重点大学计算机教材

HZ BOOKS
华章教育

信息检索系统导论

刘挺　秦兵　张宇　车万翔　等编著
哈尔滨工业大学

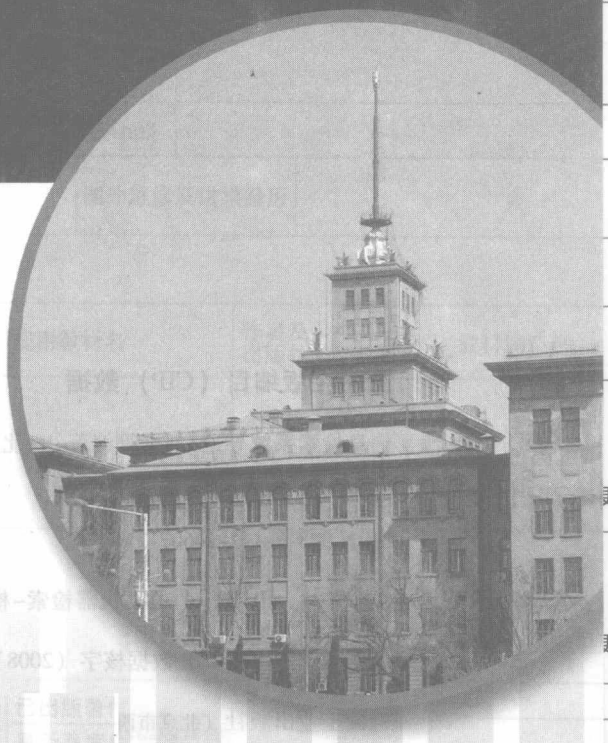
李生　主审

为教师配有电子教案

机械工业出版社
China Machine Press

重点大学计算机教材

信息检索系统导论



刘挺　秦兵　张宇　车万翔　等编著
哈尔滨工业大学

李生　主审

机械工业出版社
China Machine Press

本书对信息检索及信息检索系统的基本概念、原理、算法进行详尽介绍。主要内容包括信息检索模型、文本操作技术、文本索引和搜索技术、查询处理与 Web 检索技术、分布式信息检索、文本分类与聚类、信息过滤等，并给出 Web 信息检索的实现实例。

本书内容丰富，源于作者多年的教学及科研心得，适合作为高等院校计算机专业本科生及研究生相关课程的教材，也可作为技术人员研究信息检索与搜索引擎的参考读物。

图书在版编目（CIP）数据

信息检索系统导论/刘挺等编著. —北京：机械工业出版社，2008.8
（重点大学计算机教材）

ISBN 978-7-111-24607-7

Ⅰ. 信…　Ⅱ. 刘…　Ⅲ. 机器检索-检索系统-系统开发-高等学校-教材　Ⅳ. G354.4

中国版本图书馆 CIP 数据核字（2008）第 103376 号

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：朱　劼
三河市明辉印装有限公司印刷·新华书店北京发行所发行
2008 年 12 月第 1 版第 1 次印刷
184mm×260mm　·17 印张
标准书号：ISBN 978-7-111-24607-7
定价：35.00 元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换
本社购书热线：（010）68326294

序 言

信息检索和搜索引擎因 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被用于信息检索，而信息检索也被用于各个领域。

热门固然是一门学科兴盛的表现，每个从事研究的人都希望自己的研究领域成为热门。但热门也可能带来危险，即把信息检索当作一种时髦技术，无论适用与否都将其套用而不究其根本。对于信息检索而言，这种时髦反而是它进一步发展的障碍。

实际上，信息检索是一门复杂的学科。它的目的看似简单——找出相关的信息，却涉及计算机科学几乎所有的方面：编码、数据结构、算法……直至自然语言处理及知识的表达和应用。而在研究信息检索时，我们不得不时时思考，什么是相关信息？这个问题牵涉许多学科：信息学、数学、哲学……。要对这样一门复杂的学科有一个全面的了解，就需要一本合适的教材。一本好的教材就是一个好的引路人，而用中文编写的好的信息检索的书却为数不多。

刘挺教授和他的同事们编写的这本书正是一本能把读者引入这个领域的好书。它系统地介绍信息检索的各个方面，以及它的各种应用。从书中不但可以看到对各种古典方法及模型的描述和讨论，还可以读到对广泛使用的 Lucene 开源系统的介绍。对信息检索感兴趣的学生和学者能从这些介绍中得到莫大的帮助。

