

Ajax + Lucene

构建搜索引擎

李 刚 宋 伟 邱 哲 编著



189个代码示例

 人民邮电出版社
POSTS & TELECOM PRESS

Ajax + Lucene

构建搜索引擎

李刚 宋伟 邱哲 编著



人民邮电出版社
POSTS & TELECOM PRESS

图书在版编目 (CIP) 数据

征服 Ajax+Lucene 构建搜索引擎 / 李刚, 宋伟, 邱哲编著. —北京: 人民邮电出版社, 2006.4

ISBN 7-115-14707-8

I. 征... II. ①李... ②宋... ③邱... 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 (2006) 第 033662 号

内 容 提 要

本书分 Ajax 和 Lucene 两部分, 共 16 章, 内容涵盖了 JavaScript 的基础、层叠样式表 CSS、DOM 与 JavaScript、Ajax 工作流程、Ajax 实例、Ajax 的安全问题与性能测试、Lucene 基础、用 Lucene 来建立索引、Lucene 搜索、Lucene 的分析器、高级搜索技巧、Lucene 性能测试、Lucene 的中文问题以及 Lucene 的工具箱 SandBox。

本书最大的特点在于“理论结合实践”, 既简明扼要地讲解了理论知识, 又提供了丰富的示例程序, 力求使读者在掌握了理论知识的同时能够通过阅读示例代码加深理解。

本书适合于有一定 Web 开发基础的 Java 程序员和对 Ajax 与 Lucene 感兴趣的开发人员。

征服 Ajax+Lucene 构建搜索引擎

-
- ◆ 编 著 李刚 宋伟 邱哲
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: 24.5
字数: 526 千字
印数: 1 ~ 5 000 册
- 2006 年 4 月第 1 版
2006 年 4 月北京第 1 次印刷

ISBN 7-115-14707-8/TP · 5371

定价: 42.00 元 (附光盘)

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

PREFACE

前言

关于 Ajax

到底什么是 Ajax? 首先来看看图 1 所示的页面效果(网址是 <http://projects.backbase.com/RUI/shop.html>)。这是一个购物网站的界面, 界面左边是各种商品, 右边是一个购物车的列表。



