

<<网络信息检索>>

图书基本信息

书名：<<网络信息检索>>

13位ISBN编号：9787560623788

10位ISBN编号：7560623786

出版时间：2010-4

出版时间：西安电子科技大学出版社

作者：董守斌,袁华 编著

页数：348

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>

## 前言

随着互联网上的信息越来越丰富，人们一方面越来越相信所需要的信息能够在网上找到，另一方面也常常要为花不少时间才能找到所需的信息而烦恼。于是，搜索引擎在我们工作和生活中扮演的角色越来越活跃，关心和研究如何从网络上有效获取信息的人也越来越多起来。

“网络信息检索”一方面是亿万人每天都要进行的实践，另一方面也成为生机勃勃的研究领域。这从“全国搜索引擎与网络信息挖掘学术研讨会”近年投稿红火的情况可见一斑。

同时，在教学方面，我国一些大学纷纷开设了相关的课程，多数在研究生层次。

据我所知，华南理工大学是最早针对本科生开设这类课程的，本书作者即为其主讲教师，本书是她们几年来教学和科研实践的结晶。

读者会发现这本书是有用的其有用性，在于相比起我国先前出版的几本与网络信息检索相关题材的书籍而言，内容是最丰富的。其内容在时空上的跨度之大令人兴奋，使得这本书不仅可以作为教材，还可以作为打算进入这个领域的研发人员的入门参考。

例如第一章绪论，从网络信息的特点、信息检索的概念开始，对网络信息检索的基本含义进行了一个概要介绍，同时也概览了其发展的历史，列举了本领域知识与技术在多方面的应用。

从中读者可以感受到网络信息检索既是由来已久，也是方兴未艾的一个重要领域。

读者会发现这本书是有特色的，其最大的特色，就是和我国已经出版的几本类似的书籍相比，这本书定位在教材，而且很好地体现了这种定位。

本书的作者在网络信息检索领域工作多年，对内容的选取和篇章结构的安排颇有讲究。

在介绍技术性内容的章节，除了后面有思考题、练习题之外，其中还包含有大量举例，对于教材来说，这是很有意义的。

同时，我们还可以看到，作者不仅掌握了大量文献资料，而且在具体写作中融入了自己工作的体会，从而使得本书具有较强的感染力。

例如，将信息检索的要义概括为“两个表示，一个比较”，就很有教益，值得读者仔细体会。

## <<网络信息检索>>

### 内容概要

《网络信息检索》详细介绍了网络信息检索的原理和技术，内容包括信息检索模型、网络信息的自动获取、网络信息预处理和索引、查询语言和查询优化等。

针对网络信息检索的广泛应用，书中对搜索引擎、中文和跨语言信息检索、多媒体检索、并行和分布式信息检索、信息分类和聚类、信息提取与自动问答等重要应用的关键技术也进行了深入的探讨。

《网络信息检索》层次分明，深入浅出；既有原理阐述和理论推导，也有大量的实例分析，阐述力求系统性和科学性。

《网络信息检索》可作为高等院校计算机科学与技术、信息管理与信息系统、电子商务等专业的高年级本科生或研究生的教科书和参考书，对广大从事网络信息检索、数字图书馆、信息管理、人工智能、Web数据挖掘等研究和应用开发的科技人员也有较大的参考价值。

