

Lucene



开发权威经典

于天恩 著

- 内容精练实用，书中所列代码可以搭建一座中型搜索平台
- 完整实例再现，让读者紧跟作者构建思路
- **Lucene** 搜索引擎技术完美演绎

Lucene 搜索引擎开发权威经典

于天恩 著

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

南京金陵大学图书馆藏 李正超 黄野 吴佩孚 牛国范藏其书

内 容 简 介

这本书基于 Lucene 的当前最新版本 (2.1) 精解了 Lucene 搜索引擎的相关知识, 从基础知识到应用开发, 精练简洁, 恰到好处。

本书共包括 16 章, 分为 6 部分。第 1 部分: Lucene 基础。介绍了 Lucene 的基础知识, 包括 Lucene 的历史和发展情况、使用 Lucene 创建索引和执行搜索的基本方法以及中文分词的应用, 最后做了两个应用项目。第 2 部分: 数据解析。介绍解析不同格式数据 (如 Word、PDF 等) 的方法, 包括常用的数据解析组件、Lucene 自身的数据解析机制和 Lius 类库。第 3 部分: 索引的高级知识。介绍了 Lucene 建立索引的过程, 索引的查看和删除, 索引的同步, 索引的合并和优化等内容。第 4 部分: 搜索的高级知识。介绍使用不同的 Query 对象构建搜索请求, 使用 QueryParser 解析用户的搜索请求, 搜索结果的过滤和排序等内容。第 5 部分: Lucene 应用实例。应用本书介绍的所有知识构建一个桌面搜索引擎和一个 Web 搜索引擎。这部分作为对前面所有内容的总结。第 6 部分: Nutch 搜索引擎框架。介绍了基于 Lucene 的完整搜索引擎 Nutch。

本书适合从 Lucene 初学者到高级开发人员之间的一切 Lucene 爱好者。

图书在版编目 (CIP) 数据

Lucene 搜索引擎开发权威经典/于天恩著. —北京: 中国铁道出版社, 2008. 10

ISBN 978-7-113-09251-1

I. L… II. 于… III. 计算机网络—程序设计 IV. TP393. 09

中国版本图书馆 CIP 数据核字 (2008) 第 158174 号

书 名: Lucene 搜索引擎开发权威经典

作 者: 于天恩 著

策划编辑: 严晓舟 荆 波

责任编辑: 苏 茜

编辑部电话: (010) 63583215

编辑助理: 包 宁

封面设计: 付 巍

责任印制: 李 佳

出版发行: 中国铁道出版社 (北京市宣武区右安门西街 8 号 邮政编码: 100054)

印 刷: 北京新魏印刷厂

版 次: 2008 年 10 月第 1 版 2008 年 10 月第 1 次印刷

开 本: 787mm×1092mm 1/16 印张: 40.75 字数: 939 千

印 数: 4 000 册

书 号: ISBN 978-7-113-09251-1/TP·2995

定 价: 88.00 元 (附赠光盘)

版权所有 侵权必究

凡购买铁道版的图书, 如有缺页、倒页、脱页者, 请与本社计算机图书批销部调换。

前言

说说搜索引擎

搜索引擎这几年热起来了。

作为世界上最大、最出名的搜索引擎，Google 在很多地方发挥了重要的作用，下面是笔者的一点总结。

第一，搜索。Google 给人上网浏览资源提供了方便。

第二，营销。利用 Google 做广告，针对 Google 的搜索算法去开发网站，就可以赚取眼球经济。

第三，小软件的开发。利用 Google 提供的搜索接口，可以开发许多商用的客户端搜索软件。如 Google Earth 地图软件等。

第四，Google 为长尾理论提供了实验数据，奠定了互联网经济的理论基础，引发了“管理学界”的变革。

第五，看到了 Google 搜索的好处之后，人们更重视搜索。找人做网站的人总是会提出加个搜索功能。提供搜索服务的公司也越来越多，因为他们发现这里有利可图。甚至，连苦读圣贤书多年的大学毕业生（包括一些博士）也认为应聘搞搜索的公司是一件光彩的事情。

