实战 Python 网络爬虫

从爬虫软件开发到自己动手开发爬虫框架

从原理到实践，深入浅出，热门爬虫核心技术全掌握
涵盖丰富的爬虫工具、库、框架，十余个实战项目
资深爬虫工程师倾力奉献，入门、进阶、求职必备

黄永祥 / 著

清華大學出版社

实战 Python 网络爬虫

黄永祥 / 著

清华大学出版社
北京

内 容 简 介

本书从原理到实践，循序渐进地讲述了使用 Python 开发网络爬虫的核心技术。全书从逻辑上可分为基础篇、实战篇和爬虫框架篇三部分。基础篇主要介绍了编写网络爬虫所需的基础知识，包括网站分析、数据抓取、数据清洗和数据入库。网站分析讲述如何使用 Chrome 和 Fiddler 抓包工具对网站做全面分析；数据抓取介绍了 Python 爬虫模块 Urllib 和 Requests 的基础知识；数据清洗主要介绍字符串操作、正则和 BeautifulSoup 的使用；数据入库讲述了 MySQL 和 MongoDB 的操作，通过 ORM 框架 SQLAlchemy 实现数据持久化，进行企业级开发。实战篇深入讲解了分布式爬虫、爬虫软件的开发、12306 抢票程序和微博爬取等。框架篇主要讲述流行的爬虫框架 Scrapy，并以 Scrapy 与 Selenium、Splash、Redis 结合的项目案例，让读者深层次了解 Scrapy 的使用。此外，本书还介绍了爬虫的上线部署、如何自己动手开发一款爬虫框架、反爬虫技术的解决方案等内容。

本书使用 Python 3.X 编写，技术先进，项目丰富，适合欲从事爬虫工程师和数据分析师岗位的初学者、大学生和研究生使用，也很适合有一些网络爬虫编写经验，但希望更加全面、深入理解 Python 爬虫的开发人员使用。

图书在版编目（CIP）数据

实战 Python 网络爬虫 / 黄永祥著. —北京：清华大学出版社，2019
ISBN 978-7-302-52489-2

Ⅰ. ①实… Ⅱ. ①黄… Ⅲ. ①软件工具－程序设计 Ⅳ. ①TP311.561

中国版本图书馆 CIP 数据核字（2019）第 043080 号

责任编辑： 王金柱
封面设计： 王　翔
责任校对： 闫秀华
责任印制： 沈　露

出版发行： 清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者： 清华大学印刷厂
经　　销： 全国新华书店
开　　本： 190mm×260mm　　**印　　张：** 30.25　　**字　　数：** 774 千字
版　　次： 2019 年 6 月第 1 版　　**印　　次：** 2019 年 6 月第 1 次印刷
定　　价： 99.00 元

产品编号：082567-01

前　言

随着大数据和人工智能的普及，Python 的地位也变得水涨船高，许多技术人员投身于 Python 开发，其中网络爬虫是 Python 最为热门的应用领域之一。在爬虫领域，Python 可以说是处于霸主地位，Python 能解决爬虫开发过程中所遇到的难题，开发速度快且支持异步编程，大大缩短了开发周期。此外，从事数据分析的工程师，为获取数据，很多时候也会用到网络爬虫的相关技术，因此，Python 爬虫编程已成为爬虫工程师和数据分析师的必备技能。

本书结构

本书共分 28 章，各章内容概述如下：

第 1 章介绍什么是网络爬虫、爬虫的类型和原理、爬虫搜索策略和爬虫的合法性及开发流程。

第 2 章讲解爬虫开发的基础知识，包括 HTTP 协议、请求头和 Cookies 的作用、HTML 的布局结构、JavaScript 的介绍、JSON 的数据格式和 Ajax 的原理。

第 3 章介绍使用 Chrome 开发工具分析爬取网站，重点介绍开发工具的 Elements 和 Network 标签的功能和使用方式，并通过开发工具分析 QQ 网站。