## 书籍目录

第1章 绪论 1.1 网络信息检索概述 1.1.1 网络信息 1.1.2 信息检索 1.1.3 网络信息检索 1.2 信息检索的发展 1.2.1 手工检索 1.2.2 脱机批处理检索 1.2.3 联机检索 1.2.4 网络信息检索 1.3 网络信息检索的应用 1.3.1 搜索引擎 1.3.2 多媒体信息检索 1.3.3 话题识别与跟踪 1.3.4 信息过滤 1.3.5 问题回答 思考题 参考文献 第2章 信息检索模型 2.1 检索模型定义 2.2 布尔模型 2.3 向量模型 2.3.1 索引项权重 2.3.2 相似度量 2.3.3 计算方法 2.4 概率模型 2.5 扩展的布尔模型 2.5.1 模糊集合模型 2.5.2 扩展布尔模型 2.6 扩展的向量模型 2.6.1 广义向量空间模型 2.6.2 潜语义标引模型 2.6.3 神经网络模型 2.7 扩展的概率模型 2.7.1 推理网络模型 2.7.2 信任度网络模型 2.7.3 语言模型 2.8 小结 思考题 习题 参考文献 第3章 网络信息的自动搜集 3.1 网络信息的特点 3.1.1 Web的组成 3.1.2 Web的特点 3.2 网络信息搜集的原理 3.2.1 信息搜集的基本流程 3.2.2 遍历策略 3.2.3 页面解析 3.3 网络信息搜集的礼貌原则 3.3.1 机器人排斥协议 3.3.2 机器人元标签 3.4 高性能信息搜集 3.4.1 并行搜集 3.4.2 DNS优化 3.4.3 优先搜集策略 3.4.4 网页更新 3.4.5 网页消重 3.4.6 避免蜘蛛陷阱 3.5 专题信息搜集 3.5.1 网页的主题特性 3.5.2 专题信息搜集算法 3.6 小结 思考题 习题 参考文献 第4章 网页文本处理和索引 4.1 文本的特性 4.1.1 信息熵 4.1.2 统计定律 4.2 网页信息的特征 4.2.1 网页结构 4.2.2 网页类型 4.3 网页去噪 4.3.1 基于网页结构的方法 4.3.2 基于模板的方法 4.4 文本处理 4.4.1 词汇分析 4.4.2 排除停用词 4.4.3 词干提取 4.4.4 索引词选择 4.5 索引 4.5.1 Trie树 4.5.2 后缀树 4.5.3 签名档 4.5.4 倒排文件 4.6 小结 思考题 习题 参考文献 第5章 查询语言与查询处理 5.1 Web查询语言 5.1.1 WebSQL查询语言 5.1.2 W3QL查询语言 5.1.3 WebOQL查询语言 5.2 查询方式 5.2.1 基于关键字的查询 5.2.2 模式匹配 5.3 相关反馈 5.3.1 向量空间模型中的相关反馈 5.3.2 概率模型中的相关反馈 5.4 查询扩展 5.4.1 基于字典的简单查询扩展 5.4.2 自动局部分析 5.4.3 自动全局分析 5.5 小结 思考题 习题 参考文献 第6章 信息检索性能评价 6.1 信息检索评价指标 6.1.1 查全率和查准率 6.1.2 其他评价指标 6.2 信息检索评价基准 6.2.1 基准测试 6.2.2 TREC评测 6.2.3 Web检索评价 6.2.4 CWIRF评测 6.3 小结 思考题 习题 参考文献 第7章 搜索引擎 7.1 概述 7.1.1 发展概况 7.1.2 术语与定义 7.1.3 工作原理 7.2 链接分析 7.2.1 PageRank 7.2.2 HITS 7.2.3 算法比较 7.3 相关排序 7.3.1 Lucene检索模型 7.3.2 Nutch排序算法 7.4 大规模搜索引擎 7.4.1 体系架构 7.4.2 数据结构 7.4.3 检索算法 7.4.4 相关排序 7.5 小结 思考题 习题 参考文献 第8章 并行和分布式信息检索 8.1 并行信息检索 8.1.1 并行计算的概念 8.1.2 并行信息检索体系架构 8.1.3 并行编程 8.1.4 数据并行 8.2 分布式信息检索 8.3 元搜索引擎 8.3.1 系统架构 8.3.2 资源选择 8.3.3 文档选择 8.3.4 信息融合 8.4 P2P网络信息检索 8.4.1 P2P网络信息检索的原理 8.4.2 非结构化P2P网络信息检索 8.4.3 结构化P2P网络信息检索 8.5 小结 思考题 习题 参考文献 第9章 中文和跨语言信息检索 9.1 中文预处理 9.1.1 中文编码及转换 9.1.2 中文分词 9.2 中文信息检索 9.2.1 中文检索模型 9.2.2 中文索引 9.3 跨语言信息检索 9.3.1 基本原理 9.3.2 基于GVSM的跨语言检索 9.3.3 基于LSI的跨语言检索 9.4 小结 思考题 习题 参考文献 第10章 多媒体信息检索 10.1 基于内容的图像信息检索 10.2 图像特征提取 10.2.1 颜色特征 10.2.2 形状特征提取 10.2.3 纹理特征提取 10.3 图像相似度量 10.4 基于内容的视频信息检索 10.4.1 镜头分割 10.4.2 关键帧提取 10.5 基于内容的音频信息检索 10.6 小结 思考题 习题 参考文献 第11章 信息分类与聚类 11.1 基本知识 11.1.1 类的概念 11.1.2 对象特征描述 11.1.3 文档相似性 11.1.4 类间距离 11.2 特征描述及提取 11.2.1 特征提取 11.2.2 特征选择 11.3 聚类方法 11.3.1 划分聚类法 11.3.2 层次聚类法 11.3.3 其他聚类方法 11.4 分类方法 11.4.1 NaiveBayes算法 11.4.2 kNN算法 11.4.3 Rocchio算法 11.4.4 SVM算法 11.5 方法评测 11.5.1 聚类方法评测 11.5.2 分类方法评测 11.5.3 显著性检验 11.6 小结 思考题 习题 参考文献 第12章 Web信息抽取与问答系统 12.1 信息抽取概述 12.1.1 信息抽取的发展 12.1.2 信息抽取的评价指标 12.2 Web信息抽取 12.2.1 基于关键字的Web信息抽取 12.2.2 基于模式的Web信息抽取 12.2.3 基于样本的Web信息抽取 12.3 问答系统 12.3.1 问题分析 12.3.2 信息检索 12.3.3 答案抽取 12.6 小结 思考题 参考文献