是的，搜索很有用。但是，当你手中没有 Google 的搜索代码时，该如何搭建一个自己的搜索引擎呢？业界人士说，全新开发一套完备的企业级搜索引擎要花五年的时间。诚然，许多“业界”人士的话并不可信，他们说了一大堆很玄的话无非是为了骗外行的钱。不过，真要想做好一个搜索引擎，即使对于高手来说，没有半载也确实是做不到的。

开发搜索引擎耗费大量的时间和精力，所以有一些人开始研发独立的搜索引擎模块，并将其源代码开放，这样就可以给其他需要建立自己的搜索引擎的人提供一个基础平台。在这些开源搜索引擎模块的基础上做开发，可以节约非常多的时间和精力，大大减少了开发成本，缩短了产品投入市场的周期。而且，由于这些平台是开源的，你可以亲自检查每一行代码，修改算法和显示格式等内容，这样的搜索引擎就相当于自己写的，用起来放心。如果使用某些商业搜索模块，尽管搜索效果很好，但是，你很难知道在你点击“搜索”按钮的瞬间你的程序是否做了一些你并不想做的事情，例如，给某个陌生人发送了一个特洛伊木马。

用人家的东西，往往要受制于人家。想用得舒心，就自己去做。

写作动机

开源搜索引擎，对解决企业搜索等问题提供了可靠的二次开发平台（有的甚至不需要二次开发），大大提高了开发搜索引擎的效率，缩减了成本，好处多多。所以，需要有一些书来讲一讲如何使用开源搜索模块来提供搜索服务。目前，市面上讲这方面内容的

书有《Lucene in Action》，这书是不错的。在 2006 年的 12 月，我和出版社联系，打算来翻译这本书，可惜被别人占了先机（由此又见搜索引擎之火热）。

为了全面、系统地讲解搜索引擎技术，我决定写六大类书。

第一类：讲搜索引擎的最基础知识。包括其原理、相关术语、历史、发展和基本应用等基础知识。这是最低层次的书，是入门的书。

第二类：讲基于数据库的全文搜索机制来建立搜索引擎的基本方法。用数据库的全文搜索机制来建立搜索引擎是主流的两种做法之一。

第三类：讲基于开源搜索引擎框架来建立搜索引擎的基本方法。基于开源搜索引擎框架来建立搜索引擎是主流的两种做法之一。

这三类书，都是基础书籍，是进入搜索引擎行业的必要辅助。

第一类书，内容很少，写出来会很薄，所以我写了《做自己的搜索引擎——搜索引擎精解案例教程》，将第一类和第二类的内容合并到一起。

对于第三类书，我写了《迅速搭建全文搜索平台——开源搜索引擎实战教程》。

这样一来，搜索引擎方面所需的基础书籍就齐全了。

第四类：深入讲解某种数据库的全文检索机制或深入讲解某种开源搜索框架的实现的书籍。这样的书应该命名为：《Oracle 全文检索机制深入解析》或《Hyper Estraier 详解》等。

第五类：着眼应用，讲解基于某种开源框架或某种数据库的全文检索机制来开发一个真实的大型搜索引擎的全过程。这样的书应该命名为：《利用 Nutch 建立企业级搜索引擎》或《利用 Oracle 建立企业级搜索引擎》。这样的书要详细讲解，步骤清晰，内容可靠，目的是使读者可以直接利用里面的代码去迅速架构企业级搜索引擎，同时能够根据自己的需要对搜索引擎进行各方面的调整。这样的书还应该介绍搜索引擎服务器的软硬件环境的设置。

第六类：又回到理论。这里要讲的理论就是搜索引擎的底层数学算法，包括：分词算法、索引算法、蜘蛛抓取算法以及必要的人工智能知识，如模式识别。这类书，是最好写的，但是最不好卖的。

