

自制搜索引擎



[自制搜索引擎_下载链接1](#)

著者:[日]山田浩之

出版者:人民邮电出版社

出版时间:2016-1

装帧:平装

isbn:9787115411709

《自制搜索引擎》聚焦于Google和Yahoo!等Web搜索服务幕后的搜索引擎系统，首先讲解了搜索引擎的基础知识和原理，接着以现实中的开源搜索引擎Senna/Groonga为例，使用该引擎的源代码引导读者亲自体验搜索引擎的开发过程。这部分讲解涉及了倒排索引的制作和压缩、检索的处理流程以及搜索引擎的优化等内容。又简单介绍了一些更加专业的搜索引擎的知识和要点，为读者今后进一步学习打下了基础。本书适合所有对搜索引擎感兴趣的技术人员阅读。

作者介绍:

山田浩之（作者）

信息工程学博士。先后于日本IBM、雅虎从事分布式搜索引擎的研发工作。目前在东京大学生产技术研究所从事高性能并行数据库的研发工作。

末永匡（作者）

开源搜索引擎Senna/Groonga的开发者的。每天都在奋斗，梦想着能创建一个更加自由、更加无拘无束的，而不是一家独大的应用程序平台。

胡屹（译者）

多年从事Web开发工作。热爱编程，关注设计模式，致力于提升软件开发的质量。曾获得SCJP和PHP ZCE证书。译有《计算机是怎样跑起来的》。

目录: 第1章 搜索引擎是如何工作的 1

1-1 理解搜索引擎的构成 3

什么是搜索引擎 3

构成搜索引擎的组件 4

与搜索引擎相关的组件 5

1-2 实现了快速全文搜索的索引结构 7

全文搜索的两种方法 7

倒排索引的结构 8

倒排索引的构建方法 9

倒排索引中的术语 10

1-3 深入理解倒排索引 12

倒排索引=词典+倒排文件 12

从倒排索引中查找单词 13

将单词的位置信息加入倒排文件中 13

从倒排索引中查找短语 14

1-4 制作中文文档的倒排索引 16

分割中文句子的方法 16

权衡分割方法 17

1-5 实现倒排索引 19

实现词典 19

实现倒排文件 22

1-6 使用倒排索引进行检索 24

布尔检索 24

使用倒排索引的检索处理流程 24

关联度的计算方法 26

信息检索中的检索 27

1-7 构建倒排索引 29

使用内存构建倒排索引 29

- 使用二级存储构建倒排索引 29
- 静态索引构建和动态索引构建 32
- 1-8 准备要检索的文档 34
- 收集数据 34
- 数据规范化 35
- 第2章 准备全文搜索引擎的检索样本 37
- 2-1 全文搜索引擎wiser 39
- wiser的构成 39
- 准备用于检索的文档 40
- 2-2 安装wiser 42
- 构建wiser 42
- 启动wiser 43
- 解压缩Wikipedia的副本 44
- 2-3 运行wiser 45
- 构建倒排索引 45
- 使用倒排索引查询 46
- 比较grep和wiser的运行速度 46
- 第3章 构建倒排索引 49
- 3-1 复习有关倒排索引的知识 51
- 提取词元 51
- 为每个词元创建倒排列表 53
- 3-2 构建倒排索引 54
- 在存储器上创建倒排列表 54
- 倒排列表和倒排文件的数据结构 54
- 从源代码级别梳理倒排索引的构建顺序 56
- 进一步阅读源代码 59
- 专栏 根据实际情况设计搜索引擎（系统） 68
- 第4章 开始检索吧 71
- 4-1 检索处理的大致流程 73
- 充分理解检索处理的流程 73
- 4-2 使用倒排索引进行检索 75
- 从源代码级别梳理检索处理的流程 75
- 解读split_query_to_tokens()函数的具体实现 76
- 使用具体示例加深对检索处理流程的理解 77
- 解读函数search_docs()的实现细节 80
- 解读函数search_phrase()的实现 84
- 专栏 如何实现标签检索 88
- 第5章 压缩倒排索引 89
- 5-1 压缩的基础知识 90
- 压缩倒排索引的好处 90
- 专栏 压缩的目的 90
- 倒排索引的压缩方法 91
- 倒排文件的压缩方法 91
- 压缩的原理 94
- 5-2 实现wiser中的压缩功能 97
- 压缩功能源代码的概要 97
- 了解无需进行压缩时的操作 99
- 抓住Golomb编码的要点 101
- 解读Golomb编码中的编码处理 105
- 解读Golomb编码的解码处理 108
- 第6章 挑战wiser的优化及参数的调整 113
- 6-1 提高检索处理的效率 115
- 优化检索处理 115
- 将查询分割为无重复部分的词元序列 116

