

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

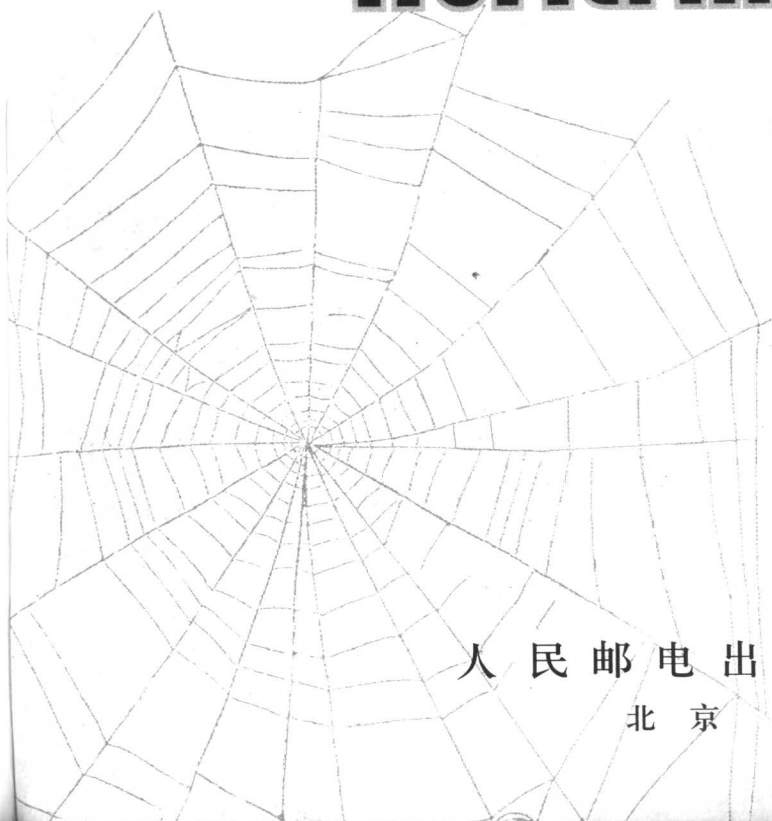
搜

索引引擎

邱哲 符滔滔 © 编著

Lucene 2.0 + Heritrix

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？

❖ 30000000 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大优化。本书的代码都是在 2.0 版本下调试通过的，这可以帮助读者了解 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，是专门为了解决 Word、Excel 和 PDF 文件如何解析这个问题而设置的。

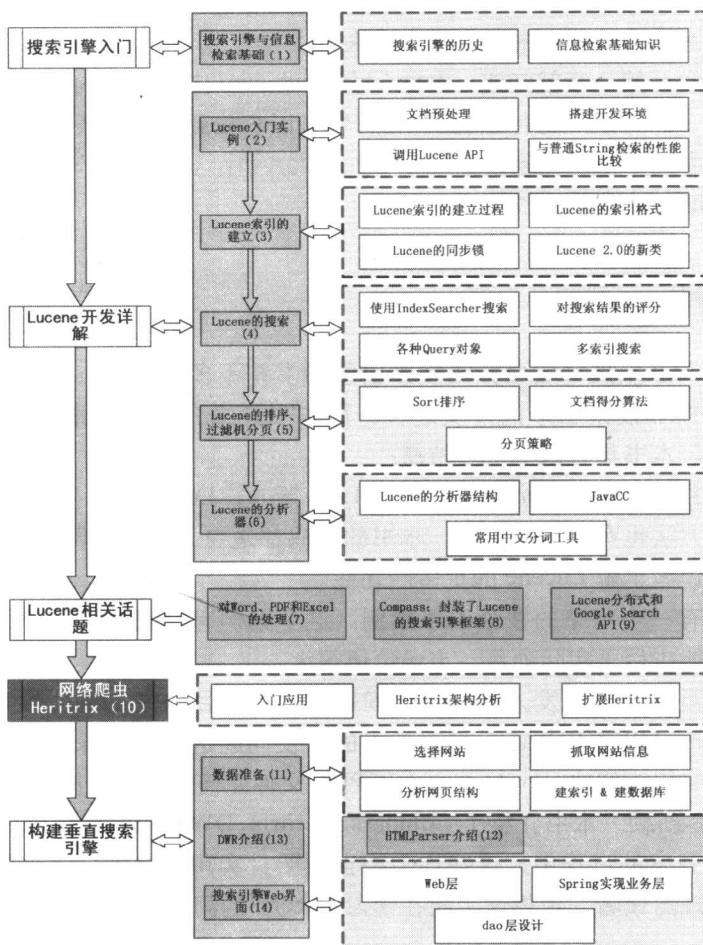
(4) 内容新颖。本书介绍了 Compass、Heritrix、DWR 和 HTMLParser 等内容。在搜索引擎开发的过程中，这些均为相当重要且实用的技术，笔者经过自身实践将它们展现给读者，希望读者能在学习 Lucene 的同时开拓视野。

