



本书全部源代码

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

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
印数: 1—4 000 册
- 2008 年 8 月第 1 版
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 1 章 搜索引擎概述	3	1.7 未来搜索技术前瞻	28
1.1 什么是搜索引擎	3	1.7.1 现状存在问题	28
1.1.1 搜索引擎与信息检索	3	1.7.2 未来发展趋势	29
1.1.2 搜索引擎的概念	4	1.8 小结	30
1.1.3 搜索引擎的使用	4		
1.1.4 搜索引擎发展历史	5	第 2 章 搜索引擎原理探秘	31
1.2 搜索引擎分类	7	2.1 解密搜索引擎原理	31
1.2.1 按照工作方式分类	7	2.1.1 搜索引擎技术框架	31
1.2.2 按照领域范围分类	8	2.1.2 网页信息抓取技术	32
1.2.3 信息类型分类	8	2.1.3 网页内容分析技术	33
1.3 主流搜索引擎	9	2.1.4 网页索引建立技术	34
1.3.1 全球著名搜索引擎	9	2.1.5 用户检索与结果排序	34
1.3.2 中文搜索引擎的发展历史	12	2.1.6 网页检索工具与接口	35
1.3.3 著名中文搜索引擎	12	2.2 网络爬虫简单实现	35
1.3.4 其他细化搜索引擎	15	2.2.1 网络蜘蛛功能需求	35
1.4 搜索引擎评价原则	15	2.2.2 网络蜘蛛实现原理	36
1.4.1 评价指标体系	15	2.2.3 网络爬虫系统结构	37
1.4.2 其他评测因素	17	2.2.4 网页采集程序设计	38
1.5 搜索引擎相关资源	17	2.2.5 网页采集程序实现	39
1.5.1 搜索引擎开源项目	18	2.2.6 程序实现存储扩展	42
1.5.2 搜索引擎研究网站	19	2.3 网页分析程序实现	44
1.5.3 搜索论坛和厂商黑板报	19	2.3.1 网页分析功能需求	44
1.6 系统运行环境准备	20	2.3.2 网页分析实现原理	45
1.6.1 Java 环境安装设置	20	2.3.3 网页分析系统结构	45
1.6.2 Tomcat 服务器安装	22	2.3.4 网页分析程序设计	46
1.6.3 Eclipse 开发环境准备	25	2.3.5 文本语素分割与过滤	49

2.4 网页索引程序实现	52	3.2 Lucene 全文检索系统部署	65
2.4.1 网页索引功能需求	53	3.2.1 下载 Lucene 系统	65
2.4.2 网页索引实现原理	53	3.2.2 Lucene 部署配置	66
2.4.3 网页索引程序设计	54	3.2.3 Lucene 测试运行	67
2.4.4 网页索引程序实现	55	3.3 Lucene 开发实例入门	69
2.5 检索程序实现	58	3.3.1 Lucene 实例功能	69
2.5.1 检索功能需求	58	3.3.2 Lucene 开发实例	70
2.5.2 检索实现原理	58	3.3.3 代码实例解析	75
2.5.3 检索程序设计	59	3.4 Nutch 开源搜索引擎部署	77
2.5.4 网页检索程序实现	59	3.4.1 Cygwin 软件安装	77
2.6 简单搜索引擎系统	61	3.4.2 Nutch 下载与安装	79
2.7 小结	62	3.4.3 Nutch 系统环境测试	80
		3.4.4 Nutch 搜索页面部署	82
第3章 开源搜索引擎入门	63	3.5 Nutch 系统调试与开发	82
3.1 开源搜索引擎简介	63	3.5.1 Eclipse 中加载 Nutch	83
3.1.1 Lucene 系统概述	63	3.5.2 Nutch 工程编译与发布	86
3.1.2 Nutch 概述	64	3.6 小结	87

第2篇 内核揭秘篇

第4章 搜索引擎数据获取	91	4.3.4 下载多个网站	103
4.1 网络蜘蛛原理	91	4.4 Nutch 互联网抓取	105
4.1.1 体系结构设计	91	4.4.1 下载列表获取	106
4.1.2 访问策略与算法	92	4.4.2 下载大量网站	106
4.1.3 效率优化与更新	93	4.5 Nutch 抓取比较	107
4.1.4 蜘蛛访问规范	93	4.6 Nutch 结果检测	111
4.1.5 开源蜘蛛简介	93	4.6.1 网页内容检索	112
4.2 Nutch 网络蜘蛛	94	4.6.2 使用 Readdb 获取摘要	113
4.2.1 Nutch 网络蜘蛛概述	94	4.6.3 使用 SegRead 读取分段	115
4.2.2 Nutch 抓取模式分类	96	4.6.4 Luke 工具使用	116
4.2.3 抓取测试站点建立	97	4.7 Nutch 配置文件解析	116
4.3 Nutch 局域网抓取	97	4.8 Heritrix 网络蜘蛛	118
4.3.1 本地下载准备	98	4.8.1 Heritrix 概述	118
4.3.2 启动下载过程	99	4.8.2 Heritrix 体系结构	119
4.3.3 下载过程解析	101	4.8.3 Heritrix 安装与使用	122

