

<<Lucene分析与应用>>

图书基本信息

书名：<<Lucene分析与应用>>

13位ISBN编号：9787111249924

10位ISBN编号：7111249925

出版时间：2008-9

出版时间：机械工业出版社

作者：吴众欣，沈家立 编著

页数：279

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>

## <<Lucene分析与应用>>

### 前言

Google被人熟知，Baidu在中国成功推广，“搜索”吸引着IT界的眼球，也吸引了更多开发者的好奇心。

于是诞生了Lucene，一个开源的全文检索API（Application Program Interface，应用程序界面）。并在LHcene的基础上，衍生出了一个全文检索引擎(Nutch)和分布式文件系统(Hadoop)。

大家一定很好奇，Google的搜索引擎是如何工作的？

采用什么样的文件系统？

提供什么样的服务？

.....我们无法得知。

Lucene与其相关的项目Nutch和Hadoop弥补了这个不足，让我们有机会了解到搜索引擎、分布式文件系统的内部工作原理。

如果介绍一个软件或者一套框架如何使用是比较容易的，但是要从源代码剖析内核，却不容易。

老吴与家立在写作期间，辗转难眠，思索如何表述才能够准确地把Lucene的设计精髓展现给读者。

最终确定通过对Lucene源代码的解说、辅以图表，并通过一些具体实例把所有源代码进行组织与剖析，完整地展示Lucene从建立索引到查询的完整过程。

并通过介绍一些Lucene的应用，和读者分享Lucene在具体项目开发中的应用环境。

同时，插入一些Lucene开发实例，抛砖引玉，试图让读者也能亲身体会Lucene本身的强大功能。

最后，为了进一步说明Lucene的应用环境，本书简单地介绍了Nutch和Hadoop。

老吴很早就开始研读Lucene的源代码，并阅读了Dong Cutting的相关论文，对Lucene的内核具有深刻的认识。

我们很想与大家分享自己的学习体会和研究成果，于是决定把它写出来，家立负责Lucene多处应用部分的写作。

Lucene是一个很活跃的开源项目，因为老吴研究得比较早，版本以1.4.3为主。

为了能够跟上Lucene的步伐，家立推荐采用了较新的1.9-2.1版本进行分析。

但是该版本的内核变化比较大，因此需要重新分析、调试、总结。

为了尽快完成，我们日日熬夜，真所谓痛并快乐着。

在此非常感谢家人的支持，朋友的鼓励。

在此，向我的爱妻张信健对我的一贯支持表示感谢！

谢谢你，我的爱人！

希望对搜索引擎内核与运行机制感兴趣的朋友阅读此书，由于时间仓促，难免有所疏漏，请读者批评指正。

吴众欣

## <<Lucene分析与应用>>

### 内容概要

本书对Lucene搜索引擎的源代码进行分析讲解，并用一些具体实例把所有源代码进行组织与剖析，完整地展示Lucene从建立索引到查询的过程。

本书通过介绍Lucene的应用，分析Lucene具体项目开发的应用环境。

最后简单地介绍了Nutch和Hadoop。

本书适用于开发搜索引擎的技术人员、Lucene爱好者等读者。

## <<Lucene分析与应用>>

### 作者简介

吴众欣，西安交通大学在读博士，主攻搜索引擎与服务组合。  
喜欢研读，头脑虽慢，滴水石穿。  
好奇心重，兴趣广泛。

## 书籍目录

前言第1章 搜索引擎与Lucene 1.1 搜索引擎与Lucene简介 1.1.1 搜索引擎分类 1.1.2 Lucene项目简介 1.1.3 其他搜索引擎开发包介绍 1.2 Lucene的系统架构 1.2.1 Lucene最简示例 1.2.2 Lucene采用的索引结构 1.2.3 Lucene软件包架构 1.3 本书的章节导航第2章 文档逻辑视图与文本分析 2.1 文档逻辑视图 2.2 Lucene的文本分析过程简介 2.3 空格解析器 (WhitespaceAnalyzer) 2.3.1 空格分词器 (WhitespaceTokenizer) 2.3.2 Token (标志) 2.4 标准解析器 (StandardAnalyzer) 2.4.1 标准分词器 (StandardTokenizer) 2.4.2 标准过滤器 2.5 打造自己的解析器 2.5.1 常用的中文分词法 2.5.2 对CJKAnalyzer的分析 2.5.3 构造自己的解析器第3章 Lucene创建索引之一 (段索引方式与倒排索引结构) 3.1 倒排结构与段索引方式 3.2 索引写入过程概述第4章 Lucene创建索引之二 (在内存中创建索引) 4.1 创建Document层面索引 4.2 写入field信息 4.3 文件倒排过程 4.4 填写postingTable 4.5 postingTable的排序过程 4.6 