

<<走进搜索引擎>>

图书基本信息

书名：<<走进搜索引擎>>

13位ISBN编号：9787121131042

10位ISBN编号：7121131048

出版时间：2011-5

出版时间：电子工业出版社

作者：潘雪峰，花贵春，梁斌 编著

页数：300

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

作者序 本书第1版出版到现在已经3年了。在这段不长的时光里，搜索引擎技术有了进一步的发展。其中比较突出的是，随着数据规模进一步增大，为提升用户体验，搜索引擎性能进一步优化；在更广泛的用户参与下，增强了基于用户行为进行效果改进的能力。这也使得本书有了改版以适应这些重大变化的必要。

基于此，本书第2版增加了搜索引擎性能调优、搜索引擎日志分析，以及基于学习进行排序优化三方面的内容，希望能让读者跟上搜索技术的发展潮流，在这一领域的前沿真切地感受到它的勃勃生机。

当前，搜索技术已经不再局限于搜索引擎本身，它所建立的一套驾驭互联网级别海量数据的架构和理念正日益扩展到整个信息技术领域。

而随着世界的日益信息化、数字化、网络化，这些理念的深远影响还会进一步显现。

这又将是一次新的科技浪潮。

时光流逝，却有如轮回。

信息技术产业，甚至整个科技界，正是在这样的浪潮更迭中不断进步。

从AT&T的有线电话到IBM的大型机，到Apple的PC机，到Intel的CPU，到Motorola的无线通信，到Microsoft的操作系统，到Cisco的路由器，到Google的搜索引擎，概莫能外。

一次次浪潮，一个个产业巨擎，终将随自己的时代而去，但它们所带来的影响却将投射在人类文明的历史上，永不消逝。

至于搜索的浪潮究竟将持续多长时间，在整个IT史上留下怎样的一笔，只有时间才能告诉人们答案。

此时此刻，置身其中，让我们打开书本，接受浪潮之巅的洗礼，走进搜索引擎。

关于本书作者 作者潘雪峰，毕业于中国科学院计算技术研究所，工学博士。研究兴趣包括多媒体内容分析、机器学习和互联网数据挖掘，现从事搜索引擎领域相关工作。

作者花贵春，目前在清华大学信息科学与技术国家实验室攻读博士学位，研究兴趣包括机器学习及其在搜索领域的应用。

作者梁斌，目前在清华大学信息科学与技术国家实验室攻读博士学位，研究兴趣包括大规模数据处理、搜索引擎和软件工程等。

致谢 作者们首先要特别感谢他们的妻子，感谢她们在繁忙的工作和学习之余，包揽了家里家外大大小小的事务，还在作者们有所懈怠的时候，从精神上给予莫大的支持和鼓励。正是她们无私的支持，才使本书得以面世。

感谢电子工业出版社计算机图书出版分社孙学瑛女士，她是推动本书完成的最为关键性的人物。她参与了此书创作的全过程，为笔者提供了有关图书市场的宝贵信息，使得本书更加面向读者。

感谢本书参考文献的作者们、搜索引擎研究界的学者们，以及为此书提出宝贵技术意见的业界同行，正是你们杰出的成就和无私的帮助，才使得本书有了写作的基础和必要。

由于作者水平有限，加之搜索领域的发展日新月异，书中不足及错误之处在所难免，敬请专家和读者给予批评指正。

潘雪峰、花贵春、梁斌 2011年2月

<<走进搜索引擎>>

内容概要

本书由搜索引擎开发研究领域三位年轻的博士生精心编写，作者们希望将自己对搜索引擎的理解和实际应用相结合，让未接触过搜索引擎原理和方法的读者也能轻松读懂该书的大部分内容。

本书在第1版的基础上，删除了搜索引擎历史等章节，并对错误和不足进行了修订和补充，同时增加了潘雪峰编写的第6章“搜索引擎日志分析”，花贵春编写的第7章“排序学习（Learning to Rank）”和梁斌编写的第8章“搜索引擎的性能调优”三个主要章节，变更的内容约占第1版的一半。

读者对象：本书作为搜索引擎原理与技术的入门书籍，面向那些有志从事搜索引擎行业的青年学生、需要完整理解并优化搜索引擎的专业技术人员、搜索引擎的营销人员，以及网站的负责人等。本书是从事搜索引擎开发的工程技术人员难得的参考书，也可作为大中专院校相关专业的教学辅导书。