图 1

当用户选中某种商品, 可使用鼠标光标拖动方式将其添加入右边的购物车中, 如图 2 所示。

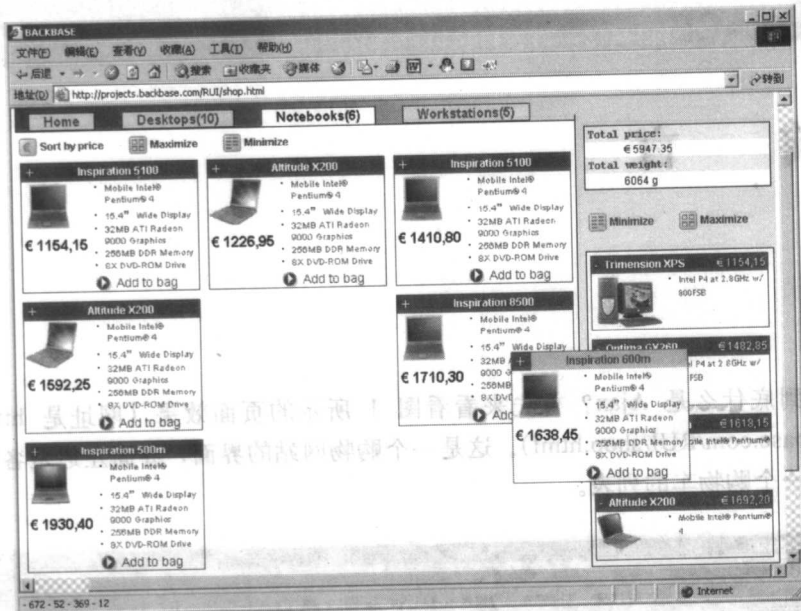


图 2

拖动完毕后，商品已经被加入了购物车的列表中，如图 3 所示。

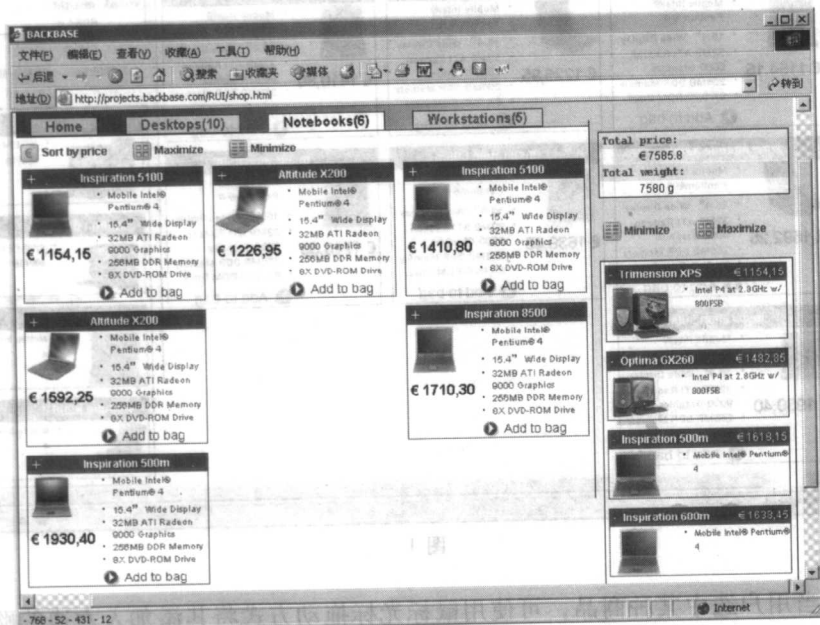


图 3

下面再举一个 Google 的例子，打开 Google Map 网站 <http://maps.google.com/>，如图 4 所示，用户可以在上面随意拖动地图，可以看到地图立刻被刷新，而不需要长时间的等待，这在以前的网站中是不可想象的。

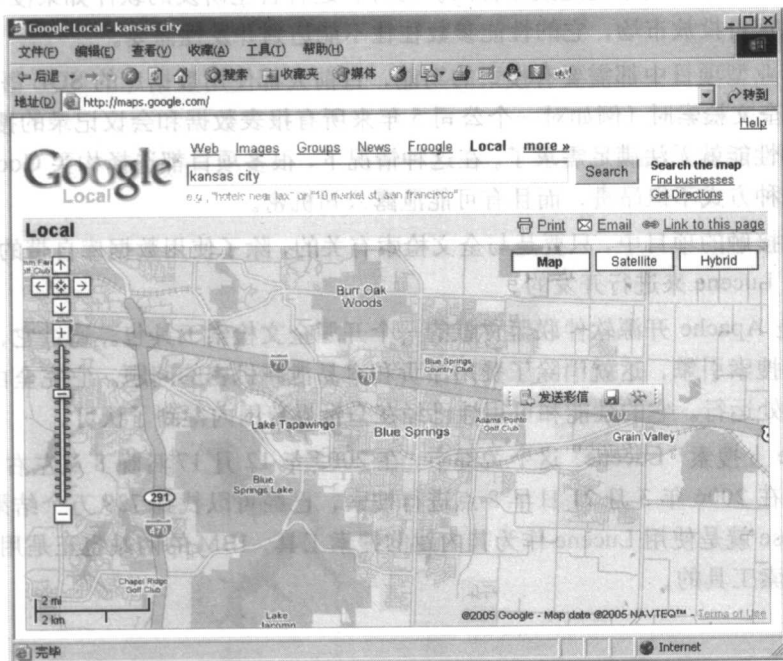


图 4

打开 Google Suggest 网站 <http://www.google.com/webhp?complete=1&hl=en>，用户输入一个关键字，输入框下就会提示出相应的建议信息。这建议信息是从 Google 的庞大数据库中读出来的，使用十分方便（注：Google Suggest 效果将会在本书的第 6 章给出实现方法）。

