HZ BOOKS
华章科技

no starch press

Webbots, Spiders, and Screen Scrapers

A Guide to Developing Internet Agents with PHP/CURL, Second Edition

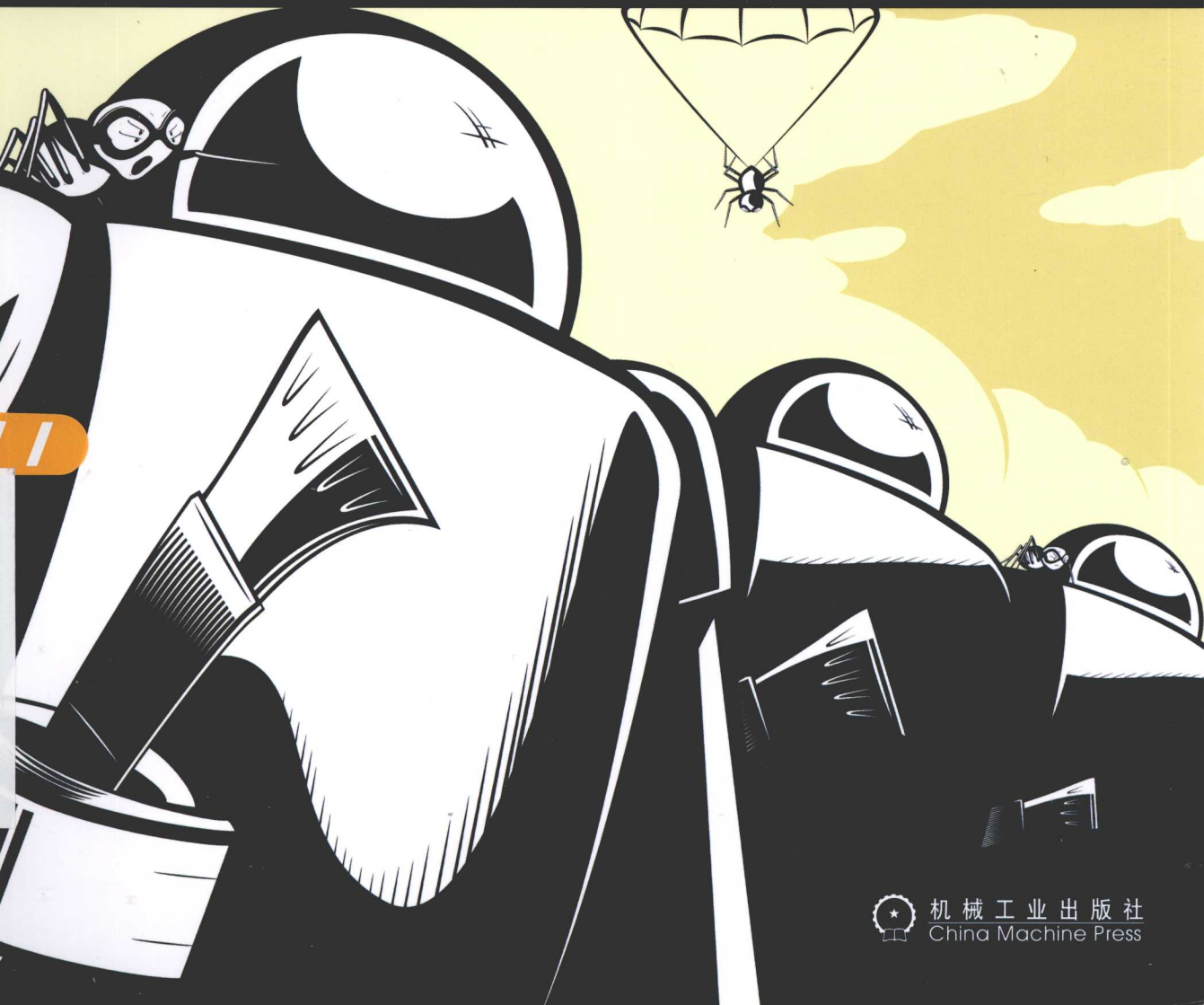
Webbots、Spiders和Screen Scrapers

技术解析与应用实践

（原书第2版）

（美） Michael Schrenk 著

张磊 沈鑫 译



机械工业出版社
China Machine Press

013932947

TP393.092
2450

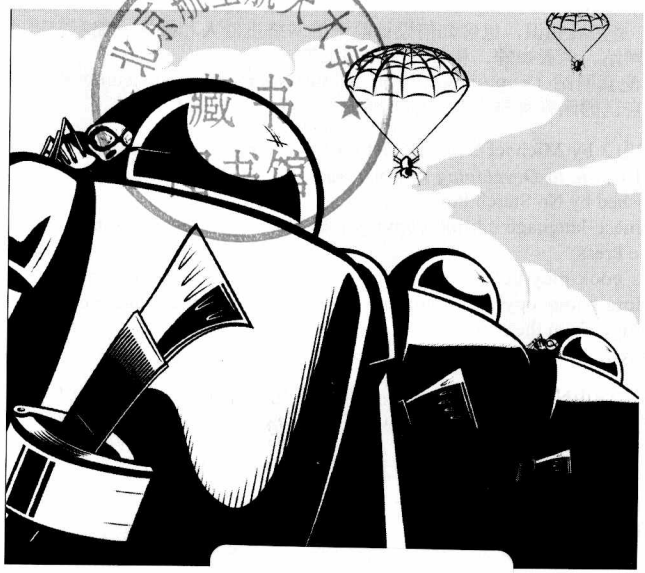
Webbots, Spiders, and Screen Sc
A Guide to Developing Internet Agents with PHP/CURL,

Webbots、Spiders和Screen Scrapers

技术解析与应用实践

（原书第2版）

（美） Michael Schrenk 著
张磊 沈鑫 译



北京航空航天大学
藏书
图书馆

北航 C1640717

TP393.092
2450

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Webbots、Spiders和Screen Scrapers：技术解析与应用实践：原书第2版 /（美）斯昆克（Schrenk，M.）著；张磊，沈鑫译. —北京：机械工业出版社，2013.3
（华章程序员书库）
书名原文：Webbots，Spiders，and Screen Scrapers：A Guide to Developing Internet Agents with PHP/CURL，Second Edition

ISBN 978-7-111-41768-2

I. W… II.① 斯… ② 张… ③ 沈… III. 网页制作工具 IV. TP393.092

中国版本图书馆CIP数据核字（2013）第047247号

版权所有 · 侵权必究
封底无防伪标均为盗版
本书法律顾问 北京市展达律师事务所

本书版权登记号：图字：01-2012-2659

本书是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，本书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。

全书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。