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。

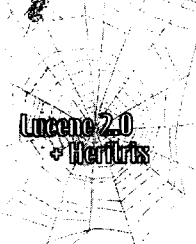
- 吴萌野，如果没有他的指导，我不可能了解 Heritrix 的使用，他的技术水平令人折服。他在一个月内“踏平”Lucene 和 Heritrix 两座大山，开发出一个比价购物的搜索引擎。
- 何进，他是一位热心的读者，也是一位 Lucene 的爱好者，Compass 的内容便是他建议加入到书中的。
- 杜华博士，他为我提供了大量的资料（那时他正在开发一个“爬虫”），更重要的是，每天深夜他在 MSN 上和我的对话给我带来了无穷的动力。有人陪伴的开发过程，要比一个人感觉好多了。
- 最为重要的是我的家人，他们一直鼓励我，让我认真地工作。他们是我坚强的后盾。

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 1 章 搜索引擎与信息检索基础	3
1.1 搜索引擎的历史	3
1.1.1 萌芽: Archie、Gopher	3
1.1.2 起步: Robot (网络机器人) 的出现与 Spider (网络爬虫)	5
1.1.3 发展: Excite、Galaxy、Yahoo 等	7
1.1.4 繁荣: Infoseek、AltaVista、Google 和 Baidu	8
1.2 信息检索系统的基本知识	11
1.2.1 什么是信息检索系统	11
1.2.2 信息检索的过程	12
1.2.3 传统查找的优点和不足	13
1.2.4 使用索引提高检索速度	14
1.2.5 倒排索引	14
1.2.6 评价信息检索系统的标准	16
1.3 Lucene 简介	16
1.4 小结	17

第二篇 Lucene 开发详解

第 2 章 Lucene 入门实例	21
2.1 实例介绍	21
2.1.1 实例说明	21
2.1.2 开发过程	21
2.2 准备工作	22
2.2.1 将文档的全角标点转换成半角标点	23
2.2.2 将大文档切分成多个小文档	25
2.2.3 预处理源文件的统一接口	27
2.3 创建 Eclipse 工程	27

2.3.1	准备工作	27
2.3.2	创建工程并引入 Lucene 的 JAR 包	29
2.3.3	运行文档预处理类	36
2.3.4	创建处理文档的索引类: IndexProcessor	37
2.3.5	创建检索索引的搜索类	40
2.4	运行效果	44
2.5	小结	45
第 3 章	Lucene 索引的建立	46
3.1	Document 逻辑文件	46
3.1.1	Lucene 的 Document	46
3.1.2	为 Document 添加多种 Field	47
3.1.3	Document 的内部实现	49
3.2	Field 的内部实现	50
3.2.1	Field 包含的类	51
3.2.2	Field 类的构造方法	52
3.3	Lucene 的索引工具 IndexWriter	54
3.3.1	IndexWriter 的初始化	54
3.3.2	向索引添加文档	56
3.3.3	限制每个 Field 中的词条的数量	58
3.4	Lucene 索引过程详解	58
3.4.1	Lucene 索引建立过程概述	59
3.4.2	使用 addDocument 方法向索引添加文档	59
3.4.3	DocumentWriter 的 addDocument 方法	61
3.4.4	文档的倒排	66
3.4.5	对 postingTable 进行排序	71
3.4.6	将 Posting 信息写入索引	73
3.5	索引文件格式	73
3.5.1	索引的 segment	73
3.5.2	.fnm 格式	73
3.5.3	.fdx 与 .fdt 格式	74
3.5.4	.tii 与 .tis 格式	75
3.5.5	deletable 格式	76
3.5.6	复合索引格式 .cfs	76
3.6	索引过程的优化	76