上面列举的网页响应速度更快、可以实现拖动等桌面应用程序的效果。普通的网页对任何提交请求，都要等待服务器的响应，然后网页才有所动作。也就是说，采用 Ajax 技术的网页具有了和 Windows 桌面很类似的功能，操作更加人性化，极大地提高了用户的体验。

如今随着 Web 2.0 概念的普及，追求更美观更人性化的页面效果成了网站开发的重中之重。Ajax 这项技术正在其中充当着重要的角色。

关于 Lucene

如今的互联网上信息量非常庞大，想要在网上获取一些有用的信息相当困难，没有搜索引擎的帮助，用户往往不知所措。

搜索引擎是一个特殊的技术行业，它的技术门槛比较高，如果完全自主研发，一个公司可能要经过几年时间的开发和完善才能建立起一套自己的搜索引擎软件，且不论这种软件的性能可靠性等技术参数，单是漫长的研发时间就会使很多商机从手中溜走。这对一些新兴的靠市场动向盈利的小公司来说是致命性的。另外，这种自主研发的软件如果没有经过严格的用户测试，就盲目投放市场，它的性能参数往往不能达到并发性需要。

在很多商业型项目中都需要有搜索的功能，目前大都使用数据库的 SQL 语言实现搜索。然而，当遇到全文检索时（例如对一个公司 5 年来所有报表数据和会议记录的搜索），数据库搜索的速度和性能就无法满足需求了。在这种情况下，很多项目都选择购买 Google 等公司的搜索服务，这种方式不但昂贵，而且有可能泄露公司机密。

在笔者所接触的项目中，只要是与全文检索有关的，除了使用数据库自带的索引功能外，其余都是使用 Lucene 来进行开发的。

Lucene 是 Apache 开源软件联盟贡献的一个开源全文检索工具包。通过它，我们可以很方便地构建起搜索引擎，这就消除了费用昂贵和容易泄露机密的问题。它完全由 Java 编写，一次编译，随处运行。它的性能和可用性已经在全世界范围内得到了认可。

在 Google 上搜索“Lucene”这个关键字，在 2005 年 12 月 17 日晚 8 点左右可以找到 112 万个结果。而在 2006 年 3 月 21 日早 8 点进行检索，已经可以找到 739 万个结果，可见其发展之快。Eclipse 就是使用 Lucene 作为其内建的搜索工具，IBM 的网站也正是用 Lucene 来构建的它站内检索工具的。

关于本书

本书由浅入深讲解理论知识，同时又配以很多示例代码。

Ajax 是一种应用于 Web 编程的技术，而 Lucene 本身是用 Java 语言编写完成的，因此本书最适合有一定 Web 开发和 Java 编程基础的人。如果读者对 Web 编程技术和 Java 编程技术不是十分清楚，建议先阅读相关资料，然后再来学习本书。

同时，Java 语言是一种面向对象的程序语言，所以读者最好具有一定的面向对象编程的理论基础。

本书分为两大部分共 16 章。

第 1 部分为 Ajax 的相关知识，共分 7 章。这部分内容包括 Ajax 入门体验、JavaScript 关键技术、CSS 表示技术、DOM 技术、XmlHttpRequest 对象和 Ajax 的工作流程、经典实例以及页面编程中的性能和安全性问题等内容，这部分内容旨在帮助读者快速掌握 Ajax 技术。

第 2 部分为 Lucene 的相关知识，共分 9 章。这 9 章的内容包括 Lucene 基本概念、入门实例、建立索引、通过索引来搜索文件、使用分析器、高级搜索技巧、性能测试、中文问题，以及 Lucene 中的一些实用工具。

