HZ BOOKS
华章教育

高等院校管理科学与工程系列
精品规划教材

INFORMATION RETRIEVAL AND PROCESSING

信息检索与处理

主 编 王知津

机械工业出版社
China Machine Press

高等院校管理科学与工程系列
精品规划教材

INFORMATION RETRIEVAL AND PROCESSING

信息检索与处理

主　编　王知津

参　编　史海燕　陈　翀　陈芳芳　赵　洪
徐　芳　韩　毅　韩正彪　朝乐门
景　璟　樊振佳

（按姓氏笔画排序）

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

信息检索与处理 / 王知津主编 . —北京：机械工业出版社，2015.5
（高等院校管理科学与工程系列精品规划教材）

ISBN 978-7-111-50383-5

I. 信… II. 王… 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（2015）第 115347 号

本书为高等院校管理科学与工程系列精品规划教材之一，供高等院校信息管理类专业学生学习信息检索专业课使用，同时兼顾了信息存储、信息检索和信息处理等方面，区别于旨在向大学生普及信息检索方法的信息检索与利用类教材。内容涉及信息检索的原理、方法、技术、系统、网络及其相关知识。全书共分 13 章，包括信息检索与信息处理、文本检索、多媒体检索、Web 检索、检索模式扩展、信息检索模型、检索结果相关反馈与优化、用户行为与交互设计、信息检索评价与试验、自动标引、自动文摘、自动分类与聚类以及智能信息处理与知识工程等。

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使用，还可作为高等院校计算机类专业师生的教学参考书。对于从事信息检索系统、数据库以及网站开发、设计的实际工作者来说，也是一本较好的参考书。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：王金强　　责任校对：殷　虹
印　　刷：北京诚信伟业印刷有限公司　　版　　次：2015 年 6 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：22.75
书　　号：ISBN 978-7-111-50383-5　　定　　价：39.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379210　88361066　　投稿热线：（010）88379007
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjg@hzbook.com

PREFACE

前 言

19 世纪下半叶，现代科学技术出现之前，科学家们为自己的研究工作搜集资料的方式是到图书馆先查找目录，再借阅图书、期刊、报纸和工具书。从 19 世纪末开始，出现了专门的文摘索引杂志，大大提高了科学家们查找资料的速度和效率。20 世纪中叶，开始出现了计算机检索系统，这一划时代的进步使信息检索从手工检索飞跃到计算机检索，经历了脱机（离线、线下）检索、联机（在线、线上）检索和国际联机检索等阶段。20 世纪 90 年代初，随着计算机技术、通信技术和网络技术的飞速发展，特别是互联网的迅速发展和广泛应用，信息检索又进入了网络检索阶段。

提起信息检索，大多数人会很自然地联想到通过搜索引擎来进行，搜索引擎似乎已成为信息检索的代名词。但事实并非如此，除了搜索引擎外，在专业人员看来，信息检索还有更为丰富的含义。诚然，“搜索引擎”中“搜索”的英文词是 search，它也有“检索”、“查找”的意思，而“信息检索”中“检索”的英文词既可以用 search 来表示，也可以用 retrieval 来表示，甚至还可以用 seeking 来表示。可见，信息检索的指向范围比搜索引擎的指向范围更为广泛，所以说，搜索引擎只是信息检索的一个方面，而不是全部。

信息检索自 20 世纪 50 年代初提出以来，历经半个多世纪的发展和建设，已经作为一门新兴的交叉学科呈现在人们面前。从学科的角度来看，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，尽管目前它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。

环顾国内外，与信息检索有关的教材可谓琳琅满目、百花齐放。仅就我国而言，目前的绝大多数信息检索教材属于“方法”类，如《信息检索与利用》，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。还有少数信息检索教材属于“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、方法、技术、系统、网络及其相关知识，带有专业性质。本教材即属于后者。

2005 年，我们曾翻译出版了《现代信息检索》（机械工业出版社）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，2011 年该书英文版第 2 版出版，2012 年该书英文版

第2版的中文版出版。尽管该书不失为当时国内外的一部佳作，但由于文化和教育背景不同，还不能完全适合我国学生。为了更加适合我国学生，我们于2009年编写出版了国家教育部“十一五”规划教材《信息检索与存储》(机械工业出版社)。2013年，机械工业出版社计划组织编写一套高质量的管理科学与工程系列精品规划教材，并邀请我们编写信息检索教材，为此，我们编写了这本《信息检索与处理》。

