MANNING

覆盖 Apache Lucene 3.0

Lucene 实战

（第2版）

Lucene IN ACTION

SECOND EDITION

〔美〕Michael McCandless
〔美〕Erik Hatcher 著
〔美〕Otis Gospodnetic

牛长流 肖宇 译

人民邮电出版社
POSTS & TELECOM PRESS

MANNING

Lucene 实战

（第2版）

Lucene IN ACTION
SECOND EDITION

〔美〕Michael McCandless
〔美〕Erik Hatcher 著
〔美〕Otis Gospodnetic
牛长流 肖宇 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Lucene实战 / (美) 麦肯德利斯 (McCandless,M.) , (美) 哈彻 (Hatcher,E.) , (美) 高斯波纳提克 (Gospodnetic,O.) 著 ; 牛长流, 肖宇译. -- 2版. -- 北京 : 人民邮电出版社, 2011.6
ISBN 978-7-115-25177-0

Ⅰ. ①L… Ⅱ. ①麦… ②哈… ③高… ④牛… ⑤肖… Ⅲ. ①互联网络一程序设计 Ⅳ. ①TP393.4

中国版本图书馆CIP数据核字(2011)第053293号

版 权 声 明

Lucene 实战（第 2 版）

◆ 著 [美] Michael McCandless Erik Hatcher Otis Gospodnetic

译 牛长流 肖宇

责任编辑 杜洁

◆ 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 电子邮件 315@ptpress.com.cn

网址 http://www.ptpress.com.cn

三河市潮河印业有限公司印刷

◆ 开本：800×1000 1/16

印张：30.5

字数：669 千字 2011 年 6 月第 1 版

印数：1－3 000 册 2011 年 6 月河北第 1 次印刷

著作权合同登记号 图字：01-2009-3793 号

ISBN 978-7-115-25177-0

定价：69.00 元

读者服务热线：(010)67132705 印装质量热线：(010)67129223

反盗版热线：(010)67171154

目录

第 1 部分 Lucene 核心

第 3 部分 案例分析

第 1 部分

Lucene 核心

本书前半部分涵盖了 Lucene 的 Java 程序包相关内容。第 1 章“初识 Lucene”是对 Lucene 的总体概述，你可据此开发一个完整的索引和搜索程序。每个后续章节都深入阐述了一个具体部分。第 2 章“构建索引”和第 3 章“为应用程序添加搜索功能”均为使用 Lucene 的第一步。第 4 章“Lucene 的分析过程”将深入到索引过程，帮助我们理解 Lucene 是如何对文本进行索引的。

通过对前 4 章的学习，你会较好地理解 Lucene 的一些基本功能。但 Lucene 真正的亮点在于搜索，所以我们将在本书前半部分结束前用两章的篇幅来介绍它，分别是：第 5 章“高级搜索技术”，这些高级搜索技术都基于 Lucene 的内置特性；第 6 章“扩展搜索”，展示了 Lucene 针对定制目的而提供的可扩展性。

第 1 章　初识 Lucene

本章要点

- 了解 Lucene
- 理解典型的搜索程序结构
- 使用基本的索引 API
- 使用搜索 API

Lucene 其实是一类强大的 Java 搜索库，它能让你很轻易地将搜索功能加入到任何程序中。近年来 Lucene 变得非常流行，同时它也是使用最为广泛的信息搜索库：它能够增强很多 Web 站点和桌面应用程序的搜索能力。尽管当初它是用 Java 编写的，由于使用太广泛，以及热心开发人员的努力，目前你已经可以自由获取大量的针对其他编程语言的 Lucene 移植版本（其中包含 C/C++、C#、Ruby、Perl、Python 以及 PHP 版本）。

