入门：引导读者快速掌握Lucene和Nutch的使用方法

揭秘：深度剖析搜索引擎内核

实战：手把手带您构建企业级搜索引擎

推荐：Web开发专家强烈推荐

本书全部源代码

Lucene + Nutch 搜索引擎开发

王学松 ◎ 编著

人民邮电出版社
POSTS & TELECOM PRESS

王学松 ◎ 编著

Lucene + Nutch 搜索引擎开发

人民邮电出版社
北京

图书在版编目（CIP）数据

Lucene+nutch 搜索引擎开发 / 王学松编著. —北京：人民邮电出版社，2008.8
ISBN 978-7-115-18216-6

Ⅰ. L… Ⅱ. 王… Ⅲ. 因特网—程序设计 Ⅳ. TP393.4

中国版本图书馆 CIP 数据核字（2008）第 077691 号

内 容 提 要

本书以 Lucene 构建搜索引擎的开发过程为主线，由浅入深，循序渐进，为读者展示如何使用 Lucene 开发自己的搜索引擎系统。全书内容包括搜索引擎概述和原理、Lucene 部署安装、Nutch 网络蜘蛛与数据获取、Lucene 索引建立、Lucene 检索与查询、搜索结果排序、文档分析器与中文分词、格式化文本分析、分布式搜索与缓存等。为便于读者理解搜索引擎快速开发过程，本书最后几章进行了应用实例的讲解，包括 Nutch 构建专题搜索、Lucene 构建企业级搜索实例以及相关的整体工程性能测试。

本书适合对搜索引擎开发有兴趣的读者阅读，包括搜索引擎开发的初学者、高等院校、信息专业学生、从事搜索开发的程序设计人员等。

Lucene+nutch 搜索引擎开发

◆ 编　　著　王学松
责任编辑　黄　焱

◆ 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
三河市潮河印业有限公司印刷

◆ 开本：800×1000　1/16
印张：29
字数：622 千字　　2008 年 8 月第 1 版
印数：1 – 4 000 册　　2008 年 8 月河北第 1 次印刷

ISBN 978-7-115-18216-6/TP

定价：59.00 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

前　言

搜索引擎技术是互联网资源导航和访问的重要手段。但是对于一般开发者而言，搜索引擎的底层开发技术过于复杂，加上各大搜索引擎厂商对核心技术严格保密，使搜索引擎开发有很高的门槛。开源搜索引擎项目 Lucence 和 Nutch 的出现改变了这一现状。使用这两个软件，普通开发者能够快速搭建搜索引擎应用。

Lucence 和 Nutch

Lucence 和 Nutch 系统使用 Java 语言开发，内部的代码和技术细节全部公开，为搜索技术提供了快速实现方法。由于采用了跨平台的开发语言，在各种开发平台的应用中有很好的适用性。

Lucence 提供了强大的全文检索功能，在桌面检索系统、网站级邮件列表、网站站内索引、企业级内部文档管理与检索、情报分析系统、知识管理系统、图书馆检索系统中都能够很好地应用，甚至在部分覆盖了上亿乃至数十亿网页的搜索引擎中也可以应用。Nutch 是 Lucence 得到广泛应用和认可之后出现的搜索引擎系统，内部使用了 Lucence 的索引管理、存储和检索技术，并进一步封装成一个真正的搜索引擎。两个系统一起完成了搜索引擎从网页下载、文本分析、索引生成、索引存储、信息检索等各个层面的应用。

本书的由来

本书编者在搜索引擎领域从事开发工作多年，非常希望有机会把自己工作中的一些积累和心得与同行共享交流。长期以来由于系统设计与开发工作繁忙，一直未能如愿。在 2007 年，有机会放下手头的工作从事知识工程方面的研究，用了一年时间完成本书的写作。本书以实例代码的形式介绍了使用 Lucene 构建搜索引擎的基本架构，力求让读者通过阅读本书，掌握使用 Lucene 开发搜索引擎的基本知识并具备一定的项目实践能力。