我们之所以把本书定名为《信息检索与处理》，主要考虑到信息检索与信息处理之间的紧密关系。信息处理是一个非常广泛而通用的概念，几乎可以涉及人类活动的任何一个领域。信息检索致力于信息的收集、加工、存储、检索、传递和利用，这与信息处理的基本过程大致相符。因此，从广义上说，信息检索的实质就是一种信息处理，只不过更加突出了存储功能和检索功能。事实上，信息检索的所有操作过程也都是在进行信息处理，与信息处理并无二致。可以说，信息检索就是检索领域中的信息处理，或者说，信息检索就是信息处理在检索领域中的具体应用和体现，信息检索所做的一切实际上都是在进行信息处理。“信息检索与处理”更加突出了信息检索中信息处理的方法和技术，特别是新方法、新技术，有助于将相关的信息处理方法和技术融入信息检索之中，这一点可以从本书的内容中得到体现。

本书分为4个部分共13章。第一部分是信息检索导论，即第1章，主要阐述信息检索的概念、原理和类型；信息处理的含义、目的、过程、内容、步骤、方式以及信息处理与信息检索的关系；信息处理系统以及信息检索系统的概念、类型和结构；信息检索的研究内容、相关领域和发展趋势。第二部分是基本信息检索，由第2～6章组成，重点介绍文本检索、多媒体检索、Web检索、检索模式扩展以及信息检索模型。第三部分是信息检索交互与评价，由第7～9章组成，着重讨论检索结果相关反馈与优化、用户行为与交互设计以及信息检索评价与试验。第四部分是信息检索自动化与智能化，由第10～13章组成，主要探讨自动标引、自动文摘、自动分类与聚类以及智能信息处理与知识工程。

本书的编写思路和大纲由王知津提出，并经集体反复讨论和修改后确定。各章的编写者及具体分工如下：王知津（第1章）、陈芳芳（第2章）、徐芳（第3章）、史海燕（第4章）、景璟（第5章）、赵洪（第6章、第10章）、韩毅（第7章）、韩正彪（第8章）、樊振佳（第9章、第11章）、陈翀（第12章）、朝乐门（第13章）。全书由王知津审稿、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，我们对本书的主要参考文献作者表示诚挚的谢意。由于篇幅所限，我们未能一一列出所有参考文献，因此，我们对未能列出的参考文献作者表示深深的歉意。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，与机械工业出版社云逸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，我们一并表示诚挚的谢意。

虽然我们尽了自己最大的努力争取把这本教材编好，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。

王知津

2015年3月12日于南开大学

SUGGESTION

教学建议

一、教学目的

1. 掌握信息检索与处理的基本知识，以及信息检索与信息处理的关系。

2. 掌握文本检索的书目记录结构、文档处理过程与原理以及文本检索技术。

3. 理解多媒体技术与检索的基本知识，掌握多媒体检索原理。

4. 掌握主要的 Web 元数据、经典 Web 检索模型以及搜索引擎和网络爬虫的结构与原理，了解各类 Web 检索系统及其在 Web 资源利用中的作用。

5. 理解跨库检索和语义检索，掌握典型的扩展检索模式的概念、原理、技术及应用。

6. 掌握信息检索模型对信息检索过程的抽象描述以及常用信息检索模型的优缺点，了解经典信息检索模型的主要类别和数学原理及其各种扩展模型的生成。

7. 理解检索策略的构建过程、优化及相关反馈评价，掌握检索结果相关反馈的原理、技术、信息过滤的概念以及信息过滤系统的构成。

8. 掌握信息检索交互模型、信息检索系统界面测评方法，了解人机交互理论、用户心智模型理论、用户界面设计的基本原则以及用户界面交互式测评理论。

9. 掌握信息检索相关性理论和信息检索评价指标体系，了解信息检索评价过程以及典型的信息检索评价实验。

10. 掌握自动标引的概念、原理和过程，了解语料库建设的标注加工规范、语料库的设计原则以及基于概率统计和语言的自动标引方法。

11. 掌握自动文摘的概念、原理和过程，了解基于统计、结构和理解的自动文摘方法。

12. 理解自动分类与聚类的原理，掌握典型的自动分类与聚类技术的基本思想以及分类器和聚类器评估方法、指标和特征选择的计算方法。

13. 掌握信息检索系统中常用的智能处理方法与技术，了解常用的智能处理技术在信息检索系统中的应用。

二、授课建议

本课程以课堂教学为主，实例讨论为辅。建议信息管理与信息系统专业的总课时数为 72 学时，非信息管理与信息系统专业的总课时数为 54 学时。各校教师可以根据实际情况适当进行调整。

