第一图书网, tushu007.com

<<迅速搭建全文搜索平台>>

图书基本信息

书名：<<迅速搭建全文搜索平台>>

13位ISBN编号：9787811231564

10位ISBN编号：7811231565

出版时间：2007-10

出版时间：清华大学

作者：于天恩

页数：287

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：http://www.tushu007.com

第一图书网, tushu007.com

前言

说说搜索引擎搜索引擎这几年热起来了。
作为世界上最大、最出名的搜索引擎，GOOgle在很多方面都发挥了重要的作用。
但是，当手中没有G00gle的搜索代码时，该如何搭建一个自己的搜索引擎呢？
业界的人士说，全新开发一套完备的企业级搜索引擎要五年的时间。
诚然，许多“业界”人士的话并不可信，不过，在搜索引擎这一块，真想要做好确实是不容易。
开发搜索引擎要耗费大量的时间和精力，所以有一些人开始研发独立的搜索引擎模块，并将其源代码开放，这样就可以给其他需要建立自己的搜索引擎的人提供一个基础平台。
在这些开源搜索引擎模块的基础上做开发，可以节约非常多的时间和精力，大大减少了开发成本，缩短了产品投入市场的周期。
而且，由于这些平台是开源的，可以亲自检查每一行代码，修改算法和显示格式等内容，这样的搜索引擎就相当于自己写的，用起来放心。
有时使用某些商业搜索模块，尽管搜索效果也很好，但是很难知道在单击“搜索”按钮的瞬间自己是否做了一些自己并不想做的事情，比如：给某个陌生人发送了一个特洛伊木马。
写这本书的动机开源搜索引擎对解决企业搜索等问题提供了可靠的二次开发平台（有的甚至不需要二次开发），大大提高了开发搜索引擎的效率，缩减了成本，好处多多。
所以，需要有一些书来介绍如何使用开源搜索模块来提供搜索服务，而目前市面上这类书籍并不多。
我编写的这本书——《迅速搭建全文搜索平台——开源搜索引擎实战教程》（以下简称《实战教程》），是《做自己的搜索引擎——搜索引擎精解案例教程》（以下简称《案例教程》）的兄弟篇，用以介绍开源搜索引擎的架构和实现。
《案例教程》和《实战教程》这两本书是非常有意义的，前者介绍搜索引擎的理论和基本应用，后者介绍在开源搜索引擎领域中如何实现搜索引擎的搭建。
有了这两本书，一个普通的程序员就可以顺利并且十分容易地掌握与搜索引擎相关的核心知识。
看过这两本书之后，就有能力深入地研究主流的开源搜索引擎的代码，之后，就成为优秀的搜索引擎工程师。
按照普通人的观点，从普通的程序员到搜索引擎工程师，这两者之间是有三级台阶的。
第一级：了解搜索引擎的原理和相关术语等基础知识。
第二级：了解现存的搜索引擎是如何运行的，懂得如何应用搜索引擎的原理去搭建搜索引擎。
第三级：认真研究一种或几种开源搜索引擎的源代码，深刻地理解其架构，从而使之成为相当于自己开发的搜索引擎。

第一图书网, tushu007.com

<<迅速搭建全文搜索平台>>

内容概要

本书作为有心进入搜索引擎业的读者的第二本基础书籍，承接其兄弟篇，讲解了开源搜索引擎的搭建过程中所要解决的基本问题，将搜索引擎这一高起点的技术讲解得清晰透彻，使其变得极为好学，没有任何神秘可言。
本书共包括5章，可以分成两个部分。
　第一部分（第1章）：建立搜索引擎的方案。
这部分用数少的文字总结建立搜索引擎的主要方案，即：常规的数据库搜索、文件搜索，基于数据库全文索引机制的搜索，利用外部非开源web搜索服务进行的搜索，以及利用开源搜索引擎实现的搜索。
　第二部分（第2--5章）：架设网络搜索引擎。
从第2章起，陆续介绍数据抓取、数据解析、建立索引和执行搜索这四项内容，它们是创建网络搜索平台所要解决的基本问题；第5章，介绍基于Hyper EStraier搜索引擎框架来搭建桌面搜索引擎和Web搜索引擎的方法，给出了相关的案例。

