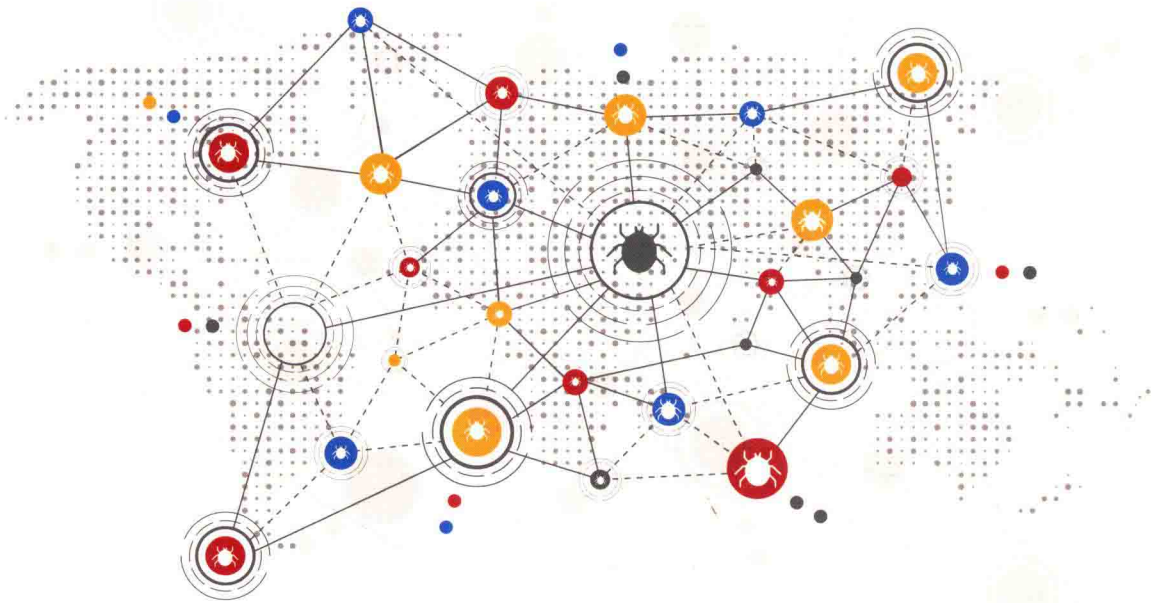
Broadview®
www.broadview.com.cn

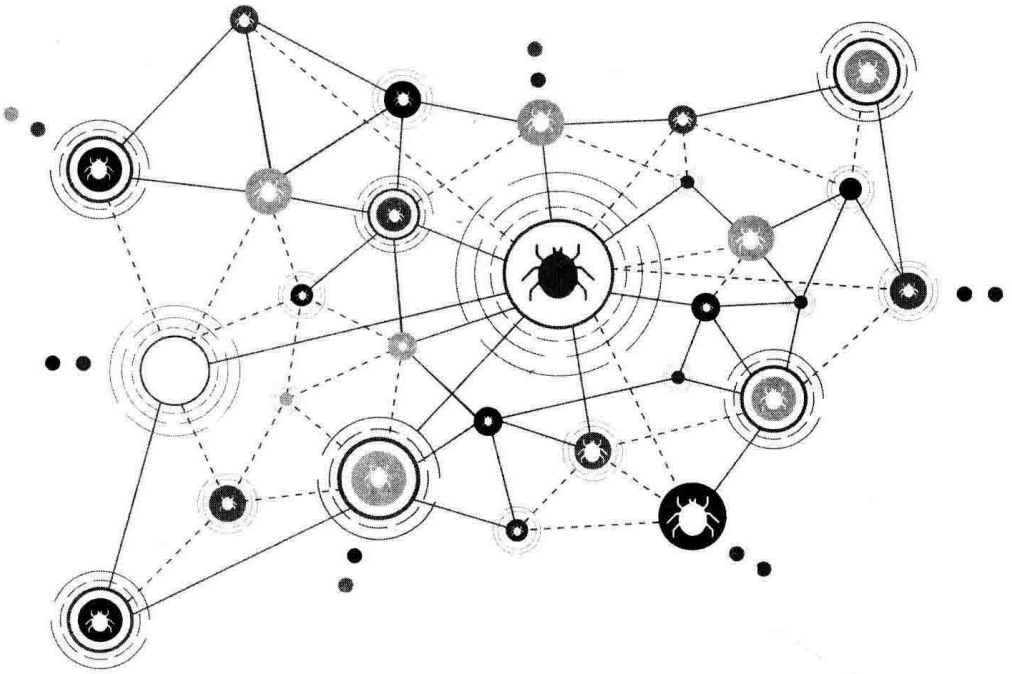


爬虫实战

从数据到产品

贺思聪◎编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn



爬虫实战

从数据到产品

贺思聪◎编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书从多个数据项目实例出发，介绍爬虫、反爬虫的各种案例，使读者了解到数据抓取和分析的完整过程。书中案例的难度由浅入深，以作者原创的代码为主，不借助现成的框架，强调在数据采集过程中的发散思维，总结攻克反爬虫的思维模式，实现以低成本的方式得到想要的数据的愿望。最后，用一个“爱飞狗”的例子，为读者展示如何从 0 到 1 地开发一个大数据产品。

图书在版编目（CIP）数据

爬虫实战：从数据到产品 / 贺思聪编著. —北京：电子工业出版社，2019.4
ISBN 978-7-121-35508-0

Ⅰ. ①爬… Ⅱ. ①贺… Ⅲ. ①数据处理 Ⅳ.①TP274

中国版本图书馆CIP数据核字(2018)第252494号

责任编辑：牛　勇
印　　刷：天津嘉恒印务有限公司
装　　订：天津嘉恒印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　印张：15.25　字数：290 千字
版　　次：2019 年 4 月第 1 版
印　　次：2019 年 4 月第 1 次印刷
定　　价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

作者简介