三、授课进度

教学内容	学习要点	课时安排	
		信息管理与信息系统专业	信息管理与信息系统相关专业
第 1 章 绪论	1. 掌握信息检索的概念、原理与类型以及信息检索语言的类型 2. 理解信息处理的含义、目的、过程、内容、步骤与方式，以及信息检索与信息处理之间的关系 3. 理解信息处理系统，掌握信息检索系统的概念、类型以及物理结构和逻辑结构 4. 了解信息检索的研究内容、信息检索的相关领域以及信息检索的发展趋势	4	4
第 2 章 文本检索	1. 以 CNMARC 数据记录为例，了解其形式及含义，掌握书目记录的结构 2. 掌握顺排文档和倒排文档的检索处理过程与原理 3. 练习并掌握布尔检索、截词检索、限制检索等文本检索技术的使用方法及技巧 4. 了解全文检索的关键技术指标以及全文数据库的结构和检索方法	12	6
第 3 章 多媒体检索	1. 了解多媒体技术的基本知识：概念、类型、特征以及数据压缩标准 2. 掌握多媒体检索的原理：图像检索、音频检索和视频检索 3. 理解多媒体数据模型：图像、音频、视频以及信息融合数据模型 4. 理解基于内容的多媒体检索：基于内容的图像、音频、视频以及多媒体融合检索	4	4
第 4 章 Web 检索	1. 掌握主要的 Web 元数据，了解关联数据等的最新进展 2. 掌握经典 Web 检索模型，熟悉垂直检索模型 3. 了解各类 Web 检索系统及其在 Web 资源利用中的作用 4. 掌握搜索引擎和网络爬虫的结构和原理 5. 了解 Web 检索的最新进展	4	4
第 5 章 检索模式扩展	1. 了解并行式检索、分布式检索、集群式检索、异构数据库检索、跨语言检索、可视化检索和语义检索 7 种扩展检索模式，掌握其概念和基本原理，理解其技术和应用 2. 了解并掌握异构数据库跨库检索的特点、原理和技术，理解异构数据集成的体系结构 3. 了解和掌握语义检索所涉及的领域，理解语义检索的 3 种类型	4	4

（续）

教学内容	学习要点	课时安排	
		信息管理与信息系统专业	信息管理与信息系统相关专业
第 6 章 信息检索模型	1. 掌握信息检索模型如何抽象描述信息检索过程以及常用的信息检索模型运用于检索过程的优缺点 2. 理解经典信息检索模型的主要类别和数学原理 3. 认识经典信息检索模型的各种扩展模型以及是如何改进的 4. 了解信息检索模型的发展，特别是互联网时代信息检索模型的现状与趋势	8	4
第 7 章 检索结果相关反馈与优化	1. 了解检索策略的构建过程及优化方法 2. 掌握检索结果相关反馈的基本原理及主要技术，了解相关反馈的评价方法 3. 掌握检索结果的全局自动扩展技术与局部自动扩展技术 4. 掌握信息过滤的基本概念及信息过滤系统的基本构成，了解文档信息过程和协同信息过滤的基本内容以及信息过滤的主要应用领域 5. 了解个性化信息检索的基本原理及实现方法以及构建用户兴趣模型的主要方法与技术	6	4
第 8 章 用户行为与交互设计	1. 掌握信息用户及其行为的相关概念 2. 了解人机交互理论和用户心智模型理论 3. 掌握信息检索的交互模型，尤其是 Saracevic 的交互式层次模型的内涵 4. 了解用户界面设计的基本原则和用户界面交互式测评的 3 个理论 5. 掌握信息检索系统界面测评的方法，能够运用该方法对常见的信息检索系统界面进行测评	6	4
第 9 章 信息检索评价与试验	1. 掌握信息检索相关性理论 2. 了解信息检索评价的过程和意义 3. 掌握信息检索评价的指标体系，重点掌握性能指标 4. 了解典型的信息检索评价实验及其重要进展	4	4
第 10 章 自动标引	1. 掌握自动标引的基本过程，理解各种标引方法的优点与缺点，并能实际综合运用 2. 了解自动标引的原理，特别是汉语自动标引的特点 3. 理解语料库建设的标注加工规范以及语料库的设计原则 4. 理解基于概率统计的自动标引和基于语言的自动标引的使用方法	6	4
第 11 章 自动文摘	1. 理解自动文摘的概念、原理及一般过程 2. 掌握基于统计的自动文摘基本原理 3. 掌握基于结构的自动文摘基本原理 4. 掌握基于理解的自动文摘基本原理 5. 了解有关自动文摘技术的发展历史和研究进展	4	4
第 12 章 自动分类与聚类	1. 理解自动分类和自动聚类的原理 2. 掌握典型分类与聚类技术的基本思想，并能运用自动分类和聚类技术解决问题 3. 掌握分类器和聚类器评估方法与指标 4. 掌握特征选择的计算方法	6	4