第一图书网, tushu007.com

<<迅速搭建全文搜索平台>>

书籍目录

第一部　分建立搜索引擎的方案　第1章　建立搜索引擎的方案　1.1　建立搜索引擎的基本方案　1.1.1　常规的数据库搜索　1.1.2　常规的文件搜索　1.1.3　基于数据库全文搜索功能的搜索　1.1.4　基于windows索引服务的全文搜索　1.1.5　四种基本方案的总结　1.2　利用商业搜索引擎接口实现的全文搜索　1.2.1　第一种基于GOogle Search API的搜索　1.2.2　第二种基于goode Search API的搜索　1.3　利用开源搜索引擎框架实现的全文搜索　小结　思考与练习第二部分　架设网络搜索引擎　第2章　数据抓取　2.1　WebLech　2.1.1　关于webLech　2.1.2　下载webLech　2.1.3　webLech的使用方法　2.1.4　使用webLech　2.2　WebSPHINX　2.2.1　关于webSPHINX　2.2.2　下载webSPHINX　2.2.3　使用websPHINx　2.3　J—Spider　2.3.1　关于J—Spider　2.3.2　下载J—spider　2.3.3　使用J—SDider　小结　思考与练习　第3章　数据解析　3.1　解析PDF文档　3.1.1　使用PDFBox解析PDF文档　3.1.2　使用Xpdf解析PDF文档　3.2　JACOB组件的使用　3.2.1　下载JACOB组件　3.2.2　JACOB的基本用法　3.3　解析word文档　3.3.1　使用textmining组件解析word文档　3.3.2　使用Java2Word组件解析Word文档　3.3.3　使用JACOB组件解析Word文档　3.4　解析Excel文档　3.4.1　使用JDBC访问Excel文档　3.4.2　使用POI组件解析Excel文档　3.4.3　使用Java Excel API解析Excel文档　3.5　解析Powerpoint，Outlook和Access等文档　3.6　解析XML文档　3.6.1　使用DOM解析XML文档　3.6.2　使用SAX解析XML文档　3.6.3　使用JDOM解析XML文档　3.6.4　使用DOM4J解析XML文档　3.6.5　把XML文档解析成纯文本　3.7　解析HTML文档　3.7.1　下载HTMLParser组件　3.7.2　HTMLParser组件的使用　3.7.3　中文问题的提出　3.7.4　网页解析的一般方法　小结　思考与练习　第4章　建立索引和执行搜索　4.1　Hyper Estraier简述　4.1.1　下载Hyper Estraier　4.1.2　安装Hyper Estraier　4.1.3　初试HyperEstmier　4.2　使用Java API　4.2.1　初试Java API　4.2.2　再试Java API　4.3　基于Hyper Estraier的应用　4.3.1　基于Hyper Estraier的桌面搜索应用　4.3.2　基于Hyper Estraier的Web搜索应用　4.4　Hyper Estraier的中文搜索　4.4.1　Hyper Estraier对中文的支持　……第5章　创建搜索引擎

第一图书网, tushu007.com

章节摘录

插图：第1章建立搜索引擎的方案本章要点本章总结了建立搜索引擎的主要方案，对开源搜索引擎的实现原理作了揭示。
1.1建立搜索引擎的基本方案如何建立搜索引擎？
基本方法有如下四种。
（1）常规的数据库搜索使用“ like ”、“ Between ”等谓词，或者数据库自带的“ instr ”等字符串函数。
基于这种原理建立的搜索引擎在数据量非常小的情况下是很有效的。
（2）常规的文件搜索常规的文件搜索就是对文件下的文件进行遍历，用搜索关键词与每个文件的内容进行对比。
这个方法可以用于少量文件的搜索。
（3）基于数据库全文搜索功能的搜索利用数据库自带的全文搜索功能，可以解决几百万条记录的数据库搜索问题，这样实现的全文搜索引擎性能是不错的。
如果能做好软硬件优化，搜索的效果就会更好。
（4）基于Windows索引服务的全文搜索使用Windows的索引服务，可以对大量文件建立起全文索引，然后执行快速的全文搜索。
考虑到知识的系统性，在这里对这四种建立搜索引擎的基本方式进行简单的回顾。

第一图书网, tushu007.com

编辑推荐

《迅速搭建全文搜索平台:开源搜索引擎实战教程》由清华大学出版社、北京交通大学出版社出版。

第一图书网, tushu007.com
<<迅速搭建全文搜索平台>>

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:http://www.tushu007.com