开发自己的

搜索引擎

Lucene + Heritrix（第2版）

◎ 邱哲 符滔滔 王学松 编著

畅销书升级 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例，可直接应用于项目
版本最新 采用了最新的Heritrix-1.14.0版、HTMLParser 1.6.0版、DWR 2.0.5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

人民邮电出版社
POSTS & TELECOM PRESS

开发自己的搜索引擎

Lucene + Heritrix（第2版）

◎ 邱哲 符滔滔 王学松 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+Heritrix / 邱哲, 符滔滔, 王学松编著. -- 2版. -- 北京 : 人民邮电出版社, 2010.1
ISBN 978-7-115-21529-1

Ⅰ. ①开… Ⅱ. ①邱… ②符… ③王… Ⅲ. ①计算机网络－程序设计 Ⅳ. ①TP393.09

中国版本图书馆CIP数据核字(2009)第182347号

内 容 提 要

本书是一本介绍搜索引擎开发的书籍，通过本书，读者可以独立构建一个企业级的搜索引擎网站。本书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。

本书是一本使用 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。

本书适合 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——Lucene+Heritrix（第 2 版）

◆ 编　著　邱　哲　符滔滔　王学松
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
北京艺辉印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：36
字数：773 千字　　2010 年 1 月第 2 版
印数：1－3 500 册　　2010 年 1 月北京第 1 次印刷

ISBN 978-7-115-21529-1

定价：69.00 元（附光盘）

读者服务热线：(010)67132692　印装质量热线：(010)67129223
反盗版热线：(010)67171154

第2版 前言

2007年初，我们编写了本书第1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1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Heritrix 1.14.0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。

（6）对于第12章，升级了内核代码版本，使用Heritrix 1.14.0版本，增加了网络爬虫Heritrix代码工程导入和配置的详细步骤。修改了具体的代码，解决由于来源网站内容变更造成的部分代码无法执行的问题。

（7）对于第13章，增加了对网页内容分析的概述和基本说明，便于读者更方便地理解相关内容。使用HTMLParser 1.60，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。

（8）对于第15章和第16章，升级核心代码为DWR 2.0.5版本。针对技术手段，从搜索引擎交互方式角度进行了分析和介绍。增加了代码部分的图例说明。

本书由邱哲、符滔滔、王学松统筹编写，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

2009年10月

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 多万篇的网页。在 Google 上查找时，可以查到 750 万个网页。不是 Google 的网页少，如果用“search engine”做关键字查找时，在 Google 中可以查找到 3 亿篇以上的网页！

再来做个有趣的实验。在 www.china-pub.com 中输入“搜索引擎”这个关键词后，只有可怜的七本书被查找了出来。

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了。可是，为什么一个在 Google 中可以查找到 3 亿网页的关键字，在国内最大的计算机网上书店中只能找到可怜的、与之相关的七本书？

300000000 网页 VS 7 本书？

本书特点

由于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。因此，可以说本书是一本详细介绍搜索引擎开发过程的图书。

（1）采用的是 Lucene 2.0。以前大家用的 1.4.3 版本，而 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，这样可以帮助读者了解 Lucene 的更多新功能。