简单易用是 Lucene 广受欢迎的关键因素之一，但是不要被这点所迷惑：后台复杂、设计先进的信息检索技术其实一直在不为人知地运行着。Lucene 是一款设计非常优秀的软件，它向用户提供了简单易用的索引和搜索 API，并屏蔽了复杂的内部实现过程。当开始使用 Lucene 时，你不必深入了解它的信息索引及检索的工作原理。同时由于 Lucene API 简单直接，你只需要学会如何使用它提供的类就可以了。再者，对于早已厌倦了臃肿软件的你而言，会惊奇地发现 Lucene 的核心 JAR 包是如此短小精悍——仅有 1MB 大小——并且不需要其他任何依赖的 JAR 包！

在本章中，我们将分析一款典型搜索程序的总体架构，以及 Lucene 在其中的使用场合。

需要指出的是，Lucene 仅仅是一个提供搜索功能的类库，所以你还需要根据实际情况自行完成搜索程序的其他模块（例如网页抓取、文档处理、服务器运行、用户界面和管理等）。在具体分析过程中，我们将首先通过一些现成的代码示例，为你展示如何使用 Lucene 进行基本的索引和搜索实现，然后简要地介绍在索引和搜索过程中需要了解的全部核心知识点。我们处在一个信息爆炸的时代，需要解决的首要问题就是具备强大的搜索能力。

NOTE Lucene 是一个快速发展的开源项目。当你读到此处时，很可能 Lucene 的很多 API 和特性都已经发生改变。本书内容基于该项目的 3.0.1 版本，由于 Lucene 版本是向后兼容的，本书所有代码示例都可以在后续 3.x 版本中编译和运行。如果在这过程中你遇到任何问题，请发送邮件至 java-user@lucene.apache.org。Lucene 社区规模巨大，他们热情并且反馈迅速，一定能够为你提供帮助。

1.1 应对信息爆炸

为了认识我们这个复杂的世界，人们发明了各种各样的方案来对信息进行分类和组织。在图书馆使用的杜威十进制分类法（Dewey Decimal System）就是层次分类方法的一个经典案例。

随着互联网的普及以及数字信息的爆炸式增长，人们已经可以足不出户地接触到海量信息。随着数据量的日益剧增，我们迫切需要采用全新的、更为动态化的方法来查找所需要的信息（如图 1.1 所示）。尽管我们可以对数据进行分门别类，但从成千上万的类别或者子类别中查找信息已经不再是一种行之有效的方法了。

如今人们需要在浩如烟海的数据中快速查找所需信息，这不仅仅体现在互联网领域中——因为台式计算机的数据存储量也随着硬盘存储能力的提高而激增。通过改变目录，展开或收起文件夹层次结构，已经不再是一种访问存储文档的有效方法。此外，人们不仅要使用计算机的原始计算功能，而且要用它进行通信交流、多媒体播放和存储等。这类应用需要计算机能够快速查找某个特定的数据片段；同时，我们还需要能够方便地查到诸如图片、视频和音频文件等各式各样的多媒体文件（Rich Media）。

我们一边要面对如此大量的数据信息，一边又在吝惜自己宝贵的时间资源，为了解决这个矛盾，我们必须找到一种灵活、自由和及时的数据查询方法，以便能够在花费最小精力代价的情况下，用这种方法快速穿越各种严格的分类界限，准确找到我们需要的信息。

为了说明在互联网和台式计算机中使用广泛的搜索应用，如图 1.1 所示是在 Google 中搜索 Lucene 的情形。图 1.2 展示了 Apple Mac OS X Finder（类似于 Microsoft Windows 资源管理器）以及它右上方显示的内嵌搜索功能。Mac OS X 音乐播放器 iTunes，同样具有内嵌搜索功能，如图 1.3 所示。

