第一图书网, tushu007.com

<<搜索引擎技术基础>>

图书基本信息

书名：<<搜索引擎技术基础>>

13位ISBN编号：9787302227960

10位ISBN编号：7302227969

出版时间：2010

出版时间：清华大学出版社

作者：刘奕群,马少平,洪涛

页数：256

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

面对浩瀚的万维网信息海洋，人类并没有如《庄子·秋水》中的河伯那样望洋兴叹、徒唤奈何，这实在是拜搜索引擎之功。
搜索引擎是人们从无远弗届、无深不入的万维网中获取信息不可或缺的手段，是人们遨游于这个海洋里孜孜以求的“探海金针”。
搜索技术也因此成为当今最热门的研究热点之一，为信息检索、数据挖掘、自然语言处理等众多领域所共同关注。
与世界上其他国家的发展路径有所区别，中国搜索引擎的发展一直坚持本土化的道路，一方面，确保了数以亿计的中文网民获取互联网信息过程的便利；另一方面，也确立了中文信息处理技术在世界范围的影响，这是与以百度、搜狗、搜搜等为代表的一系列“国产”搜索引擎的技术创新和产业发展分不开的。
技术创新和产业发展都需要优秀人才的支撑。
培养对搜索技术具有比较深刻理解的计算机专业高端人才是中文搜索引擎乃至信息处理产业发展的迫切需求。
然而，搜索引擎属于比较新的研究方向，其核心技术研发与知识体系演化的速度很快，如何从纷繁复杂的产品及其功能中凝炼出搜索引擎人才真正需要的知识与技能，是相关教学工作开展中面临的重要问题。
鉴于搜索引擎发展过程中融合了学术界与产业界两方面的创新成果，我们认为，解决这一问题也需要大学与搜索引擎企业的共同努力。
作为这方面的一个积极探索，清华大学计算机系和百度公司从2009年春季起开始合作开设“搜索引擎技术基础”课程，希望为相关人才培养贡献绵薄之力。
课程受到了清华大学同学的欢迎与好评，也激励了不少同学尝试开展搜索引擎方面的研究与创新。
清华大学的刘奕群博士、马少平教授与百度公司的洪涛先生、刘子正先生合作完成的这本书就是该课程的教材。
作为为数不多的搜索引擎技术中文教科书之一，该教材系统评价了搜索引擎技术与产业发展的概况，对搜索引擎领域得到广泛应用的各种核心算法和应用模式进行了阐述与探讨。
“鸳鸯绣出从君看，更把金针度与人”。
相信每一位对搜索引擎感兴趣的学生和学者都能通过学习或参考此书而有所收获。

第一图书网, tushu007.com

内容概要

这是一本关于搜索引擎的教科书，它从研究实践者的角度介绍了搜索引擎的相关技术及其产业，并试图协助读者成为搜索引擎领域的局内人。
与传统的将搜索引擎作为信息检索系统实现的一个特殊实例的做法不同，作者试图把搜索引擎作为一个独立的研究课题，从纷繁复杂的互联网数据现象和搜索引擎工作案例中提炼知识点，对现代商业搜索引擎的体系结构、运行原理、运营机制和核心算法进行总结和讲解。
　　本书是清华大学计算机系与百度公司合作在清华大学开设的“搜索引擎技术基础”课程的教材，适合作为高等院校信息科学技术、图书馆学等相关专业本科生与研究生相关课程的教材，也可作为相关领域技术人员与搜索引擎技术爱好者的参考资料。

作者简介

刘奕群，2003年本科毕业于清华大学计算机系并免试推荐直接攻读博士学位，2007年获博士学位后留校任教至今，目前在清华大学计算机系教授“搜索引擎技术基础”与“搜索引擎产品设计与实践”等课程。
主要从事与搜索引擎技术相关的互联网应用研究工作，包括网络信息检索、网络用户行为分析、网络产品性能评价等。
发表相关领域学术论文40余篇，申请专利7项，并与百度公司、搜狐公司、微软亚洲研究院等单位开展多项搜索引擎技术方面的合作研究。
马少平，1982年本科毕业于清华大学计算机系，1984年获清华大学计算机系硕士学位后留校任教，1991-1992年在日本学习，1997年获清华大学计算机系博士学位，1998年晋升为教授，1999年聘为博士生导师。
现任清华大学智能技术与系统国家重点实验室主任、中国人工智能学会常务理事、知识工程专业委员会副主任、中国中文信息学会理事、信息检索与内容安全专业委员会副主任。
主要从事智能信息处理方面的研究工作，包括汉字识别、文本信息检索、图像信息检索、中文古籍的数字化与检索等。
洪涛，1986年和1989年先后获得北京大学计算机学士学位和心理学硕士学位，1995年在纽约州立布法罗大学计算机系取得博士学位。
长期从事自然语言处理、搜索引擎／信息检索、互联网广告技术、数据挖掘、模式识别和金融数据分析建模等方面的研发工作。

