

# 《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

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：[www.tushu000.com](http://www.tushu000.com)

# 《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。

# 《Webbots、Spiders和Scr》

## 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。

## 书籍目录

译者序 前言 第一部分 基础概念和技术 第1章 本书主要内容 1.1 发现互联网的真正潜力 1.2 对开发者来说 1.2.1 网络机器人开发者是紧缺人才 1.2.2 编写网络机器人是有趣的 1.2.3 网络机器人利用了“建设性黑客”技术 1.3 对企业管理者来说 1.3.1 为业务定制互联网 1.3.2 充分利用公众对网络机器人的经验不足 1.3.3 事半功倍 1.4 结论 第2章 网络机器人项目创意 2.1 浏览器局限性的启发 2.1.1 聚合并过滤相关信息的网络机器人 2.1.2 解释在线信息的网络机器人 2.1.3 个人代理网络机器人 2.2 从疯狂的创意开始 2.2.1 帮助繁忙的人解脱 2.2.2 自动执行，节省开支 2.2.3 保护知识产权 2.2.4 监视机会 2.2.5 在网站上验证访问权限 2.2.6 创建网上剪报服务 2.2.7 寻找未授权的Wi-Fi网络 2.2.8 跟踪网站技术 2.2.9 让互不兼容的系统通信 2.3 结论 第3章 下载网页 3.1 当它们是文件，而不是网页 3.2 用PHP的内置函数下载文件 3.2.1 用fopen（）和fgets（）下载文件 3.2.2 用file（）函数下载文件 3.3 PHP/CURL库介绍 3.3.1 多种传输协议 3.3.2 表单提交 3.3.3 基本认证技术 3.3.4 cookie 3.3.5 重定向 3.3.6 代理名称欺诈 3.3.7 上链管理 3.3.8 套接字管理 3.4 安装PHP/CURL 3.5 LIB\_http库 3.5.1 熟悉默认值 3.5.2 使用LIB\_http 3.5.3 了解更多HTTP标头信息 3.5.4 检查LIB\_http的源代码 3.6 结论 第4章 基本解析技术 4.1 内容与标签相混合 4.2 解析格式混乱的HTML文件 4.3 标准解析过程 4.4 使用LIB\_parse库 4.4.1 用分隔符分解字符串：split\_string（）函数 4.4.2 提取分隔符之间的部分：return\_between（）函数 4.4.3 将数据集解析到数组之中：parse\_array（）函数 4.4.4 提取属性值：get\_attribute（）函数 4.4.5 移除无用文本：remove（）函数 4.5 有用的PHP函数 4.5.1 判断一个字符串是否在另一个字符串里面 4.5.2 用一个字符串替换另一个字符串中的一部分 4.5.3 解析无格式文本 4.5.4 衡量字符串的相似度 4.6 结论 4.6.1 别相信编码混乱的网页 4.6.2 小步解析 4.6.3 不要在调试的时候渲染解析结果 4.6.4 少用正则表达式 第5章 使用正则表达式的高级解析技术 5.1 模式匹配——正则表达式的关键 5.2 PHP的正则表达式类型 5.2.1 PHP正则表达式函数 5.2.2 与PHP内置函数的相似之处 5.3 从例子中学习模式匹配 5.3.1 提取数字 5.3.2 探测字符串序列 5.3.3 字母字符匹配 5.3.4 通配符匹配 5.3.5 选择匹配 5.3.6 分组和范围匹配的正则表达式 5.4 与网络机器人开发者相关的正则表达式 5.4.1 