（2）本书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。

（3）着重解决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。

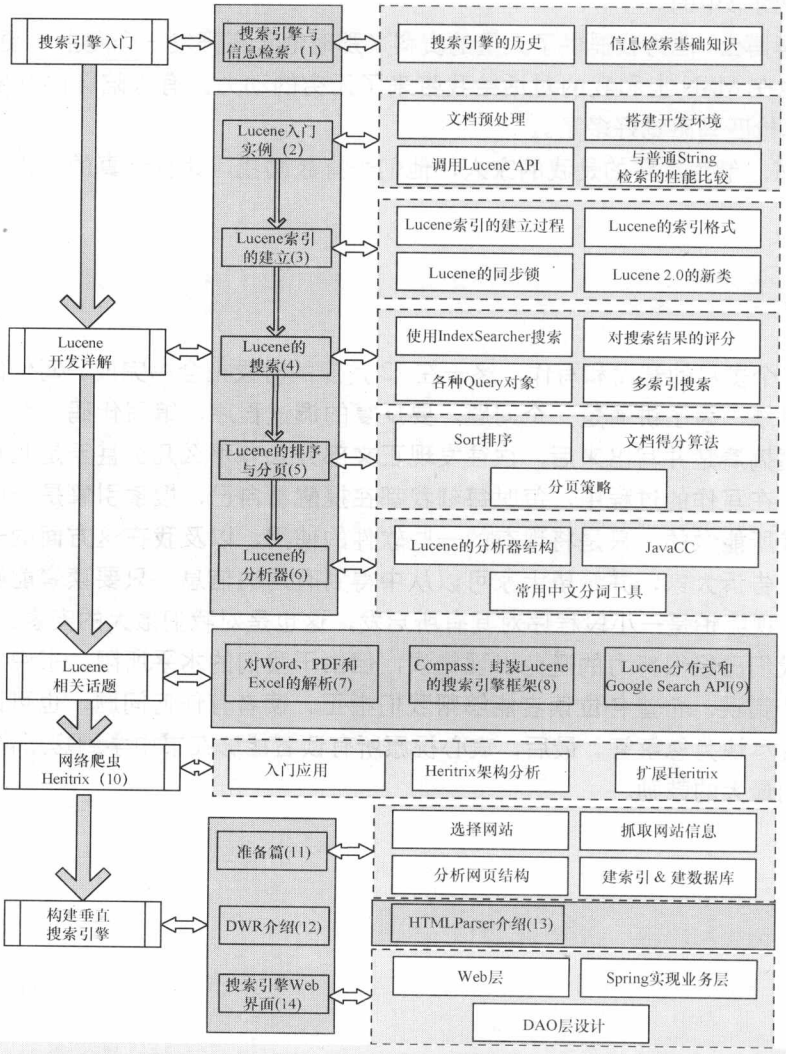
（4）内容新颖、实用。本书介绍了 Compass、Heritrix、DWR 和 HTMLParser 等内容。在搜索引擎开发的过程中，这些均为相当重要且实用的技术，笔者经过自身实践将其展现给读者，希望能让读者在学习 Lucene 的同时开拓视野。

售后服务

我们为本书开通了专用的 BLOG，网址是 http://lucenebook.spaces.live.com/，读者可以直接同我们交流，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。读者可以随时同我们联系。

本书的内容

本书内容非常丰富，细节可以参考目录。在这里，笔者给出全书的结构图，让读者有一个总体的认识，其中深色背景，并且后面有（1）、（2）字样的，表示相应的章节。



感谢

在写这段前言的时候，ZZQ 正在忙着校对我们的稿件，要向他表示感谢，因为他是我见过的最为认真负责的审稿人。同时，这本书的写作过程中，还要感谢以下几个人。

- 吴萌野：可以说，如果没有他的指导，我不可能了解 Heritrix 的使用，它的技术水平我真是服了，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，也是一个 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，要比一个人单枪匹马感觉好多了。
- 当然，最为重要的是我的家人，他们一直鼓励我，让我认真的工作，是我坚强的后盾。

结束语

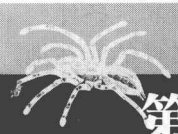
经过 6 个多月的开发和写作，终于在 12 月底的时候把全书完成。写作的过程是艰苦的，有时候，为了讲清楚一个问题，要反复的调试程序，编写代码。而当花了大量精力把问题搞清楚并写出来后，往往发现正文部分只有那么几页甚至是几行，但是我想说的是，在写作的过程中，每时每刻我都在提醒着自己，搜索引擎是一门博大精深的学科，我所能做的，只是将现有的一些软件的使用，以及我在这方面的一些微薄经验拿出来，告诉大家，并希望大家可以从中得到有用的信息。只要读者能够在阅读的过程中，发现哪怕是一小段程序对其有所启发，这也是对我们最大的安慰。

虽然我们已经很努力的避免出现错误，但由于我们的水平所限，书中不可避免的会出现一些错误，希望各位读者能够帮我们指正。读者有任何问题，也可以反馈给我们，我们会尽快为您解答。最后，衷心祝愿所有读者能够在书中学到您所需要的知识，这是对我们最大的鼓励。

邱哲 符滔滔

2007 年 1 月 31 日

目　录



第 1 章　搜索引擎与信息检索

Google 的巨大成功让整个世界都把眼光投入到搜索引擎这个领域中。一夜间，各种各样的搜索服务席卷而来，从最初的 Google、Yahoo 到现今的 Baidu、MSN、中搜、Sogou 等，搜索引擎的品牌越来越多，服务也越来越丰富。同时，伴随着 Web2.0 的疯狂普及，网络信息的膨胀速度成指数急速增长，各种各样的网站都需要为其加入检索功能，以满足用户的需要。另外，在企业级应用的市场上，全文信息检索的需求也一直在增加，各种文档处理、内容管理软件都需要加入全文检索的功能。

