第一图书网, tushu007.com

<<搜索引擎>>

图书基本信息

书名：<<搜索引擎>>

13位ISBN编号：9787030342584

10位ISBN编号：7030342585

出版时间：2012-5

出版时间：科学出版社

作者：李晓明、闫宏飞、王继民

页数：330

字数：472750

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

内容概要

搜索引擎：原理、技术与系统（第二版）系统介绍了互联网搜索引擎的工作原理、实现技术及系统构建方案。
全书分三篇共13章。
上篇介绍搜索引擎的基本原理和技术，讲述一个小型简单搜索引擎实现的具体细节；中篇详细讨论了大规模分布式搜索引擎系统的设计要点及其关键技术；下篇结合“中国Web信息博物馆”和“中国互联网数字资源财富库藏”的实践经验，介绍了构建大规模Web历史网页和非网页仓储系统的技术和方法，以及中文网页的自动分类与聚类、开放域问题系统的构建等。

搜索引擎：原理、技术与系统（第二版）层次分明，由浅入深，上篇和中篇涉及内容提供了源代码下载地址；既有深入的理论分析，也有大量的实验数据和程序，具有学习和实用双重意义。

搜索引擎：原理、技术与系统（第二版）可作为高等院校计算机科学与技术、软件工程、信息管理与信息系统、电子商务等专业的研究生或高年级本科生的教学参考书和技术资料；对广大从事网络技术、Web站点管理、数字图书馆、Web挖掘等研究和应用开发的科技人员有很高的参考价值；书中提供了大量源代码，除了用于构建搜索引擎之外，对于学习编程，提高编程技巧，以及实现一个大规模应用开发也有一定的参考价值。

书籍目录

目录第二版前言第一版前言第一章 引论第一节 搜索引擎的概念第二节 搜索引擎的发展历史第三节 一些著名的搜索引擎第四节 小结上篇 Web搜索引擎基本原理和技术第二章 Web搜索引擎工作原理和体系结构第一节 基本要求第二节 网页搜集第三节 预处理第四节 查询服务第五节 体系结构第六节 小结第三章 Web信息的搜集第一节 概述一、超文本传输协议二、一个小型搜索引擎系统第二节 网页搜集一、定义URL类和Page类二、与服务器建立连接三、发送请求和接收数据四、网页信息存储的天网格式第三节 多道搜集程序并行工作一、多线程并发工作二、控制对一个站点并发搜集线程的数目第四节 如何避免网页的重复搜集一、记录未访问、已访问URL和网页内容摘要信息二、域名与IP的对应问题第五节 搜集信息的类型第六节 小结第四章 对搜集信息的预处理第一节 索引网页库第二节 网页编码识别一、基本而重要的概念二、常用字符编码三、常用字符编码算法四、字符的输入和显示五、编码识别第三节 中文自动分词第四节 分析网页和建立倒排文件第五节 小结第五章 信息查询服务第一节 检索的定义第二节 查询服务的实现一、结果集合的形成二、查询结果显示第三节 小结中篇 对质量和性能的追求第六章 可扩展搜集子系统第一节 天网系统概述和集中式搜集系统结构一、天网系统结构二、集中式搜集系统第二节 利用并行处理技术高效搜集网页的一种方案一、节点间URL的划分策略二、关于性能的讨论三、性能测试和评价四、系统的动态可配置性设计第三节 天网分布式搜集系统第四节 对Deep Web的认识一、Deep Web的成因二、搜索Deep Web的方法第五节 小结第七章 