本书特色

(1) 实例丰富。本书包括了大量的代码片断，并都进行了详细的注释。在第 1 部分中，专门用一章来讲解完整的 Ajax 范例，以便帮助读者更好地理解相关技术的综合运用。另外，在第 2 部分中，每一章都给出了大量的例子来对 API 的使用进行演示；同时，在“性能测试”一章中也专门给出一个有关建立索引的较为完整的工具，以便于读者进行学习和参考。

(2) 引导式学习。本书的各个章节都是从零开始讲起，然后不断地扩展需求，提出新的问题，然后讲解问题的解决方法。

(3) 程序员本色。由于作者本身就是一线开发人员，因此最了解程序员的需求。

(4) 及时跟进最新动态。Ajax 与 Lucene 的发展迅猛，我们将及时跟踪它们的最新动向，请读者随时关注我们的 Blog。读者可以直接同我们交流。同时，一些在本书光盘中并未选录的资料也将被放置在我们的 Blog 中，地址是 <http://blog.sina.com.cn/u/1197616372>，读者可以方便获取到。

注意

- 在本书即将完成之际，Lucene 的网站 <http://lucene.apache.org> 发布了其最新版本 1.9。而本书所有的实例均使用其上一个版本 1.4.3，建议读者学习时使用 1.4.3 版本。
- 本书中所有的代码标号仅是为正文中引用方便，配套光盘中的代码均按章节存储在相应的目录下。

作者

2006 年 4 月

PREFACE

光盘使用说明

本书提供了本书所有案例的源代码，在每个文件夹下都有相应的说明文件，读者可以参考使用。

1. 运行环境

操作系统：Windows 2000

Java 平台：JDK 1.4

Server：Tomcat 5.0

开发环境：Eclipse 3.1

2. 光盘内容

(1) Ajax 目录

在该目录中提供了本书第 1~7 章的全部源代码，具体的配置可以参考相应文件夹下的“说明.txt”文件。具体文件对应结构如下。

Ch01：本章包括两个工程 Register 和 DynaCheckRegister，其中工程 Register 是传统的 Web 注册应用程序，工程 DynaCheckRegister 是使用 Ajax 技术的动态 Web 注册程序。

Ch02：本章共包含 30 多个程序，这 30 多个程序主要讲述 JavaScript 的基本用法。

Ch03：本章所有的程序讲解了 CSS 样式的基本用法。程序的文件名和在书中的编号一致。

Ch04：本章所有的程序讲解了使用 JavaScript 访问表单和 XML 的方法。程序的文件名和在书中的编号一致。

Ch05：本章讲述了动态注册程序 DynaCheckRegister，其中 src 目录用来存放服务器端的 java 程序，dynareg 目录是用来部署的模块，里面包括使用到的页面文件。

Ch06：本章讲解了 4 个例子，都在 Ajax 工程目录下。GoogleSuggest 和 Slider 两个例子较为复杂，都用到了数据库驱动包 msbase.jar、mssqlserver.jar 和 msutil.jar 来连接 SQL Server 数据库，读者需要根据书中的介绍到微软公司的官方网站下载这 3 个 JAR 文件。这两个程序

都用到了数据库 C3C，可以直接将目录下的 C3C.dat 文件还原，方法是先在 SQL Server 的企业管理器中建立一个名为 C3C 的数据库文件，然后再在其上进行还原。

Ch07：可以直接将目录下的 ajax 目录放在 tomcat 的 webapps 目录下使用，在测试第二个程序时，需按书中的介绍做相应调整。

(2) Lucene 目录

该目录下包括两个 Eclipse 的工程，其中 LuceneHelloWorld 工程包括了从第 10 章~第 15 章的源代码，它们分别以 Java 的包形式存放于文件夹中。例如，第 10 章的源码就存放在路径为 \LuceneHelloWorld\ch10 的目录下，依此类推。具体内容如下。

