第一图书网, tushu007.com

<<信息检索系统导论>>

图书基本信息

书名：<<信息检索系统导论>>

13位ISBN编号：9787111246077

10位ISBN编号：7111246071

出版时间：2008-12

出版时间：机械工业出版社

作者：刘挺 等编著

页数：257

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：http://www.tushu007.com

第一图书网, tushu007.com

前言

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。

第一图书网, tushu007.com

内容概要

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。

第一图书网, tushu007.com

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。

书籍目录

序前言作者简介教学建议第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 1.1.2 信息检索的基本流程 1.1.3 与信息检索相关的学科 1.2 信息检索的研究内容 1.2.1 信息检索要解决的问题 1.2.2 信息检索中的基础研究课题 1.2.3 信息检索中的关键技术 1.2.4 信息检索中的应用研究 1.3 信息检索的历史、现状与未来 1.3.1 信息检索的历史 1.3.2 信息检索的现状与未来 1.4 本书结构 本章小结 思考练习第2章 信息检索模型 2.1 信息检索模型的定义和分类 2.1.1 信息检索模型的定义 2.1.2 信息检索模型的分类 2.2 布尔模型 2.2.1 布尔模型的定义 2.2.2 布尔模型示例 2.3 向量空间模型 2.3.1 向量空间模型的定义 2.3.2 常见相似度计算方法 2.3.3 向量空间模型与布尔模型的比较 2.4 概率模型 2.4.1 概率模型的定义 2.4.2 概率模型的优缺点 2.5 扩展布尔模型 2.5.1 扩展布尔模型简介 2.5.2 基本模糊集合模型 2.5.3 扩展模糊集合模型 2.6 统计语言模型 2.6.1 语言模型简介 2.6.2 数据稀疏和平滑 2.6.3 基于语言模型的检索模型 2.6.4 基于语言模型的信息检索模型的优缺点分析 2.7 隐性语义索引模型 2.7.1 隐性语义索引 2.7.2 隐性语义索引模型原理 2.7.3 隐性语义索引实例 2.7.4 隐性语义索引模型的特点 2.8 基于本体论的模型 2.8.1 本体论的概念 2.8.2 描述本体的语言 2.8.3 本体的构造 2.8.4 常用的本体库简介 2.8.5 本体论在信息检索系统中的应用 本章小结 思考练习 参考文献第3章 信息检索系统的评价 3.1 引言 3.2 性能评价指标 3.2.1 准确率和召回率 3.2.2 单值评价方法 3.2.3 一些特殊的评价方法 3.2.4 其他测度方法 3.3 国外信息检索评测 3.3.1 TREC评测 3.3.2 NTCIR评测 3.3.3 CLEF评测 3.4 国内信息检索评测 3.4.1 863信息检索评测 3.4.2 SEWM中文Web评测 3.5 信息检索评价的研究 3.5.1 现有研究成果介绍 3.5.2 今后的研究问题与趋势 本章小结 思考练习 参考文献第4章 文本操作技术 4.1 引言 4.2 英文词法分析 4.2.1 断词 4.2.2 词干提取 4.3 中文词法分析 4.3.1 最大匹配法 4.3.2 歧义词切分 4.3.3 未登录词识别 4.3.4 分词系统介绍 4.3.5 语料及评测 4.4 相关资源 4.4.1 停用词表 4.4.2 词典资源 4.5 英文拼写检查 4.5.1 形态还原 4.5.2 词语相似度计算 本章小结 思考练习 参考文献第5章 文本索引和搜索 5.1 引言 5.2 倒排文件 5.2.1 倒排文件简介 5.2.2 倒排文件的使用 5.2.3 倒排文件的建立 5.2.4 倒排文件的维护 5.2.5 倒排文件的压缩 5.2.6 倒排文件性能分析 5.3 词汇表的存取 5.3.1 排序数组 5.3.2 B树 5.3.3 Trie树 5.4 后缀数组 5.4.1 后缀数组的构造 5.4.2 后缀数组的使用 5.4.3 后缀数组的分析 5.5 签名文件 5.5.1 签名文件的构造 5.5.2 签名文件的使用和维护 5.5.3 签名文件的分析 5.6 文本搜索技术 5.6.1 BF算法 5.6.2 KMP算法 5.6.3 BM算法 5.6.4 精确模式匹配算法的选择 本章小结 思考练习 参考文献第6章 查询处理技术 6.1 引言 6.2 查询构造方法 6.2.1 单一词查询 6.2.2 上下文查询 6.2.3 布尔查询 6.3 相关反馈与查询重构 6.3.1 向量空间模型中的反馈与查询重构 6.3.2 概率模型中的反馈与查询重构 6.3.3 布尔模型中的反馈与查询重构 6.3.4 相关反馈的评价 6.4 自动查询扩展技术 6.4.1 查询扩展的全局分析方法 6.4.2 查询扩展的局部分析方法 6.4.3 基于词典库的查询扩展 6.5 交互式查询扩展 6.6 查询处理的发展趋势 本章小结 思考练习 参考文献第7章 Web检索技术 7.1 引言 7.2 Web检索的工作流程及系统结构 7.2.1 工作流程 7.2.2 系统结构 7.3 Web数据的采集 7.3.1 Web数据采集系统的工作原理 7.3.2 Web数据采集系统的相关概念及协议 7.3.3 Web数据采集系统的基本结构 7.3.4 Web数据采集系统的分类 7.4 网页的预处理 7.4.1 网页去重 7.4.2 正文提取 7.5 相关性排序系统 7.5.1 早期的相关性排序技术 7.5.2 链接分析技术 7.5.3 多特征融合的相关性排序算法 7.6 Web检索系统的其他模块 本章小结 思考练习 参考文献第8章 分布式信息检索 8.1 引言 8.2 分布式信息检索系统体系结构 8.3 文档集合的划分 8.4 文档集合的选择 8.4.1 文档集合的表示 8.4.2 集合选择算法 8.4.3 文档集合选择算法的评价 8.5 检索结果的合并 本章小结 思考练习 参考文献第9章 Web信息检索实践 9.1 引言 9.2 利用Lucene建立索引 9.2.1 在Lucene中建立索引的主要步骤 9.2.2 基本索引程序 9.2.3 深入控制Lucene索引过程 9.2.4 与索引相关的并发问题 9.3 利用Lucene进行搜索 9.3.1 IndexSearcher 9.3.2 Hits 9.3.3 Query与QueryParser 本章小结 思考练习 参考资源第10章 文本分类与聚类 10.1 引言 10.2 文本分类 10.2.1 文本分类概述 10.2.2 文本分类的过程 10.2.3 分类算法 10.2.4 文本分类的评估指标 10.2.5 相关评测和相关资源 10.3 文本聚类 10.3.1 文本聚类概述 10.3.2 层次聚类 10.3.3 基于划分的聚类 10.3.4 基于密度的方法 10.3.5 自组织映射 10.3.6 基于模型的方法 10.3.7 文本聚类结果的描述 11.3.8 文本聚类的评价方法 本章小结 思考练习 参考文献第11章 信息过滤技术 11.1 引言 11.2 信息过滤的概念及

