《Webbots、Spiders和Scr》

图书基本信息

书名：《Webbots、Spiders和Screen Scrapers》

13位ISBN编号：9787111417682

10位ISBN编号：7111417682

出版时间：2013-5

出版社：斯昆克 (Michael Schrenk)、张磊、 沈鑫 机械工业出版社 (2013-05出版)

作者：斯昆克

页数：282

《Webbots、Spiders和Scr》

内容概要

《Webbots、Spiders和Screen Scrapers:技术解析与应用实践(原书第2版)》共31章，分为4个部分：第一部分（1～7章），系统全面地介绍了与Webbots、Spiders、Screen Scrapers相关的各种概念和技术原理，是了解和使用它们必须掌握的基础知识；第二部分（8～16章），以案例的形式仔细地讲解了价格监控、图片抓取、搜索排名检测、信息聚合、FTP信息、阅读与发送电子邮件等9类常见机器人的设计与开发方法，非常具备实战指导意义；第三部分（17～25章），总结和归纳了大量的高级技巧，包括蜘蛛程序的设计方法、采购机器人和秒杀器、相关的密码学、认证方法、高级cookie管理、如何计划运行网络机器人和蜘蛛、使用浏览器宏抓取怪异的网站、修改iMacros，等等；第四部分（26～31章）是拓展知识，包含如何设计隐蔽的网络机器人和蜘蛛、编写容错的网络机器人、设计网络机器人青睐的网站、消灭蜘蛛、相关的法律知识等。

作者简介

作者：（美国）斯昆克（Michael Schrenk） 译者：张磊 沈鑫 Michael Schrenk，资深网络安全专家和软件开发专家，在网络机器人领域有15年的研究和开发经验，他的工作足迹从美国的硅谷到莫斯科，曾服务于BBc、政府机构、世界500强等很多企业和机构，积累了丰富的实战经验。他还是著名培训讲师和演讲嘉宾，曾多次在国际顶级安全技术大会Defcon上发表演讲，深受欢迎。 张磊，资深系统架构设计者，曾任职于雅虎软件研发（北京）有限公司，从事Sponsored Search离线算法模块的开发工作，之后任职于爱奇艺，曾负责数据平台、广告数据算法、搜索系统、推荐系统等多个技术团队。 沈鑫，资深网络工程师，热衷于网络安全技术的应用与实践，乐于技术分享，曾任“ 52破解论坛 ”版主、“ 赏金论坛 ”加壳与脱壳区版主。

《Webbots、Spiders和Scr》

书籍目录

章节摘录

版权页： 插图： 处理大规模数据 读者将很快发现，网络机器人能够收集大规模的数据。一个简单的自动网络机器人或者蜘蛛程序，即便在几个月里面每天只运行一次，所收集的数据仍然是巨大的。由于没有人拥有无限的存储空间，管理所收集和存储的数据就变得非常重要。在本章，作者将描述如何处理网络机器人收集到的数据，然后研究如何减小存储的数据量。 7.1 组织数据 组织网络机器人下载的数据是需要规划的。不管是利用设计好的文件结构，还是利用关系数据库，其结果都要满足应用程序想要解决特定问题的需求。例如，如果数据主要是由文本组成，被许多人访问，或者需要快速获取或支持检索的能力，那么关系数据库是比较好的选择，因为关系数据库具备这些功能。另一方面，如果要存储许多图片、PDF文件或者word文件，在结构化的文件系统上进行存储是比较好的选择。你甚至可以创建一个混合系统，使用结构化目录系统存储媒体文件，而其索引存储在关系数据库里面。 7.1.1命名规范 虽然并不存在一个绝对“ 正确 ”的方式来组织数据，但在存储网络机器人产生的数据时有不少坏方法。多数错误来自于给网络机器人收集的数据采用了非描述性的或者有歧义的名字。因此，你的设计必须包括命名规范，使其唯一标识文件、目录以及数据库属性。要提前给数据命名，在规划阶段就做，而不是边收集数据边命名。要一直采用支持数据结构增长的命名方式。例如，如果一个房地产方向的网络机器人使用“ 居民楼 ”来命名资产，那么随着应用扩展到土地、办公楼或者商业的时候，将变得难以维护。给数据进行命名升级可能会非常繁琐，因为这些名字在代码和文档里会多次引用。 命名规范可以强制实施任意你喜欢的规则，但是请考虑如下指导原则： 必须强硬地实施每一个命名标准，否则这些标准将不会成为标准。 通常用基于对象的类型来命名是比较好的，而不是基于对象本身是什么。

《Webbots、Spiders和Scr》

编辑推荐

《Webbots、Spiders和Screen Scrapers:技术解析与应用实践(原书第2版)》是Webbots（网络机器人）、Spiders（蜘蛛）、Screen Scrapers（抓屏器）领域的权威著作，在国际安全领域被广泛认可，是资深网络安全专家15年工作经验的结晶。不仅全面而详细地解析了Webbots、Spiders和Screen Scrapers的技术原理和高级技巧，而且以案例的方式讲解了9种常用网络机器人的设计和开发方法，可操作性极强。除了有丰富的理论和实践内容外，还介绍了商业用途的思路，不厌其烦地告诫开发者如何开发出遵纪守法且不干扰网络的具有建设性的网络机器人。

《Webbots、Spiders和Scr》

精彩短评

1、270页分了31章，每一张能有多少篇幅？把目录弄下来，去baidu搜索都好过这本破书。
2、入门还可以，感觉只是小学水平
3、书的实用性强在于作者的函数库适合解决简单的采集任务，并提供复杂任务的解决之道。书的后半部分翻译略显糟糕。
4、这真是一本让人无比失望的书。可毕竟副标题是“技术”，是“应用”，是“实践”，我为什么要寄望得到更多内涵？
5、代码适用性低，不过开了眼界
6、超级好的一本书啊，学完这本书我就打算毕业课设设计一个搜索网站了，后台搜索引擎用lucene，网络爬虫就自己解决了，哈哈
7、代码和内容有点旧了。篇幅不大，分了很多章，导致各部分都只能泛泛而谈。不过想看php爬虫的也许可以看看吧。
8、感觉把相对容易的部分写书了，难的付费学习？以书中的内容基本完成不了稍微难点的模拟登陆

精彩书评

1、很久以来，我一直都对网络机器人比较感兴趣，曾经也对抢票插件等等有很高的兴致，但无奈资料太少，自己一直也没有搞明白。这本书是个及时雨，遇到这本书令我有说不出的开心。书中不仅有原理，而且还有相当多的实践，代码也比较完整，非常适合独自研究。里面有很多比较诱人的东西，比如解码加密网站、管理cookie、相关信息变更后采用邮件通知、在网店里自动购买&抢购&秒杀等等，非常详尽，实用性非常高。补：读完这本书之后，马上就来了个让我一展身手的机会，系主任带着研究生做一个大数据的项目，但是缺乏原始数据，需要一个爬虫能够从网页上没日没夜的抓取数据，研究生里面没人懂网络爬虫技术，那时候系主任正带着我做毕业设计，问我会不会，我当然会啊，于是这个项目的数据都是靠我这个本科生用两个晚上写的php爬虫抓取的，后来快要毕业了，系主任还让我给研究生讲了讲网络爬虫的设计方法，那个开心自豪啊。

版权说明