Ch10：第 10 章内的所有代码，主要包括如何通过 Lucene 建立索引的示例。

Ch11：第 11 章内的所有代码，主要包括如何使用 Lucene 搜索索引的示例。

Ch12：第 12 章内的所有代码，主要包括了对 Lucene 的分析器进行测试的示例。

Ch13：第 13 章内的所有代码，主要包括了对 Lucene 的一些高级搜索方式的代码示例。

Ch14：第 14 章内的代码，另外还有一部分在 LucenePerformance 工程内。LucenePerformance 工程包括了第 14 章性能测试中的例子的源代码。

Ch15：第 15 章内的代码，包括了对 Lucene 中文进行测试的代码示例。

3. 运行方法

(1) 对于所有 Eclipse 工程，均可在 Eclipse 环境中导入后直接运行。

(2) 对于所有 Web 应用，可将 Web 应用的目录直接复制到 Tomcat 的 webapps 目录下直接运行。

(3) 对于一些演示 Javascript 或 DOM 的用法的 HTML 文件，可以直接用 IE 打开运行。如果要看源代码，可以使用记事本或 UltraEdit 等工具查看。

(4) 对于使用到数据库连接的 Google Suggest 例子和 Slider 例子，需先安装 SQL Server，然后还原光盘中所提供的数据库文件。

4. 疑难解答

由于 Ajax 和 Lucene 涉及的知识比较庞杂，读者在调试的时候可能会遇到各种问题，可以给作者发 E-mail: ptbook@163.com，或者到作者的 Blog 上提问: <http://blog.sina.com.cn/u/1197616372>。

目 录

第 1 章 来点 Ajax 的感觉	1
1.1 两个例子	1
1.1.1 百度地图	1
1.1.2 Google Suggest	2
1.1.3 体会 Ajax	3
1.2 一个传统的 Web 注册程序	5
1.2.1 运行效果预览	5
1.2.2 注册 Servlet 的代码	6
1.2.3 相关 JSP 页面代码	8
1.3 使用 Ajax 实现的 Web 注册程序——DynaCheckRegister	10
1.3.1 运行效果预览	10
1.3.2 注册 Servlet 的代码	11
1.3.3 相关 JSP 页面代码	13
1.4 两种技术的比较	16
1.4.1 效果	17
1.4.2 涉及到的知识	17
1.5 总结	17
第 2 章 JavaScript 基础	18
2.1 浏览器对象模型	18
2.1.1 查看浏览器信息——Navigator 对象	19
2.1.2 分割浏览器窗口——Frame 对象	20
2.1.3 访问加载在窗口中的文档的 URL——location 对象	22
2.1.4 查看访问过的网址列表——history 对象	23
2.1.5 取得计算机的各种显示特性——screen 对象	25
2.2 函数	26
2.3 与用户交互的基本方法	30

2.3.1	提示对话框——alert()方法	30
2.3.2	确认信息对话框——confirm()方法	31
2.3.3	响应用户输入对话框——prompt()方法	32
2.4	事件响应与编程	33
2.4.1	响应鼠标事件	35
2.4.2	响应键盘事件	36
2.4.3	控制事件——event 对象	36
2.5	JavaScript 核心对象	39
2.5.1	数组对象	39
2.5.2	Math 对象	41
2.5.3	String 对象	42
2.6	面向对象的 JavaScript	43
2.6.1	JavaScript 中的对象	43
2.6.2	使用 JSON 创建数组和对象图	45
2.6.3	构造函数	47
2.6.4	使属性和函数可以关联到对象——prototype	49
2.6.5	JavaScript 对象反射	50
2.6.6	方法和函数	52
2.6.7	Ajax 的事件处理和函数环境	54
2.7	胖客户端 (Rich Client)	56
2.7.1	什么是胖客户端	56
2.7.2	Ajax 解决的问题	56
2.8	总结	58
第 3 章	层叠样式表 CSS	59
3.1	使用 CSS 的原因	59
3.1.1	样式与内容的分离	60
3.1.2	可读性	60
3.1.3	一次性写入	60
3.1.4	CSS 的版本与浏览器的兼容性	60
3.2	CSS 基础	61
3.2.1	样式表的基础	61
3.2.2	内部样式表	63
3.2.3	外部样式表	64
3.2.4	创建和使用样式类	67