因此，搜索功能可以说是无处不在！所有主流操作系统都内嵌了搜索功能。对于

Mac OS X 系统来说，它的突出特性在于，它集成的索引和搜索功能涵盖了所有文件类型，包括具有大量元数据的文件类型，比如电子邮件、通讯录等[1]。

图 1.1 在互联网上利用 Google 进行搜索

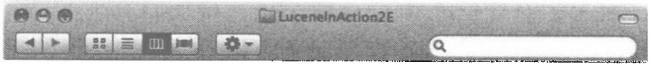

图 1.2 Mac OS X Finder 及其内嵌的搜索器

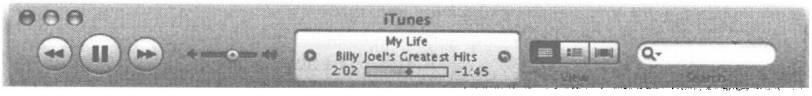

图 1.3 Apple's iTunes 嵌入直观的搜索功能

不同的人都在为解决同一个问题而奋斗着——信息量太过庞大——从而采取不同的方法来处理它。一些人致力于不断推出新的用户界面，一些人则是推出智能助理，还有一些人致力于开发复杂的搜索工具或者类似于 Lucene 的搜索库来解决这个问题。在进入示例代码之前，我们先从总体上介绍一下 Lucene 是什么，Lucene 能够做什么，以及它的发展历程。

1.2 Lucene 是什么

Lucene 是一款高性能的、可扩展的信息检索（IR）工具库。信息检索是指文档搜索、文档内信息搜索或者文档相关的元数据搜索等操作。Lucene 能够融入到你的应用程序

[1] 本书作者 Erik 和 Mike 对 Apple 相关的技术都颇感兴趣。

中，以增加搜索功能。它是一款以 JAVA 实现的成熟、自由、开源的软件项目，也是 Apache 软件基金（Apache Software Foundation）中的一个项目，并且基于 Apache 软件许可协议授权。因此，Lucene 在近年来已经成为最受欢迎的开源信息检索工具库。

NOTE 本书中，我们将一直使用信息检索（或它的英文缩写 IR）这个术语来描述 Lucene 这一类型的搜索工具。人们通常认为信息检索库就是搜索引擎，其实两者是有区别的，我们不能混淆信息检索库与 Web 搜索引擎这两个概念。

使用 Lucene 后你会很快发现，它为你提供了一套简单而强大的核心 API，并且在使用它们时你不必深入理解全文索引和搜索机制。你只需要掌握 Lucene 中少数几个类就可以将它集成到你的应用程序中了。由于 Lucene 只是一个 Java 类库，对于不同的索引和搜索内容它是通用的，相对于其他大量的搜索程序来说，这是个很大的优势。Lucene 的设计紧凑而简单，能够很容易地嵌入桌面应用程序当中。

在 Lucene 的核心 JAR 包之外，有大量的扩展模块，它们提供一些附加功能。其中一些功能对于几乎所有搜索程序来说都是至关重要的，譬如 spellchecker 和 highlighter 模块。这些模块位于一个我们称之为 contrib 的单独区域，本书会多次提及这类 contrib 模块。这些模块的数量太大，所以我们将在第 8 章和第 9 章分两章来详细讲解它们！

Lucene 的站点链接为 http://lucene.apache.org/java，在这里你可以更详细地了解 Lucene 的最新状况。主要有：新手教程、Lucene 最近所有版本 API 的 Java 帮助文档、问题跟踪系统、版本下载链接以及 Lucene 的维基链接 http://wiki.apache.org/lucene-java，该维基链接包含了大量的由 Lucene 社区更新和维护的网页。

很可能你在使用 Lucene 的时候自己却并不知道！因为 Lucene 所提供的搜索功能目前正以令人难以置信的速度大量应用于各种场合：网飞公司（NetFlix）、掘客（Digg）、MySpace 社交网站、LinkedIn 专业网络社区、联邦快递（Fedex）、苹果公司（Apple）、特码捷票务公司（Ticketmaster）、SalesForce.com 网站、大不列颠百科全书光盘、日蚀集成开发环境（Eclipse IDE）、梅奥医学中心（the Mayo Clinic）、New Scientist 杂志社、Atlassian 软件公司（JIRA）、Epiphany 浏览器、麻省理工学院在线课件（OpenCourseWare）和数字空间系统（DSpace）、Hathi Trust 数字图书馆、Akamai 公司的前端运算（Edge Computing）平台等。或许你的名字也即将出现在这个列表中！Lucene 维基页面的技术支持分页有更多的用户列表。

