开发自己的搜索引擎

Lucene 2.0 + Heritrix

讲解如何使用Lucene和Heritrix，Web开发专家强烈推荐
一步一步带领您亲手构建企业级搜索引擎网站

附超值案例

邱哲 符滔滔 ◎ 编著

人民邮电出版社
POSTS & TELECOM PRESS

开发自己的搜索引擎

Lucene 2.0 + Heritrix

邱哲 符滔滔 ◎ 编著

人民邮电出版社
北京

图书在版编目（CIP）数据

开发自己的搜索引擎：Lucene 2.0+Heritrix / 邱哲，符滔滔编著.
—北京：人民邮电出版社，2007.6
ISBN 978-7-115-16000-3

Ⅰ. 开… Ⅱ. ①邱…②符… 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（2007）第 040773 号

内 容 提 要

本书是一本针对搜索引擎开发的书籍。通过学习本书，读者可以独立构建出一个企业级的搜索引擎网站。本书详细讲解了搜索引擎与信息检索基础，Lucene 入门实例，Lucene 索引的建立，使用 Lucene 进行搜索，排序，过滤和分页，Lucene 的分析器，对 Word、Excel 和 PDF 格式文档的处理，Compass 搜索引擎框架，Lucene 分布式和 Google Search API，爬虫 Heritrix，HTMLParser，DWR 等内容。最后综合使用所讲述的技术，构建了一个典型的垂直搜索系统，该系统具有很强的商业实用价值。

本书是一本介绍如何使用 Lucene 2.0 和 Heritrix 来构建搜索引擎的书。通过对相关 API 和源代码的分析，力求使读者在掌握应用的基础上能够深入其核心，自行扩展和开发相应组件，开发出更有创意的搜索引擎产品。

本书适合从事计算机软件开发的人员阅读，同时也可以作为搜索引擎爱好者的入门书籍。阅读本书需要具备 Java 语言基础。

开发自己的搜索引擎——Lucene 2.0+Heritrix

◆ 编　　著　邱　哲　符滔滔
责任编辑　屈艳莲

◆ 人民邮电出版社出版发行　　北京市崇文区夕照寺街 14 号
邮编　100061　　电子函件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京顺义振华印刷厂印刷
新华书店总店北京发行所经销

◆ 开本：800×1000　1/16
印张：33.75
字数：662 千字　　2007 年 6 月第 1 版
印数：1－5 000 册　　2007 年 6 月北京第 1 次印刷

ISBN 978-7-115-16000-3/TP

定价：65.00 元（附光盘）

读者服务热线：(010) 67132692　印装质量热线：(010) 67129223

作者简介

◎ 邱哲

北京理工大学硕士，现为某公司技术经理，主要从事欧美软件外包开发。在J2EE方面有4年的开发经验，在搜索引擎与“爬虫”方面有3年的开发经验，著有《征服Ajax+Lucene构建搜索引擎》一书。

◎ 符滔滔

北京理工大学硕士，曾就职于Eskalate和IBM，有7年Java开发经验，主要从事J2EE和搜索引擎方面的研究。

前　言

背景

搜索，这两个字无疑是当今互联网业界最为流行的字眼之一。在 Baidu 上输入“搜索引擎”这个关键字，可以找到 3000 多万条目；在 Google 上查找时，可以查到 750 万条目。不是 Google 的条目少，当用“search engine”作关键字查找时，在 Google 中可以查找到 3 亿多条目。

再来做个有趣的实验。当在 www.china-pub.com 中输入“搜索引擎”这个关键字后，只有可怜的 7 本书被查找了出来。

从上大学开始，我就知道，www.china-pub.com 应该是国内最大的计算机网上书店之一了。可是，为什么在 Google 中可以查找到 3 亿多条目的关键字，在计算机网上书店中只能找到区区 7 本书？

❖ **300000000 网页 vs 7 本书？**

本书特点

目前市面上从技术层面介绍搜索引擎的书并不多；即使有，也大多停留在理论阶段，而非搜索引擎的开发过程。而本书详细介绍了搜索引擎开发过程，实用性很强。本书具有以下一些特点：

（1）采用最新的 Lucene 2.0。相比 1.4.3 版本，Lucene 2.0 重写了很多 API，内部的实现方法也有了很多优化。本书的代码都是在 2.0 版本下调试通过的，这样可以帮助读者了解 Lucene 的更多新功能。

（2）配有一个完整的搜索引擎案例。这个案例有很强的实用价值，只需稍加修改，就能应用于实际项目，市场价值很大。

（3）着重解决令开发人员头痛的问题。本书的目的是指导项目实践，因此没有罗列各个 API 的用法，而是对常见的开发问题进行深入探讨。比如本书的第 7 章，是专门为解决 Word、Excel 和 PDF 文件如何解析这个问题而设置的。