3.6.1	合并因子 mergeFactor	76
3.6.2	maxMergeDocs	78
3.6.3	minMergeDocs	78
3.7	索引的合并与索引的优化	78
3.7.1	FSDirectory 与 RAMDirectory	78
3.7.2	使用 IndexWriter 来合并索引	79
3.7.3	索引的优化	81
3.8	从索引中删除文档	82
3.8.1	索引的读取工具 IndexReader	83
3.8.2	使用文档 ID 号来删除特定文档	86
3.8.3	使用 Field 信息来删除批量文档	88
3.9	Lucene 的同步问题	89
3.9.1	为什么要进行同步以及 Lucene 的同步法则	90
3.9.2	commit.lock 与 write.lock	90
3.10	Lucene 2.0 的新类: IndexModifier 类	91
3.11	小结	92
第 4 章	Lucene 的搜索	93
4.1	使用 IndexSearcher 进行搜索	93
4.1.1	初始化 IndexSearcher	93
4.1.2	IndexSearcher 最简单的使用	94
4.1.3	IndexSearcher 的多种 search 方法	95
4.2	Hits 类详解	97
4.2.1	Hits 类的公有接口	97
4.2.2	效率分析	98
4.2.3	Hits 内部的缓存	100
4.2.4	Hits 类的工作原理	103
4.3	对搜索结果的评分	104
4.3.1	文档与词条的向量空间	104
4.3.2	Lucene 的文档得分算法	105
4.4	构建各种 Lucene 内建的 Query 对象	108
4.4.1	toString 查看原子查询	109
4.4.2	查询重写与权重	109
4.4.3	TermQuery 词条搜索	110
4.4.4	BooleanQuery 布尔搜索	111

4.4.5	RangeQuery 范围搜索	119
4.4.6	PrefixQuery 前缀搜索	122
4.4.7	PhraseQuery 短语搜索	125
4.4.8	MultiPhraseQuery 多短语搜索	128
4.4.9	FuzzyQuery 模糊搜索	133
4.4.10	WildcardQuery 通配符搜索	137
4.4.11	SpanQuery 跨度搜索	138
4.5	第三方提供的 Query 对象: RegexpQuery	146
4.6	通过 QueryParser 转换用户关键字	148
4.6.1	词条的定义	149
4.6.2	QueryParser 初始化	149
4.6.3	改变 QueryParser 默认的布尔逻辑	150
4.6.4	短语和 QueryParser	151
4.6.5	FuzzyQuery 和 QueryParser	152
4.6.6	通配符与 QueryParser	153
4.6.7	查找指定的 Field	153
4.6.8	RangeQuery 与 QueryParser	157
4.6.9	QueryParser 和 SpanQuery	158
4.7	多 Field 搜索与多索引搜索	159
4.7.1	多域搜索 MultiFieldQueryParser	159
4.7.2	MultiSearcher 在多个索引上搜索	161
4.7.3	ParallelMultiSearcher: 多线程搜索	164
4.7.4	Searchable 和 RMI	167
4.8	小结	168
第 5 章	排序、过滤和分页	170
5.1	相关度排序	170
5.1.1	使用 Score 进行自然排序	170
5.1.2	Searcher 的 explain 方法	172
5.1.3	通过改变 boost 值来改变文档的得分	173
5.2	使用 Sort 来排序	177
5.2.1	Sort 简介	177
5.2.2	SortField	178
5.2.3	按文档得分进行排序	179
5.2.4	按文档的内部 ID 号来排序	182