4.9 小结	125	5.6.2 索引数据库记录	173
第5章 搜索引擎信息索引	126	5.6.3 索引优化与合并	176
5.1 文档索引原理	126	5.7 Nutch 中的 Lucene 索引	178
5.1.1 索引概述	126	5.8 小结	180
5.1.2 索引基本结构	127	第6章 搜索引擎查询处理	181
5.1.3 倒排索引原理	128	6.1 信息查询原理	181
5.1.4 索引分类	128	6.1.1 信息查询概述	181
5.1.5 高性能索引	129	6.1.2 查询基本流程	182
5.2 Lucene 索引器	129	6.1.3 查询结果显示	183
5.2.1 Lucene 索引介绍	129	6.1.4 高性能查询	183
5.2.2 Lucene 索引结构	130	6.2 Lucene 查询概述	184
5.2.3 多文件索引结构	133	6.2.1 Lucene 查询操作基础	184
5.2.4 复合索引结构	134	6.2.2 Lucene 查询实例入门	184
5.3 Lucene 索引实例	135	6.2.3 查询工具 IndexSearcher 类	187
5.3.1 索引创建代码解析	135	6.2.4 查询封装 Query 类	188
5.3.2 索引创建器 (IndexWriter)	137	6.2.5 查询分析器 QueryParser 类	188
5.3.3 索引管理器 (IndexReader)	138	6.2.6 查询结果集 Hits 类	191
5.3.4 索引修改器 (IndexModifier)	139	6.3 Lucene 基本查询	192
5.3.5 索引分析器 (Analyzer)	140	6.3.1 Lucene 查询 Query 对象	192
5.4 Lucene 索引操作	141	6.3.2 最小项查询 TermQuery	195
5.4.1 添加文本文件索引	141	6.3.3 区间范围搜索 RangeQuery	198
5.4.2 创建 Lucene 增量索引	143	6.3.4 逻辑组合搜索 BooleanQuery	202
5.4.3 使用索引项删除文档	144	6.3.5 字符串前缀搜索 PrefixQuery	205
5.4.4 使用编号删除文档	145	6.3.6 短语搜索 PhraseQuery	208
5.4.5 压缩文档编号	147	6.3.7 模糊搜索 FuzzyQuery	211
5.4.6 索引文档更新	148	6.3.8 通配符搜索 WildcardQuery	214
5.5 Lucene 索引高级特性	151	6.3.9 位置跨度搜索 SpanQuery	217
5.5.1 选择索引域类型	151	6.4 Lucene 高级查询	224
5.5.2 索引参数优化	162	6.4.1 索引内存检索	224
5.5.3 使用磁盘索引	165	6.4.2 多关键字跨域检索	226
5.5.4 使用内存索引	166	6.4.3 多检索器跨索引检索	228
5.5.5 同步与锁机制	168	6.5 Nutch 中的 Lucene 查询	230
5.6 Lucene 高级应用实例	169	6.6 小结	231
5.6.1 创建本地搜索的索引	169		

