# <<开发自己的搜索引擎>>

## 图书基本信息

书名：<<开发自己的搜索引擎>>

13位ISBN编号：9787115215291

10位ISBN编号：7115215294

出版时间：2010-1

出版时间：人民邮电出版社

作者：邱哲，符滔滔，王学松 著

页数：562

## 前言

2007年初，我们编写了本书第l版，该书曾经连续数周占据互动出版网计算机类畅销书排行榜首。转眼间两年过去了，我收集了很多读者的建议和问题，针对书中的细节进行了调整，推出了本书的第2版。

第2版主要在以下方面进行了改进。

（1）对第l版中语言进行了优化，以使行文更加流畅，便于阅读，同时对一些表达模糊的地方进行了改写。

（2）针对读者提出的一些问题，进行了勘误。

（3）对书中的大部分Visio图片进行了重绘，看起来更美观。

（4）对书中涉及的软件进行全面梳理，对版本进行了更新，主要包括以下两部分。

· Lucene采用了2．0稳定版，正文中的范例都是使用它来实现的。目前Lucene已经发布到了2．4版，但是这个版本还没有大规模的商业化应用，存在很多不稳定因素，因此笔者只在附录中介绍了其核心功能。

· 网络爬虫采用了Heritrix1．14．0版本，包括后面的大案例都是使用这个版本重新实现的。

（5）去除了第9章中已经不再使用的Google的searchAPI部分。

# <<开发自己的搜索引擎>>

## 内容概要

《开发自己的搜索引擎：Lucene+Heritrix(第2版)》是一本介绍搜索引擎开发的书籍，通过《开发自己的搜索引擎：Lucene+Heritrix(第2版)》，读者可以独立构建一个企业级的搜索引擎网站。
《开发自己的搜索引擎：Lucene+Heritrix(第2版)》讲解了搜索引擎与信息检索基础，Lucene入门实例，索引的建立，使用Lucene来搜索，排序，分析器，对Word、Excel和PDF格式文档的解析，Compass搜索引擎框架，Lucene分布式，爬虫Heritrix，HTMLParser，DWR等内容。
最后综合这些技术，构建了一个典型的垂直搜索系统，具有很强的商业实用价值。
《开发自己的搜索引擎：Lucene+Heritrix(第2版)》是一本使用Lucene和Heritrix来讲解搜索引擎构建的书，通过对API和源代码的分析，力求使读者在应用的基础上，能够深入其核心，自行扩展和开发相应组件，发挥想象力，开发出更具有创意的搜索引擎产品。
《开发自己的搜索引擎：Lucene+Heritrix(第2版)》适合Java程序员和从事计算机软件开发的编程人员阅读，同时也可以作为搜索引擎爱好者的入门书籍。

# <<开发自己的搜索引擎>>

## 作者简介

邱哲，北京理工大学软件工程硕士。
现为Eskalate．com公司技术经理，同时负责开发人员招聘工作一-主要从事欧美软件外包开发，曾承接多家美国本土公司项目，在J2EE方面有7年的开发经验。
曾经编写了《souts Web设计与开发大全》、《开发自己的搜索引擎——Lucerie 2．0+Heritrix》。
王学松，博士。
曾任职知名互联网搜索引擎公司，担任高级软件工程师、研发经理等职位，参与大型搜索引擎开发多年。
开发完成亿级网页的互联网科技类信息垂直搜索引擎系统，完成中文搜索引擎的页面下载与分析、大规模索引建立、分类聚类技术、高并发检索和Web高速访问技术开发。
目前从事海量信息挖掘、语义网搜索引擎和基于内容图像检索的研究和开发。

# <<开发自己的搜索引擎>>

## 书籍目录

## 编辑推荐

销书升级，原书是国内第一本讲解搜索引擎开发的畅销书　　超值，提供了价值上万元的大型数码产品搜索引擎开发案铡，可直接应用于项目　　版本最新，采用了最新的Heritrix-1.140版、HTMLParser1.6.0版、DWR2.0.5版　　实践性强，用案例的方式讲解，便于读者实践　　注重原理讲解，提供了结构框图和流程图，讲解搜索引擎的原理　　《开发自己的搜索引擎：Lucene+Heritrix(第2版)》在第1版的基础上做了以下改变：　　重新组织了实例中开发前期的准备内容，涉及信息来源准备、基本开发环境准备、工程总体框架和配置信息管理等。

升级了内核代码版本，使用Heritrix-1．14．0版本，并增加了网络爬虫Heritrix代码工程导人和配置的详细步骤。

升级了实例代码，解决了因来源网站内容变更而导致的部分代码无法执行的问题。

增加了对网页内容分析的概述和基本说明，便于读者理解相关代码和内容。

HTMLParser升级到1．6．0版本。
为了适应网站代码的修改，变动了其中的正则分析代码和网页解析代码。

更新了实例中与文档组织和存储相关的内容，并根据所分析网站页面内容的变化，升级了其中的部分代码。

升级交互篇和用户Web界面中的核心代码为DWR2．0．5版本，并针对目前的主流技术，对搜索引擎交互方式进行分析和介绍，还增加了代码的部分图例说明。

## 版权说明