本书还有一个配套网站（http://www.WebbotsSpidersScreenScrapers.com）提供了书中所有的示例代码，方便读者设计自己的网络机器人并在这个网址上进行实践性操作。

Copyright © 2012 by Michael Schrenk. Title of English-language original: *Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL, Second Edition*, ISBN 978-1-59327-120-6, published by No Starch Press.

Simplified Chinese-language edition copyright © 2013 by Beijing Huazhang Graphics & Information Co., China Machine Press.

No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system, without permission, in writing, from the publisher.

All rights reserved.

本书中文简体字版由No Starch Press授权机械工业出版社在全球独家出版发行。未经出版者书面许可，不得以任何方式抄袭、复制或节录本书中的任何部分。

机械工业出版社（北京市西城区百万庄大街22号 邮政编码 100037）
责任编辑：吴 怡
三河市杨庄长鸣印刷装订厂印刷
2013年5月第1版第1次印刷
186mm×240mm · 18.75印张
标准书号：ISBN 978-7-111-41768-2
定 价：69.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

译 者 序

在战争之中若能做到知己知彼，则能百战而不殆；在诡谲多变的商业竞争中也是一样，总是掌握信息更多的一方更容易取胜。如今互联网已成为最重要的信息源，但是从互联网上获得竞争优势的公司仍然少得可怜。不知道仍有多少老板苦于不了解行业信息，以至于有许多决策是在信息黑暗之中拍脑袋做出的。这本书正是希望教会读者如何从互联网上获取各种数据，仅使用PHP语言和相关的库函数，简单而又能自动化。

有一些编程基础和HTTP基础就可以顺畅地阅读本书了，了解过PHP则会更好一些。抓取网页、提取数据、管理数据、模拟表单请求、模拟登录、管理cookie等，这些网络机器人的基本问题都能够从本书中找到答案。当你能综合这些技能用于获取互联网上的信息，你会猛然发现自己掌握了一个多么强大的工具，自己已经成为公司的一股核心竞争力了。

还有一点需要特别说明的是，网络机器人不是一种可以随意使用的技术。我们在网络上的行为需要有礼貌，不能侵犯隐私，不能随意使用有版权的数据。关于这一点作者在书中有多次交代，也希望读者能够重视和遵守。