搜索引擎正处于它的兴旺时期，但这并不意味着我们已经成功地解决了它的问题。它正在扩展到人们生活的各个角落而成为一种必需的工具。由此带来的新的问题尚待我们去解决。毫无疑问，这本书为解决这些问题提供了一个很好的基础，而它的系统性的介绍和深入浅出的描述，也将使它成为适合本科生及研究生的教材之一。

聂建云

加拿大蒙特利尔大学教授

前言

信息检索这个术语同时被情报科学领域和计算机科学领域所使用。在情报科学领域，信息检索主要是指如何使用文献检索工具查找资料，例如查询科学引文数据库等；在计算机科学领域，信息检索是指信息检索算法及软件系统的研究和开发，比如建立索引的方法，检索结果的排序算法等。概括地说，前者是对工具的使用，后者是工具的研制，二者大不相同，以往出版的以“信息检索”为题的书籍大多是情报领域的，而本书属于计算机领域，为此作者在书名中特别增加了“系统”一词，以示区别，而本书中出现的“信息检索”一词均指计算机领域的信息检索，这一点请读者特别注意。从这一定位上说，本书适合计算机专业、软件专业以及情报管理专业中偏重于计算机系统开发的本科生、研究生阅读，也可供信息检索领域的相关学者参考。

近年来，随着以 Google、百度为代表的搜索引擎公司的崛起，越来越多的青年才俊对 Internet 搜索技术产生了浓厚的兴趣，很多同学选修信息检索课程，剖析开源的搜索引擎代码，申请加入与搜索相关的研究室，有的同学毕业后加盟与搜索技术相关的企业，有的则开始创业，开发新型的搜索引擎。这些充满激情、才华横溢的大学生们迫切地希望了解搜索引擎的奥秘，掌握搜索技术的架构和算法思想。

本书正是为满足读者对搜索技术的渴望而编写的，不过本书并不直接讲述如何搭建一个搜索引擎，而是侧重介绍搜索引擎背后的理论和算法。事实上，搜索引擎是一种网络上的应用，它的基础在“信息检索”。搜索引擎是随着 Internet 而发展起来的，只有十几年的历史，而信息检索伴随着计算机而兴起，有几十年的历史。信息检索方面长期的理论储备和技术积淀，为今天搜索引擎的蓬勃发展奠定了基础。掌握了信息检索的基础理论和技术，才能更深刻地理解搜索引擎的内涵，把握其千变万化中不变的本质。

八年来，笔者在哈尔滨工业大学为研究生和本科生讲授信息检索课程，同时从事信息检索方面的研究。在教学与科研实践中，深感需要一本完整而系统地介绍信息检索的教材，为此在以往教案的基础上经过反复删改补充完成了此书。本书一共有 12 章，可以分为四个部分，第一部分是第 1 ~ 3 章，介绍信息检索的基础内容，包括绪论、模型和评价；第二部分是 4 ~ 6 章，全面介绍信息检索系统中的各项技术，包括查询处理、文档处理、索引和检索技术；第三部分是第 7 ~ 9 章，讲述检索中的一些高级话题，包括 Web 检索、分布式检索以及开源的搜索代码等；第四部分是第 10 ~ 12 章，主要介绍信息检索应用，分别介绍分类和聚类、信息过滤和自动问答技术。本书试图从基础到应用，从理论到实践，从经典算法到最新的研究成果全面地介绍信息检索系统中的核心技术。由于笔者的研究背景所限，本书专注于文本检索，而没有涉及图像、语音、视频等多媒体检索。

本书是在哈尔滨工业大学信息检索研究室十余位老师和同学的共同努力下完成的。参加本书编写的还有：高立琦、刘桂平、张志辉、马金山、孙军、龚诚、郑伟、陈儒、陈毅恒、洪宇、

张志昌，还有毕业后一直在中科院计算所工作的张刚。此外，刘怀军、祝惠佳、赵妍妍、林建国等人参加了校对工作。秦兵老师担当了信息检索课程的主要教学任务，她也是这本书的主要组织者，为本书的成稿付出了大量的心血。笔者们的老师李生教授在百忙中担任了本书的主审，同时本书也得到了李生教授主持的国家自然科学基金重点项目“下一代信息检索研究”（编号60736044）的资助。