有了这六类书，搜索引擎的教材体系就完备了。

读者手中这本《Lucene 搜索引擎开发权威经典》是第四类书的第一本，也是我写的第三本讲解搜索引擎技术的图书。

这本书用来精讲 Lucene 搜索引擎技术，写这本书要在学术和实用开发之间做出权衡，我的出发点明确：一切以实用为主，适当兼顾学术。这本书没办法对 Lucene 进行面面俱到的讲解，不是我做不到，是不能那么做。我讨厌写“大全”。我宁愿只写个五六百页，写得重点突出，写得精彩实用。

眼下，关键问题是把技术以尽可能好懂的方式介绍出来，从而有利于读者学习，有利于解决工程项目中的实际问题。

关于前辈篇

这本书属于第四类书，是第四类书的第一本。他的前辈篇是：《做自己的搜索引擎——搜索引擎精解案例教程》（简称《案例教程》）和《迅速搭建全文搜索平台——开源搜索引擎实战教程》（简称《实战教程》）。

《案例教程》介绍了搜索引擎的历史及当前的发展状况，和搜索引擎相关的公司、技术以及搜索引擎对人类生活的影响。讲解了搜索引擎的原理和相关技术。介绍了搜索引擎的基本构造方法，实现索引建立和搜索的基本算法。在基于数据库的全文检索方面，介绍了通常采用数据库的全文索引服务时搜索引擎的基本搭建方式。对于 Windows 索引服务，专门稍微细致地讲了一些。它适合于作为学习搜索引擎的基础和入门教程，里面的数据库全文搜索部分对于希望利用数据库实现搜索引擎的人大有益处。

《实战教程》在《案例教程》的基础上，首先总结了建立搜索引擎的主要方案，即常规的数据库搜索、文件搜索，基于数据库全文索引机制的搜索，利用外部非开源 Web 搜索服务进行的搜索，以及利用开源搜索引擎实现的搜索。之后，陆续介绍了数据抓取、数据解析、建立索引和执行搜索这四项内容。这四项，是创建网络搜索平台所要解决的基本问题。最后，介绍了基于 Hyper Estraier 搜索引擎框架来搭建桌面搜索引擎和 Web 搜索引擎的方法，给出了相关的案例。这本书适合于作为学习开源搜索引擎的基础和入门教程。讲解开源搜索引擎的原理和实现方案，是以 Hyper Estraier 作为例子的。

现在，这本书在其前辈篇的基础上落笔，所以，对搜索引擎的相关原理和技术规范都略去不提，如果读者在这些方面（如分词、索引、排序等）遇到问题，请参考前辈篇。

正是由于本书落笔于其前辈篇，所以，我才可以把这本书写得精辟！早就想写一本 Lucene 了，就是怕写成空中楼阁，所以才决定写成一个系列，这样，先打基础，再一层一层地往上盖，给读者的感觉就不唐突了。写这本书不是为了抒情，所以，读者能学会才是目的，读者说好才是真的好。

写作过程

本书的写作过程是很轻松的，因为目的明确，不需要考虑细枝末节，只要理清主干就行，所以动笔之后一气呵成，大约用了 30 个小时就完成了初稿，那是 2007 年六一儿童节的时候，我那几天的心情真是不错，很爽朗。

书稿刚刚完成的时候是不能交给出版社的，我得从写书的热情中拔出来，热情容易使人出成绩，也容易使人犯错误。

我得换换脑子，试着用外行的眼光去看这本书，试着将自己变成一个完全不懂 Lucene 的人，以这种心态去重新审视这本书，这本身就是与读者交流的一次模拟。

我也找了许多人给这本书挑毛病，实事求是地讲，没挑出什么大毛病。作为“权威”，毕竟与“大全”不同，前者着重于精度，后者着重于广度。得到了许多人的认可，我很高兴。

这本书没有在 2007 年出版，原因比较复杂，涉及许多问题，但究其根本是我的健康问题。有时候，只为了几天的休息，我不得不将一件事情向后推迟一年甚至两年。

能在 2008 奥运年出版这本书，我很高兴。正应了那句俗语：来得早，不如来得巧。

本书特点

写书，应以教育为目的。我的书内容以实践为主，并不深挖理论。教育是目的，写书是方法，方法服从于目的。书，不见得要写得深透。书的好坏要由书面向的读者群来评价。我写书的原则是：实用。