本书从开始到第二部分结束，由张磊翻译；第三部分开始到最后，由沈鑫翻译。吴怡编辑仔细校对了全书，向她表示感谢。

如果读者在阅读的过程中遇到技术方面的疑问，欢迎发邮件给webbotsbook@gmail.com，沈鑫和我将会尽力给予解答。由于译者的水平和精力有限，译稿之中难免出现疏漏之处，希望读者能给予谅解，并给我们发邮件指正问题。勘误表将发布在本书中文版官方网站上，并不断更新：www.webbotsbook.com。

张 磊

2013年元宵夜，于北京

前　言

作者接触万维网是从接触浏览器开始的。当时使用的第一款浏览器是Mosaic，由Eric Bina和Marc Andreessen所开创。Andreessen后来与人共同创办了Netscape（网景公司）和Loudcloud公司。

1995年接触万维网后不久，作者就非常好奇，如此丰富多彩的互联网具有如此简洁的浏览器。浏览器不仅仅是一款使用万维网的工具，它仿佛就是万维网本身，它是新的电视机！正如电视机使用简单的频道旋钮和音量旋钮遥控器来远程操控一样，浏览器利用超链接、书签和后退按钮隐藏互联网的复杂性。

老派的“客户端－服务器”技术

当作者认识到并不需要使用浏览器来查看网页的时候，伟大的发现时刻来临了。作者发现了Telnet，这款20世纪80年代早期就开始用来与网络上的计算机进行通信的程序，也可以用来下载网页。作者发现浏览器的背后并不神秘，下载网页与现有的从网络计算机上请求信息的方法并无二致。

突然间，万维网成了不需要浏览器也能够理解的东西，它就是一个熟悉的“客户端-服务器”架构。在这个架构里，简单的客户端访问远程服务器上的文件，区别在于这里的客户端即是浏览器，服务器发送网页供浏览器来展示。

浏览器唯一的革命性在于，不像Telnet，它让每个人用起来都很方便。易用性和过度扩张的内容规模使浏览器迅速获得了大众的认可。浏览器的出现使互联网的用户由科学家和计算机程序员过渡为普通大众，这些人并不理解计算机网络是如何工作的。遗憾的是，普通人不明白“客户端-服务器”协议的简单性，所以对浏览器的依赖便愈发严重。他们不知道还存在其他的方式来使用万维网，并且这些方式很可能是更有趣的。

作为一个程序员，作者意识到如果能够使用Telnet来下载网页，那么作者也可以通过编写程序来做同样的事情。如果作者愿意，还可以编写自己的浏览器！又或者，可以编写自动化的代理程序（指Webbot、Spider、ScreenScrapers，即网络机器人、蜘蛛程序以及抓屏器）来解决浏览器所不能解决的问题。

浏览器的问题

浏览器的基本问题是，它们是手动的工具。浏览器只是下载并展示网页——即便你已经

此为试读，需要完整PDF请访问：www.ertongbook.com

看到它包含的信息，你仍然需要判断它是不是你想要的；或者说你必须判断要不要去点击一个链接打开另一个网页。更糟糕的是，浏览器本身不会思考。当有重要的网络事件发生的时候它不能通知你。当然它也不能预知你的行为，例如自动填写表格、购物、或者为你下载文件。为了实现这些功能，需要网络机器人的自动功能和智能。一旦发现浏览器固有的局限性，你就会看到，网络机器人开发者的机遇来到了，并且有着无限的前景。

本书主要内容

本书阐明浏览器的局限性，并探索如何利用网络机器人来打破这些局限性。本书通过示例脚本程序和项目来介绍如何设计并编写网络机器人程序。本书还解答下面这些关于设计的问题：

- 网络机器人项目的创意是从哪里来的？
- 怎样能够在享受网络机器人带来的乐趣的同时又不惹麻烦？
- 是否有可能写出来能不声不响地运行而不被人发现的网络机器人？
- 编写健壮的、容错的、当网络内容发生改变也不会出问题的网络机器人的技巧是什么？

从作者犯过的错误中学习

作者从事编写网络机器人、蜘蛛程序以及抓屏器的工作已经有15年的时间了。在这个过程中作者犯了许多错误。因为网络机器人有能力对网站做出非常规的请求，系统管理员可能无法将网络机器人的请求和黑客攻击者的行为区分开来。还好，作者还没犯过让自己需要对簿公堂的错误，但是也招致了威胁电话、恐吓邮件，以及非常尴尬的时刻。让人快乐的是，作者从这些境遇中学到了很多，面对一个愤怒的系统管理员这样的事情已经是很久很久之前了。通过阅读作者的故事，从作者的错误中学习，读者可以避免许多痛苦。