提取电话号码 5.4.2 下一步学习什么 5.5 何时使用正则表达式 5.5.1 正则表达式的长处 5.5.2 模式匹配用于解析网页的劣势 5.5.3 哪个更快，正则表达式还是PHP的内置函数 5.6 结论 第6章 自动表单提交 6.1 表单接口的反向工程 6.2 表单处理器、数据域、表单方法和事件触发器 6.2.1 表单处理器 6.2.2 数据域 6.2.3 表单方法 6.2.4 多组件编码 6.2.5 事件触发器 6.3 无法预测的表单 6.3.1 JavaScript能在提交之前修改表单 6.3.2 表单HTML代码通常无法阅读 6.3.3 cookie在表单里不存在，却会影响其操作 6.4 分析表单 6.5 结论 6.5.1 不要暴露身份 6.5.2 正确模拟浏览器 6.5.3 避免表单错误 第7章 处理大规模数据 7.1 组织数据 7.1.1 命名规范 7.1.2 在结构化文件里存储数据 7.1.3 在数据库里存储文本数据 7.1.4 在数据库里存储图片 7.1.5 用数据库，还是用文件系统 7.2 减小数据规模 7.2.1 保存图片文件的地址 7.2.2 压缩数据 7.2.3 移除格式信息 7.3 生成图片的缩略图 7.4 结论 第二部分 网络机器人项目 第8章 价格监控网络机器人 8.1 目标网站 8.2 设计解析脚本 8.3 初始化以及下载目标网页 8.4 进一步探讨 第9章 图片抓取网络机器人 9.1 图片抓取网络机器人例子 9.2 创建图片抓取网络机器人 9.2.1 二进制安全下载过程 9.2.2 目录结构 9.2.3 主脚本 9.3 进一步探讨 9.4 结论 第10章 链接校验网络机器人 10.1 创建链接校验网络机器人 10.1.1 初始化网络机器人并下载目标网页 10.1.2 设置页面基准 10.1.3 提取链接 10.1.4 运行校验循环 10.1.5 生成URL完整路径 10.1.6 下载全链接路径 10.1.7 展示页面状态 10.2 运行网络机器人 10.2.1 LIB\_http\_codes 10.2.2 LIB\_resolve\_addresses 10.3 进一步探讨 第11章 搜索排名检测网络机器人 11.1 搜索结果页介绍 11.2 搜索排名检测网络机器人做什么工作 11.3 运行搜索排名检测网络机器人 11.4 搜索排名检测网络机器人的工作原理 11.5 搜索排名检测网络机器人脚本 11.5.1 初始化变量 11.5.2 开始循环 11.5.3 获取搜索结果 11.5.4 解析搜索结果 11.6 结论 11.6.1 对数据源要厚道 11.6.2 搜索网站对待网络机器人可能会不同于浏览器 11.6.3 爬取搜索引擎不是好主意 11.6.4 熟悉Google API 11.7 进一步探讨 第12章 信息聚合网络机器人 12.1 给网络机器人选择数据源 12.2 信息聚合网络机器人举例 12.2.1 熟悉RSS源 12.2.2 编写信息聚合网络机器人 12.3 给信息聚合网络机器人添加过滤机制 12.4 进一步探讨 第13章 FTP网络机器人 13.1 FTP网络机器人举例 13.2 PHP和FTP 13.3 进一步探讨 第14章 阅读电子邮件的网络机器人 14.1 POP3协议 14.1.1 登录到POP3邮件服务器 14.1.2 从POP3邮件服务器上读取邮件 14.2 用网络机器人执行POP3命令 14.3 进一步探讨 14.3.1 电子邮件控制的网络机器人 14.3.2 电子邮件接口 第15章 发送电子邮件的网络机器人 15.1 电子邮件、网络机器人以及垃圾邮件 15.2 使用SMTP和PHP发送邮件 15.2.1 配置PHP发送邮件 15.2.2 使用mail（）函数发送电子邮件 15.3 编写发送电子邮件通知的网络机器人 15.3.1 让合法的邮件不被过滤掉 15.3.2 发送HTML格式的电子邮件 15.4 进