这本书包含了必要的理论，但以实践为主。所讲的理论都不是纸上谈兵，都是可以立即付诸实践进行工程应用的。代码可以直接拿出来用（只是不要忘了输入信息验证等基本的安全检查）。

书中含有大量的案例，由浅入深。浅，并不从“什么是程序设计”开始，所以读者需要具备一些编程的基础知识才能看懂我的书。深，并没有深到“只可意会，无法言传”的地步，所以读者不需要担心无法看懂我的书。

当然，我在书中的确阐述了一些思想，这是因为：我以为教育不仅是要传知识，更重要的是要传思想。李阳的《疯狂英语》，胡敏的《新魅力英语》都是在给人以信念、生活的动力，实为著书的典范。尽管我的书使得一些读者可以迅速找到工作，迅速赚钱，发家致富，但这些是小的方面。大的方面是思想。知识有价，思想无价。一本书，如果能够启迪读者的思维，使读者找到自己人生的方向，那便是功德无量了。希望读者可以体会到。

写一本好书，不知能改善多少人的生活，给多少人带来生活的希望和动力。以下几点可以成为你看这本书的理由：

第一，搜索引擎很热，构建搜索引擎很有用，有前途，有“钱”途。这本书可以让你轻松掌握搭建搜索引擎平台所需的核心知识并且轻松搭建起自己的搜索引擎平台。

第二，这本书讲解详细，你想要的代码全部提供。即使不修改一行代码，你也可以建立起一个中型的搜索平台了。这点非常符合“懒人”的需求。

第三，网上有些文档，多半是英文的，还有其他国家语言的。许多人外文不太好，看起来困难。我可以做那些文档的翻译工作，如果有时间的话。这本书，至少比直接翻译外国文档要强一些，而且是中国人写的，看起来会很轻松。

第四，即便网上所有的文档都是中文写的，我依然认为买一本书来看比在网上浏览要好得多。人要爱护眼睛，软件工程师尤其是这样。通常我都会把网上的资料打印下来看，我不知道你是否也有这样的习惯。如果你也这么做，你就会发现，买这本书和打印网上文档相比，可以让你节约很多钱。而且，我写的书，比网上的 API 文档还是强一些的。

谅解和支持

这本书是我用心写成的。

写书的人用心写书，未必能写到最好，但往往可以比自己不用心时写得好。一个本

来就可以把书写得很好的人用心去写书就可以把书写得更好，甚至写成经典。看书的人也是一样，用心学，好书更有用。

我喜欢读书，读过很多书，好书对我来说是享受，它能减少我的思考时间，哪怕只有几分钟，也是相当值得的。做人做事得将心比心，我写的每一本书都是用心的，用心，尽力做好。这样就可以节约你们的时间。

我尽心尽力地写了这本书，从章节的安排到案例的编写，都是经过仔细揣摩的。我希望做到最好。然而，没有最好，只有更好。

我不能确定这里面没有错误，但我很希望多次的检查可以使里面的错误减到最少。即使有错，也应该是文字错误或者排版错误，不会有涉及程序设计或算法的原则性错误。

我希望你说这是一本好书，但我更希望你告诉我这本书的缺点和不足。因为我知道这里一定有不足。我不能因为自己用了心，就恬然地认为自己做到了完美。问心有没有愧是一回事，做得好不好是另一回事。前者，我说了算；后者，你们说了算。

关于读者在这本书中发现的任何问题，错字也好，排版格式也好，我都很希望你能告诉我，以帮助我改良它。而且，你有可能收到意想不到的神秘礼物，或许是我珍藏十几年的树叶书签，或许是我亲手叠的纸飞机，或许是我写的诗词……总之，有钱买不到！

