

<<搜索引擎与信息检索教程>>

图书基本信息

书名：<<搜索引擎与信息检索教程>>

13位ISBN编号：9787508453941

10位ISBN编号：7508453948

出版时间：2008-4

出版时间：水利水电出版社

作者：袁津生 等编著

页数：278

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>

<<搜索引擎与信息检索教程>>

内容概要

随着搜索引擎技术的发展和不断完善，越来越多的人开始对搜索引擎原理和技术进行研究，越来越多的人喜欢上了搜索引擎。

本书从教学的角度出发，全面地阐述了搜索引擎的技术和信息检索技术，包括：搜索引擎的基本原理与技术、搜索引擎的数据结构和搜索引擎的爬虫、信息获取与信息检索技术、分类与聚类技术以及Web信息检索技术。

本书适合高等院校计算机科学与技术专业及相关专业的高年级学生和研究生阅读参考，也适合相关领域的工程技术人员参阅。

书籍目录

前言第1章 搜索引擎概述 1.1 搜索引擎的概念 1.2 搜索引擎的发展史 1.3 搜索引擎的分类 1.4 搜索引擎的信息检索模型 1.5 建立搜索引擎的关键技术 1.6 中文搜索引擎的发展趋势 1.7 主要搜索引擎介绍 1.7.1 谷歌(Google)搜索 1.7.2 雅虎(Yahoo)搜索 1.7.3 百度(Baidu)搜索 1.7.4 天网搜索 1.8 小结 思考题第2章 搜索引擎的工作原理 2.1 搜索引擎的基本结构及工作原理 2.2 网页的搜集 2.3 网页内容的提取 2.4 查询服务 2.5 小结 思考题第3章 信息检索的模型 3.1 经典模型 3.1.1 布尔模型 3.1.2 向量模型 3.1.3 概率模型 3.2 代数模型 3.2.1 广义向量空间模型 3.2.2 神经网络模型 3.3 其他概率模型 3.3.1 贝叶斯网络 3.3.2 推理网络模型 3.3.3 信任度网络模型 3.4 小结 思考题第4章 文本操作 4.1 文本预处理 4.1.1 文本的词法分析 4.1.2 中文分词技术 4.1.3 无用词汇的删除 4.1.4 词干提取技术 4.1.5 索引词条的选择 4.1.6 词典 4.2 文本聚类 4.2.1 文本聚类算法 4.2.2 文本聚类中的相关概念 4.2.3 特征空间的降维处理 4.3 文本压缩 4.3.1 基本概念 4.3.2 统计方法 4.3.3 字典方法 4.3.4 倒排文档压缩 4.4 小结 思考题第5章 文本信息检索技术 5.1 顺排文档检索 5.1.1 表展开法 5.1.2 逻辑树展开法 5.1.3 其他顺排文档检索算法 5.2 倒排文档检索 5.2.1 倒排文档的检索 5.2.2 倒排文档的建立 5.2.3 逆波兰表达式 5.2.4 检索指令表的生成 5.2.5 检索实施 5.3 布尔检索 5.4 加权检索 5.4.1 检索词加权检索 5.4.2 词频加权检索 5.4.3 标引加权检索 5.5 全文检索 5.5.1 全文检索的技术指标 5.5.2 全文检索的实现 5.5.3 全文检索效率的提高 5.6 超文本检索 5.6.1 超文本技术概述 5.6.2 超文本的功能及结构 5.6.3 超文本检索的优缺点 5.7 分布式信息检索 5.7.1 分布式检索的查询协议 5.7.2 分布式检索系统的结构 5.7.3 分布式信息检索模式 5.7.4 分布式检索资源选择 5.8 分布式数据库查询技术 5.8.1 分布式数据库的基本概念 5.8.2 利用C#实现分布式数据库查询 5.8.3 基于.NET Remoting的查询技术 5.8.4 基于DCOM的分布式查询技术 5.8.5 基于JDBC的查询技术 5.8.6 基于Servlet的查询技术 5.8.7 基于CORBA的查询技术 5.8.8 基于Agent的查询技术 5.9 小结 思考题第6章 信息检索评价 6.1 相关性 6.1.1 相关性的特征 6.1.2 相关性研究类别 6.1.3 相关性模型 6.2 信息检索性能评价 6.2.1 信息检索系统的有效性 6.2.2 评价指标 6.2.3 Web检索系统性能评价 6.3 信息检索领域的相关组织和会议 6.4 小结 思考题第7章 文本分类与聚类 7.1 分类与聚类介绍 7.1.1 文本分类 7.1.2 文本聚类 7.1.3 文本分类的算法 7.1.4 文本聚类的算法 7.1.5 自动分类与自动聚类 7.1.6 文本分类的评测方法与指标 7.1.7 文本聚类的评测方法与指标 7.2 常用文本分类方法 7.2.1 文本分类的问题 7.2.2 kNN分类算法 7.2.3 NB分类算法 7.2.4 决策树分类算法 7.2.5 Rocchio分类算法 7.2.6 支持向量机分类算法 7.2.7 特征选择分类算法 7.2.8 文本分类系统的实现 7.3 常用文本聚类方法 7.3.1 层次聚类算法 7.3.2 分割聚类算法 7.3.3 基于密度的聚类算法 7.3.4 基于网格的聚类算法 7.3.5 基于模型的聚类算法 7.4 小结 思考题第8章 Web信息检索技术 8.1 Web信息处理的基本技术 8.1.1 Web信息的基本特点 8.1.2 Web信息的表现方式 8.1.3 Web信息系统结构 8.1.4 网络信息资源的组织与管理 8.2 Web数据挖掘 8.2.1 Web挖掘流程 8.2.2 Web挖掘的分类及现状 8.2.3 Web数据挖掘和Web信息检索的区别 8.3 Web信息检索的关键技术 8.3.1 文档搜集 8.3.2 文档预处理 8.3.3 索引数据库的建立 8.3.4 相似度计算与排序方法 8.4 搜索引擎的基本结构 8.4.1 搜索引擎的结构分类 8.4.2 网页收集模块 8.4.3 网页索引模块 8.4.4 查询模块 8.4.5 用户界面 8.4.6 搜索引擎的主要指标及分析 8.5 搜索引擎的数据结构 8.5.1 存储结构 8.5.2 信息库 8.5.3 文本索引 8.5.4 词典 8.5.5 采样表 8.5.6 前向索引 8.5.7 后向索引 8.6 搜索引擎爬虫 8.6.1 网络爬虫 8.6.2 深度优先策略 8.6.3 广度优先策略 8.6.4 不重复抓取策略 8.6.5 网页抓取优先策略 8.6.6 网页重访策略 8.6.7 网页抓取提速策略 8.6.8 Robots协议 8.6.9 网页内容提取技术 8.7 元搜索引擎 8.7.1 元搜索引擎的基本构成 8.7.2 常用元搜索引擎介绍及其分类 8.7.3 与独立搜索引擎的比较 8.7.4 主要技术指标及分析 8.8 小结 思考题第9章 搜索引擎开发技术 9.1 实例简介 9.1.1 搜索引擎的体系结构 9.1.2 网页搜集 9.1.3 网页预处理 9.1.4 查询服务 9.2 环境搭建与配置 9.2.1 idk 1.6的安装与配置 9.2.2 eclipse的安装与配置 9.2.3 Tomcat的安装与配置 9.2.4 Heritrix的安装与配置 9.3 网页搜集的实现 9.3.1 扩展Heritrix 9.3.2 抓取网页 9.4 预处理的实现 9.4.1 原始网页的处理 9.4.2 建立索引——Lucene 9.5 提供查询服务 9.5.1 搜索引擎架构设计 9.5.2 后台设计和实现 9.5.3 页面设计和实现 9.5.4 部署到Tomcat 9.6 小结 实验参考文献