## 章节摘录

插图：网络信息是指通过互联网可以利用的各种信息资源的总和。

随着互联网的迅速发展，网络信息作为一种新型的信息资源，发挥着越来越重要的作用。

与传统的非网络信息资源相比，网络环境下的信息资源具有以下几个方面的特点：（1）网络信息内容丰富。

互联网已经成为全球最大的信息资源基地，同时其信息资源的增长十分迅速。

在互联网上几乎可以获得任何领域的信息，其内容涉及政治、经济、文化、科学和娱乐等各个方面，涵盖社会科学、自然科学、人文科学和工程技术等各个领域。

（2）网络信息变化频繁。

在互联网上，信息地址、信息链接和信息内容经常处于变动之中，信息资源的更换和消亡更是无法预测。

因而，网络信息时时刻刻处在变化和发展之中。

（3）网络信息结构复杂。互联网对网络信息资源本身的组织管理尚未形成完全统一的标准和规范，网络信息呈全球化分布结构，信息资源物理地存储在世界不同地区各种不同类型的服务器上。

因此，在信息的组织和检索方面比较复杂。

（4）网络信息格式多样。

网络信息的媒体形式多种多样，包括文本、图形、图像、声音和视频等，各种类型的媒体信息都有多种不同的信息描述格式，例如文字信息的格式有HTML、TXT、PDF、DOC等格式；图像信息的格式有BMP、GIF、JPG等格式，因此网络信息格式呈现多样化。

（5）网络信息价值差异。

由于网络信息的发布具有很大的自由度和随意性，且缺乏必要的质量控制和管理机制，因而，网络信息资源的价值差异较大，既有较大参考价值的有用信息，也有毫无用处的垃圾信息，甚至还有不少有害的信息，可谓良莠不齐。

因此，如何评价、选择和过滤信息成为网络信息组织和检索的重要任务。

编辑推荐

《网络信息检索》：新世纪计算机类本科规划教材

#### 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>