我的邮箱（yutianen@163.com）将真诚地用来为读者答疑解惑，同时接受一切读者的反馈意见。需要说明的是，我的时间非常少，如果不能及时查收或回复你的邮件，希望谅解。我若看到了你的邮件，就一定会认真回复的。

希望这本书在你开发搜索引擎的路上有帮助，那便是我的欣慰。

一些勉励

艺多不压身。

学技术很有用，可以让人心里踏实。只要技术活着，我们就可以靠技术活着。技术死了，我们再去学新技术，从而维持自己活着。这是一种与时俱进。

只要肯钻研，没有学不通的技术。

人类若不灭亡，而且人性保持不变的话，技术就无穷无尽。只要有兴趣，有动力，就可以不停地学，锲而不舍，最终把所有的技术都学会。

生而知之者，上也；学而知之者，次也；困而学之者，又其次也。

“人事有可陷者，亦有不可陷者”，陷于技术，不值得。

悟性，因人而异，不可强求。欲望，因人而异，不应执着。——于天恩

学技术是简单的事情，做工程师是简单的事情，做科学家也是简单的事情，这些事情都比当官轻松得多。

但简单的事情未必好办，因为简单是相对的。不过有一点是肯定的，“世上无难事，只怕有心人”。即便是我也不认为自己可以做成每件事情，但我会对我想做的事情用心，尽力，做到问心无愧。

诚然，人的天资很重要，非常重要，相当重要。但有几句话要提醒读者：“勤能补拙

是良训，一分辛劳一分才”，“只要有恒心，铁杵磨成针”。天资是天给的，不是人定的。但在天资之外，你还有很多事情可以做，通过勤奋、坚持、刻苦，纵然超不过天才，你也定然能成为超越常人的人才，这不是也很好吗？

幸福在哪里？在精心的耕耘中，在艰苦的劳动里。
努力吧！用晶莹的汗水去酿造成功，用辛勤的双手去编织绚丽的人生！

衷心感谢

石志国，他的《ASP 精解案例教程》是我编程的开端，提起编程，总会想起。想起2004年的初冬在寒风凛冽的雨夜读书的事情，想起那时陪伴我的歌，不禁眼泪潸然。

对哈工大天萌联合的一切成员表示感谢！那些曾跟我在一起的朋友，我会记得你为我泡的每一杯咖啡和茶。那些始终保持独立的朋友，我也祝愿你们会有更加辉煌的未来。

于天恩

2008年8月于天人居

第 1 部分 Lucene 基础 (第 1~5 章)

作为本书的第一部分,笔者在这里使用五章的篇幅介绍了 Lucene 的基础知识。包括 Lucene 的历史和发展情况、使用 Lucene 创建索引和执行搜索的基本方法以及中文分词的应用,最后做了两个应用项目来说明使用 Lucene 建立搜索引擎的详细过程。