写入field名字文件 (.fnm文件) 4.7 写入field信息文件 (.fdt, .fdx文件) 4.8 写入频率与位置文件 (.frq与.prx文件) 4.9 TermVector方式写入索引 (.tvf, .tvd与.tvx文件) 4.10 字典文件 (.tis与.tii文件) 4.11 写入规格化文件第5章 Lucene创建索引之三 (索引合并过程) 5.1 document层面的合并过程 5.2 field与term的合并过程 5.2.1 field信息合并过程 5.2.2 term信息合并过程 5.2.3 合并norm信息 5.3 Lucene索引采用的压缩算法 5.3.1 front coding (端部编码) 5.3.2 variable-byte coding (变长字节编码) 5.3.3 delta-coding或delta-encoding 5.4 小结第6章 Lucene查询过程之一 (查询模型与引擎预热) 6.1 查询模型 6.1.1 向量模型 6.1.2 布尔模型 6.1.3 Lucene的评分 (score) 方式 6.2 查询简单示例 6.3 引擎预热 6.3.1 获得并打开索引文件 6.3.2 获得segment信息 6.3.3 FSDirectory打开索引过程 6.3.4 获得field信息 6.3.5 获得term信息第7章 Lucene查询过程之二 (查询解析与语法) 7.1 构建查询解析器 (QueryParser) 7.2 Lucene的查询语法 7.2.1 项 (Term) 查询 7.2.2 域 (Field) 7.2.3 词条查询 (Term Modifiers) 7.2.4 布尔操作符 (Boolean Operator) 7.2.5 组合查询 (Grouping) 7.2.6 针对field的组合查询 (Field Grouping Field) 7.2.7 Escaping Special Character (转义字符) 7.3 Lucene查询语法树的构建过程 7.3.1 过程分析 7.3.2 语法树分析实例第8章 Lucene查询过程之三 (相似度匹配与算法分析) 8.1 查询与相似度计算 8.1.1 查询器 (Searcher) 的查询过程 8.1.2 查询语句的权重计算 8.1.3 获得topK个document 8.2 Lucene查询算法分析 8.2.1 相似度计算简单实例 8.2.2 线性相似度计算 8.2.3 基于倒排索引的相似度计算 8.2.4 Lucene的相似度计算第9章 Lucene索引与查询全程示例 9.1 实例描述 9.2 建立索引过程 9.2.1 选择文档中建立索引的field 9.2.2 选择field录入方式 9.2.3 生成segment文件 9.2.4 生成fields文件 9.2.5 posting文件 9.2.6 合并segment index生成index文件 9.2.7 合并后的文件关系 9.3 查询过程第10章 Lucene的常用应用场景分析 10.1 对大型XML文档集合的检索 10.1.1 都柏林文件介绍 10.1.2 XML分析器介绍 10.1.3 Lucene在大型XML文件中的应用 10.2 MuhiSearcher的应用 10.2.1 MultiSearcher的应用 10.2.2 ParallelMuhiSearcher的应用第11章 利用Lucene构建分布式搜索引擎 11.1 分布式文件系统和Hadoop 11.1.1 Hadoop文件系统体系结构 11.1.2 系统交互过程: 单一NameNode方式 11.1.3 系统组件描述 11.2 Nutch简单剖析 11.3 体验Nutch附录A TestIndexWriterMerging附录B TestDocumentWriter与DocHelper

章节摘录

1.1搜索引擎与Lucene简介从最初的图书检索到链接查询，对图片、多媒体的搜索，直至现在的人肉搜索，搜索引擎作为信息融合平台将万千世界带到你的周围，让你触手可得，悄悄改变着你的生活，同时也可能将你暴露于众目睽睽之下。

有心人可能会考虑它背后的机理，以体味搜索引擎给我们的生活带来的变化。

现今的商业搜索引擎还是Google一家独大，微软也提供了MSN搜索引擎，但技术与经验积累还不够。

而百度则是中文搜索中的佼佼者。

在商业搜索引擎中，核心技术与外部世界之间隔着一扇沉重的大门，幸好开源社区常常会将这扇门撬开些许缝隙，让我们能窥得冰山一角。

Lucene与MG4J正是开源搜索引擎项目，虽然“代码之前，了无秘密”，但是冰山一角也非轻易窥得。本章我们也是浮光掠影地谈一谈搜索引擎，力图能给大家一些新的信息，以便从多个角度来认识搜索引擎与Lucene。

<<Lucene分析与应用>>

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>