5.2.5	按一个或多个 Field 来排序	183
5.2.6	改变 SortField 中的 Locale 信息	190
5.3	搜索的过滤器	191
5.3.1	过滤器的基本结构	191
5.3.2	一个简单的 Filter: 建立索引	192
5.3.3	一个简单的 Filter: 打印索引文档信息	194
5.3.4	一个简单的 Filter: 安全级别与过滤器代码	196
5.3.5	一个简单的 Filter: 在搜索时应用过滤器	197
5.3.6	一个简单的 Filter: 总结	198
5.3.7	按范围过滤 RangeFilter	199
5.3.8	在结果中查询 QueryFilter	202
5.3.9	缓存结果: CachingWrapperFilter	205
5.4	翻页问题	206
5.4.1	依赖于 session 的翻页	206
5.4.2	多次查询	207
5.4.3	缓存 + 多次查询	207
5.4.4	缓存 + 多次查询 + 数据库	208
5.5	小结	208
第 6 章	Lucene 的分析器	209
6.1	分析	209
6.1.1	分词	209
6.1.2	Lucene 的分析器的结构	210
6.1.3	Lucene 的分析器的实现	212
6.2	Lucene 与 JavaCC	213
6.2.1	JavaCC 简介	214
6.2.2	JavaCC 为 Lucene 提供的分析器脚本	214
6.2.3	Lucene 的标准分析器	218
6.2.4	标准过滤器: StandardFilter	220
6.2.5	大小写转换器: LowerCaseFilter	221
6.2.6	忽略词过滤器: StopFilter	221
6.3	分析器的进阶	222
6.3.1	再看 StandardAnalyzer 中的管道过滤器结构	222
6.3.2	长度过滤器: LengthFilter	223
6.3.3	PerFieldAnalyzerWrapper	223

6.3.4 其他	224
6.4 对中文的分析	224
6.4.1 现有的中文分词方式简介	225
6.4.2 中科院的分词软件和 JE 分词	227
6.5 小结	232

第三篇 Lucene 相关话题

第 7 章 对 Word、Excel 和 PDF 的处理	235
7.1 使用 PDFBox 处理 PDF 文档	235
7.1.1 PDFBox 的下载	235
7.1.2 在 Eclipse 中配置	236
7.1.3 使用 PDFBox 解析 PDF 内容	237
7.1.4 运行效果	238
7.1.5 与 Lucene 的集成	239
7.2 使用 xpdf 来处理中文 PDF 文档	241
7.2.1 xpdf 的下载	241
7.2.2 配置	242
7.2.3 提取中文	243
7.2.4 运行效果	246
7.3 使用 POI 来处理 Excel 和 Word 文件格式	246
7.3.1 对 Excel 的处理类	247
7.3.2 ExcelReader 的运行效果	251
7.3.3 POI 中 Excel 文件 Cell 的类型	252
7.3.4 对 Word 的处理类	254
7.4 使用 Jacob 来处理 Word 文档	256
7.4.1 Jacob 的下载	256
7.4.2 在 Eclipse 中配置	256
7.5 小结	258
第 8 章 Compass: 封装了 Lucene 的框架	259
8.1 Compass 简介	259
8.1.1 Compass 的下载	259
8.1.2 Compass 的代码片断	260
8.2 Compass 的初始配置	261
8.2.1 Compass 的配置文件	261

8.2.2	将索引存放于内存中	262
8.2.3	使用 JDBC 来存储索引	262
8.2.4	使用连接池来存储索引	263
8.2.5	加载 compass.cfg.xml 文件	264
8.3	域模型的配置	265
8.3.1	实体代码	265
8.3.2	实体关系	271
8.3.3	实体 Book 的配置文件	271
8.3.4	通用元数据定义文件 (.cmd.xml)	272
8.3.5	Author 和 Article 的配置文件	276
8.4	使用 Compass 来建立索引	278
8.4.1	索引代码	278
8.4.2	对象关系图和运行结果	280
8.5	使用 Compass 来搜索	281
8.5.1	使用 find() 方法搜索	281
8.5.2	CompassHits 类型	282
8.5.3	CompassHit 类型	283
8.5.4	使用 Lucene 语法来查找	284
8.6	配置 Analyzer 和 Optimizer	286
8.7	小结	287
第 9 章	Lucene 分布式和 Google Search API	288
9.1	Lucene 与分布式	288
9.1.1	什么是 GFS	288
9.1.2	为 Lucene 提供分布式的几点设想	289
9.2	Google 的 Search API	291
9.2.1	搭建环境	292
9.2.2	构建搜索类	292
9.2.3	设置查询时的参数和查询语法	295
9.2.4	运行测试	296
9.3	小结	297
第四篇	网络爬虫 Heritrix	301
第 10 章	无比强大的网络爬虫 Heritrix	301
10.1	Heritrix 使用入门	301