（续）

教学内容	学习要点	课时安排	
		信息管理与信息系统专业	信息管理与信息系统相关专业
第 13 章 智能信息处理与知识工程	1. 掌握信息检索系统中常用的智能处理方法与技术 2. 理解常用智能处理技术在信息检索系统中的应用模式 3. 了解自然语言处理技术在智能信息检索中的应用现状与发展趋势 4. 了解机器学习技术在智能信息检索中的应用现状与发展趋势 5. 了解自动问答系统在智能信息检索中的应用现状与发展趋势 6. 了解 Web 信息挖掘在智能信息检索中的应用现状与发展趋势 7. 了解知识工程在智能信息检索中的应用现状与发展趋势	4	4
实例与讨论：建议教师在讲授过程中尽量增加实例教学，每章安排一次讨论，在教师的指导下，由学生准备并进行实例分析，讨论时间由教师灵活掌握，总的时间已经包括在各章之中，没有单列。建议教师尽量安排一些实践性的作业练习，并将作业练习与分析讨论结合起来		—	—
课时总计		72	54

说明：本书的总授课学时数仅作为参考，对于不同类型的学校和专业，可以适当调整课时，以适应授课对象的特点。对于授课内容，教师也可以根据实际情况进行增删改。建议增加学生的分析讨论，增加学生动手操作的作业，有助于学生理解与接受本书知识点。对于非信息管理与信息系统专业的，可以根据本专业情况调整授课时数，以便有利于学生进行学习。课程讨论是必要的，也是有效果的，因此建议加强这一环节，从而调动学生学习的积极性和主动性，促进学生理解、掌握和运用相关知识点。

CONTENTS

目　录

第三部分　信息检索交互与评价

第四部分　信息检索自动化与智能化

PART 1

第一部分

信息检索导论

第 1 章　绪论

CHAPTER 1

第1章 绪　论

教学目的与要求

本章的主要内容包括信息检索的概念、原理与类型，信息检索语言；信息处理的含义、目的、过程、内容、步骤与方式，信息检索与信息处理之间的关系；信息处理系统，信息检索系统的概念、类型、物理结构和逻辑结构；信息检索的研究内容及相关领域，信息检索的发展趋势等。重点掌握信息检索与处理的基本知识，了解信息检索系统的概况以及信息检索的研究内容、相关领域和发展趋势。

引言

20多年前，“信息检索”还只是信息检索专业领域里使用的一个专门术语，广大信息用户都不熟悉甚至很少听说过，并且还带有一点“神秘”的色彩。事实上，很久以来，人们在学习、工作和生活的各个活动领域里，每时每刻都在需要信息和利用信息，对信息的需求一天都没有停止过。在那个时期，人们之所以还不熟悉信息检索，一方面是因为信息的数量还不是特别多，人们通过对印刷型文献资料的“手翻眼看”的简单操作基本上就可以满足自己的信息需求；另一方面，信息检索的专门方法和手段还不普及，绝大多数信息用户对于信息检索的知识还了解甚少。因此，严格意义上的检索操作都不是用户亲自进行的，而是由专职检索人员代替完成的。

20世纪90年代以后，随着信息数量的爆炸性增长，人们的信息需求越来越多、越来越复杂、越来越迫切。与此同时，随着计算机、通信和网络技术的飞速发展，特别是互联网的触角延伸到世界的各个角落，成为家喻户晓、人人必备的大众工具，从而使信息检索也发生了翻天覆地的巨大变化。如果说20多年前大多数信息用户还不知信息检索为何物的话，那么今天再提起“信息检索”已经不是什么新鲜事了，它已经成为大多数人耳熟能详的常用术语。

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。目前，信息检索已经逐渐形成了包括自己的理论、方法、技术和应用领域在内的完整的学科体系，信息检索作为一门新兴的交叉学科呈现在人们面前。尽管这个领域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。20多年前，信息管理与信息系统在我国的迅速兴起和大力推广，为信息检索提供了新的发