基于关键词的通用搜索技术已经发展到了一定的高度，但这只是拉开了 Internet 信息处理的序幕，更为广阔的市场需求和研究空间正在我们面前展开。在搜索方面，垂直搜索、个性化搜索、多语言搜索、移动搜索、问答式搜索、社区化搜索等代表了未来的趋势；在文本挖掘方面，信息的抽取与聚合、实体关系挖掘、意见挖掘与情感倾向性分析、针对各种类型文本的多层次多角度分类等很多以往只在学术论文中提到的内容开始展现出实用价值。希望本书能够起到为国内信息检索领域多铺一块垫脚石的作用，帮助更多的读者提高对信息检索的兴趣，加深对信息检索的了解，加入到信息检索的研发队伍中来。

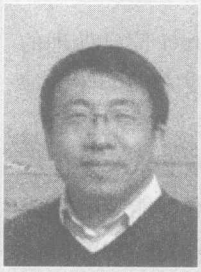
由于作者水平有限，书中疏漏在所难免，敬请读者批评指正。

作　者

2008 年 7 月

于哈尔滨工业大学

作者简介



刘 挺 教授，博士生导师。哈尔滨工业大学计算机研究所副所长，信息检索研究室主任。国家863“中文处理”重点项目总体组专家。中国中文信息学会理事，信息检索专委会副主任，计算语言学专委会委员，《中文信息学报》编委。中国计算机学会中文信息技术专委会委员，YOCSEF委员。曾任IJCNLP、AIRS等国际会议的程序委员会委员，以及全国信息检索会议NCIRCS的程序委员会主席，JSCL的多届委员等。主要研究方向为信息检索和自然语言处理，主持多项国家、部委、国际合作、企业合作等科研项目，在相关领域发表论文60余篇。

秦 兵 哈尔滨工业大学计算机学院教授，硕士生导师，2005年获哈尔滨工业大学计算机应用专业博士学位。中国计算机学会高级会员，中文信息学会会员。主要研究方向：自然语言处理、信息检索、文本挖掘等。目前承担信息检索等课程的教学工作，参加和承担多项国家自然科学基金、国家863计划、省部委及企业合作项目。在国内外刊物和会议上发表论文多篇。

张 宇 哈尔滨工业大学计算机科学与技术学院副教授，硕士生导师，中国计算机学会高级会员，中文信息学会会员。2000年在哈尔滨工业大学获得计算机应用技术专业的博士学位。曾在卡尔顿大学（加拿大，渥太华）进修。目前主要研究方向为话题的检测与跟踪、自动问答、个性化信息检索。负责或参与完成国家自然科学基金、863项目、省部委项目以及企业合作项目10余项，在国内外期刊、会议上发表学术论文30余篇。

车万翔 在哈尔滨工业大学计算机科学与技术学院获得计算机应用技术专业工学博士学位。曾于2005年获微软学者。现任哈尔滨工业大学计算机学院讲师。曾任第三届学生计算语言学会议副主席。主要研究方向：自然语言处理。承担和参加多项国家、省部委、企业合作项目。在相关领域重要期刊和学术会议发表论文20余篇。