<<走进搜索引擎>>

书籍目录

第1章 引言1

1.1 搜索引擎概述2

1.1.1 目录式搜索引擎2

1.1.2 全文搜索引擎3

1.1.3 元搜索引擎 (Meta-Search Engine) 3

1.2 搜索引擎的主要需求3

1.2.1 快4

1.2.2 全4

1.2.3 准4

1.2.4 稳5

1.2.5 省5

1.3 搜索引擎的4大系统6

1.3.1 搜索引擎的体系结构6

第2章 搜索引擎的下载系统8

2.1 爬虫的发展历史9

2.1.1 世界上第1个爬虫9

2.1.2 爬虫的发展历程9

2.2 万维网及其网页分析9

2.2.1 蝴蝶结型的万维网10

2.2.2 万维网的直径12

2.2.3 万维网的规模及变化特征12

2.2.4 网页的特征13

2.3 有关爬虫的基本概念13

2.3.1 爬虫13

2.3.2 种子站点14

2.3.3 URL14

2.3.4 Backlinks14

2.4 网页抓取原理14

2.4.1 telnet和wget14

2.4.2 从种子站点开始逐层抓取15

2.4.3 不重复抓取?略19

2.4.4 网页抓取优先策略25

2.4.5 网页重访策略26

2.4.6 Robots协议30

2.4.7 其他应该注意的礼貌性问题31

2.4.8 重要性网页优先抓取策略32

2.4.9 抓取提速策略 (合作抓取策略) 34

2.5 网页库38

2.6 下载系统回顾及未来发展41

参考文献42

第3章 搜索引擎的分析系统44

3.1 知识准备45

3.1.1 HTML语言45

3.1.2 锚文本 (anchor text) 45

3.1.3 半结构化数据 (semi-structured data) 45

<<走进搜索引擎>>

- 3.2 信息抽取及网页信息结构化45
 - 3.2.1 网页结构化的目标46
 - 3.2.2 建立HTML标签树48
 - 3.2.3 通过投票方法得到正文52
 - 3.2.4 网页结构化过程回顾55
- 3.3 网页查重56
 - 3.3.1 网页查重技术发展历史56
 - 3.3.2 网页查重实现方法58
- 3.4 中文分词61
 - 3.4.1 什么是中文分词61
 - 3.4.2 通过字典实现分词61
 - 3.4.3 基于统计的分词方法65
- 3.5 PageRank67
 - 3.5.1 PageRank的来由68
 - 3.5.2 PageRank的基本想法68
 - 3.5.3 PageRank的计算公式69
 - 3.5.4 PageRank的计算方法73
- 3.6 分析系统结构图76
- 参考文献77
- 第4章 搜索引擎的索引系统79
 - 4.1 知识准备80
 - 4.1.1 信息80
 - 4.1.2 索引80
 - 4.1.3 倒排索引、倒排表、临时倒排文件、最终倒排文件80
 - 4.1.4 其他概念81
 - 4.2 全文检索81
 - 4.3 文档编号82
 - 4.3.1 编号的本质82
 - 4.3.2 文档编号的方法83
 - 4.3.3 游程编码84
 - 4.4 倒排索引87
 - 4.4.1 经典的倒排索引87
 - 4.4.2 正排索引(前向索引) 88
 - 4.4.3 倒排索引90
 - 4.5 数据规模的估计92
 - 4.5.1 齐普夫法则92
 - 4.5.2 布尔检索模型下的索引规模估计94
 - 4.6 涉及存储规模的一些计算97
 - 4.6.1 正排表与倒排表的合并97
 - 4.6.2 多个临时倒排文件的归并100
 - 4.6.3 倒排索引分布式存储103
 - 4.6.4 倒排文件缓存106
 - 4.6.5 倒排索引词典统计信息的计算106
 - 4.7 倒排索引文件的创建过程107
 - 4.7.1 创建倒排表107
 - 4.7.2 计算统计信息109
 - 参考文献110