1.2.1 Lucene 能做些什么

人们初次接触到 Lucene 时，很容易将它和一些即用型程序搞混淆，比如文件搜索程序、网页搜索器以及网站搜索引擎等。其实这并不是 Lucene 的真面目：Lucene 只是一个软件类库，或者一个工具箱，而并不是一个完整的搜索程序。Lucene 专注于文本索引和搜索功能，并且运行效果非常不错。Lucene 能够让你的应用程序在不用了解复杂的

索引和搜索实现的情况下，通过调用它的一个简单易用的 API，就能够按照固定规则来进行事务处理。这时你的整个程序将围绕 Lucene 这个核心来运行。

很多完整的搜索程序其实都是建立在 Lucene 核心之上的。如果你正在寻找一些成型的网页搜索程序、文档处理程序以及搜索引擎，可以选择 Lucene 维基页面在其技术支持分页所列出的一些现成的应用程序。

Lucene 允许你向自己的应用程序中添加搜索功能。Lucene 能够把你从文本中解析出来的数据进行索引和搜索。Lucene 并不关心数据来源、格式，甚至不关心数据的语种，只要能把它转换为文本格式即可。也就是说你可以索引和搜索存储在文件中的如下数据：远程 Web 服务器上的网页、本地文件系统中的文档、简单的文本文件、Word 文档、XML 文档、HTML 文档或者 PDF 文档，或者其他能够从中提取文本信息的数据格式。

同样，你也可以利用 Lucene 来索引存储在数据库中的数据，以给你的用户提供一些其他数据库所不具备的诸如全文搜索等功能。一旦你的应用程序集成了 Lucene，用户就可以进行诸如 `+George +Rice -eat-pudding`、`Apple-pie+Tiger`、`animal:monkey AND food:banana` 等有着复杂查询条件的搜索。有了 Lucene，你可以为电子邮件信息、归档邮件列表、即时聊天信息以及维基（Wiki）页面等信息进行索引和搜索。下面让我们来回顾一下 Lucene 的历史。

1.2.2 Lucene 的历史

Lucene 最初是由 Doug Cutting 编写的[1]。当时其工具包可以从 SourceForge 的 Lucene 主页下载。在 2001 年 9 月，Lucene 加入 Apache 软件基金会的 Jakarta 家族并开始提供高质量 Java 开源软件产品；2005 年该项目一跃成为顶级的 Apache 项目。目前，Lucene 包含大量的子项目，具体可以通过 http://lucene.apache.org 了解。本书主要包含 Lucene 项目的 Java 子项目，详见 http://lucene.apache.org/java，但大多数人一般将这个 Java 子项目称为 Lucene 项目。

自那以后，每个 Lucene 版本的发布都使得这个项目备受关注，吸引着更多的用户和开发人员的眼球。截止 2010 年，最新的 Lucene 版本号是 3.0.1。表 1.1 列出了 Lucene 的版本发布历史。

表 1.1 Lucene 的版本发布历史

版　本	发布日期	里程碑
0.01	2000 年 3 月	在 SourceForge 网站第一次发布开源版本
1.0	2000 年 10 月	
1.01b	2001 年 7 月	在 SourceForge 网站最后一次发布
1.2	2002 年 6 月	在 Apache Jakarta 第一次发布
1.3	2003 年 12 月	加入复合索引文件格式，增强了 QueryParser、远程搜索、Token 定位、扩展站的评分 API 等

[1] Lucene 是 Doug 妻子的中名，同时这也是她外祖母的姓。