教学建议

教学内容	学习要点及教学要求	课时安排 计算机专业及软件专业
第1章 绪论	• 掌握信息检索的概念，了解信息检索与其他相关领域的关系 • 掌握信息检索系统的结构 • 了解信息检索的发展趋势 • 了解信息检索的难点	2
第2章 信息检索模型	• 掌握三种经典的模型：布尔模型、向量空间模型和概率模型 • 重点掌握与信息检索密切相关的语言模型 • 掌握隐性语义索引模型，熟悉其基本思想和原理 • 了解扩展布尔模型和基于本体论模型的概念和基本原理	4～6
第3章 信息检索系统的评价	• 掌握信息检索系统性能评价中常用的评价指标，包括准确率、召回率以及一些单值评价方法 • 了解国外的信息检索评测会议，包括 TREC、NTCIR 以及 CLEF • 了解国内的信息检索评测会议，包括 863 信息检索评测以及 SWEM 中文 Web 评测及方法	2
第4章 文本操作技术	• 了解英文断词技术 • 掌握英文词干提取技术 • 掌握汉语分词技术，包括歧义词消解、未登录词识别、实用分词系统的架构 • 掌握信息检索中的停用词处理技术 • 了解英文的拼写检查，包括形态还原技术、词汇相似度计算技术	2
第5章 文本索引和搜索	• 掌握倒排文件索引的概念，高效实现和使用方法，以及性能分析等 • 了解排序数组、B 树和 Trie 树三种快速的词汇表存取方法 • 掌握后缀数组索引的概念，实现和使用方法，以及性能分析等 • 掌握签名文件索引的概念，实现和使用方法，以及性能分析等 • 掌握 BF、KMP 和 BM 三种单模式匹配算法	4～6
第6章 查询处理技术	• 了解查询构造的基本方法，包括单一词查询、上下文查询和布尔查询 • 掌握相关反馈技术，分别介绍向量空间模型、概率模型和布尔模型中相关反馈技术的应用 • 掌握自动查询扩展技术，包括全局分析和局部分析的方法 • 了解交互式查询扩展技术	2～4

（续）

教学内容	学习要点及教学要求	课时安排 计算机专业及软件专业
第7章 Web检索技术	• 了解Web检索的特点 • 了解Web检索技术的类别 • 掌握Web检索技术的工作原理 • 掌握Web搜索引擎的体系结构 • 掌握Web搜索引擎系统的各个子系统及相关技术	4～6
第8章 分布式信息检索	• 掌握分布式信息检索的基本概念及相关问题 • 了解分布式信息检索的体系结构 • 掌握分布式信息检索的文档集合划分问题 • 掌握常用的分布式信息检索的集合选择算法 • 了解分布式信息检索中的检索结果合并	2～4
第9章 Web信息检索实践	• 掌握Web信息检索的体系结构 • 掌握利用Lucene为文本集合建立索引 • 掌握利用Lucene为用户提供搜索服务	2～4
第10章 文本分类与聚类	• 了解文本分类和聚类的定义以及应用 • 掌握文本的表示和特征提取以及权重计算的基本方法 • 掌握文本分类基本流程以及相关机器学习算法 • 掌握文本聚类的基本算法及评价方法	4～6
第11章 信息过滤技术	• 了解信息过滤的研究背景，包括信息过滤的概念及特点、信息过滤与其他自然语言处理技术的区别与联系、信息过滤领域的核心问题 • 掌握信息过滤研究的体系结构、组成以及评测体系 • 掌握基于内容的信息过滤策略，包括特征抽取、基于规则的信息过滤方法和基于统计的信息过滤方法 • 了解基于协作的信息过滤策略，包括基于用户、基于模型和基于项目的协作过滤研究	2～4
第12章 问答系统	• 了解各种问答系统的特点以及与其他种类问答系统的区别 • 掌握基于常问问题集问答系统的实现方法，特别是问句的相似度计算与问答对的索引建立 • 掌握基于大规模文档集的问答系统的简单实现方法 • 了解现有问答技术评测会议的评测方法	2
	教学总学时建议	32～48

注：本科生教学可选择第1～9章及第11章，研究生教学则可选择第1～8章及第10～12章。

目　录

第1章 绪 论

1.1 信息检索简介

1.1.1 信息检索的概念和处理对象

什么是信息检索呢？概括地说，信息检索就是从非结构化的信息集合中找出与用户需求相关的信息。相应的，信息检索系统就是用来实现信息检索功能的计算机软件系统。

这里要强调的是，与数据库系统处理的结构化信息不同，信息检索系统处理的是“非结构化信息”。什么是“非结构化信息”呢？一篇新闻就是一条非结构化信息，新闻中会出现一些人名、地名、机构名等实体，以及这些实体之间的关系（比如某人是某地区某机关的负责人），还有与这些实体相关的事件（比如某人访问了某地）。但是这些人、事、物、关系和事件并不像关系数据库的二维表中存放的信息那样，被精确地分割并严格地存放在合适的字段或记录中。这种在现实世界中自然存在的模糊而带有歧义且没有经过规格化的信息被称为“非结构化的”（unstructured）信息。