掌握网络机器人技术

读者将学习到编写多种网络机器人的技能，包括：

- 按部就班地下载整个网站。
- 解码加密的网站。
- 解锁需要认证的网页。
- 管理cookie。
- 解析数据。
- 编写蜘蛛程序。
- 管理网络机器人生成的大规模数据。

利用已有的脚本

本书用了几个代码库，可以使读者编写网络机器人、蜘蛛程序和抓屏器更加容易。这些库里的函数和声明为本书所用的大部分示例脚本提供了基础。使用这些库将节约读者的时间，因为它们做了底层的工作，仅把更高层次的规划和开发工作留给读者。所有这些库都可以从本书的网站上下载到。

关于本书的官方网站

这本书的官方网站（http://www.WebbotsSpidersScreenScrapers.com）是一个附加资源。书中所有的示例项目尽可能地使用这个网站作为抓取目标，或者作为资源供读者的网络机器人下载或采取其他操作。这些目标提供了一成不变的环境供读者锻炼编写网络机器人的技巧。有一个可控的学习环境是重要的，因为不管我们多么努力，当目标网站变化的时候，网络机器人都可能会失败。知道目标网站不会改变会让调试变得更为简单。

这个网站上也有一些链接指向相关的网站，包括白皮书、图书更新信息、以及读者能够和其他网络机器人开发者进行沟通的一个区域（请看图1）。在这个网站上，读者还可以获取本书中用到的所有代码库。

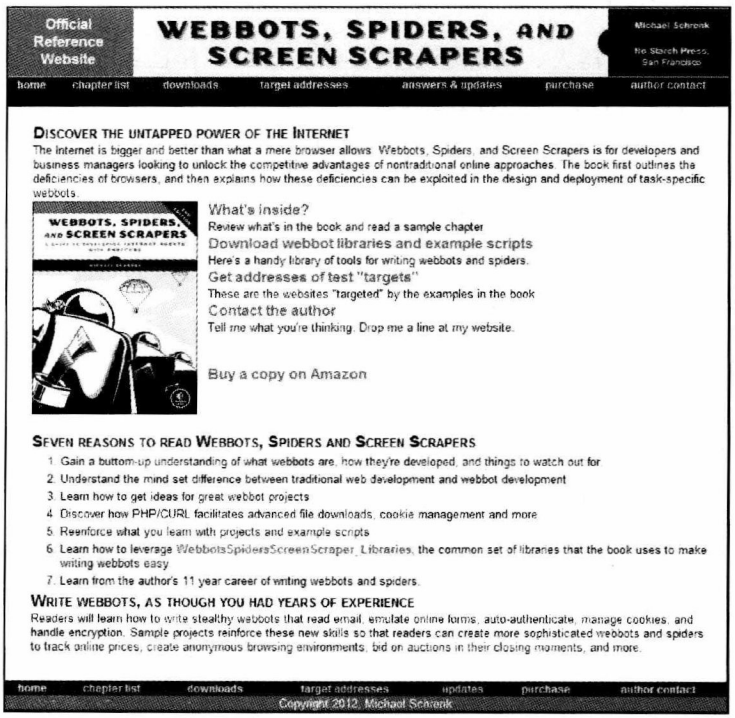


图1　本书的官方网站

关于本书代码

本书中的大部分代码是纯PHP。然而，有时候PHP和HTML在同一个脚本中混用，并且很多情况下它们是在同一行里面。在这些情况下，PHP代码会以粗体的样式与HTML代码区分开来，如代码清单1所示。

代码清单1　以粗体形式把PHP代码与HTML代码区分开

```
<h1>Coding Conventions for Embedded PHP</h1>
<table border="0" cellpadding="1" cellspacing="0">
<tr>
<th>Name</th>
      <th>Address</th>
  </tr>
  <? for ($x=0; $x<sizeof($person_array); $x++)
      {     ?>
      <tr>
           <td><? echo person_array[$x]['NAME']?></td>
                <td><? echo person_array[$x]['ADDRESS']?></td>
      </tr>
<?     }   ?>
</table>
```

读者可以以个人目的使用本书中的任意脚本，只要读者不会再次发布它们。如果读者使用了本书中的脚本，你要为它的使用和执行担负所有责任，并且不得在任何情况下售卖或开发相关的产品。然而，如果读者改进了一份脚本或者开发了全新的（相关的）脚本，那么非常欢迎通过本书的网站与网络机器人社区分享。