<<走进搜索引擎>>

- 第5章 搜索引擎的查询系统112
 - 5.1 知识准备113
 - 5.1.1 什么是信息熵113
 - 5.1.2 检索和查询的区别115
 - 5.1.3 检索词和查询词的区别115
 - 5.1.4 自动文本摘要 (Automatic Text Summarization) 116
 - 5.2 网页信息检索116
 - 5.2.1 早期的检索模型116
 - 5.2.2 向量空间模型 (Vector Space Models) 118
 - 5.2.3 关键词权重的量化方法TF/IDF122
 - 5.2.4 搜索引擎采用的检索模型125
 - 5.2.5 多文档列表求交计算127
 - 5.2.6 检索结果排序132
 - 5.2.7 堆排序132
 - 5.3 中文自动摘要137
 - 5.3.1 自动摘要的发展历史137
 - 5.3.2 自动摘要的含义和实现137
 - 5.4 生成搜索结果页142
 - 5.4.1 生成搜索结果页142
 - 5.5 搜索结果页的缓存144
 - 5.6 推测用户查询意图145
 - 5.6.1 查询分类146
 - 5.6.2 推测信息类、事物类的查询意图147
 - 5.7 查询系统的当前热点和发展方向147
 - 5.7.1 查询系统的当前热点148
 - 5.7.2 查询系统的发展方向148
 - 参考文献149
- 第6章 搜索引擎日志分析150
 - 6.1 简介151
 - 6.1.1 人机交互的记录—?日志151
 - 6.1.2 分析搜索引擎日志的意义153
 - 6.1.3 本章的主要内容154
 - 6.2 知识准备155
 - 6.2.1 二分图模型 (Bipartite Model) 155
 - 6.2.2 图模型(graphical model)156
 - 6.2.3 LDA (Latent Dirichlet Allocation) 模型158
 - 6.2.4 随机游走 (Random Walk)159
 - 6.2.5 小结160
 - 6.3 查询日志分析161
 - 6.3.1 查询日志的内容161
 - 6.3.2 查询词频统计162
 - 6.3.3 查询串提示 (Suggestion) 163
 - 6.3.4 命名实体 (Named Entity) 类别识别165
 - 6.3.5 小结167
 - 6.4 点击日志分析167
 - 6.4.1 点击日志的内容168
 - 6.4.2 查询串提示 (Suggestion) 再分析169

<<走进搜索引擎>>

- 6.4.3 查询和结果类别属性传递170
- 6.4.4 搜索结果相似性度量171
- 6.4.5 查询结果排序172
- 6.4.6 点击数据的稀疏性174
- 6.4.7 小结176
- 6.5 隐私问题177
 - 6.5.1 日志的两面性177
 - 6.5.2 日志的安全使用179
 - 6.5.3 小结179
- 6.6 本章总结180
- 参考文献180
- 第7章 排序学习 (Learning to Rank) 183
 - 7.1 排序概述184
 - 7.2 传统的排序模型186
 - 7.2.1 查询相关的排序模型186
 - 7.2.2 查询无关的排序模型188
 - 7.3 排序学习简介以及研究现状190
 - 7.3.1 排序学习简介190
 - 7.3.2 排序学习问题的研究现状191
 - 7.4 排序学习模型的应用实例192
 - 7.5 排序学习方法的框架194
 - 7.5.1 参数设置194
 - 7.5.2 排序学习方法的框架195
 - 7.6 评测数据集196
 - 7.6.1 LETOR数据集196
 - 7.6.2 Microsoft Learning to Rank数据集197
 - 7.6.3 Yahoo Webscope数据集198
 - 7.7 排序学习模型简介198
 - 7.7.1 实例199
 - 7.7.2 Pointwise方法199
 - 7.7.3 Pairwise方法204
 - 7.7.4 Listwise方法207
 - 7.7.5 3种排序方法的对比210
 - 7.8 排序学习模型性能比较211
 - 7.8.1 评测方法211
 - 7.8.2 排序模型性能的比较215
 - 7.9 排序学习的研究方向217
 - 7.9.1 标准标注的自动构建217
 - 7.9.2 排序特征217
 - 7.9.3 半监督学习/主动学习218
 - 7.9.4 查询相关的排序模型218
 - 7.9.5 利用用户行为特征218
 - 7.10 总结219
 - 参考文献219
- 第8章 搜索引擎的性能调优223
 - 8.1 系统调优概述224
 - 8.2 瓶颈识别225

<<走进搜索引擎>>

8.3 涉及CPU的优化方法226

8.3.1 上下文切换问题 (context switching) 227

8.3.2 中断和轮询228

8.3.3 CPU的Affinity问题229

8.3.4 流水线问题229

8.4 涉及内存的优化方法235

8.4.1 概述235

8.4.2 对换区236

8.4.3 cache line240

8.4.4 false sharing问题245

8.4.5 内存的锁问题247

8.4.6 内存库的使用257

8.5 涉及磁盘的优化方法262

8.5.1 磁盘IO的调度262

8.5.2 其他常见磁盘参数调优264

8.5.3 磁盘读写方式265

8.5.4 文件缓存问题267

8.5.5 5分钟法则269

8.6 涉及网络的优化方法271

8.6.1 搜索首页，结果页提速方法271

8.6.2 Web server的架构选择274

参考文献284

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>