10.1.1	下载和运行 Heritrix	301
10.1.2	在 Eclipse 里配置 Heritrix 的开发环境	304
10.1.3	创建一个新的抓取任务	308
10.1.4	设置抓取时的处理链	310
10.1.5	设置运行时的参数	312
10.1.6	运行抓取任务	314
10.1.7	Heritrix 的镜像存储结构	318
10.1.8	终止抓取或终止 Heritrix 的运行	319
10.2	Heritrix 的架构	320
10.2.1	抓取任务 CrawlOrder	320
10.2.2	中央控制器 CrawlController	321
10.2.3	Frontier 链接制造工厂	324
10.2.4	用 Berkeley DB 实现的 BdbFrontier	329
10.2.5	Heritrix 的多线程 ToeThread 和 ToePool	332
10.2.6	处理链和 Processor	335
10.3	扩展和定制 Heritrix	338
10.3.1	向 Heritrix 中添加自己的 Extractor	339
10.3.2	定制 Queue-assignment-policy 的两个问题	343
10.3.3	定制 Queue-assignment-policy 继承 QueueAssignmentPolicy 类	344
10.3.4	扩展 FrontierScheduler 来抓取特定的内容	344
10.3.5	在 Prefetcher 中取消 robots.txt 的限制	346
10.4	小结	347

第五篇 构建垂直搜索引擎

第 11 章	搜索引擎综合实例：准备篇	351
11.1	实例简介以及实现途径	351
11.1.1	选择网站	352
11.1.2	太平洋电脑网和网易手机频道	352
11.1.3	分析网站内容并准备抓取清单	353
11.1.4	从下拉列表获得手机品牌首页	356
11.1.5	解析手机品牌页面	359
11.2	在 Heritrix 中为 pconline 开发抓取所需的定制类	361
11.2.1	保存所有产品的页面和图片	362
11.2.2	不保存其他无关页面	362

11.2.3 开始抓取	364
11.3 在 Heritrix 中为网易手机频道开发抓取所需的定制类	365
11.3.1 分析网易手机频道	365
11.3.2 设计抓取代码	368
11.4 在 Eclipse 中创建工程结构	373
11.4.1 下载插件	373
11.4.2 在 Eclipse 中配置插件	374
11.4.3 创建工程	375
11.4.4 设置工程的 Context	376
11.4.5 设定源代码存放和输出路径	377
11.4.6 添加 Java 代码	379
11.4.7 添加 Jar 包	380
11.4.8 创建 JSP 文件	381
11.4.9 工程整体结构一览	383
11.5 设定配置文件及其相关类	385
11.5.1 系统属性配置文件	385
11.5.2 封装配置文件	385
11.6 产品详细信息文件格式	387
11.7 解析网页信息的基类 Extractor	389
11.8 太平洋电脑网手机产品页面 Extractor	393
11.9 pconline 产品信息运行效果测试	397
11.9.1 编写测试函数	397
11.9.2 执行测试	398
11.10 网易手机频道的产品信息运行效果	401
11.11 构建产品信息词库	404
11.12 数据库与索引结构	407
11.12.1 定义 Product 类	407
11.12.2 确定数据库与索引的结构	409
11.13 数据库处理和索引处理	411
11.13.1 对数据库进行操作	412
11.13.2 对索引进行操作	414
11.14 调用数据库处理类和索引处理类	415
11.15 运行	420
11.16 小结	422