现实世界中存在着大量的非结构化信息，除文本外，还有图像、图形、语音、视频等多媒体信息。本书不讨论多媒体检索，而是专注于文本检索，因此本书中所涉及的检索对象默认为文本。文本又有各种各样的类型，如网页、邮件、博客、论坛上的帖子、聊天记录、短信等，不同类型的文本有不同的特点，比如论坛上的帖子往往非常口语化，存在大量的别称、省略语等现象，给检索带来很大的挑战。

要处理好非结构化文本，就要尽可能地从非结构化信息中找出一些结构来。所谓的“非结构信息”并非真的没有结构，只是其结构不是显式存在的，而是隐含的，要找出其中的结构需要运用由浅到深的各类文本处理技术。比如，中文分词技术就可以把词语从句子中分割出来，而隐性语义分析（LSA）技术则从词汇－文档关系的深入挖掘中发现文本的深层结构。

1.1.2 信息检索的基本流程

上面介绍了信息检索的处理对象——“非结构化信息”，那么信息检索的基本流程是怎样的呢？在信息检索的定义中，除“非结构化”以外，我们还会用到以下几个关键词：“信息集合”、“用户需求”、“相关信息”和“找出”，这一串关键词已经刻画出了检索的流程。大家都有使用搜索引擎的经验，在百度上查找信息不外乎以下几步：输入若干关键词（用户需求），搜索引擎（信息检索系统）从网（信息集合）上中“找出”包含这些关键词的若干网页（相关信息），这就是用户体验到的检索流程。不过，为了响应用户的检索需求，信息检索系统需要事先做一些准备工作，这两项准备工作就是信息的采集与加工。

由于 Internet 上的信息浩如烟海，如果信息检索系统在某个用户发出了检索请求后再去 Internet 上找答案，则根本无法在有限的时间内返回结果。信息检索系统的实际做法是先进行信息采集，把信息源的信息拷贝到本地，构成待检索的信息集合。信息源的种类非常多，可能是纸质的图书，可能是一台电脑里的 DOC、PPT 或 PDF 等格式的电子文档，也可能是某个企业内部

的文件等。在当前的信息检索领域，人们最关注的信息源还是Internet，因为Internet上的信息是开放的，信息量非常巨大，而且还在不断更新。在Internet上采集信息的软件被称为爬虫（crawler）或蜘蛛（spider），也称作网络机器人（robot）。爬虫在Internet这个巨大的迷宫中搜索前进，每访问一个网页就把其中的内容传回本地服务器。由于网上信息的格式种类繁多，因此需要进行必要的编码方式的转换或文档格式的转换等。同时，网上存在大量的垃圾页面，需要清理，网页内还会有导航条、广告等与内容无关的信息，也需要通过网页分析将之去除，方便后续处理。

信息加工最主要的任务就是为采集到本地的信息编排索引，为查询做好准备。在传统的图书编目工作中，图书管理员需要对书籍进行分类、标引，并撰写摘要，这些工作称为二次文献。信息加工的过程和图书编目的过程类似，但全部由计算机自动完成。由于网上的信息太多，鱼龙混杂，需要一定的技术手段来判别哪些信息具有可信性、权威性。

在信息采集与加工阶段之后，就进入到用户能够参与的检索过程中了。用户输入查询式，查询可能是单个查询词，也可能是几个关键词的逻辑组合，或者是自然语言的问句。信息检索系统接收该查询，转换为查询的机内表示形式，然后在索引表中快速搜索，找到与用户的需求最匹配的若干文档，按照一定准则排序，将一部分结果返回给用户，请用户对系统返回的检索结果进行浏览。如果用户能够把他对返回结果中各个网页相关性的判断反馈给检索系统，检索系统就能够更准确地理解用户的要求，重新给出一批更有可能满足用户需求的文档，这个过程叫做**相关反馈**（relevance feedback）。用户用什么形式表达需求，怎样理解用户的需求，怎样计算用户需求与文档之间的相关度以及怎样呈现检索结果是检索过程中的关键环节。