在这样的背景下，搜索引擎的技术迅速发展。各种讨论搜索的文章、杂志、论文铺天盖地，论坛和博客上的帖子也是层出不穷。一时间，搜索技术成为最为热门的技术之一。

不过，搜索引擎技术并非是一种大众技术，从其出现开始，就一直是一种高门槛的技术，它的后台包括学术领域的众多先进思想和设计，其涉及的学科包括自然语言处理、人工智能、离散数学、排列组合、编译原理等。因此设计一个性能良好，并且实用性强的搜索引擎并非易事。

本书不研究上述多种学科与搜索引擎的关联理论，但是读者了解和掌握搜索引擎技术的方方面面，会对阅读有很大的帮助。因此，在本章中，将带领读者了解一下搜索引擎和信息检索的基础知识、发展历史、现今状况等内容。

1.1 搜索引擎的历史

在互联网发展的最初阶段，网站的数量相对较少，信息查找比较容易。随着互联网爆炸性地发展，用户很难找到所需的资料。这时，搜索引擎的需求就出现了，一些为满足大众信息检索需求的专业搜索网站也就应运而生。

1.1.1 萌芽：Archie、Gopher

1. Archie

事实上，搜索引擎的的诞生追溯到 1990 年，在加拿大蒙特利尔（Montreal）的麦吉尔大学，一个学生制作了一个自动索引互联网上匿名 FTP 网站文件的程序。

这个能够自动索引互联网上匿名 FTP 网站文件的程序，被他们称为 Archie。Archie 是

互联网上用来查找文档的自动搜索服务工具，这些文档的标题必须满足特定条件。

通常，为了从匿名 FTP 服务器上下载一个文件，必须知道这个文件的所在地，同时必须知道这个匿名 FTP 服务器的地址，及文件所在的目录名。Archie 可以帮助用户在遍及全世界的千余个 FTP 服务器中寻找文件。Archie Server 又被称作文档查询服务器。用户只要给出所要查找文件的全名或部分名字，文档查询服务器就会指出在哪些 FTP 服务器上的哪个路径下存放着这样的文件。使用 Archie 进行查询的前提是，输入要查找的文件名或部分文件名，知道某个或几个 Archie 服务器的地址。

如今，提供 Archie 服务的网站已经很少了，笔者在 Google 上查找了一下，链接到了一个波兰的网站，仍在提供着 Archie 服务，如图 1-1 所示。有兴趣的读者可以上去一看。

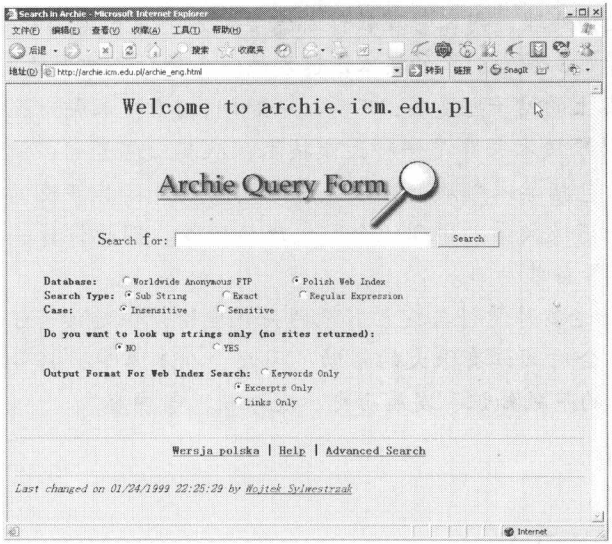


图 1-1　一个 Archie 网址

从概念上讲，Archie 的工作十分简单。每隔一段时间，一个特殊的程序连到每一个已知的匿名 FTP 主机，然后下载所有公共文件的完整目录表。这些表存储于 Internet Archives Database（Internet 档案数据库）中。当用户要求 Archie 检索一个文件时，所要进行的工作就是对该数据库进行检索。

2. Gopher

受其启发，美国明尼苏达大学的一个学生 Mark McCahill，于 1991 年发明了一种叫 Gopher 的搜索协议。Gopher 的命名来自于这所学校的吉祥物。这种协议与 Archie 最大的不同是，Archie 仅能够索引网络上的文件，而 Gopher 却可以对网页也进行索引。同时，另外两个程序 Veronica 和 Jughead 用来对以 Gopher 格式进行索引的文件进行检索。Veronica（Very Easy Rodent-Oriented Net-wide Index to Computerized Archives）是指非常方