主要研究内容 11.2.1 信息过滤的概念和主要特点 11.2.2 信息过滤与信息检索、信息抽取以及分类等研究的区别 11.2.3 信息过滤系统的分类体系 11.3 信息过滤系统的结构及评价 11.3.1 信息过滤系统的组成 11.3.2 信息过滤系统的评价 11.4 基于内容的信息过滤 11.4.1 信息过滤中应用的统计模型 11.4.2 信息过滤中应用的文本分类方法 11.5 协作过滤 11.5.1 基于用户的协作过滤 11.5.2 基于模型的协作过滤 11.5.3 基于项目的协作过滤 本章小结 思考练习 参考文献第12章 问答系统 12.1 引言 12.2 问答系统的发展历程 12.3 问答系统的种类 12.3.1 问答系统分类方法 12.3.2 自然语言的数据库问答系统 12.3.3 对话式问答系统 12.3.4 基于常问问题集的问答系统 12.3.5 基于大规模文档集的问答系统 12.3.6 阅读理解系统 12.3.7 基于知识库的问答系统 12.4 基于常问问题集的问答系统实现 12.4.1 候选问题集的建立 12.4.2 句子相似度计算 12.5 基于大规模文档集的问答系统实现 12.5.1 问答的任务与系统实现流程 12.5.2 问题分析 12.5.3 相关文档检索 12.5.4 句段检索 12.5.5 答案抽取 12.5.6 问答结果的答案评测及其面对的问题和困难 本章小结 思考练习 参考文献

章节摘录

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。

编辑推荐

《信息检索系统导论》特点：理论联系实际，介绍了用Lucene实现Web信息检索的实例。
融入了作者的科研实践心得，对相关的前沿技术也有所涉及。
每章后都附有大量的参考文献，并提供思考题供读者进行深入研究。
《信息检索系统导论》为授课教师提供电子课件，请登录华章网站下载。

第一图书网, tushu007.com

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:http://www.tushu007.com