- 6-2 禁用短语检索 119
- 分析对2字符的字符串进行检索时的行为 119
- 分析对3字符的字符串进行检索时的行为 120
- 6-3 改变检索结果的输出顺序 122
- 作为检索结果排序核心的指标 122
- 按照文档大小降序排列的检索结果 124
- 专栏 排名欺诈 128
- 6-4 让1个字符的查询也能检索出结果 29
- 获取以特定字符开头的词元的列表 129
- 合并检索到的结果 131
- 专栏 如何实现相似文档的检索 131
- 6-5 调整控制倒排索引更新的缓冲区容量 133
- 确认由缓冲区容量的差异带来的不同效果 133
- 用sar命令分析负载 134
- 6-6 调整只有英文字母的词元的分割方法 135
- 如何避免用英文单词检索时准确率下降的问题 135
- 如何判断某字符是否属于索引对象 135
- 修改负责分割词元的函数 136
- 6-7 确认压缩的效果 138
- 观察Golomb编码的效果 138
- 对比压缩启用前后的索引大小 138
- 专栏 避免滥用全文搜索引擎 139
- 第7章 为今后更加深入的学习做准备 141
- 7-1 wiser没能实现的功能 143
- 倒排索引之外的全文搜索索引 143
- 高效处理大规模数据的存储器 143
- 利用缓存提高检索的速度 143
- 使用各种各样的压缩方法 144
- 优化搜索结果的排名 144
- 调整准确率和召回率 145
- 降低检索结果排序处理的负载 147
- 并行处理 147
- 结合对属性的筛选过滤 148
- 分面搜索 148
- 专栏 时延和吞吐量 149
- 7-2 全文搜索引擎Groonga的特点 150
- 通过词元的部分一致检索提升召回率 150
- 使用内存映射文件 151
- 片段 152
- 专栏 宣传活动的重要性 152
- 7-3 实现出考虑到用户意图的搜索引擎 153
- 引入停用词 153
- 应对词素解析的错误 153
- 专栏 断句错误 154
- 处理全角字符和半角字符 155
- 对查询进行归一化 156
- 留意布尔检索的解析过程 156
- 通过词素解析器适当地解析查询 157
- 对错误的输入进行修正 157
- 输入补全 158
- 建议用户检索相关的关键词 159
- 7-4 收集、提取文档时的要点 160
- 制作爬虫时的处理要点 160
- 在提取文本时需要处理的要点 163

Appendix 附录 165
A-1 深度话题 166
近几年的压缩方法 166
动态索引构建 169
分布式索引 174
A-2 wiser中的文本提取和存储 178
用于处理XML的2种API——DOM和SAX 178
提取文档的标题和正文 179
掌握状态的迁移 182
构建文档数据库 187
后记 191
· · · · · (收起)

[自制搜索引擎_下载链接1](#)

标签

搜索引擎

计算机

编程

Programming

软件开发

计算机科学技术

IT

search

评论

书中大部分的内容都了解过，可能有些了解的比较浅薄，虽然书中介绍的也不多，但是

感觉根据提示可以搜索获取更多。应该是译者把很多场景本地化了，想着是一个日本作者的书，看着各种本土化的案例，总是感觉有些怪。

第一章讲概念和原理，后面是代码注释，没细读。译者把书中的例子都替换为中文的对应例子，便于理解。

很快过了一遍，需要看的时候，就不时翻翻，挺好的书，通俗易懂，IT书籍日韩的思路讲解都不错，清晰，翻译的水准也比英文版高，英文看原版比较多。

谷歌的时代，早就该结束了。

概念讲得清晰，但讲的东西并不多。

开头讲述倒排索引的部分比较清晰，后面的编程部分草草略过～

Lucene学习前了解下搜索引擎

我们还需要努力，加油，有恒！

了解搜索引擎中倒排索引的构建和查询的实现的核心

对搜索引擎做了一个非常简单的介绍，但受限与书本的页数，没有介绍到比较深入的内容。虽然书本比较小，但依旧能够吸收到不少知识。

非常好的一本书，翻译也不错，该书应算简洁经典奇作了。有评论说没有wiser安装和详细讲解，那应该是看书不仔细，看网页不仔细，基础欠缺的原因。注意wiser的源码是要在linux或者mac的机器上编译运行的。

前言里说有，结果到了出版社的网站上看，根本找不到，书里面的wiser介绍也非常简略，根本没有安装的相关信息，只是提到它会用到什么库。

学习搜索引擎非常好的入门书，直接开阔你的思维，跟着代码走一遍更是搞

感谢 胡屹 的签名赠书

[自制搜索引擎 下载链接1](#)

书评

在此之前还看过《这就是搜索引擎：核心技术详解》，以及翻阅过图书馆一两本搜索相关教科书。其他书，都是部分概念讲得还行，但是极少有真正的代码讲解。该书经典之处，一是行文思路逻辑清晰，编排合理。搜索原理、倒排索引核心与实现、检索排序、搜索的优化，以及还有哪...

本书主要讲解了：如何从零开始，编写一个基于“倒排索引”（Inverted Index）的“全文检索引擎”（FullText Search Engine）。最终的成果，相当于一个迷你版的Apache Lucene。（注意：是自制Lucene，而不是调用Lucene）
《自制搜索引擎》详细书评：[\[Blog链接\]](#) 书籍附带的迷你全...

1. 第一章“搜索引擎是如何工作的”是我感觉收获最多的一部分，把倒排说明说得比较清楚，通过一些插图能容易理解倒排索引对于搜索的核心作用； 2. 本书是通过C语言来实现一个小型的搜索引擎，代码量比较少而且都比较清楚，对于实现一个搜索引擎的理解是很有帮助的； 3. 另外还...

[自制搜索引擎_下载链接1](#)