网页净化与消重第一节 网页净化与元数据提取一、DocView模型二、网页的表示三、提取DocView模型要素的方法四、模型应用及实验研究第二节 网页消重算法一、消重算法二、算法评测第三节 小结第八章 高性能检索子系统第一节 检索系统基本技术一、系统设计与结构二、索引创建三、检索过程第二节 适于查询的网页索引结构一、倒排索引结构二、平面位置索引第三节 倒排索引压缩一、倒排索引压缩技术二、词典与倒排表的压缩第四节 索引剪枝一、静态索引剪枝方法二、动态索引剪枝方法第五节 混合索引技术一、混合索引的原理二、混合索引的实现第六节 倒排文件缓存机制一、倒排文件缓存二、负载特性三、缓存策略的选择第七节 小结第九章 相关排序与系统质量评估第一节 传统IR的相关排序技术第二节 链接分析与相关排序一、链接分析二、Web查询模式下的新信息第三节 相关排序的一种实现方案一、形成网页中词项的基本权重二、利用链接的结构三、收集用户反馈信息四、计算最终的权重第四节 信息检索技术评估一、信息检索技术评估指标二、TREC和CWIRF信息检索评估三、搜索引擎技术评估第五节 小结下篇 Web信息资源的组织与应用服务第十章 大规模Web历史网页仓储系统的构建第一节 国外Web历史网页保存现状一、Internet Archive二、PANDORA三、其他相关Web保存项目第二节 中国Web信息博物馆的系统设计一、Web InfoMall的设计目标二、Web InfoMall的体系结构第三节 历史网页的存储一、数据的组织二、存储结构三、数据管理与压缩四、存储性能第四节 数据访问一、PageID的索引二、URL的索引三、数据服务四、性能与优化第五节 网页的格式保存第六节 小结第十一章 大规模Web非网页信息仓储系统的构建第一节 网络资源库藏相关工作一、Ibiblio二、Internet Archive三、Wikimedia四、中国互联网数字资源财富库藏第二节 CDAL系统概况第三节 CDAL系统设计一、系统体系结构二、可扩展的存储组织方案第四节 网络资源描述信息获取一、Ontology概述二、描述信息获取机制三、改进查询的方法四、改进排序的方法第五节 基于局部聚类思想的共现词汇算法一、基本定义二、FDC共现词汇算法第六节 小结第十二章 中文网页自动分类与聚类第一节 文档自动分类算法的类型第二节 实现中文网页自动分类的一般过程第三节 影响分类器性能的关键因素分析一、实验设置二、训练样本三、特征选取四、分类算法五、截尾算法六、中文网页分类器的设计方案第四节 天网目录导航服务一、问题的提出二、天网目录导航服务的体系结构三、天网目录的运行实例第五节 文本聚类方法一、文本聚类的一般过程二、文本间相似性的度量三、常用聚类算法四、聚类结果的评估五、搜索引擎返回结果的聚类第六节 小结第十三章 开放域问答系统第一节 概述一、问答系统的历史二、著名开放域问答系统介绍三、开放域问答系统的通用体系结构第二节 问句的分析一、问句中的指代消解二、问句分类三、问句主题提取第三节 文档和段落检索一、检索模型的选用二、查询生成三、查询结果排序四、增强索引的功能第四节 答案提取和验证模块一、生成候选答案集合二、答案提取第五节 问答系统的改进方法一、问答系统中外部资源的利用二、寻找特殊类问题的解决方案三、通过系综方法构建问答系统第六节 问答系统的评测一、TREC问答系统评测二、问答系统评测指标第七节 实例:天

