,同时又具有快速、准确的计算能力和较高的可靠性,计算机化的信息检索对于具有大海捞针式苦恼的手工检索来说,无疑拥有着无法比拟的优越性。因此可以说,计算机与信息检索的结合,开辟了信息检索发展的新纪元。

表 1-1 手工检索与计算机化检索的对比

	手工检索	计算机化检索
信息集合	文字型检索工具 (书本式、卡片式)	机读数据库
需求集合	文字型检索课题 (自然语言表示)	形式化表示的提问式
匹配选择	眼看、手翻、脑子判断	计算机程序