《Lucene+nutch搜索引擎开发》

图书基本信息

书名：《Lucene+nutch搜索引擎开发》

13位ISBN编号：9787115182166

10位ISBN编号：7115182167

出版时间：2008-8

出版社：人民邮电出版社

作者：王学松

页数：452

《Lucene+nutch搜索引擎开发》

内容概要

《Lucene+nutch搜索引擎开发》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，《Lucene+nutch搜索引擎开发》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。

《Lucene+nutch搜索引擎开发》

书籍目录

《Lucene+nutch搜索引擎开发》

《Lucene+nutch搜索引擎开发》

章节摘录

第1篇　入门篇　　第1章　搜索引擎概述　　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。

《Lucene+nutch搜索引擎开发》

编辑推荐

《Lucene+nutch搜索引擎开发》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。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。　　互联网搜索的使用水平可以反映全民的信息处理能力，几年前有研究发现美国用户比欧洲用户的互联网使用水平领先半年左右，主要是根据谁搜索时平均使用的关键词的个数多。中文用户的搜索使用水平相对于西文用户目前仍然处于比较初级的阶段，而中文网站搜索功能的缺失也是一个重要的因素。　　网站拥有了较多内容后，最先会考虑基于目录的内容分类，以解决信息快速定位的问题，随着内容量的进一步增加，很多内容在发表之后就很快被湮没，成为“信息孤岛”，而不断加深的目录结构也会让用户逐渐失去耐心，这时，关键词检索的优势就体现出来了：　　关键词检索可以让处于“信息孤岛”状态的内容以一种更直接的方法提供给用户；和基于目录/分类的树形结构不同，基于关键词检索还可以让内容之间实现网状的关联结构，从而大大提高信息的引用密度。　　基于传统数据库的关键词检索由于性能问题让很多网站放弃了搜索功能，问题的解决归根结底还是需要一个全文引擎。而Lucene开源引擎的出现让这种原来被少数公司掌握的技术得到了迅速的普及，这里应该再次感谢引擎的核心贡献者Doug Cutting先生，同时也希望有更多的中文开发人员能积极投入：到Lucene的相关项目开发中去，尤其在中文和其他亚洲双字节语言处理方面的问题。　　Lucene也是我学习的第一个Java程序，当初是通过jdb一行行dubug了解其中的原理和机制的，非常高兴有这样一本专门的参考书出现，它无疑会为开发人员了解并更快掌握全文检索技术节省大量的时间。

《Lucene+nutch搜索引擎开发》

精彩短评

1、大量的重复无用代码....
2、选择性的看了部分。对lucene中的信息索引，查询处理，结果排序及中文分词有了大致的认识。版本太老了，基本上就是api的介绍。而且有的函数已经被取消。
适合入门。
http://www.xmind.net/m/Dec6/
3、讲的比较系统，但附带的代码有点问题
4、写的不是太好
5、国内垃圾书几大技术特点：1、某教授带领多人合著的；2、代码采用中文注释的；3、不注重代码排版的；4、采用Windows系统，图片都是Windows弹出窗口和对话框的；5、官话连篇的
6、作为入门书，对搜索引擎的了解。
7、lucene 太老了。nutch开发讲的太少
8、还没看完。。。。
9、倒排索引结构，boolean模型，向量模型。
10、由于搜索引擎开发门槛在那里，作者的实践经验在那里，虽然写的还不如百度百科，虽然在08年的前沿知识在现在已经陈旧，但还是给3星安慰一把。关注了推荐的博客、网站，其他没有帮助。耗时：40min
11、不能与时俱进啊，很多API都过时了唉。看来还得自己多动手才是。
12、读研时从图书馆借来粗粗翻过，建议还不如直接看官网的帮助文档更有用。
13、对我等菜鸟帮助挺大的，如果是高手的话没必要看直接源代码走起

《Lucene+nutch搜索引擎开发》

精彩书评

1、最近做的东西有相关nutch和lucene的内容，其实这本书貌似nutch的东西没有讲很多，版本也比较老了，还不如直接网上搜索来的快，lucene倒是讲了很多，不过基本都是api的介绍，可能这样看起来比直接看文档舒服点。原理的方面也是基本的介绍了下，说的不多也不太深入。感觉如果要用lucene和nutch做东西就直接网上找答案吧，毕竟已经更新好几代了，这书里面的有些函数都被取消了。买回来当本入门书籍就可以了。
2、买了有段时间了，最近刚读完，觉得还好吧，挺系统的。没有具体调试过上面的代码，不过看书主要看原理，代码也不那么重要。
3、买了这本书，直接看这几天一直困惑自己的中文分词~~前面介绍了一大段中文分词的基本概要，和lucene的分析器后面nutch的分析器只是简单的介绍了几个类，nutch中文分词只用了200字左右。书中也没用很系统的介绍nutch如何实现中文分词，~~后面的案例也只是简单的单字切分。另外nutch最出彩的插件也没提到过~~

《Lucene+nutch搜索引擎开发》

章节试读

1、《Lucene+nutch搜索引擎开发》的笔记-第65页

这书写的还不错，就是到目前为止版本太老了，没有更新。期待作者的更新版本。

《Lucene+nutch搜索引擎开发》

版权说明