贺思聪，毕业于电子科技大学，在ThoughtWorks任高级咨询师，担任海外交付团队技术负责人、架构师。

具有十余年产品研发经验，涉及澳洲矿业相关数字产品研发、保险业网站技术架构、光网络设备软件研发、医疗AR/VR、机械控制、加工仿真模拟等领域。

精通大规模重构实践、测试驱动开发，熟悉微服务架构及架构实践，熟悉敏捷项目开发管理流程和相关实践，具有丰富的敏捷团队管理经验。

出版编辑联络：安 娜
微信&QQ：80303489
邮箱：anna@phei.com.cn

前 言

智能设备（如智能手环、百度音箱、扫地机器人等）的普及使收集个人数据变得非常容易。机器性能的提高使得分析、使用数据变得更加自动化。大量的数据结合强大的计算性能，使数据从量变到质变的过程极短，我们的导航早已不再是傻傻地按照既有的策略规划行驶路线，而是一直在向“老司机”学习，不断更新算法，从而带来更精准的预测。

在这个时代，数据就是新一代的资源。我们的身边充满了数据流。我们既是数据流的生产者，也是数据流的消费者。对个人而言，如果能够合理地识别、收集、分析、利用这些数据，就能够在我们做决策时给出一些新的想法。例如，在 GitHub 上一个非常有效的比特币高频交易的源代码，其作者在 2016 年年底到 2017 年 1 月这段时期内，用 6000 元的初始资金赚到了 25 万元。他所利用的就是对比特币这种新交易手段交易数据的洞察，利用机器自动收集分析行情并进行自动化交易。为了解决“什么时候买机票最便宜”的问题，我通过长达两年的数据抓取，收集到上百亿条机票价格数据并进行数据分析及可视化，最后形成了一个名为“爱飞狗”的产品。爱飞狗可将近期各平台的历史价格展示给用户，让不对称的价格信息变得更加透明化。通过对这些数据进行分析，我们可以掌握国内航空公司机票票价变化规律。基于人的经验，在机器学习的帮助下，我的这套方法可以对国内的航班价格提供较为准确的预测，为用户的出行节约成本。