第 1 章 Lucene 简介 3

本章简要介绍 Lucene 的发明者、发展史和当前状况。介绍了下载 Lucene 的方法和一个体验 Lucene 的简单案例。

1.1 认识 Lucene 3

1.1.1 Lucene 是什么 3

1.1.2 Lucene 的作者 3

1.1.3 Lucene 的历史 3

1.1.4 Lucene 的现在 4

1.2 Lucene 体验实例 4

1.2.1 下载 Lucene 4

1.2.2 编写辅助类 6

1.2.3 最简单的搜索引擎 10

1.2.4 索引器的开发 11

1.2.5 索引器的运行 18

1.2.6 搜索器的开发 20

1.2.7 搜索器的运行 27

小结 28

第 2 章 创建索引 29

本章介绍使用 Lucene 建立搜索引擎的第一步:创建索引。对 IndexWriter、Document、Field 三个类进行了主要介绍。

2.1 创建索引的基本方式 29

2.1.1 理解创建索引的过程 29

2.1.2 创建 Field 31

2.1.3 创建 Document 31

2.1.4 创建 IndexWriter 31

2.2 创建索引实例	32
2.2.1 简单索引	32
2.2.2 复杂一点的索引	35
2.2.3 为文件创建索引	39
2.2.4 为某一目录下的所有文件创建索引	42
小结	46
第3章 执行搜索	47
本章介绍在已有索引的基础上使用 Lucene 执行搜索的方法。主要讲解 IndexSearcher、Term、Query 和 Hits 四个类的用法。	
3.1 执行搜索的基本方式	47
3.1.1 创建搜索器对象: IndexSearcher	47
3.1.2 封装搜索条件: 使用 Term 和 Query 对象	47
3.1.3 执行搜索	48
3.1.4 提取搜索结果: 了解 Hits 对象	48
3.1.5 提取搜索结果: 了解 Document 对象	49
3.1.6 提取搜索结果: 了解 Field 对象	49
3.2 执行搜索实例	49
3.2.1 简单搜索	49
3.2.2 分词问题	52
小结	53
第4章 中文分词	55
为了使 Lucene 支持中文搜索, 中文分词器被开发了出来。本章介绍主流的中文分词器在 Lucene 中的使用。	
4.1 说说分词	55
4.1.1 分词的方法	55
4.1.2 做分词器	56
4.2 Lucene 的分词器	56
4.2.1 二分法分词器	56
4.2.2 Lucene 自带的中文分词器	62
4.2.3 NGram 分词器的原理和用法	68
4.2.4 JE 分词器的原理和用法	69
4.2.5 IK 分词器的原理和用法	75
4.2.6 其他分词器	87
小结	87

第 5 章 Lucene 搜索引擎开发基础实例89

本章综合前面各章的内容, 利用 Lucene 的索引和搜索机制构建两个基本的应用程序。
一个桌面文件搜索引擎和一个基于 Web 的搜索引擎。

5.1 概述.....	89
5.2 桌面搜索引擎: 索引器.....	89
5.2.1 索引器简介.....	89
5.2.2 索引器设计.....	89
5.2.3 索引器实现.....	91
5.2.4 实例效果.....	99
5.3 桌面搜索引擎: 搜索器.....	101
5.3.1 搜索器简介.....	101
5.3.2 搜索器设计.....	101
5.3.3 搜索器实现.....	101
5.3.4 实例效果.....	109
5.4 Web 搜索引擎: 搜索器.....	111
5.4.1 搜索器简介.....	111
5.4.2 搜索器界面.....	111
5.4.3 搜索器实现.....	112
5.4.4 搜索器部署.....	115
5.4.5 实例效果.....	117
小结.....	118

第 2 部分 数据解析 (第 6~8 章)

从这部分起, 笔者将逐步介绍 Lucene 搜索引擎的高级知识, 每学完一部分读者实力增强将立竿见影, 在数据分析这部分笔者将介绍解析不同格式数据的方法, 这样我们就可以从不同格式的文件中提取出纯文本, 从而可用 Lucene 为之建立索引。

第 6 章 常用的数据解析组件121

搜索引擎需要从二进制文件中提取出纯文本, 就需要一些解析器。本章介绍解析常见的二进制文件的方法, 同时对格式复杂的 HTML 和 XML 文件的解析方法也进行讲解。

6.1 常见二进制文档的解析.....	121
6.1.1 解析 PDF 文档.....	121
6.1.2 解析 Word 文档.....	131
6.1.3 解析 Excel 文档.....	137
6.2 XML 文档的解析.....	145
6.2.1 使用 DOM4J 解析 XML 文档.....	146
6.2.2 XML 文档标准解析器.....	149

