<<搜索引擎零距离>>

图书基本信息

书名：<<搜索引擎零距离>>

13位ISBN编号：9787302201472

10位ISBN编号：7302201471

出版时间：2009 年5月

出版时间：清华大学出版社

作者：王亮

页数：394

前言

搜索引擎是指因特网上专门提供查询服务的一类网站，这些网站通过网络搜索软件（又称为网络搜索机器人）或网站登录等方式，收集因特网上大量网站的页面，经过加工处理后建库，从而能够对用户提出的各种查询作出响应，提供用户所需的信息。
用户的查询途径主要包括自由词、全文检索、主题词检索、分类检索及其他特殊信息的检索f企业、人名、电话黄页等）。
本书中所记述的种种理论与知识，是笔者在多年的搜索引擎开发过程中的积累与沉淀，其中既涉及了Google和Baidu这种类型的通用搜索引擎，也涉及了垂直搜索引擎的开发技术。
垂直搜索引擎概念的提出，就是针对某一特定领域、某一特定人群或某一特定需求提供的有一定价值的信息和相关服务。
可以简单的说成是搜索引擎领域的行业化分工。
众多专业性网站、行业网站独立服务于互联网的成功，恰恰证明了互联网的格局应该是多方面的。
通用搜索引擎的性质，决定了其不能满足特殊领域和特殊人群的精准化信息需求服务。
市场需求多元化决定了搜索引擎的服务模式必将出现细分，以针对不同行业提供更加精确的行业服务模式。
可以说通用搜索引擎的发展为垂直搜索引擎的出现提供了良好的市场空间，势必将出现垂直搜索引擎在互联网中占据部分市场的趋势，也是搜索引擎行业细分化的必然趋势。

<<搜索引擎零距离>>

内容概要

《搜索引擎零距离：基于Ruby+Java搜索引擎原理与实现》的内容中，既有教科书式的理论阐述，也有“七天入门”式的实例解析，还有《Linux内核情景分析》风格的细致的代码分析，甚至还有一些英语文献翻译，从初学者到有一定经验的搜索引擎开发人员，各个层次的读者都能找到一些适合自己阅读的章节。

随着网络信息资源的急剧增长，人们越来越多地关注如何快速有效地从海量的网络信息中，抽取出潜在的、有价值的信息，使之有效地在管理和决策中发挥作用。

搜索引擎技术解决了用户检索网络信息的困难，目前搜索引擎技术正成为计算机科学界和信息产业界争相研究、开发的对象。

《搜索引擎零距离：基于Ruby+Java搜索引擎原理与实现》的作者是一位资深的搜索引擎开发人员，书中对数据获取（网络信息挖掘）与数据检索（搜索引擎）两个方面作了深入的介绍。

《搜索引擎零距离：基于Ruby+Java搜索引擎原理与实现》首先提出了一套“网络数据挖掘”的完整理论，并给出一个实际的智能爬虫系统，通过理论与实际的完整呈现，使读者能够对“网络数据挖掘”有一个比较具体的认识，然后介绍了一个专用程序语言IRS，并给出了这个语言的编译器以及虚拟机的实现方法。

《搜索引擎零距离：基于Ruby+Java搜索引擎原理与实现》还通过对多个开源搜索引擎项目抽丝剥茧的细致分析，引出搜索引擎的一些基本原理与开发方法，并介绍了一个商业化搜索引擎的实例。

《搜索引擎零距离：基于Ruby+Java搜索引擎原理与实现》的最后还结合一个Java框架介绍了一些软件设计思想。

《搜索引擎零距离：基于Ruby+Java搜索引擎原理与实现》涉及网络数据挖掘、搜索引擎原理、编译原理、数据库原理、正则表达式、软件工程、设计模式、Ruby语言、HTTP协议等计算机科学与技术的知识，适合搜索引擎开发人员作为参考，也适合有一定计算机基础的读者阅读，以扩展视野。

作者简介

王亮，主持或参与过多个大中型索引擎开发与运营，具有丰富的搜索引擎算法理论知识与实际开发运营经验。
曾任职于爱立信、Smarter.com、上海网村、上海迈众，2009年创立上海睿驿信息技术有限公司并任CEO，致力于提供搜索引擎相关的产品和服务。