1.1.3 与信息检索相关的学科

信息检索是一门多学科交叉的应用技术学科。信息检索的对象包括文字、图片、音频、视频等，信息检索需要利用各类媒体处理技术（比如自然语言处理、图像处理、语音处理、视频处理等）对信息进行加工，找出一定的结构，为检索提供方便。信息检索常常要面对海量数据，普通台式机的处理能力远远不够，并行/分布式处理技术在这个领域大有用武之地。数据库和数据挖掘被用来解决结构化数据的检索与知识发现问题，它们已取得的成果对于文本检索与文本挖掘都有直接的借鉴作用。知识管理、情报学、社会学等偏重管理和人文的学科从不同的角度使用信息检索技术并从中获益。

- **自然语言处理**

自然语言处理是利用计算机技术处理语言信息的学科，其目标是让计算机能够“理解”人类的语言——自然语言。对于信息检索来说，仅仅停留在处理表层文本信息是远远不够的，字符层面的匹配与相似度计算并不能帮助计算机理解待检索文本的“含义”，也不能深入理解用户的检索意图，检索出的结果非常有可能偏离用户的需求。要提高检索系统自身的智能化水平，以及检索系统人机交互界面的自然度，就需要不断地将自然语言处理结合到文本信息检索中来。

- **分布式计算**

Internet构成了人类历史上最大的信息平台，拥有海量的数据。面对巨大的文本数据、大量的检索请求和用户对于检索时间的苛刻要求，信息检索的效率成为一个亟待解决的问题，依靠单台计算机不可能完成这样的任务，必须依靠分布式信息检索技术才能解决。事实上，几乎所有实用的大型搜索系统都采用了分布式的体系结构来解决信息检索中的效率问题。

- **数据库**

数据库和信息检索俨然一对姐妹。与信息检索不同，数据库的处理对象是结构化信息。数

据库技术已经有比较完备的理论基础，而信息检索技术的经验性比较强，理论基础相对薄弱，需要进一步借鉴数据库中的一些成熟理论。信息检索中的信息抽取技术旨在把非结构化数据转化为结构化数据，以数据库形式存放，这样，一些信息检索问题就可以转化为数据库查询问题了。

- **数据挖掘**

数据挖掘一般是针对数据库进行的，借鉴到信息检索中就成为文本挖掘。面向非结构化数据的文本挖掘，将帮助用户对 Internet 上庞杂的信息进行综合分析，找出这些信息背后所蕴含的规律和趋势，找出事情的本质，提升搜索技术的内涵。此外，对用户日志进行数据挖掘能够从总体上观察分析用户的行为，也能够针对每个个体用户的需求提供个性化服务。

- **情报学**

情报学是研究情报的产生、传递、利用规律和用现代化信息技术与手段，使情报流通过程、情报系统保持最佳效能状态的一门科学。它帮助人们充分利用信息技术和手段，提高情报产生、加工、存储、流通、利用的效率。信息检索和情报学有紧密的历史渊源，情报学的理论对信息检索系统的设计仍有指导意义。

- **社会学**

社会学研究社会发展中的现象和规律。随着搜索引擎技术的使用越来越广泛，社会学家通过对众多用户使用搜索引擎的行为（比如浏览了哪些网页，输入了哪些查询词等）进行分析和统计来研究社会心理和行为的状态和趋势，比如时尚流行、语言变化、使用习惯等。

1.2 信息检索的研究内容

在这一节中，我们首先分析信息检索要解决的问题，然后介绍信息检索领域的一些基础研究课题和应用研究的课题。

1.2.1 信息检索要解决的问题

- **处理海量数据量**

传统的信息检索主要应用在图书馆检索、图书资料、情报检索等方面，数据量在千兆级以下，利用单机或者几台微型计算机就可以完成处理。随着 Internet 的发展，信息检索技术面对的数据量越来越大。很多传统的方法都不能直接应用在海量数据的处理上，否则时间开销将会变得非常大。这就给传统的算法提出了新的要求。如何解决在海量数据的情况下时间与空间上的矛盾，是信息检索技术研究的一个方向。

- **评价检索**

要想进一步发展和改善信息检索系统的性能就离不开评测。目前，如何评价信息检索的结果仍是一个困扰科学家的问题。这主要是因为信息检索结果具有很强的主观性。同一个检索结果对于不同需求的用户来说，价值很有可能是不一样的。即使是同一个用户，信息需求也可能是随时发生变化的。在这样的情况下，很难做到客观评价；没有客观的评判标准，就很难做到最优化。尽管如此，研究人员还是设计出了一些评测指标来检验信息检索系统的优劣。