本书的特点

1．体系完整，内容新颖

本书采用最新版本进行讲解，内容全面，涉及 Lucene 搜索系统的安装、开发和提高，既能指导新手快速入门，又能为有经验的读者提高进阶的能力。本书覆盖了搜索

引擎应用开发的方方面面，如信息下载、文本分析、索引建立、信息检索等，对高性能搜索系统也有相关的描述。

2．注重实效，工程实用

本书以实际工程应用为主线，从实际应用需求、基本结构、具体代码实现和应用效果展开讲解，重视技术的应用。所有实例代码都经过调试和测试，保证代码可用。选择垂直搜索和专题搜索等实际工程应用实例，并做适当的简化、抽象。

3．通俗易懂，条理清晰

本书对复杂的技术内容进行整理和梳理，把复杂的搜索技术原理，以高度概括和通俗的语言进行描述，便于读者理解。层次化展开每个功能点，方便不同层面读者选择读取。

4．图例丰富，步骤详尽

采用大量的直观结构图和原理图，生动、形象地展示枯燥的信息检索技术问题，方便读者阅读。对 Lucene 和 Nutch 系统安装和环境设置，按照实际操作给出完整的过程，便于快速学习使用，并给出代码实例的开发和测试步骤，方便读者自己测试和调试。

5．提供学习光盘和学习社区

为了方便读者学习，本书配套光盘提供实例代码、开发包以及部分测试数据。另外本书有专门的学习和讨论社区（www.rzchina.net），帮助大家解决阅读中遇到的问题，并为搜索引擎爱好者提供一个互相交流的场所。

本书的内容

本书共分为 3 篇 12 章，具体内容如下。

第 1 篇入门篇，第 1 章至第 3 章。

第 1 章：从搜索引擎概念出发，介绍搜索引擎的基本结构和发展历史，并对全球知名搜索引擎和国内著名搜索引擎进行展示和评述，此外本章还介绍了基本的开发工具和测试环境。

第 2 章：通过对搜索引擎原理的分析和探讨，揭示了搜索引擎的基本实现方法。通过 Java 程序，对简化的搜索引擎模型进行了原理分析并实现设计代码。实现的代码没有考虑复杂的搜索引擎应用环境，但是其中的代码稍加改造就可以用于数据抓取、文本分析等基本工作。

第 3 章：介绍了主流开源搜索引擎 Lucene 及其派生项目 Nutch 的历史。讲解了 Lucene 全文检索系统的部署和简单应用示例的开发过程，讨论了 Nutch 开源搜索引擎系统的部署和代码加载、调试过程。

第 2 篇内核揭秘篇，第 4 章至第 10 章。

第 4 章：介绍网络蜘蛛的原理，详细分析了网络蜘蛛的工作机制和体系结构。用

下载实例来展示 Nutch 系统的网络蜘蛛的强大功能。简单介绍了运行环境、测试和检索方法，从整体上给读者提供丰富的信息。

第 5 章：介绍搜索引擎和全文检索的索引原理和机制。从全文检索的原理出发，通过实例讲述如何创建 Lucene 索引、如何管理索引。对索引的目录结构和内部结构进行分析，介绍了 Lucene 提供的高级功能和属性。最后分析了 Nutch 如何使用 Lucene 的索引功能创建搜索引擎。

第 6 章：介绍全文检索的查询和检索原理，介绍搜索引擎和全文检索的用户查询的原理和机制。通过实例讲解如何访问 Lucene 索引、如何确定不同需求的查询检索。通过对查询的主要功能类的分析，介绍了 Lucene 提供的高级检索查询功能和属性。最后介绍了 Nutch 如何使用 Lucene 的查询检索功能完成搜索引擎的用户服务。

第 7 章：从全文检索的排序算法原理出发，介绍搜索引擎和全文检索的结果排序原理和机制。通过实例讲述如何控制 Lucene 检索结果的排序，通过对 Lucene 排序公式的分析，介绍了排序的原理和内部结构，分析了 Nutch 使用链接分析的排序因素。