第7章 搜索引擎结果排序 232**7.1 搜索引擎文档排序原理** 232

7.1.1 传统检索排序技术 232

7.1.2 向量模型排序局限 233

7.1.3 搜索引擎相关性排序 234

7.1.4 链接分析 PageRank 原理 235

7.1.5 搜索引擎排序流程 236

7.2 Lucene 检索排序 236

7.2.1 Lucene 相关性因素 236

7.2.2 Lucene 相关排序流程 237

7.2.3 Lucene 排序计算体系 237

7.2.4 Lucene 排序控制方法 240

7.3 文档 Boost 加权排序 240

7.3.1 Lucene 中 Boost 介绍 241

7.3.2 Boost 值全文档排序 241

7.3.3 Boost 值文档域排序 244

7.3.4 BoostingTermQuery 排序 246

7.4 Sort 对象检索排序 248

7.4.1 Sort 对象概述 248

7.4.2 Sort 对象相关性排序 249

7.4.3 Sort 对象文档编号排序 252

7.4.4 Sort 对象独立域排序 254

7.4.5 Sort 对象联合域排序 257

7.4.6 Sort 对象逆向排序 260

7.5 Lucene 相关性公式 261

7.5.1 Lucene 评分结果分析 262

7.5.2 Lucene 排序公式 264

7.5.3 其他动态排序因子 265

7.6 Lucene 自定义排序 266

7.6.1 自定义排序比较接口 266

7.6.2 自定义排序接口类实例 267

7.6.3 自定义排序结果测试实例 269

7.6.4 自定义排序测试结果 272

7.7 Nutch 中的结果排序 272

7.7.1 Nutch 排序因素 273

7.7.2 Nutch 链接分析 273

7.7.3 Nutch 相关度计算 274

7.8 小结 275**第8章 文档分析器与中文分词** 276**8.1 文档分析与中文分词原理** 276

8.1.1 文档分析预处理概述 276

8.1.2 文档分析基本流程 276

8.1.3 中文分析处理中的分词 277

8.2 Lucene 分析器内核原理 278

8.2.1 Lucene 分析器原理 278

8.2.2 Analysis 包简介 280

8.2.3 Analyzer 类的组合结构 282

8.2.4 JavaCC 构造分析器 282

8.2.5 StopAnalyzer 内核代码分析 283

8.2.6 StandardAnalyzer 内核
代码分析 284**8.3 Lucene 分析器应用模式** 285

8.3.1 使用默认分析器建立索引 285

8.3.2 使用多种分析器建立索引 287

8.3.3 使用分析器检索查询 288

8.4 Lucene 主要分析器应用实例 291

8.4.1 停用词分析器 StopAnalyzer 291

8.4.2 标准分析器 StandardAnalyzer 294

8.4.3 简单分析器 SimpleAnalyzer 297

8.4.4 空格分析器 WhitespaceAnalyzer 298

8.4.5 关键字分析器 KeywordAnalyzer 300

8.5 TokenStream 分词器内核分析 301

8.5.1 Tokenizer 分词器 301

8.5.2 标准分词器 StandardTokenizer 302

8.5.3 字符分词器 CharTokenizer 303

8.5.4 空格分词器
WhiteSpaceTokenizer 304

8.5.5 字母分词器 LetterTokenizer 304

8.5.6 小写分词器 LowerCaseTokenizer 305

8.6	TokenStream 过滤器内核分析	305	9.3.5	Filter 模式搜索链接提取	334
8.6.1	TokenFilter 过滤器	306	9.3.6	Lexer 模式遍历文档	336
8.6.2	标准过滤器 StandardFilter	306	9.4	PDF 文档分析	337
8.6.3	停用词过滤器 StopFilter	307	9.4.1	常用的 PDF 处理包	338
8.6.4	小写过滤器 LowerCaseFilter	308	9.4.2	PDFBox 安装配置	338
8.6.5	长度过滤器 LengthFilter	309	9.5	PDFBox 应用实例	339
8.6.6	词干过滤器 PorterStemFilter	310	9.5.1	PDFBox 提取文档内容	339
8.7	Lucene 中文分词	310	9.5.2	PDFBox 文档内容索引	342
8.7.1	中文分词基本原理方法	311	9.6	Office 文档分析	346
8.7.2	StandardAnalyzer 分析器 中文处理	312	9.6.1	常用 Office 文档处理包	346
8.7.3	CJKAnalyzer 中文分析器	313	9.6.2	使用 POI 安装与配置	346
8.7.4	ChineseAnalyzer 中文分析器	315	9.6.3	POI 原理与接口介绍	348
8.7.5	IK_CAnalyzer 中文分析器	316	9.7	POI 分析 Office 文档实例	348
8.7.6	中科院 ICTCLAS 中文分词	318	9.7.1	POI 处理 Excel 文档	348
8.7.7	JE 中文分词	318	9.7.2	POI 处理 Word 文档	352
8.7.8	中文分词问题	320	9.8	XML 文档分析	354
8.8	Nutch 分词和预处理	321	9.8.1	主流 XML 文档分析器	355
8.8.1	Nutch 分析器	321	9.8.2	JDOM 分析器安装配置	356
8.8.2	Nutch 中文分词	324	9.8.3	xerces 分析器安装配置	358
8.9	小结	324	9.9	XML 解析应用实例	359
			9.9.1	使用 JDOM 分析 XML 文档	359
			9.9.2	使用 xerces 分析 XML 文档	361
			9.10	Nutch 文档处理	364
			9.11	小结	364
第9章	搜索引擎文本分析	325	第10章	分布式搜索与缓存	365
9.1	非结构化文本简介	325	10.1	分布式检索与缓存	365
9.1.1	非结构化文本概述	325	10.1.1	分布式搜索引擎现状	365
9.1.2	非结构化文本检索	325	10.1.2	分布式搜索引擎原理	366
9.2	HTML 文档分析	326	10.1.3	搜索引擎缓存现状	367
9.2.1	主流 HTML 文档分析器	326	10.1.4	搜索引擎缓存原理	367
9.2.2	HTMLParser 安装配置	327	10.2	Nutch 与分布式检索	368
9.2.3	HTMLParser 的框架结构	328	10.2.1	Google 分布式文件系统	368
9.3	HTMLParser 应用实例	329	10.2.2	MapReduce 系统介绍	369
9.3.1	HTMLParser 功能模式	329	10.2.3	Hadoop 分布式文件系统	370
9.3.2	HTMLParser 内容解析方式	329			
9.3.3	Visitor 模式正文解析	330			
9.3.4	Filter 模式简单链接提取	332			