掌握获取信息的能力使我们能够站在更高的角度识别一些规律。例如，在求职的过程中，大量的公司信息很难进行分辨，即便是某些 APP 提供了很多的筛选功能，但仍无法满足我们分析的需求。再如，大量的房产信息淹没在海量数据中，跟踪这些数据的变化或许能够发现一些规律或结论。在这样一个数据丰富的时代，每个人都应该学习一些从数据采集到数据分析的综合技能。

本书从基础知识出发，通过丰富的案例，详细介绍数据抓取和分析的整个过程，帮助读者构建相关能力。

本书不同于大多数介绍爬虫的技术书，不仅讲述如何进行数据抓取，而且通过丰富的案例讲解抓取数据的思路，介绍数据分析、可视化的方法，以及如何根据数据分析结果，开发一个应用，以求为读者提供一个从采集数据到应用数据的完整视角。本书以介绍技术思路为主，不会详细介绍一些特别基础的知识点，例如，Python的基础知识、软件包的安装操作等，所以需要读者自行查阅一些相关资料。另外，由于移动应用、网站等更新速度非常快，当阅读到本书时，可能书中介绍的一些方法已经发生了变化，读者可以自行研究，把知识灵活地运用到实践中。

特别声明

本书仅限于讨论爬虫技术，书中展示的案例只是为了读者更好地理解抓取的思路和操作，达到防范信息泄露、保护信息安全的目的，请勿用于非法用途！严禁利用本书所提到的技术进行非法抓取，否则后果自负，本人和出版商不承担任何责任。

读者服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **提交勘误**：您对书中内容的修改意见可在 提交勘误 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方 读者评论 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/35508

目 录

第 1 章

基础知识

1.1 什么是爬虫

爬虫是“网络爬虫”的简称，在百度百科上网络爬虫的定义是：

网络爬虫（又被称为网页蜘蛛、网络机器人，在 FOAF 社区中，经常被称为网页追逐者），是一种按照一定的规则，自动抓取互联网信息的程序或者脚本。另外一些不常使用的名字还有蚂蚁、自动索引、模拟程序或者蠕虫。

我们最常接触的网络爬虫是百度、搜搜、谷歌（Google）等公司的搜索引擎，这些搜索引擎通过互联网上的入口获取网页，实时存储并更新索引。搜索引擎的基础就是网络爬虫，这些网络爬虫通过自动化的方式进行网页浏览并存储相关的信息。

在互联网的早期，大多数网站都是静态网站，没有大量的图片，更没有大量视频素材，那时的网络爬虫只要能处理静态的 HTML 网页。随着互联网内容的不断丰富，搜索文字、图片，甚至视频都成了最基本的需求。

技术上，AJAX、单页面应用、HTML5 等网页技术的发展淘汰了单纯的静态页面爬虫，类似搜索引擎的通用爬虫能够加载并处理好动态页面已成为基本的要求。另外，由于过于复杂的动态网页会给搜索引擎带来很大的困扰，因此大多数网站会在动态网站的基础上，通过 SEO（Search Engine Optimization，搜索引擎优化）对部分想要被搜索到的信息引入静态网页，以便能够更方便地被搜索引擎搜索到（进而容易被普通用户搜索到）。

近几年，互联网开始朝着移动应用的方向发展。海量的信息从移动端生产并消费，遗憾的是，搜索引擎通常并不能触及这些信息。例如，抖音等短视频 APP 中的视频，目前还不能在百度等搜索引擎搜索到；淘宝的商品信息也无法在常规搜索引擎中搜索到（只能在淘宝的 APP 中搜索到），等等。由于这些信息无法通过网页搜索到，因此搜索引擎不适合解决此类问题。在商务上，厂商之间可以通过合作的方式对移动应用中的内容进行查询，例如，搜狗就能搜索到微信公众号的信息。在技术上，可以开发定向爬虫抓取页面信息，再对其中的数据进行处理，例如各种比价网站收集价格信息的过程等。