3.3 常用的样式表属性	68
3.3.1 字体属性	68
3.3.2 颜色和背景属性	70
3.3.3 文字属性	72
3.3.4 边框属性	73
3.3.5 其他	75
3.4 总结	77
第 4 章 DOM 与 JavaScript	78
4.1 DOM 文档对象模型	78
4.1.1 什么是 DOM	78
4.1.2 DOM 的层次结构	79
4.2 Document 对象	80
4.2.1 Document 对象的属性	81
4.2.2 document 对象的方法	83
4.3 JavaScript 访问 HTML 表单	85
4.3.1 (表单) form 基础	85
4.3.2 表单对象	87
4.3.3 使用单行文本框对象	90
4.3.4 使用密码框	92
4.3.5 使用多行文本框	93
4.3.6 使用单选框	95
4.3.7 使用复选框	97
4.3.8 使用下拉菜单	98
4.3.9 使用链接对象	100
4.3.10 使用 image (图像) 对象	102
4.3.11 DOM 事件处理	104
4.4 JavaScript 访问 XML	106
4.4.1 第一个简单的例子	107
4.4.2 使用 JavaScript 读取 XML 文档	108
4.5 总结	112
第 5 章 Ajax 工作流程	113
5.1 概述	113
5.2 XMLHttpRequest 对象	114
5.2.1 认识 XMLHttpRequest 对象	114

5.2.2	XMLHttpRequest 对象的方法	116
5.2.3	XMLHttpRequest 对象的属性	117
5.2.4	使用 XMLHttpRequest 对象向服务器端发送请求	117
5.3	监视 response 的状态	119
5.3.1	回调函数——监视浏览器的状态	119
5.3.2	操纵返回的数据	120
5.4	一个完整的实例	121
5.5	详解 DynaCheckRegister 注册程序	125
5.5.1	页面脚本	125
5.5.2	实现原理	129
5.6	总结	134
第 6 章	Ajax 实例	135
6.1	提示等待	135
6.1.1	原理分析	135
6.1.2	客户端代码总体说明	136
6.1.3	客户端代码——创建主框架	136
6.1.4	客户端代码——实现函数处理	136
6.1.5	客户端代码——创建数据显示的页面	138
6.1.6	服务器端代码——处理请求	139
6.2	级联下拉框	140
6.2.1	原理分析	141
6.2.2	客户端代码说明	141
6.2.3	服务器端代码说明	144
6.3	Google suggest 的例子	146
6.3.1	原理分析	147
6.3.2	客户端代码总体说明	148
6.3.3	客户端代码——创建用户进行查询的页面	149
6.3.4	客户端代码——网页界面显示样式的设计	150
6.3.5	客户端代码——程序运行脚本的实现	152
6.3.6	服务器端代码总体说明	163
6.3.7	服务器端代码——创建实体	163
6.3.8	服务器端代码——创建数据库连接	164
6.3.9	服务器端代码——实现业务逻辑处理	165
6.3.10	服务器端代码——实现请求处理	166