一步探讨 15.4.1 使用回复邮件剪裁访问列表 15.4.2 使用电子邮件作为你的网络机器人运行的通知 15.4.3 利用无线技术 15.4.4 编写发送短信的网络机器人 第16章 将一个网站转变成一个函数 16.1 编写一个函数接口 16.1.1 定义函数接口 16.1.2 分析目标网页 16.1.3 使用describe\_zipcode ( ) 函数 16.2 结论 16.2.1 资源分发 16.2.2 使用标准接口 16.2.3 设计定制的轻量级“Web服务” 第三部分 高级设计技巧 第17章 蜘蛛 17.1 蜘蛛的工作原理 17.2 蜘蛛脚本示例 17.3 LIB\_simple\_spider 17.3.1 harvest\_links ( ) 17.3.2 archive\_links ( ) 17.3.3 get\_domain ( ) 17.3.4 exclude\_link ( ) 17.4 使用蜘蛛进行实验 17.5 添加载荷 17.6 进一步探讨 17.6.1 在数据库中保存链接 17.6.2 分离链接和载荷 17.6.3 在多台计算机上分配任务 17.6.4 管理页面请求 第18章 采购机器人和秒杀器 18.1 采购机器人的原理 18.1.1 获取采购标准 18.1.2 认证买家 18.1.3 核对商品 18.1.4 评估购物触发条件 18.1.5 执行购买 18.1.6 评估结果 18.2 秒杀器的原理 18.2.1 获取采购标准 18.2.2 认证竞拍者 18.2.3 核对拍卖商品 18.2.4 同步时钟 18.2.5 竞价时间 18.2.6 提交竞价 18.2.7 评估结果 18.3 测试自己的网络机器人和秒杀器 18.4 进一步探讨 18.5 结论 第19章 网络机器人和密码学 19.1 设计使用加密的网络机器人 19.1.1 SSL和PHP内置函数 19.1.2 加密和PHP/CURL 19.2 网页加密的简要概述 19.3 结论 第20章 认证 20.1 认证的概念 20.1.1 在线认证的类型 20.1.2 用多种方式加强认证 20.1.3 认证和网络机器人 20.2 示例脚本和实践页面 20.3 基本认证 20.4 会话认证 20.4.1 使用cookie会话的认证 20.4.2 使用查询会话进行认证 20.5 结论 第21章 高级cookie管理 21.1 cookie的工作原理 21.2 PHP/CURL和cookie 21.3 网络机器人设计中面临的cookie难题 21.3.1 擦除临时性cookie 21.3.2 管理多用户的cookie 21.4 进一步探讨 第22章 计划运行网络机器人和蜘蛛 22.1 为网络机器人配置计划任务 22.2 Windows XP任务调度程序 22.2.1 计划网络机器人按日运行 22.2.2 复杂的计划 22.3 Windows 7任务调度程序 22.4 非日历事件触发器 22.5 结论 22.5.1 如何决定网络机器人的最佳运行周期 22.5.2 避免单点故障 22.5.3 在计划中加入变化性 第23章 使用浏览器宏抓取怪异的网站 23.1 高效网页抓取的阻碍 23.1.1 AJAX 23.1.2 怪异的JavaScript和cookie行为 23.1.3 Flash 23.2 使用浏览器宏解决网页抓取难题 23.2.1 浏览器宏的定义 23.2.2 模拟浏览器的终极网络机器人 23.2.3 安装和使用iMacros 23.2.4 创建第一个宏 23.3 结论 23.3.1 宏的必要性 23.3.2 其他用途 第24章 修改iMacros 24.1 增强iMacros的功能 24.1.1 不使用iMacros脚本引擎的原因 24.1.2 创建动态宏 24.1.3 自动装载iMacros 24.2 进一步探讨 第25章 部署和扩展 25.1 一对多环境 25.2 一对一环境 25.3 多对多环境 25.4 多对一环境 25.5 扩展和拒绝服务攻击 25.5.1 简易的网络机器人也会产生大量数据 25.5.2 目标的低效 25.5.3 过度扩展的弊端 25.6 创建多个网络机器人的实例 25.6.1 创建进程 25.6.2 利用操作系统 25.6.3 在多台计算机上分发任务 25.7 管理僵尸网络 25.8 进一步探讨 第四部分 拓展知识 第26章 设计隐蔽的网络机器人和蜘蛛 26.1 设计隐蔽网络机器人的原因 26.1.1 日志文件 26.1.2 日志监控软件 26.2 模拟人类行为为实现隐蔽 26.2.1 善待资源 26.2.2 在繁忙的时刻运行网络机器人 26.2.3 在每天不同时刻运行网络机器人 26.2.4 不要在假期和周末运行网络机器人 26.2.5 使用随机的延迟时间 26.3 结论 第27章 代理 27.1 代理的概念 27.2 虚拟世界中的代理 27.3 网络机器人开发者使用代理的原因 27.3.1 使用代理实现匿名 27.3.2 使用代理改变位置 27.4 使用代理服务器 27.4.1 在浏览器中使用代理 27.4.2 通过PHP/CURL使用代理 27.5 代理服务器的类型 27.5.1 公共代理 27.5.2 Tor 27.5.3 商业代理 27.6 结论 27.6.1 匿名是过程，不是特性 27.6.2 创建自己的代理服务 第28章 编写容错的网络机器人 28.1 网络机器人容错的类型 28.1.1 适应URL变化 28.1.2 适应页面内容的变化 28.1.3 适应表单的变化 28.1.4 适应cookie管理的变化 28.1.5 适应网络中断和网络拥堵 28.2 错误处理器 28.3 进一步探讨 第29章 设计受网络机器人青睐的网站 29.1 针对搜索引擎蜘蛛优化网页 29.1.1 定义明确的链接 29.1.2 谷歌轰炸和垃圾索引 29.1.3 标题标签 29.1.4 元标签 29.1.5 标头标签 29.1.6 图片的alt属性 29.2 阻碍搜索引擎蜘蛛的网页设计技巧 29.2.1 JavaScript 29.2.2 非ASCII内容 29.3 设计纯数据接口 29.3.1 XML 29.3.2 轻量级数据交换 29.3.3 简单对象访问协议 29.3.4 表征状态转移 29.4 结论 第30章 消灭蜘蛛 30.1 合理地请求 30.1.1 创建服务协议条款 30.1.2 使用robots.txt文件 30.1.3 使用robots元标签 30.2 创造障碍 30.2.1 选择性地允许特定的网页代理 30.2.2 使用混淆 30.2.3 使用cookie、加密、JavaScript和重定向 30.2.4 认证用户 30.2.5 频繁升级网站 30.2.6 在其他媒体中嵌入文本 30.3 设置陷阱 30.3.1 创建蜘蛛陷阱 30.3.2 处理不速之客的方法 30.4 结论 第31章 远离麻烦 31.1 尊重 31.2 版权 31.2.1 请善用资源 31.2.2 不要纸上谈兵 31.3 侵犯动产 31.4 互联网法律 31.5 结论 附录A PHP/CURL参考 附录B 状态码 附录C 短信网关

## 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也有不少坏方法。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

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:[www.tushu000.com](http://www.tushu000.com)