网开放域问答系统第八节 小结参考文献附录 术语图目录图1-1 2012年3月在Google上检索“伊拉克战争”的结果图1-2 2012年3月在Open Directory上检索“伊拉克战争”的结果图2-1 搜索引擎示意图图2-2 搜索引擎三段式工作流程图2-3 搜索引擎的体系结构图3-1 TSE搜索引擎界面图3-2 TSE查询结果页面图3-3 TSE网页快照页面图3-4 TSE系统结构图3-5 Web信息的搜集图3-6 Sockets和端口图3-7 通过Socket建立连接图4-1 网页预处理系统结构图4-2 原始网页库中的记录格式图4-3 索引网页库算法图4-4 字符的输入和显示流程图4-5 GB2312,Big5和GBK字符编码分布图4-6 正向减字最大匹配算法流程图4-7 切词算法流程图4-8 分析网页与建立倒排文件流程图4-9 过滤网页中非正文信息算法图4-10 正向索引表记录格式图4-11 由正向索引建立反向索引图5-1 信息查询的系统结构图5-2 基本检索算法图5-3 动态摘要算法图5-4 用户查询日志的记录格式图6-1 天网系统概貌图6-2 搜集系统的主控结构图6-3 协调进程工作算法图6-4 分布式Web搜集系统结构图6-5 负载方差图6-6 并行搜集系统与集中式搜集系统的性能对比图6-7 分布式系统效率图6-8 URL两阶段映射图6-9 天网分布式搜集系统P_Arthur体系结构图6-10 人才招聘网站首页图7-1 用DocView模型提取的网页要素图7-2 净化后的网页图7-3 HTML Tree结构图7-4 内容块权值传递过程图7-5 有主题网页DocView模型生成过程图7-6 计算网页特征项权值的算法图7-7 正文段落识别过程图7-8 基于anchor text的超链选取算法图7-9 网页净化前后分类效果对比图7-10 查全率随选取关键词个数的变化图8-1 检索系统集成框架结构图8-2 天网WWW检索分布式系统构架图8-3 倒排索引结构示意图图8-4 按块组织的倒排链的结构图8-5 位置索引的结构图8-6 CLPS结构示意图图8-7 倒排链中文档号之间的d-gaps分布图图8-8 不同文档号分配下平均每个查询对应文档号序列的压缩大小图8-9 不同压缩算法对文档号的解压速度图8-10 不同文档号分配下平均每个查询对应词频序列的压缩大小图8-11 不同压缩算法对词频的解压速度图8-12 平均每个查询对应的位置信息需要的存储空间图8-13 索引剪枝方法的分类图8-14 MAXSCORE算法的示例图8-15 WAND算法选择候选文档的过程图8-16 基于最大块索引的支点文档号的选择示例图8-17 Interval-Base剪枝方法中文档子区间划分的示例图8-18 SAAT方法处理查询处理模式及分数累加器数量的变化图8-19 当前支持高效SR+IR剪枝的索引结构图8-20 扩展词典树结构示例图8-21 扩展词典匹配查找算法图8-22 搜索引擎检索系统缓存结构图8-23 文档数据访问对象大小分布图8-24 I/O与PAGE序列序号-频度分布图8-25 I/O与PAGE序列时间间隔分布图8-26 I/O和PAGE序列中唯一模式串图9-1 Inktomi提供的几种搜索引擎技术的比较图9-2 词典在系统中的地位图9-3 新词学习图9-4 网页的互联结构示意图9-5 信息获取技术评估的“森林”图9-6 查准率和召回率基础定义图示图9-7 查准率和召回率例子图9-8“省事的”11点标准召回率例子图9-9 实践中召回率例子图9-10 实际中的44个查询词的评价统计表和P-R图图9-11 测试集在检索评估中的角色图9-12 帮助判断相关结果页面的计算机辅助程序入口图9-13 帮助判断相关结果页面的计算机辅助程序操作界面图10-1 Web InfoMall体系结构图10-2 网页数据的分割图10-3 Web InfoMall的存储结构图10-4 网页的引用压缩示意图图11-1 CDAL提供的资源访问方式图11-2 CDAL系统结构图图11-3 基于Ontology的网络资源描述信息获取图11-4 概念的属性及其词汇扩展(以电影类资源为例)图11-5 获得描述信息的改进排序算法图11-6 网络资源描述信息展示图12-1 自动文档分类算法的分类图12-2 中文网页自动分类的一般过程图12-3 中文网页分类器的工作原理图图12-4 WebSmart——一个网页实例集搜集和整理工具图12-5 一种中文网页的分类体系图12-6 Macro-F1值随样本数的变化图12-7 Micro-F1值随样本数的变化图12-8 CHI、IG、DF、MI的比较(Macro-F1)图12-9 CHI、IG、DF、MI的比较(Micro-F1)图12-10 kNN与NB分类结果的比较图12-11 k的取值对分类器质量的影响(Marco-F1)图12-12 k的取值对分类器质量的影响(Micro-F1)图12-13 兰式距离法与欧式距离法对12个不同类别的分类情况图12-14 基于层次模型的kNN与基本kNN的比较图12-15 RCut和SCut截尾算法的比较图12-16 天网目录的体系结构图12-17 天网目录导航服务图12-18 文本聚类的一般过程图12-19 层次聚类实例图12-20 k-均值算法进行文本聚类的过程图12-21 搜索结果聚类系统Carrot2图13-1 START系统界面图13-2 Ask Jeeves查询结果图13-3 问答系统的通用体系结构图13-4 天网开放域系统的体系结构表目录表4-1 网页索引文件表4-2 URL索引文件表6-1 SOIF数据描述表6-2 SOIF具体语法表6-3 参照序列,假设节点数为2表7-1 类别编号对照表表7-2 消重实验结果表7-3 当N=10、δ=0.01时5种算法的查全率和准确率表7-4 考察δ的取值对算法3和4的影响表7-5 分段签名算法的时间复杂度及性能表7-6 基于关键词的各算法的时间复杂度及性能(N=10, δ=0.01)表8-1 MTF对序列进行转换的过程表8-2 对包含100万词条的词典使用不同编码所需要的空间表8-3 平均每个查询对应词频链的空间大小(文档号按URL序分配)表8-4 不同索引的组织结构及其支持的查询处理方式表8-5 数据集基本统