6.3.11 性能问题	168
6.4 Slider (滑块)	168
6.4.1 原理分析	168
6.4.2 客户端代码总体说明	169
6.4.3 客户端代码——创建页面的框架	170
6.4.4 客户端代码——实现程序的脚本	172
6.4.5 服务器端代码总体说明	179
6.4.6 服务器端代码——定义实体	179
6.4.7 服务器端代码——创建与数据库的连接	180
6.4.8 服务器端代码——业务逻辑处理	181
6.4.9 服务器端代码——客户端请求处理	183
6.5 总结	184
第 7 章 Ajax 的安全问题	185
7.1 JavaScript 和浏览器的安全性	185
7.1.1 JavaScript 修改设置	185
7.1.2 IFrame 的问题	186
7.2 JavaScript 的性能问题	189
7.2.1 循环	189
7.2.2 局部变量和全局变量	189
7.2.3 不使用 eval	189
7.2.4 减少对象查找	189
7.2.5 字符串连接	189
7.2.6 类型转换	190
7.2.7 使用直接量	191
7.2.8 字符串遍历操作	191
7.2.9 高级对象	191
7.2.10 插入 HTML	191
7.2.11 下标查询	191
7.2.12 创建 DOM 节点	191
7.3 总结	192
第 8 章 Lucene 基础	193
8.1 信息获取与搜索引擎	193
8.1.1 信息获取	193
8.1.2 搜索引擎的发展与分类	194

8.2	Lucene 的历史	195
8.2.1	什么是全文检索与全文检索系统	195
8.2.2	什么是 Lucene	196
8.2.3	Lucene 的发展和现状	197
8.2.4	使用 Lucene 能做什么	197
8.2.5	谁在使用 Lucene	198
8.3	建立索引和搜索	198
8.3.1	索引	198
8.3.2	搜索	200
8.3.3	倒排索引	202
8.4	总结	203
第 9 章	第一个“搜索引擎”	204
9.1	初识 Lucene	204
9.1.1	下载 Lucene	204
9.1.2	Lucene 的包结构	204
9.2	为本机的文件创建索引	209
9.2.1	选择要建立索引的文件和索引存放的位置	209
9.2.2	建立索引	209
9.3	试试搜索的效果	213
9.3.1	运行搜索	213
9.3.2	如何进行搜索——LuceneSearch 类	213
9.3.3	修改测试	215
9.4	值得改进的地方	216
9.4.1	思路	216
9.4.2	其他	217
9.5	总结	217
第 10 章	用 Lucene 来建立索引	218
10.1	索引建立的过程	218
10.1.1	提取文本	218
10.1.2	构建 Document	219
10.1.3	分析并建索	219
10.2	Lucene 的文档格式	219
10.2.1	文档 (Document)	219
10.2.2	字段 (Field)	223

10.3 索引的添加——IndexWriter 类	227
10.3.1 初始化	227
10.3.2 向索引添加文档	228
10.3.3 调整性能参数	231
10.3.4 限制 Field 长度	232
10.4 Lucene 的索引文件格式简述	234
10.4.1 段 (Segment)	235
10.4.2 索引文件	235
10.4.3 复合索引格式	236
10.4.4 索引转换	236
10.5 索引的存放位置——FSDirectory 与 RAMDirectory	236
10.5.1 在磁盘中存放——FSDirectory	237
10.5.2 在内存中存放——RAMDirectory	238
10.5.3 索引的合并	239
10.6 从索引中删除文档	241
10.6.1 删除索引的工具——IndexReader 类	241
10.6.2 删除与反删除某个特定的文档	242
10.6.3 按字段 (Field) 来删除	244
10.7 Lucene 的索引优化	245
10.7.1 为什么要优化索引	245
10.7.2 优化索引的方法——optimize()	245
10.8 Lucene 索引的同步机制	246
10.8.1 同步法则	246
10.8.2 Lucene 的索引“锁”	246
10.9 总结	247
第 11 章 Lucene 搜索	248
11.1 搜索的流程	248
11.1.1 初始化 Lucene 的检索工具 IndexSearcher	248
11.1.2 构建 Query	249
11.1.3 搜索并处理返回结果	249
11.2 搜索与结果	249
11.2.1 检索工具——IndexSearcher 类	249
11.2.2 检索结果——Hits	254
11.3 Lucene 的评分机制	260