<<搜索引擎技术基础>>

书籍目录

第1章 为什么要关注搜索引擎 1.1 互联网上最重要的应用系统 1.2 人类历史上最大规模的信息集散平台 1.3 学术界重要的技术研发平台 1.4 经济领域能够盈利的“生意”第2章 搜索引擎的基本概念与发展历史 2.1 互联网与万维网的发展 2.2 英雄辈出：搜索引擎的发展历史回顾 2.3 搜索引擎的定义与运行原理概述 2.4 总结：我们能够从历史中学到什么？
参考文献第3章 搜索引擎性能评价 3.1 搜索引擎评价与Cranfield评价体系 3.2 查询样例集合构建 3.2.1 查询样例集合构建中的真实性 3.2.2 查询样例集合构建中的代表性 3.2.3 查询样例集合构建中信息需求表述的完整性 3.3 正确答案集合构建 3.4 搜索引擎评价指标 3.5 搜索引擎性能评价的新进展 参考文献第4章 搜索引擎体系结构概述 4.1 数据抓取子系统的主要功能与性能需求 4.1.1 及时性 4.1.2 全面性 4.1.3 高效性 4.2 内容索引子系统的主要功能与性能需求 4.2.1 内容索引子系统的主要功能 4.2.2 倒排索引结构 4.2.3 内容索引子系统的性能需求 4.3 内容检索子系统的主要功能与性能需求 4.3.1 内容检索子系统与文本信息检索系统 4.3.2 内容检索子系统的相关性需求 4.3.3 内容检索子系统的查询理解需求 4.3.4 内容检索子系统的效率需求 4.4 链接结构分析子系统的主要功能与性能需求 4.4.1 基于链接结构分析评价数据质量 4.4.2 基于链接结构分析扩展文档描述 4.4.3 链接结构分析子系统的效率需求 4.5 搜索引擎体系结构设计理念 参考文献第5章 数据抓取子系统设计及核心算法 5.1 抓取系统的基本架构 5.2 数据抓取涉及的网络协议 5.2.1 URL规范 5.2.2 HTTP协议 5.2.3 User-Agent 5.2.4 robots协议 5.3 网页抓取技术 5.3.1 网页抓取的基本过程 5.3.2 基于异步I/O模型的抓取器 5.3.3 抓取压力控制 5.3.4 对URL重定向的支持 5.3.5 对HTTPS协议的支持 5.4 链接选取策略 5.4.1 爬虫的抓取方式 5.4.2 抓取优先级策略 5.4.3 网页的重访策略 5.4.4 链接去重策略 5.5 网页存储技术 5.5.1 分布式哈希存储系统 5.5.2 基于BigTable的网页存储系统 参考文献第6章 内容索引子系统设计及核心算法第7章 内容检索子系统设计及其核心算法第8章 链接结构分析子系统设计及核心算法第9章 万维网数据质量评估第10章 万维网垃圾网页识别第11章 搜索引擎广告技术第12章 中文搜索引擎的现状与未来

章节摘录

插图：对于搜索引擎系统而言，内容索引子系统的性能需求可以概括为：充分利用系统资源和高效完成索引服务。
一方面，内容索引子系统通常是搜索引擎中耗费硬件资源最多的一个子系统，前文中提到的“索引规模战争”之所以在相当一段时间内被作为搜索引擎系统性能水平的主战场，也是因为索引规模直接关系到搜索引擎系统软硬件设计水平的高低。
以索引规模作为搜索引擎系统性能的试金石尽管有些偏颇，但也不无道理；另一方面，索引服务的效率也是搜索引擎重点关注的性能指标，搜索引擎如果要在用户可以接受的时间之内返回结果，首先就需要内容索引子系统能够在尽量短的时间内把与用户查询词对应的索引项加以返回，以便内容检索子系统进行相似度计算使用。
在提高系统资源的利用率方面，在内容索引子系统设计中重点考虑如何在保存尽量多有用信息的基础上减少系统所需的磁盘存储资源。
索引建立、更新过程中重点需要进行的是磁盘写操作，而索引查询过程中重点需要进行的是磁盘读操作。
索引建立、更新的时间效率只需要与数据抓取子系统的运行效率相适应即可，由于网络带宽低于硬盘访问速度，因此这方面的时间效率要求相对较低。
由于用户查询是在线实时进行，而内容检索子系统的运算大都在内存中完成，因此索引查询的时间效率要求较高，而大规模磁盘读写也往往成为搜索引擎提供高效在线服务的主要瓶颈。
同时，由于搜索引擎需要的存储系统规模异常庞大，涉及的存储介质同样种类繁杂、数量庞大，这些介质在面临大规模读写时也难免会出现硬件问题。

编辑推荐

《搜索引擎技术基础》特色：面对浩瀚的万维网信息海洋，搜索引擎为人们提供了一条获取所需信息的捷径，而百度等中文搜索引擎在商业和技术领域的成功也使得越来越多的国内读者对搜索行业产生兴趣。
作者长期从事搜索引擎领域的相关研究与工程实践，并实际参与了多个中文搜索引擎设计与实现的全过程，《搜索引擎技术基础》从研究实践者的角度介绍了搜索引擎的相关技术及其产业，并引领读者成为搜索引擎领域的局内人。
《搜索引擎技术基础》最大的特色是将清华大学计算机系在搜索技术方面的研究成果与百度公司在搜索应用领域积累的丰富经验融合进内容中，向读者展示大规模商用搜索引擎的工作原理、核心技术与运营方式。
清华大学计算机系主任孙茂松教授与百度公司董事长兼首席执行官李彦宏先生分别为《搜索引擎技术基础》作序。
《搜索引擎技术基础》也是清华大学计算机系与百度公司合作在清华大学开设的“搜索引擎技术基础”课程的教材。

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