10.2.4 Nutch 分布式文件系统	372	10.3.2 Lucene 索引服务器	378
10.2.5 Nutch 分布式检索概述	372	10.4 Nutch 与搜索缓存	381
10.2.6 Nutch 分布式检索器	375	10.5 开源系统缓存系统	383
10.3 Lucene 分布式检索	376	10.6 小结	384
10.3.1 Socket 通信基础	376		

第 3 篇 实战篇

第 11 章 Nutch 专题搜索引擎实例

11.1 专题搜索需求分析	387
11.1.1 专题搜索功能需求	387
11.1.2 专题搜索用例分析	388
11.2 构建 Nutch 基础搜索引擎	389
11.2.1 Nutch 搜索功能分析	390
11.2.2 信息下载功能测试	390
11.2.3 Nutch 基础 Web 检索	391
11.2.4 Web 用户页面修改	393
11.3 专题搜索系统设计	395
11.3.1 系统框架设计	396
11.3.2 选择开发工具组件	397
11.4 专题关键词管理	397
11.4.1 专题关键词策略	398
11.4.2 关键词存储设计	398
11.4.3 关键词管理程序	400
11.5 专题资源发现	403
11.5.1 专题网页链接发现	403
11.5.2 专题资源网站提取	406
11.6 专题信息下载	407
11.6.1 批量信息下载	407
11.6.2 信息自动下载	410
11.7 专题信息分析与索引	412
11.7.1 网页信息分析	413
11.7.2 创建索引	413
11.8 检索辅助功能	414

第 12 章 Lucene 实现企业搜索实例

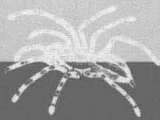
12.1 企业搜索需求分析	425
12.1.1 企业搜索需求概述	425
12.1.2 企业搜索用例分析	426
12.2 企业级搜索系统设计	427
12.2.1 系统框架设计	427
12.2.2 Lucene 检索框架	428
12.3 企业级搜索系统设计	428
12.3.1 创建 Lucene 工程	429
12.3.2 全文检索索引生成	429
12.3.3 全文检索检索页面	433
12.4 数据引擎设计	438
12.4.1 数据库数据管理	438
12.4.2 非结构化文档	439
12.5 企业信息索引	442
12.5.1 数据索引建立	443
12.5.2 信息检索代码	447
12.5.3 检索 Web 代码	449
12.5.4 检索结果测试	451
12.6 小结	452

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。

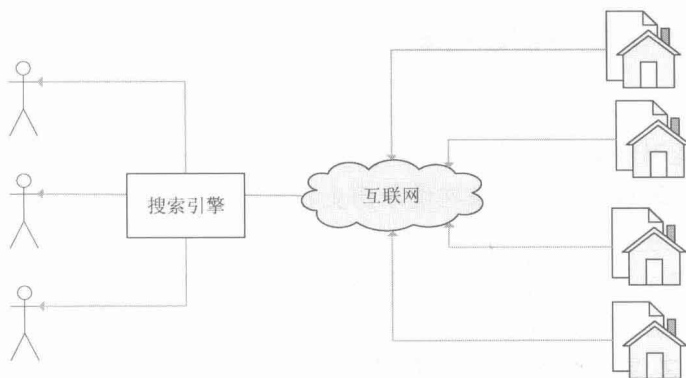


图 1.1 搜索引擎信息导航功能图

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