<<自己动手写搜索引擎>>

图书基本信息

书名：<<自己动手写搜索引擎>>

13位ISBN编号：9787121096402

10位ISBN编号：7121096404

出版时间：2009-11

出版时间：电子工业出版社

作者：罗刚

页数：353

前言

15在中国，随着互联网从城市到农村的普及，搜索引擎对日常生活产生越来越大的影响。
例如，笔者自己一般每天就有15个左右的问题需要求助于搜索引擎。
从04年开始笔者也从数据库相关软件开发转入搜索引擎相关开发工作。
Google20世纪末，在美国国家科学基金会的支持下，斯坦福大学的两个学生在他们的教授指导下开始了一个数字图书馆项目。
后来，他们创建了Google公司，开创了通过互联网搜索技术共享人类信息的新纪元。
Google通过网络广告取得了巨大的商业回报，到现在仍然是世界500强企业中赢利能力最强的公司之一。
NASDAQ证券交易市场的最高股价是Google公司的股票。
搜索引擎开发成为一项极有含金量的技术。
Web开始写作《自己动手写搜索引擎》这本书以前，已经有一些介绍搜索理论或者搜索开发工具的图书，但是往往表现出来的是纯粹的理论推导和公式定理，或者仅仅是现成开源软件的介绍、分析和使用，并没有介绍其理论依据。
有的读者是数学专业的博士，对于相关的数学模型一看就明白，但对于算法实现可能仍然缺少经验。
有的读者是培训学校毕业的学生，可能对Web开发框架和软件工具的使用很熟悉，但缺少理论基础和深入创新的能力。
本书的一个特点在于前面是原理介绍，接着是具体的代码实现。
不仅讲解抽象的知识，更重要的是把知识转化成具体软件应用的过程也展示出来。
Lucene《自己动手写搜索引擎》是猎兔企业搜索开发团队的软件产品研发和项目实践的经验汇总。
感谢Lucene，它把搜索引擎开发工作变成了广大程序员都能够参与的游戏。
所以本书选用Lucene来全方位展现一个商用级别的搜索解决方案。
中文分词当前仍然是实现中文搜索的热门话题之一。
本书重点介绍了中文分词的相关理论和代码实现，以及在搜索引擎中实用中文分词等细节。
本书用简单的例子介绍了搜索引擎完整的实现过程，同时也没有忽略一些经典的算法实现。
该书适合需要具体实现搜索引擎的程序员使用，对于自然语言处理等相关研究人员也有一定参考价值，同时猎兔搜索团队也已经开发出以本书为基础的专门培训课程。
本书附带光盘中的代码经过了详细的注释。
为了帮助初学者更容易地了解程序的功能，经过笔者的精心整理后，每个主要变量和每行主要的执行程序都加上了注释，前后对比图如下所示。

<<自己动手写搜索引擎>>

内容概要

《自己动手写搜索引擎》是猎兔企业搜索开发团队的软件产品研发和项目实践的经验汇总。
《自己动手写搜索引擎》全方位展现出一个商用级别的Lucene搜索解决方案，主要包括爬虫、自然语言处理和搜索实现部分。
爬虫部分介绍了网页遍历方法和从网页提取主要内容的方法。
自然语言处理部分包括了中文分词从理论到实现以及在搜索引擎中的实用等细节。
其他自然语言处理的经典问题与实现包括：文档排重、文本分类、自动聚类、语法解析树、拼写检查、拼音转换等理论与实现方法。
在实现搜索方面，《自己动手写搜索引擎》用简单的例子介绍了完整的搜索实现过程，覆盖了从索引库的设计和索引库与数据库的同步到搜索用户界面设计与实现。
搜索用户界面包括实现布尔逻辑查询、按区间范围查询、搜索结果按日期排序等。
《自己动手写搜索引擎》还进一步介绍了搜索排序的优化方法。
最后以基于Lucene的搜索服务器Solr为例，展示了Lucene的最新应用方法。

作者简介

罗刚，猎兔搜索（http：／／www．lietu．com）创始人，当前猎兔搜索在北京和上海均设有研发部。
带领猎兔搜索技术开发团队先后开发出猎兔中文分词系统、猎兔智能垂直搜索系统以及网络信息监测系统等，实现互联网信息的采集、过滤、搜索和实时监测。

<<自己动手写搜索引擎>>

书籍目录

章节摘录

1995年，两个年轻的学生Larry Page和Sergey Brin在一点上达成了共识——从大量数据中检索信息是计算系统面临的最大的挑战之一。
1996年，他们创建了一个叫做BackRub的搜索引擎。
这个搜索引擎后来叫做Google。
1998年，Page和Brin在Larry的大学宿舍创立Google公司的第一个数据中心。
2000年，Google开始成为全球最大的搜索引擎一直到现在。
2005年，Google股票市值超过1000亿美元。
公司首次登陆华尔街时，“让世界更美好”便是它们阐明的目标之一。
除了搜索引擎，集成AJAX技术的Gmail邮箱和Google地图创造了更好的Web界面视觉效果。
和Gmail邮箱集成的聊天工具Gtalk也得到广泛好评。
和MSN在客户端保存历史聊天记录不同，Gtalk聊天工具可以在服务器端保存并搜索聊天记录。
可以浏览整个地球的Google Earth也带给人们一种全新的体验。
你想起什么大脑没有记住的知识了吗？
用搜索引擎吧。
它往往不会让你失望。
事实上这正是Google的创始人设想的。
Google创始人Brin在一次用户大会上讲道：“如果你想搜就能搜，几乎像拥有第二个大脑，那就妙极了。
”就在Brin（一直是两人中站在前台的那个）说在兴头儿上的时候，Page在不声不响地露面之后，便带着“共谋者般的微笑”走出了房间。

<<自己动手写搜索引擎>>

编辑推荐

首次揭示商业级搜索引擎实现秘密。
业内知名开发团队倾情奉献。
引领Lucene开发技术升级。
《自己动手写搜索引擎》特色： 根据猎兔搜索开发团队多年搜索和自然语言处理研发经验，将国内搜索工程项目实践结合当前流行理论实现搜索技术。
全方位展现出一个商用级别的Lucene搜索解决方案，主要包括网络爬虫、中文分词和搜索实现等。
用深入浅出的方式介绍了隐马尔可夫模型等流行理论。
相关的算法全部利用Java语言实现。

版权说明