6.3	HTML 文档的解析.....	151
6.3.1	下载 HTMLParser 组件.....	151
6.3.2	网页编码问题.....	153
6.3.3	网页解析的一般方法.....	154
6.3.4	常见的解析要求.....	162
6.4	集成的数据解析器.....	172
6.4.1	文本文件解析器.....	172
6.4.2	集成解析器.....	174
6.5	改良的搜索引擎.....	177
6.5.1	改良的索引器.....	177
6.5.2	测试搜索器.....	186
	小结.....	193
第 7 章	Lucene 自身的数据解析方法.....	195
Lucene 本身也提供了一些数据解析方法, 比较简单, 本章概要介绍了这些方法。		
7.1	处理 PDF 文档——LucenePDFDocument.....	195
7.2	处理纯文本文件——TextDocument.....	195
7.3	综述.....	199
	小结.....	199
第 8 章	Lius 类库.....	201
Lius 类库是一套为 Lucene 开发的辅助类库, 它可以辅助建立 Lucene 索引并且对索引进行管理, 本章介绍了使用 Lius 类库解析数据的相关方法。		
8.1	初识 Lius.....	201
8.1.1	Lius 简介.....	201
8.1.2	Lius 的功能.....	202
8.1.3	下载 Lius.....	202
8.1.4	Lius 的目录结构.....	204
8.1.5	一个简单的例子.....	208
8.2	借助 Lius 解析普通数据.....	209
8.2.1	取我所需, 为我所用.....	209
8.2.2	解析 Word.....	210
8.2.3	解析 Excel.....	211
8.2.4	解析 PDF.....	213
8.2.5	解析 PowerPoint.....	215
8.2.6	解析 RTF.....	217
8.2.7	解析 TXT.....	218
8.2.8	解析 XML.....	220
8.3	借助 Lius 解析 HTML 数据.....	221

8.3.1 Lius 解析 HTML 的方式	221
8.3.2 JTinyHtmlIndexer 的使用	222
8.3.3 NekoHtmlIndexer 的使用	224
小结	225

第 3 部分 索引的高级知识 (第 9~10 章)

在这一部分, 笔者将介绍和 Lucene 索引相关的一些稍微深入一点的知识, 这部分知识将有助于读者较为完善地建立和管理索引, 读者将在这一部分学到: Lucene 建立索引的过程、索引的查看和删除、索引的同步、索引的合并和优化等内容。

第 9 章 索引的建立和优化

本章介绍 Lucene 建立索引的相关知识, 包括: Lucene 的索引机制、建立索引的详细过程、索引优化的主要方法。

9.1 索引建立的过程	229
9.1.1 Lucene 的索引机制	229
9.1.2 文本的分析	230
9.2 索引文件的生成	244
9.2.1 生成索引文件的过程	244
9.2.2 索引文件的格式	247
9.3 索引的优化	248
9.3.1 索引优化的本质	248
9.3.2 复合式索引格式	248
9.3.3 调整索引优化参数	259
9.3.4 内存缓冲器与索引合并	264
9.3.5 限制每个 Field 的词条数量	274
9.3.6 索引本身的优化	276
9.3.7 查看索引的过程	279
小结	285

第 10 章 索引的管理

本章介绍了管理 Lucene 索引的一些方法, 包括查看索引、更新索引和删除索引。同时介绍了同步索引的相关知识以及使用 Luke 和 Limo 软件来管理索引的方法。

10.1 查看索引的信息	287
10.1.1 使用 IndexWriter 类读取索引的相关信息	287
10.1.2 使用 IndexReader 及其子类读取索引的相关信息	289
10.2 删除索引中的文档	296
10.2.1 使用 IndexReader 从索引中删除文档	296
10.2.2 使用 IndexModifier 从索引中删除文档	304

10.3	更新索引中的文档	307
10.3.1	更新索引中的单个文档	307
10.3.2	批量更新索引中的文档	310
10.4	索引的同步	314
10.4.1	Lucene 的并发访问规则	314
10.4.2	线程安全性	314
10.4.3	索引锁机制	315
10.5	使用 Luke 管理索引	315
10.5.1	下载 Luke	315
10.5.2	使用 Luke	317
10.6	使用 Limo 管理索引	321
10.6.1	下载 Limo	322
10.6.2	使用 Limo	324
小结		324

第 4 部分 搜索的高级知识（第 11~12 章）

本部分的知识将有助于读者较为灵活地处理用户的搜索请求以及显示搜索结果，读者将在这部分学到：使用不同的 Query 对象构建搜索请求、使用 Query Parser 解析用户的搜索请求、搜索结果的过滤和排序等内容。