计信息表9-1 新词学习对检索准确率的影响表9-2 影响权值的HTML标签表9-3 补偿因子定义表表9-4 2004中文Web信息检索评测提交结果表9-5 主题提取表9-6 导航搜索表9-7 用户查询信息类别表10-1 网页存储性能(个/秒)表10-2 网页访问性能(个/秒)表11-1 几个网络资源库藏系统的特征表11-2 CDAL中的资源分布表12-1 样本集中类别及实例数量的分布情况表表12-2 kNN和NB算法的分类质量和分类效率比较表12-3 欧式距离与兰式距离的比较表12-4 基于层次模型的kNN与基本kNN的比较表12-5 RCut和SCut截尾算法的比较表12-6 一个分类器的设计方案表13-1 问题分类体系结构及TREC问答任务中问题的分布表13-2 天网开放域系统在TREC2005中的表现

章节摘录

版权页： 插图： 第二节 网页搜集 搜索引擎这样一个软件系统应该是何种工作方式？
如果说软件系统是工作在某个数据集合上的程序的话，这个软件系统操作的数据不仅包括内容不可预测的用户查询，还要包括在数量上动态变化的海量网页，并且这些网页不会主动送到系统来，而是需要由系统去抓取。
首先，我们考虑抓取的时机：事先还是即时。
我们都有经验，在网络比较畅通的情况下，从网上下载一篇网页大约需要1秒钟左右，因此如果在用户查询的时候即时去网上抓来成千上万的网页，一个个分析处理，和用户的查询匹配，不可能满足搜索引擎的响应时间要求。
不仅如此，这样做的系统效益也不高（会重复抓取太多的网页）；面对大量的用户查询，不可能想象每来一个查询，系统就到网上“搜索”一次。
因此我们看到，大规模搜索引擎服务的基础应该是一批预先搜集好的网页（直接或者间接）。
这一批网页如何维护？
可以有两种基本的考虑。
定期搜集，每次搜集替换上一次的内容，我们称之为“批量搜集”。
由于每次都是重新来一次，对于大规模搜索引擎来说，每次搜集的时间通常会花几周。
而由于这样做开销较大，通常两次搜集的间隔时间也不会很短（如早期天网的版本大约每3个月来一次，Google在一段时间曾是每隔28天来一次）。
这样做的好处是系统实现比较简单，主要缺点是“时新性”（freshness）不高，还有重复搜集所带来的额外带宽的消耗。
增量搜集，开始时搜集一批，往后只是：①搜集新出现的网页；②搜集那些在上次搜集后有过改变的网页；③发现自从上次搜集后已经不再存在了的网页，并从库中删除。
由于除新闻网站外，许多网页的内容变化并不是很经常的（有研究指出50%网页的平均生命周期大约为50天（Choetal.2000，Cho2002》，这样做每次搜集的网页量不会很大（例如，我们在2003年初估计中国每天有30万～50万变化了的网页），于是可以经常启动搜集过程（如每天）。
30万网页，一台PC机，在一般的网络条件下，半天也就搜集完了。
这样的系统表现出来的信息时新性就会比较高，主要缺点是系统实现比较复杂；这种复杂还不仅在于搜集过程，而是还在于下面要谈到的建索引的过程。
上面讲的是系统网页数据库维护的基本策略。
在这两种极端的情况之间也可能有一些折中的方案，J.Cho博士在这方面做过深入的研究（Choetal.2000，Cho2002），他根据一种网页变化模型和系统所含内容时新性的定义，提出了相应优化的网页搜集策略。
其中一个有趣的结论是：在系统搜集能力一定的情况下，若有两类网页（如“商业”和“教育”），它们的更新周期差别很大（如“商业”类网页平均更新周期是“天”，而“教育”类网页平均更新周期是“月”），则系统应该将注意力放在更新慢的网页上（Choetal.2000），以使系统整体的时新性达到比较高的取值。
在具体搜集过程中，如何抓取一篇篇的网页，也可以有不同的考虑。
最常见的一种是所谓“爬取”：将Web上的网页集合看成是一个有向图，搜集过程从给定起始URL集合S（或者说“种子”）开始，沿着网页中的链接，按照先深、先宽或者某种别的策略遍历，不停地从S中移除URL，下载相应的网页，解析出网页中的超链接URL，看是否已经被访问过，将未访问过的那些URL加入集合S。
整个过程可以形象地想象为一个蜘蛛（spider）在蜘蛛网（Web）上爬行（crawl）。
后面我们会看到，真正的系统其实是多个“蜘蛛”同时在爬。

编辑推荐

《搜索引擎：原理、技术与系统（第2版）》保留了第一版上篇的大部分内容，即搜索引擎的基本原理，过去这么些年并没有什么变化；删除了第一版中的第九，第十二和十三章，增加了第十，第十一和十三章，分别介绍基于搜索引擎技术开发并从2002年一直运行至今的“中国web信息博物馆”、“中国数字财富库藏”及开放域问答系统。
同时，较大幅度修订了第一版中的部分小节内容。
《搜索引擎：原理、技术与系统（第2版）》分三篇共13章，内容包括引论、Web搜索引擎工作原理和体系结构、web信息的搜集、对搜集信息的预处理、信息查询服务等。

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