<<搜索引擎零距离>>

书籍目录

章节摘录

插图：第1章 网页数据挖掘1.1 网页数据挖掘定义数据挖掘（Data Minin9，DM），是从存放在数据库、数据仓库或其他信息库中的大量数据中提取或“挖掘”有趣知识的过程。
随着网络的不断发展，因特网目前已成为一个巨大的、分布广泛的和全球性的信息服务中心。
从海量的网络信息中寻找有用的知识，早已成为人们的迫切需求。
各种类似Google、Baidu等的搜索引擎也层出不穷，Web数据挖掘的应用在现实中不断体现。
Web数据挖掘建立在对大量的网络数据进行分析的基础上，采用相应的数据挖掘算法，在具体的应用模型上进行数据的提取、数据筛选、数据转换、数据挖掘和模式分析，最后做出归纳性的推理、预测客户的个性化行为以及用户习惯，从而帮助决策和管理，减少决策的风险。
Web数据挖掘涉及多个领域，除数据挖掘外，还涉及计算机网络、数据库与数据仓储、人工智能、信息检索、可视化、自然语言理解等技术。
1.2 Web数据挖掘面临的问题Web的巨大、分布广泛和内容多样使得目前的Web数据挖掘面临着众多问题和挑战。
首先，对有效的数据仓库和数据挖掘来说，Web上的数据过于庞大。
而且，Web上的数据具有极强的动态性，不仅数量增长快而且更新十分迅速。
但是面对如此大量的Web信息，却有调查表明：99%的Web信息对于99%的用户是无用的。
这样看来，面对网络上形形色色的用户群体，许多由Web搜索引擎所检索到的资料将会被淹没。

后记

从2000年进入大学计算机专业，开始编写Pascal语言程序算起，我已经在程序设计领域耕耘了8年。
大学时伴随我最多是的VC 6．0，阅读着各色的C++书籍，编写自己的一个个C++项目，各种各样的算法、理论、编程思想，经过自己的理解消化之后，从书本进入自己的头脑。
这些计算机科学与软件工程中的知识，在后来的工作中有意无意地被使用到，正是因为对《编译原理》的熟悉，当遇到网页信息挖掘这样一个复杂项目的时候，我设计了IRS语言，使用这个语言中各种灵活的表达式、函数、流程控制来实现对高度灵活复杂的网页信息挖掘需求。
2004年，我在一家商品搜索引擎公司工作，与同事们一起用C、Pefl、PHP语言为这个系统设计各种功能模块。
这段时间中，我开始了解搜索引擎系统的概念、算法等方方面面。
2005年，我在一家无线搜索引擎公司接受了一个任务：用Java语言独立实现一个工业级别的搜索引擎。
由于预算的有限与时间的紧迫，我向开源软件寻求灵感与帮助。
深入研究了L,ucene和。
Nutch之后，我发现它们刚好能符合我的需求，基于这两个系统，我开始架构自己的搜索引擎系统UUSE。
在以后的若干年中，Lucene的版本从1．4升级到了2．3，功能与性能不断改善，Nutch中的“分布式计算与文件系统”模块独立成为Hadoop项目，性能和可扩展性也有了长足的进步。
2006年，Apache发布了开源的Solr项目，这使得架设一个中小规模的搜索引擎变得比较容易。

<<搜索引擎零距离>>

编辑推荐

《搜索引擎零距离:基于Ruby+Java搜索引擎原理与实现》特色：国内垂直搜引擎的扛鼎之作；集开源搜索引擎之大成，融会贯通，自成一体；无线搜索引擎核心技术零距离接触；Web信息挖掘专用程序设计语言，语法标准首次发布；垂直爬虫专业并行虚拟机核心技术展示；多年商业搜索引擎开发运营经验之提炼总结；真实的中型分布式搜索引擎开发案例全景展现；最新Java前沿技术在经典计算机理论上的优秀应用；专业信息检索理论与商业搜索需求的完美结合；Java软件工程设计模式最佳实践。

<<搜索引擎零距离>>

版权说明