第 11 章 搜索请求的构建和解析 327

本章介绍复杂搜索请求的构建方式，主要包括组合关键词搜索、关键词的解析、多索引搜索和多线程搜索。

11.1	构建多种搜索请求	327
11.1.1	词条搜索	327
11.1.2	组合搜索	330
11.1.3	范围搜索	344
11.1.4	前缀搜索	350
11.1.5	短语搜索	353
11.1.6	多短语搜索	360
11.1.7	模糊搜索	364
11.1.8	通配符搜索	370
11.1.9	正则表达式匹配搜索	373
11.1.10	综述	376
11.2	解析搜索请求	376
11.2.1	QueryParser 的基本使用	377
11.2.2	使用 QueryParser 解析多个关键词	380
11.3	高级搜索	387

11.3.1 多字段搜索	387
11.3.2 多索引搜索	397
11.3.3 多线程搜索	402
小结	404
第 12 章 搜索结果的处理和显示	405
本章介绍使用 Lucene 处理搜索结果的方法, 包括提取搜索结果、过滤搜索结果、对搜索结果进行排序以及搜索结果的高亮显示。	
12.1 提取搜索结果	405
12.1.1 Hits 对象的方法	405
12.1.2 对性能的说明	410
12.2 过滤搜索结果	413
12.2.1 简单的结果过滤	414
12.2.2 利用 Filter 类实现过滤	421
12.2.3 QueryFilter 的使用	421
12.2.4 PrefixFilter 的使用	432
12.2.5 RangeFilter 的使用	439
12.2.6 ChainedFilter 的使用	446
12.2.7 CachingWrapperFilter 的使用	455
12.2.8 FilteredQuery 的使用	458
12.2.9 综述	462
12.3 搜索结果的排序	462
12.3.1 按照文档得分排序	462
12.3.2 自定义排序规则	477
12.3.3 让系统决定如何排序	479
12.3.4 按照索引中的文档编号排序	483
12.3.5 按照文本字段排序	487
12.3.6 指定字段的数据类型	491
12.3.7 按多个字段排序	496
12.3.8 综述	504
12.4 搜索结果的高亮显示	504
12.4.1 关于高亮显示	504
12.4.2 Lucene 高亮处理的基本方式	506
12.4.3 Lucene 高亮处理的基本设置	511
12.4.4 为高亮显示设置新的格式	515
小结	519

第5部分 lucene 应用实例（第13~14章）

本部分会将前面介绍过的内容融合起来，构造一个实用的桌面搜索引擎和一个 Web 搜索引擎。你应该认真去体会构建搜索引擎的全过程，然后模仿着去做自己的搜索引擎。

第13章 桌面搜索引擎开发实例 523

本章介绍一个桌面搜索引擎，综合前面所学的一切内容给出一个完备的案例。

13.1 桌面搜索引擎简介	523
13.2 桌面搜索引擎索引器	524
13.2.1 索引器简介	524
13.2.2 索引器设计	524
13.2.3 解析器实现	525
13.2.4 获取文件列表	536
13.2.5 索引器实现	538
13.2.6 实例效果	545
13.3 桌面搜索引擎搜索器	547
13.3.1 搜索器简介	547
13.3.2 搜索器设计	548
13.3.3 搜索器实现	548
13.3.4 实例效果	556
小结	557

第14章 Web 搜索引擎开发实例 559

本章介绍了一个 Web 搜索引擎案例，在前面案例的基础上前进了一层，目的是为了综合一下前面所学的知识。

14.1 Web 搜索引擎简介	559
14.2 Web 搜索引擎索引器	559
14.2.1 索引器简介	559
14.2.2 索引器设计	559
14.2.3 解析器实现	561
14.2.4 获取文件列表	572
14.2.5 索引器实现	574
14.2.6 实例效果	581
14.3 Web 搜索引擎搜索器	583
14.3.1 搜索器简介	583