另外，示例脚本是以教学为目的的。这些脚本可能不会反映最高效的编程方法，因为编写它们的首要目的是为了可读性。

注意：本书使用的代码库受W3C代码监控和许可（http://www.w3.org/Consortium/Legal/2002/copyright-software-20021231），并能够从本书的网站上下载到。软件的维护也在该网站上进行。如果读者对这份代码做出了有意义的贡献，请到网站查看你的改进可能会以何种方式成为下一个发行版的一部分。本书所展示的软件示例受本书的版权保护。

阅读本书的要求

要理解本书所讲的技术，需要了解HTML和基本的互联网工作原理，仅有一点点计算机网络经验的编程新手也可以阅读本书。然而还有一条要强调，本书不会教你编程方法或者互联网协议TCP/IP的原理。

硬件

读者不需要准备精密的硬件设备来开始编写网络机器人程序，一台二手电脑也很可能已经

具备了运行书中示例的最低要求。如下任意一款设备都足以使用本书中的示例项目和数据：

- 一台使用Windows XP、Windows Vista或者Windows 7操作系统的个人电脑。
- 任意基于Linux、UNIX或者FreeBSD的计算机。
- 一台运行OS X（或者更新的系统）的苹果电脑。

电脑需要充足的存储空间，特别是当读者计划编写蜘蛛程序或者自助网络机器人的时候。如果下载过多的文件，会消耗掉所有可用资源（特别是硬盘驱动器）。

软件

为了方便，书中的软件示例使用PHP[㊀]、cURL[㊁]和MySQL[㊂]。所有这些软件都可以在网站上免费下载。除了免费，这些软件包还具有良好的可移植性，能够在多种计算机和操作系统下良好地运行。

注意：*如果读者计划使用本书中的示例脚本，需要具有PHP的基本知识，并会编程。*

互联网连接

有互联网连接是很方便的，但并不是必需的。如果读者没有网络连接，可以通过安装Apache[㊃]来创建自己的局域网（在一个私有网络里的一到多个网站服务器）。如果连这个也无法做到，可以使用本地的文件作为程序操作的目标。然而，这两种方式都不如编写使用在线网络连接的网络机器人更加有趣。另外，如果没有网络连接，读者将无法访问在线资源，而这些资源对你的学习是很有帮助的。

声明（这很重要）

不管读者开发什么，你都必须为自己开发程序的行为负责。从技术的角度讲，一个有益的网络机器人与一个破坏性的网络机器人并无区别，主要的区别在于开发者的意图（以及把脚本调试得怎么样）。因此，是使用本书的知识去做遵纪守法的具有建设性的事情，还是用它去做一些会惹来麻烦甚至违法犯罪的事情，这完全取决于读者。如果读者真的干了什么坏事，可别打电话去找作者。

请参考第31章去了解如何编写符合道德的网络机器人。第31章会给你一些帮助信息，但不是法律咨询。如果读者有这方面的问题，请在付诸实践之前找位律师谈谈。

㊀ http://www.php.net.

㊁ http://curl.haxx.se.

㊂ http://www.mysql.com.

㊃ http://www.apache.org.

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第一部分
基础概念和技术

大多数介绍网站开发的书籍都阐述如何创建网站，这本书教导开发者如何将已有网站合并、改编并自动运营，来满足具体需求。

读者可能已经具有从计算机科学领域学到的经验，用来开发网络机器人、蜘蛛和抓屏器。但是，如果读者已经熟悉这本书里的一些概念，开发网络机器人可能要求以不同的背景来看待这些技术。也就是说，即便你很有经验，仍然建议你阅读整本书。

如果读者还不具有这些领域的经验，那也不要紧，本书前7章将介绍设计开发网络机器人的基础知识。这将成为本书后面讨论的其他项目和高级话题的基础。

第一部分介绍网络自动化的概念并探索驾驭网络资源的基本技术。

第 1 章　本书主要内容

该章探索编写网络机器人的乐趣，以及为什么网络机器人开发是一个高收入的职业，会有很多的发展机会。

第 2 章　网络机器人项目创意

我们已经被误导，认为不得不接受网站的现状。这主要是因为浏览器不允许我们做别的事情。然而，如果仔细考虑你到底想要做什么，而不是浏览器允许做什么，你将以全新的视角去看待你所钟爱的网络资源。从该章你将学到网络浏览器受很多局限的困扰，以及这些局限可能会如何触发念头去开发属于自己的网络机器人项目。

第 3 章　下载网页

该章介绍PHP/CURL，一个开源的库，可以简单地下载网页——即使目标网页使用了诸如转发、加密、认证、cookie等高级技术。

此为试读，需要完整PDF请访问：www.ertongbook.com