章节摘录

第7章 搜索引擎概述Internet上的信息量之大、范围之广、用户之多都比以往任何时候表现得突出，然而如何从浩瀚的信息海洋中得到所需要的信息就显得更加重要。

网络搜索引擎的出现从某种程度上解决了这个问题，它是目前比较有效的网上信息获取方法，多数网上用户使用搜索引擎来获得所需的信息。

据CNNIC的统计，用搜索引擎搜索仅次于电子邮件的应用。

目前，网上比较有影响的搜索工具中，中文的有：Google、百度（Baidu）、北大天网、爱问（iask）、雅虎（Yahoo）、搜狗（Sogou）等搜索引擎；英文的有：Yahoo、AltaVista、Excite、Infoseek、Lycos、Aol等。

另外还有专用搜索引擎，例如，专门搜索歌曲和音乐的；专门搜索电子邮件地址、电话与地址和公众信息的；专门搜索各种文件的FTP搜索引擎等。

本章主要介绍搜索引擎的概念、搜索引擎的发展史、搜索引擎的分类以及一些著名的搜索引擎。

1.1 搜索引擎的概念搜索引擎并不真正搜索互联网，它搜索的实际上是预先整理好的网页索引数据库，真正意义上的搜索引擎，通常指的是收集了Internet上几千万到几十亿个网页并对网页中的每一个词（即关键词）进行索引，建立索引数据库的全文搜索引擎。

当用户查找某个关键词的时候，所有在页面内容中包含了该关键词的网页都将作为搜索结果被搜出来。

在经过复杂的算法进行排序后，这些结果将按照与搜索关键词的相关度高低依次排列。

现在的搜索引擎已普遍使用超链分析技术，除了分析索引网页本身的内容，还分析索引所有指向该网页的链接的URL、Anchor Text，甚至链接周围的文字。

所以，有时候，即使某个网页A中并没有某个词，比如“信息检索”，但如果网页B中有链接“信息检索”指向这个网页A，那么用户搜索“信息检索”时也能找到网页A。

而且，如果有越多网页的“信息检索”链接指向网页A，那么网页A在用户搜索“信息检索”时也会被认为更相关，排序也会越靠前。

搜索引擎的原理可以分为四步：从Internet网上抓取网页、建立索引数据库、在索引数据库中搜索排序、对搜索结果进行处理和排序。

（1）从Internet上抓取网页。

利用能够从Internet上自动收集网页的Spider系统程序，自动访问Internet，并沿着任何网页中的所有URL爬到其他网页，重复这过程，并把爬过的所有网页收集回来。

（2）建立索引数据库。

由分析索引系统程序对收集回来的网页进行分析，提取相关网页信息（包括网页所在URL、编码类型、页面内容包含的关键词、关键词位置、生成时间、大小、与其他网页的链接关系等），根据一定的相关度算法进行大量复杂的计算，得到每一个网页针对页面内容中及超链接中每一个关键词的相关度（或重要性），然后用这些相关信息建立网页索引数据库。

编辑推荐

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>