第 8 章：介绍搜索引擎和全文检索的预处理原理和分词机制。通过对 Lucene 和 Nutch 的分析器、分词器、过滤器的介绍，说明索引和检索的字符串在搜索引擎内部是如何处理的。通过原理和实现的分析，讲解 Lucene 分析器的应用以及如何实现自定义的分析器。

第 9 章：介绍搜索引擎和全文检索的非结构化文本处理方法，提供了每种文档的多个可选开源开发组件。通过对 PDF、DOC、XML 等不同格式文本的实例分析，为开发者提供了一系列的文本分析解决方案。读者可以灵活运用各个组件，构造一个完整的分析器。

第 10 章：介绍搜索引擎和全文检索的分布式结构方法，介绍了使用分布式文件系统及 Nutch 的分布式检索，以及使用松耦合的服务方式实现 Lucene 远程多机检索的方法。针对 Nutch 的缓存机制进行分析，并介绍了现在主流的 Java 语言 Cache 系统。

第 3 篇实战篇，第 11 章至第 12 章。

第 11 章：分析、设计、实现了一个专题搜索引擎，从需求开始分析，确定了模块和功能。使用 Nutch 开发工具，依次完成了专题搜索的各模块功能。通过实例运行和测试，展示了开源搜索引擎 Nutch 和 Lucene 的强大功能和搜索引擎的开发过程。

第 12 章：设计实现了一个企业级搜索引擎。从需求开始分析，确定了模块和功能。使用 Lucene 全文检索工具，完成企业级搜索引擎的基本功能。通过加载实例，展示了开源全文检索系统 Lucene 的强大功能。

适合的读者

- 搜索引擎初学者。

- 高校信息专业学生。
- 从事搜索开发的程序设计人员。
- 希望了解搜索技术的编程爱好者。
- 熟悉 Java 语言的各类开发人员。

本书作者

本书由王学松主笔编写，其他参与编写和资料整理的人员有曹秩倩、陈轮、董世星、陈能技、陈向辉、宫垂刚、陈鑫玮、陈垚光、程高伟、戴敏梅、邓尉、董国栋、董霖、段毅、方擎、高德波、龚小鹏、陈衍卿、韩雷、郝红旗、何俊斌、陈东、贺文婧、侯利军、胡诗群、胡添、扈新波、常建功、华剑锋 、黄洁娴 、博奎、孙健等。在此对大家的辛勤工作表示感谢。

由于水平有限、编写时间仓促，书中难免存在疏漏和不足之处，恳请广大读者提出宝贵意见。本书责任编辑的联系方式是 huangyan@ptpress.com.cn，欢迎来信交流。

编　者

2008 年 7 月

目　录

第1篇　入门篇

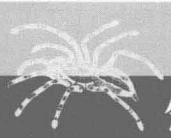
第 2 篇 内核揭秘篇

第 3 篇 实战篇

搜索引擎开发

Lucene+Nutch

第1篇 入门篇



第1章 搜索引擎概述

在近年来搜索引擎迅猛发展，百度2005年在纳斯达克成功上市，Google在全球市场突飞猛进。搜索引擎得到了前所未有的关注。实际上在此之前，人们在工作中已经大量使用搜索引擎，完成信息查找的工作。本章将对搜索引擎的应用现状、发展前景等进行介绍。

1.1 什么是搜索引擎

搜索引擎是一款特别的软件系统，能够从互联网上自动搜集信息，并为用户提供查询服务。搜索引擎对原始文档进行了一系列的整理和处理。用户的查询结果是搜索引擎按照某种规则计算获得的。搜索引擎为网民提供了资源查找和导航的有效手段。

1.1.1 搜索引擎与信息检索

搜索引擎并不是一个完全创新的系统，而是借鉴了以往全文检索系统和网络软件系统开发而成的。搜索引擎采用了以往产品的很多技术和思路，尤其是继承了很多信息检索系统的技术和方法。互联网搜索引擎在继承历史技术的同时，针对互联网信息处理的特点，开发出了互联网信息查找工具。