- **处理多源信息**

现实世界的信息载体是丰富多样的，从文字到图片，从声音到视频，都记录着丰富的信息。单就文本而言，还存在着语言差异，同样的信息可以用中文和英文等多种语言来表达。如何将多种多样的信息在计算机内加以统一表示，如何使不同类型的信息相互融合、补充是信息检索需要解决的问题。

1.2.2 信息检索中的基础研究课题

- **信息检索理论与形式模型**

信息检索系统的设计与实现涉及多方面的内容，比如数据格式、数据结构、算法实现复杂度、操作系统的支持、分布式等。在研究过程中，如果将这些因素都考虑进去，则很难集中精力分析。信息检索模型的研究建立在抽象信息检索过程的基础上，采用四元组的形式模型表示一个信息检索系统。形式模型由查询逻辑表示、文档逻辑表示、系统表示和排序函数组成。采用形式模型研究信息检索的优点是，能够关注问题的核心和本质，忽略实现上的细节。在形式模型的指导下，提出了布尔逻辑模型、向量空间模型、概率模型、扩展布尔逻辑模型、生成式统计语言模型、判别式概率模型等信息检索模型，促进了信息检索技术的发展。

- **信息检索系统的体系结构**

信息检索是应用性很强的学科，它的研究不仅仅停留在理论上，在系统实现上也有大量问题值得关注。需要考虑的问题主要有系统的性能、数据的存储（例如是否压缩、存储格式等）、系统扩展性、系统的体系结构以及检索的有效性。信息检索系统的性能主要体现在检索的速度和响应用户的时间。对于检索系统的用户来说，检索速度肯定是越快越好。而对于系统来说，速度由多个部分共同决定（如查询处理、索引、检索、排序等），需要分析哪部分是关键部分，然后采取相应的策略提高速度。只有各部分的性能达到相互匹配，整体的性能才会呈现最优的状态。此外，采用一些机制，如缓存、调整体系结构（如开发并行性）、增加硬件投入等，都可以提高速度。信息检索体系结构主要研究如何提高检索系统的并行性，例如，通过采用分布式存储结构和检索结构提高检索性能。在衡量检索系统的优劣时，除了要关注体系结构和性能，扩展性也是一个不能忽视的方面。良好的扩展性会为进一步发展带来便利。但是一味追求扩展性，并不总是给系统带来便利，相反还可能牺牲系统性能。从工程实践角度来看，综合采用内存和外部存储的多级缓存、分布式群集和负载均衡技术是信息检索技术发展的重要方面。

- **内容表示**

在信息检索分布化和网络化的趋势下，信息检索系统的开放性和集成性要求越来越高，这要求能够检索和整合不同来源和结构的信息，包括支持各种格式化文件，如 TEXT、HTML、XML、RTF、MS Office、PDF、PS2/PS、MARC、ISO2709 等标准和格式；支持多语种信息的检索；支持结构化数据、半结构化数据及非结构化数据的统一处理；和关系数据库检索的无缝集成以及其他开放检索接口的集成等。XML 格式是一种标准、通用的数据交换格式，已经渗透到了 Internet 的很多领域，关于结构化文档 XML 的信息交换、提取、处理、查询的研究也日益受到重视。目前，已经有人提出了许多面向 XML 的查询语言，这些查询语言一般基于路径和树模式。特别是在信息检索理论中，已经在探讨 XML 文档处理的索引技术，以期达到内容和结构的双重检索。

- **信息检索评价方法和评测数据**

信息检索评价通常指的是对信息检索的性能进行评价，一般采用单值评价的方式，比如准确率（precision rate）和召回率（recall rate）来衡量信息检索系统的性能。搜索引擎也通常采用类似的方法。准确率是检索出的相关文档数与检索出的文档总数的比率，衡量检索系统（搜索引擎）的查准率；召回率是检索出的相关文档数和文档库中所有的相关文档数的比率，衡量检索系统（搜索引擎）的查全率。特别的，对于搜索引擎系统来讲，因为没有一个搜索引擎系统能够搜集到所有的 Web 网页，所以召回率很难准确计算。目前的搜索引擎系统都非常关心精度。除了准确率和召回率之外，还有 R 准确率、MAP（Mean Average Precision）、P@10 等多