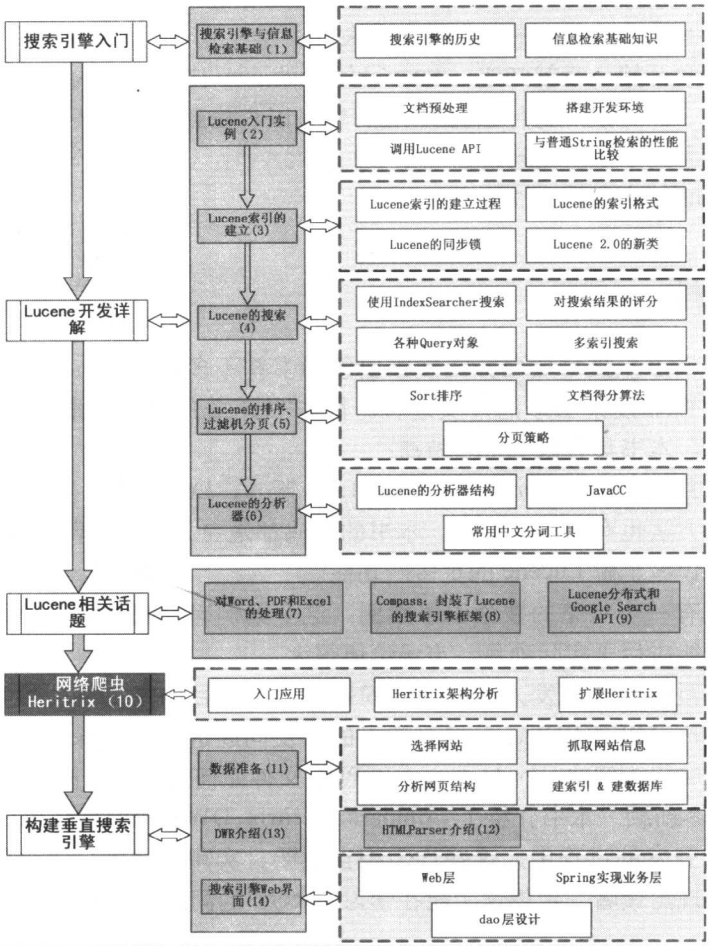
（4）内容新颖。本书介绍了 Compass、Heritrix、DWR 和 HTMLParser 等内容。在搜索引擎开发的过程中，这些均为相当重要且实用的技术，笔者经过自身实践将它们展现给读者，希望读者能在学习 Lucene 的同时开拓视野。

联系我们

我们为本书开通了专用的 BLOG，网址是 http://lucenebook.spaces.live.com/，读者可以直接与我们交流，共同学习和提高。另外，我们还为本书提供了专门的联系邮箱，luceneheritrix@163.com。读者可以随时与我们联系。

本书的内容简介

本书内容非常丰富，具体细节可以参考目录。下面给出全书的结构图，让读者有一个总体的认识。其中深色背景并且标有(1)、(2)等字样的标题表示相应的章节。



感谢

在写这段前言的时候，ZZQ（网名）正在忙着审阅我们的稿件，在此向他表示感谢，是他鼓励我们写这本书，并给予了我们很大的帮助。同时，还要感谢以下几个人。

- 吴萌野，如果没有他的指导，我不可能了解 Heritrix 的使用，他的技术水平令人折服。他在一个月内“踏平” Lucene 和 Heritrix 两座大山，开发出一个比价购物的搜索引擎。
- 何进，他是一位热心的读者，也是一位 Lucene 的爱好者，Compass 的内容便是他建议加入到书中的。
- 杜华博士，他为我提供了大量的资料（那时他正在开发一个“爬虫”），更重要的是，每天深夜他在 MSN 上和我的对话给我带来了无穷的动力。有人陪伴的开发过程，要比一个人感觉好多了。
- 最为重要的是我的家人，他们一直鼓励我，让我认真地工作。他们是我坚强的后盾。

结束语

经过 6 个多月的开发和写作，终于完成了书稿。写作的过程是艰苦的，有时候，为了讲清楚一个问题，要反复地调试程序、编写代码。而当花了大量精力把问题弄清楚并写出来后，往往发现正文部分只有那么几页甚至是几行。但是我想说的是，在写作的过程中，每时每刻我都在提醒着自己，搜索引擎是一门博大精深的学科，我所能做的，只是将现有的一些软件的使用方法以及我在这方面的一些微薄经验告诉大家，并希望大家可以从中得到有用的信息。只要读者能够在阅读的过程中，发现哪怕是一小段程序对自己有所启发，也是对我们的最大安慰。

虽然我们已经很努力地完善书稿，但由于我们的水平所限，书中不可避免地会出现一些不足之处，希望读者能够帮助我们指正。一旦出现这样的问题，烦请您将具体情况说明发到 luceneheritrix@163.com，我们将在再版时改正，同时会附上您的名字，以表示感谢。同时，读者有任何问题，也可以发送到该邮箱，我们会尽快为您解答。最后，衷心祝愿所有读者能够在本书中学到您所需要的知识。

本书由邱哲、符滔滔组织编写，同时参与编写的还有王石、熊英、付京周、

袁福庆、张杰、赵显琼、卜庆玲、常利、冯曼菲、匡妍娜、雷成健、李小波、刘浩然、刘会神、王晓悦、马震、齐志华、韩延峰、舒军、孙大林、孙佳楠、王辉、王沛等人，在此一并表示感谢。

编　者

2007 月 4 日

目　录

第一篇　搜索引擎入门

第二篇　Lucene 开发详解

第五篇 构建垂直搜索引擎