在搜索引擎诞生之前，完成检索任务的是信息检索系统。信息检索系统（Information Retrieval System）是一种文本处理工具，包括了信息存储、组织、表现、查询、存取等5个主要部分。信息检索通常是对文本信息的检索，其核心是文本信息的索引和查询。从历史上看，信息检索经历了手工检索和计算机检索两个主要阶段。

- 手工检索阶段是信息检索的早期阶段。手工检索来源于图书情报的索引和查找，应用在图书馆的参考咨询和文摘索引工作中。到20世纪40年代，信息检索经过近一个世纪的发展，已经成为图书馆不可或缺的工具和用户服务项目。此时的信息检索更多的是手工的编目工作。
- 计算机检索阶段是利用计算机实现自动化处理的阶段。计算机检索在20世纪60～80年代形成并发展。随着计算机技术在信息检索领域的逐步应用，计算机与信息检索理论紧密结合，形成完整的信息检索系统。信息检索的发展得到了信息处理技术、通信技术、计算机技术和数据库存储技术的推动。目前的信息检索在各领域高速发展，得到了广泛的应用。

目前信息检索技术发展到了网络化、智能化、个性化的阶段。信息检索处理的对象不

仅包含封闭、稳定、集中管理的信息内容，还包括开放、动态、更新快、分布式的网络信息。信息检索的用户也由原来的情报专业人员扩展到普通用户。普通用户没有专门的检索理论背景，要求信息检索系统提供更方便的检索方式、更人性化的检索结果。

从实现技术来看，全文信息检索技术是搜索引擎的技术基础，有一定的相似性。在处理的对象和内容上，信息检索系统与互联网搜索引擎有很大差异。互联网信息搜索和全文信息检索有很多不同，这些差别主要体现在处理的原始信息格式、数据量、结果的排序算法、用户访问的实时性、系统安全性要求上。

1.1.2 搜索引擎的概念

自 1994 年基于 Web 的搜索引擎出现以来，搜索引擎便得到了极大的发展。搜索引擎解决了海量互联网资源的快速定位和检索，在人们的日常生活和工作中发挥了越来越大的作用。搜索引擎实际上是一款网络化的软件系统，能够提供强大的检索功能，为普通用户提供互联网资源的查询和导航。从普通用户角度来看，搜索引擎通常可以接受检索查询词或短语，返回跟这些查询相关的网页标题和摘要信息。

搜索引擎的基本作用是对网站和网页内容进行信息导航，如图 1.1 所示。

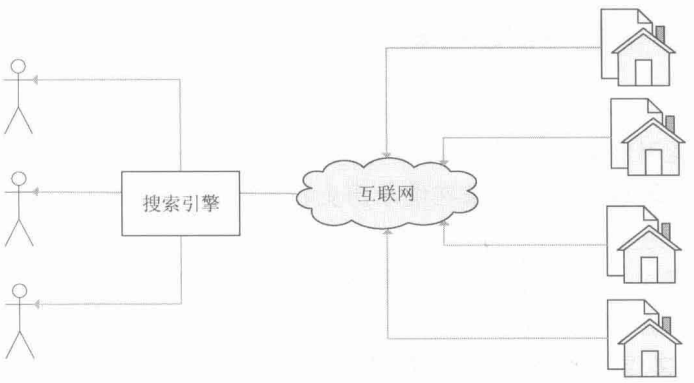


图 1.1　搜索引擎信息导航功能图

搜索引擎已经成为互联网上非常重要的网络导航服务工具。搜索引擎搜集互联网中的资源和信息，发现新的网站和网页，经过抓取和分析，存储相应的信息副本。在此基础上，进一步对信息进行理解、提取、组织和处理，并为用户提供检索服务，从而起到信息导航的目的。

1.1.3 搜索引擎的使用

搜索引擎是一个基本的互联网资源导航工具，能够从庞大的互联网中帮助用户找到目