定向爬虫

定向爬虫抓取的是特定的信息，它获取信息的方式多种多样，存储及分析的方式也随着应用不同而不同。信息既可能存在于网页之中，又可能存在于各种移动端应用中。

对于定向的静态网页的抓取，我们会分析出需要抓取的网站，然后根据链接的内容、关键字等信息决定下一个网页的抓取。这种网页的抓取极具针对性，只需找到遍历的方式方法即可。通用的爬虫框架能够减轻一部分开发工作，但自己针对特定需求写一个专用的爬虫也不难。

在丰富的移动资源中存在大量的有用信息，并且绝大多数应用都采取了前后端分离的架构设计。前端调用后端的 API，后端会为前端提供结构化或者半结构化的数据（一般是 JSON 格式），所以通过分析数据来源的 API，可以模拟调用这些 API 来获取信息。结构化、半结构化的数据非常有利于数据处理，这也给数据存储、处理、分析带来了很大的便利。

1.2 数据获取渠道

数据获取的渠道多种多样，当一条路走不通时，可以试试其他渠道。常见的渠道有如下四类。

1. 网站

很多非移动专享的应用都有自己的网站。网站可分为 PC 端和移动端两种，很多时候，它们对于反爬虫行为的防范并不一致。网站对应的调试工具很多，可以更加方便地进行破解。但很多网站都有许多防范措施，导致分析、抓取信息的成本较高。

2. 手机 APP

某些应用则只有手机 APP 版本，完全没有提供网站，典型的 APP 有“摩拜单车”“立刻出行”等。针对这些应用，我们可以使用抓包工具抓取 APP 的流量并对请求进行分析。mitmdump 等工具可以拦截流量，再结合 Appium 模拟用户的单击行为，从而进行自动化的数据获取。

3. 小程序

越来越多的应用都开发了微信小程序或支付宝小程序。小程序提供了快捷的访问路径，也为数据获取提供了新的途径。小程序多用 JavaScript 书写，因此可以方便地进行反编译。在技术上，由于小程序相对比较新，很多在方面设计欠佳，导致加密签名的方法可以被反编译出来，从而顺利获取 API 的访问方式。对此，小程序开发者可以考虑登录验证的方式，即对访问进行验证，从而阻止非预期的高频访问。

4. 搜索引擎

搜索引擎上保存了大量的网站信息，如果不能从网站直接抓取，则可以考虑抓取搜索引擎的快照页面。快照页面中的信息有可能不是最新的，但可以在一定程度上帮助我们获取数据。例如，“天眼查”的新界面中已经隐藏了一些信息，但百度的快照中依然存在一些旧的信息（例如联系电话、邮箱等），通常这些信息不会轻易变更。当然搜索引擎也有一些防抓取的措施，以及页面访问数量等限制，我们可以通过缩小关键字范围，或使用搜索工具限制时间等措施减少页面的总量。

1.3 抓包分析工具

1. Charles

Charles 是一款跨平台的抓包软件，它本质上是一款代理软件，当应用将流量转发到 Charles 时，它能够对数据包进行拦截、分析以及修改，从而达到分析网络流量的目的。它能够支持任何允许设置网络代理的软件，支持的代理类型包括 HTTP 代理、HTTPS 代理、Socks5 代理。

在浏览器上设置 Charles 和普通的代理设置并无差异，在此不再赘述。在移动端配置时，由于 Android 的某些应用会忽略系统的全局代理，所以 Charles 无法获得流量。这时候就要借助 Postern 这款软件进行流量的转发。Postern 会模拟出一个 VPN 来拦截系统的所有流量，并转发到 Charles 中。Postern 可以自定义规则，选择性地将流量转发到 Charles 中，从而过滤一些无用的信息。

Charles 还提供了编辑请求、生成 curl 命令等非常实用的功能，在后面的例子中将会介绍。

需要注意的是，Android 7 及 iOS 的系统中引入了 SSL Pinning 技术，因此无法抓取到一些 HTTPS 的请求。SSL Pinning 会检查客户端的证书是否和服务端的证书相匹配，如果不匹配则断开连接。由于 Charles 属于代理软件，可以认为是中间人攻击软件，因此解密 SSL 时需要安装 Charles 的证书才能解密在 Android 7 之前的版本，手机上安装了 Charles 的证书后，客户端验证证书链时认为证书是匹配的，从而可以建立连接。