展契机。信息检索与信息管理的结合，使其成为信息管理的重要组成部分；而信息检索与信息系统的融合，又使两者进一步完善。

作为本书的绪论，本章主要介绍信息检索与处理的基本知识，包括信息检索的概念、原理与类型，以及检索语言、检索策略、检索技术与方法。作为本书的一个特色，特别强调信息检索和信息处理的相互融合，因此，对信息处理的含义、目的、过程、内容、步骤、方式以及信息检索与信息处理的关系等也做了相应的介绍。在信息检索系统方面，主要介绍信息检索系统的概念、类型以及物理结构和逻辑结构。此外，作为一门学科，还要介绍信息检索的研究对象、内容、方法、任务、相关领域以及信息检索的产生、发展与未来趋势等。

1.1 信息检索概述

1.1.1 信息检索的概念

“信息检索”（information retrieval, IR）一词最早是由美国学者 C. N. Mooers 于 1950 年使用的，[㊀㊁]随后在学术和实践领域得到广泛的应用。

在我国，20 世纪 90 年代前，信息检索仅限于图书情报领域，并被称为“情报检索”，还未被世人熟知。随着信息爆炸时代的到来，人们信息需求的愿望越来越强烈，而随着计算机和互联网的迅速普及，人们越来越认识到“检索”的重要性，越来越多的人通过检索满足自己的信息需求。在检索越来越深入人心，人们越来越离不开检索的情况下，“信息检索”逐渐取代了“情报检索”。

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，信息检索日臻完善。信息检索这一概念首先假设，包含相关信息的文献或其他形式的记录已经按照某种有助于检索的顺序组织起来，而信息检索就是对每个有检索意义的信息项进行表示、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并不是一件轻而易举的事情。例如，我们在万维网（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信息的所有网页（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过去三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，利用目前的 Web 搜索引擎界面，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。取而代之的是，用户必须首先将这些信息需求转换为搜索引擎（或 IR 系统）能够处理的查询式或查询（query）。这种转换以其最普遍的形式生成一组关键词或检索词，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 年代前，知道“信息检索”这个术语的人还不多。随着互联网的形成、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，使用这个术语的不同人

㊀ Mooers, C.N.The theory of digital handling of non-numerical information and its implications to machine economics[C]//Proceedings of the Meeting of the Association for Computing Machinery.Rutgers University, 1950.

㊁ Mooers, C.N.Information Retrieval Viewed as Temporal Signaling[C]//Proceedings of the International Congress of Mathematicians.1950:1, 572-573.

有不同的理解和解释，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”看作“信息存储与检索”的简称。因此，所谓信息检索，包括存储和检索两个过程，即信息存储和信息检索。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；信息检索是指当需要这些信息时，再把它们从存放的地方查找和提取出来。可见，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，更要知道如何存储，因为如何存储决定了如何检索。

第二类是狭义的，对于数量庞大的广大信息用户来说，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 信息检索原理

如上所述，广义的信息检索包括存储和检索两个过程。存储过程的实质是对信息进行标引，以形成信息特征标识，为检索过程提供入口和路径。检索过程的实质是对提问（从用户的信息需求中提炼出来）进行标引，以形成提问特征标识，然后按照存储过程所提供的入口和路径，从信息集合中查获与提问标识相符合的信息子集。可见，检索过程是存储过程的相反过程或逆过程。

在现实中，把用户的复杂信息需求与近乎无限的信息集合进行直接的比较和匹配是不现实的，取而代之的可行方式是对两者的简约代表进行比较和匹配，即间接比较和匹配。因而，信息检索原理的实质就是提问特征标识与信息特征标识的比较和匹配，这种比较和匹配代表着信息需求与信息集合之间的比较和匹配。比较和匹配的结果，如果两者一致，则检索命中或检索成功；如果两者不一致，则检索未命中或检索未成功。从用户的角度来看，信息检索原理的核心是用户所使用的检索词或者由检索词和运算符所组成的检索式与数据库中的检索词及其逻辑关系之间的比较和匹配机理。

从集合论的观点来看，检索过程是对信息集合进行选择或划分的过程，选择或划分的依据就是一系列检索条件。由于存储过程和检索过程都不具有唯一性，所以对于同一个信息需求或检索课题来说，检索方式也是多种多样的。

广义信息检索的基本原理如图 1-1 所示。

在存储过程中，专门负责信息检索系统和数据库建设的人从各种各样的信息来源中，搜集有用信息，对有用信息进行主题内容分析，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（如检索词表）把分析出来的概念转换成该系统或数据库所采用的词语（在自然语言检索系统中，直接使用自然语言，而不需要转换），再按照一定的存储规则和方式将这些有用信息组织成可供检索的数据库，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各种各样的信息需求，为了检索并获取自己所需要的信息，他必须对自己的需求进行主题内容分析，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（如检索词表）把分析出来的概念转换成该系统或数据库所采用的词语（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