绕过 SSL Pinning 的方法有：

- 使用 Android 7 以下版本的手机，这是最为简单有效的方案。
- 破解 Android 7 以上的手机并进行 root，安装 Xposed 框架，然后安装 JustTrustMe 进行破解；或者对 root 过的手机使用 Frida 结合 Universal Android SSL Pinning Bypass with Frida 脚本。

2. Packet Capture

Packet Capture 是 Android 系统上一款好用的抓包软件，它无须对手机进行 root，

即可方便地查看流量的细节，因为它可以模拟成一个 VPN 对应用程序的请求进行抓包。与 Charles 相比，Packet Capture 有以下不同：

- 只能在手机上抓取、查看，处理起来不是很方便。常用来做快速抓包，或判定请求的类型和参数等。
- 不能修改网络的流量。
- 能够针对特定的 APP 进行流量拦截，这样可以减少一些软件的后台通信对抓包的干扰。
- 以 VPN 的形式提供服务，可以抓取设置代理后无法工作的软件。

3. mitmproxy

mitmproxy 是用 Python 和 C 开发的一款中间人代理软件。与 Charles 类似，mitmproxy 可在终端下运行，并且可以用来拦截、修改、重放和保存 HTTP/HTTPS 请求。与 Charles 不同的是，mitmproxy 可以利用 Python 脚本进行定制化的操作。通常来讲，我们会用 Charles 进行一系列分析，在需要拦截、修改、保存请求时再使用 mitmproxy 工具及其脚本。

1.4　爬虫和反爬虫的斗争

1. 常见的方法

在抓取对方网站、APP 应用的相关数据时，经常会遇到一系列的方法阻止爬虫。一方面是为了保证服务的质量，另一方面是保护数据不被获取。常见的一些反爬虫和反反爬虫的手段如下。

（1）IP 限制

IP 限制是很常见的一种反爬虫的方式。服务端在一定时间内统计 IP 地址的访问次数，当次数、频率达到一定阈值时返回错误码或者拒绝服务。这种方式比较直接简单，但在 IPv4 资源越来越不足的情况下，很多用户共享一个 IP 出口，典型的如“长城宽带”等共享型的 ISP。另外手机网络中的 IP 地址也是会经常变化的，如果对这些 IP 地址进行阻断，则会将大量的正常用户阻止在外。

对于大多数不需要登录就可以进行访问的网站，通常也只能使用 IP 地址进行限制。比如“Freelancer 网站”，大量的公开数据可以被访问，但同一个 IP 地址的访问是有一定的限制的。针对 IP 地址限制非常有效的方式是，使用大量的“高匿名”代理资源。这些代理资源可以对源 IP 地址进行隐藏，从而让对方服务器看起来是多个 IP 地址进行访问。另一种限制方式是，根据业务需要，对国内、国外的 IP 地址进行单独处理，进而对国外的高匿名代理进行阻断，例如使用海外的 IP 地址访问“天眼查网站”则无法访问。

（2）验证码

验证码是一种非常常见的反爬虫方式。服务提供方在 IP 地址访问次数达到一定数量后，可以返回验证码让用户进行验证。这种限制在不需要登录的网页界面比较常见，它需要结合用户的 cookie 或者生成一个特殊标识对用户进行唯一性判断，以防止同一个 IP 地址访问频率过高。验证码的存在形式非常多，有简单的数字验证码、字母数字验证码、字符图形验证码，网站也可以用极验验证码等基于用户行为的验证码。针对简单验证码，可以使用打码平台进行破解。这种平台通过脚本上传验证的图片，由打码公司雇用的人工进行识别。针对极验验证等更复杂的验证码，可以尝试模拟用户的行为绕过去，但通常比较烦琐，难度较大。谷歌所用的验证码更为复杂，通常是用户端结合云端进行手工打码，但会带来整体成本较高的问题。

要想绕过这些验证码的限制，一种思路是在出现验证码之前放弃访问，更换 IP 地址。ADSL 拨号代理提供了这种可能性。ADSL 通过拨号的方式上网，需要输入 ADSL 账号和密码，每次拨号就更换一个 IP 地址。不同地域的 IP 地址分布在多个地址段，如果 IP 地址都能使用，则意味着 IP 地址量级可达千万。如果我们将 ADSL 主机作为代理，每隔一段时间主机拨号一次（换一个 IP），这样可以有效防止 IP 地址被封禁。这种情况下，IP 地址的有效时限通常很短，通常在 1 分钟以下。结合大量的 ADSL 拨号代理可以达到并行获取大量数据的可能。如果网站使用了一些特殊的唯一性的标识，则很容易被对方网站识别到，从而改进反爬虫策略，面对这种情况，单独切换 IP 地址也会无效。遇到这种情况，必须要搞清楚标识的生成方式，进而模拟真实用户的访问。

（3）登录限制

登录限制是一种更加有效的保护数据的方式。网站或者 APP 可以展示一些基础

的数据，当需要访问比较重要或者更多的数据时则要求用户必须登录。例如，在天眼查网站中，如果想要查看更多的信息，则必须用账号登录；“知乎”则是必须在登录后才能看到更多的信息。登录后，结合用户的唯一标识，可以进行计数，当访问频度、数量达到一定阈值后即可判断为爬虫行为，从而进行拦截。针对“登录限制”的方法，可以使用大量的账号进行登录，但成本通常比较高。

针对微信小程序，可以使用 wx.login()方法，这种方式不需要用户的介入，因而不伤害用户的体验。小程序调用后会获取用户的唯一标识，后端可以根据这个唯一标识进行反爬虫的判断。

（4）数据伪装

在网页上，我们可以监听流量，然后模拟用户的正常请求。mitmproxy 等工具可以监听特定网址的访问（通常是 API 的地址），然后将需要的数据存储下来。基于 Chrome Headless 的工具也可以监听到流量并进行解析。在这种情况下，某些网站会对数据进行一些伪装来增加复杂度。例如，在某网站上展示的价格为 945 元，在 DOM 树中是以 CSS 进行了一些伪装。要想得到正确的数值，必须对 CSS 的规则进行一些计算才行，某网站上展示的价格如图 1-1 所示。

图 1-1　某网站上展示的价格

该网站使用特殊的字体对数据进行了伪装。例如，3400，对应显示的是 1400，如图 1-2 所示。如果能够找到所有的字体对应的关系，则可以逆向出正确的价格。

某电影网站使用特殊的字符进行数据隐藏，这种不可见的字符会增加复杂度，但还是可以通过对应的 UTF-8 字符集找到对应关系，从而得到正确的值，如图 1-3 所示。

图 1-2　3400 显示为 1400

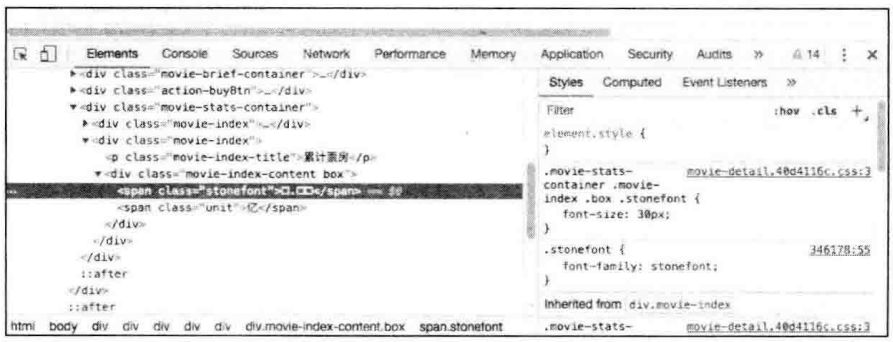


图 1-3　网站用特殊字符进行伪装