第 4 章主要介绍 Fiddler 抓包工具的原理和安装配置，Fiddler 用户界面的各个功能及使用方法。

第 5 章讲述了 Urllib 在 Python 2 和 Python 3 的变化及使用，包括发送请求、使用代理 IP、Cookies 的读写、HTTP 证书验收和数据处理。

第 6 章~第 8 章介绍 Python 第三方库 Requests、Requests-Cache 爬虫缓存和 Requests-HTML，包括发送请求、使用代理 IP、Cookies 的读写、HTTP 证书验收和文件下载与上传、复杂的请求方式、缓存的存储机制、数据清洗以及 Ajax 动态数据爬取等内容。

第 9 章介绍网页操控和数据爬取，重点讲解 Selenium 的安装与使用，并通过实战项目“百度自动答题”，讲解了 Selenium 的使用。

第 10 章介绍手机 App 数据爬取，包括 Appium 的原理与开发环境搭建、连接 Android 系统，并通过实战项目“淘宝商品采集”，介绍了 App 数据的爬取技巧。

第 11 章介绍 Splash、Mitmproxy 与 Aiohttp 的安装和使用，包括 Splash 动态数据抓取、Mitmproxy 抓包和 Aiohttp 高并发抓取。

第 12 章介绍验证码的种类和识别方法，包括 OCR 的安装和使用、验证码图片处理和使用第三方平台识别验证码。

第 13 章讲述数据清洗的三种方法，包括字符串操作（截取、查找、分割和替换）、正则表达式的使用和第三方库 BeautifulSoup 的安装以及使用。

第 14 章讲述如何将数据存储到文件，包括 CSV、Excel 和 Word 文件的读取和写入方法。

第 15 章介绍 ORM 框架 SQLAlchemy 的安装及使用，实现关系型数据库持久化存储数据。

第 16 章讲述非关系型数据库 MongoDB 的操作，包括 MongoDB 的安装、原理和 Python 实现

MongoDB 的读写。

第 17 章至第 21 章介绍了 5 个实战项目，分别是：爬取 51Job 招聘信息、分布式爬虫——QQ 音乐、12306 抢票爬虫、微博爬取和微博爬虫软件的开发。

第 22 章至第 25 章介绍了 Scrapy 爬虫框架，包括 Scrapy 的运行机制、项目创建、各个组件的编写（Setting、Items、Item Pipelines 和 Spider）和文件下载及 Scrapy 中间件，并通过实战项目“Scrapy+Selenium 爬取豆瓣电影评论”、“Scrapy+Splash 爬取 B 站动漫信息”和“Scrapy+Redis 分布式爬取猫眼排行榜”、“爬取链家楼盘信息”和“QQ 音乐全站爬取”，深入讲解了 Scrapy 的应用和分布式爬虫的编写技巧。

第 26 章介绍爬虫的上线部署，包括非框架式爬虫和框架式爬虫的部署技巧。

第 27 章介绍常见的反爬虫技术，并给出了可行的反爬虫解决方案。

第 28 章介绍爬虫框架的编写，学习如何自己动手编写一款爬虫框架，以满足特定业务场景的需求。

本书特色

循序渐进，涉及面广：本书站在初学者的角度，循序渐进地介绍了使用 Python 开发网络爬虫的各种知识，内容由浅入深，几乎涵盖了目前网络爬虫开发的各种热门工具和前瞻性技术。

实战项目丰富，扩展性强：本书采用大量的实战项目进行讲解，力求通过实际应用使读者更容易地掌握爬虫开发技术，以应对业务需求。本书项目经过编者精心设计和挑选，根据实际开发经验总结而来，涵盖了在实际开发中所遇到的各种问题。对于精选项目，尽可能做到步骤详尽、结构清晰、分析深入浅出，而且案例的扩展性强，读者可根据实际需求扩展开发。

从理论到实践，注重培养爬虫开发思维：在讲解过程中，不仅介绍理论知识，注重培养读者的爬虫开发思维，而且安排了综合应用实例或小型应用程序，使读者能顺利地将理论应用到实践中。

特色干货，倾情分享：本书大部分内容都来自作者多年来的编程实践，操作性很强。值得关注的是，本书还介绍了爬虫软件和爬虫框架的开发，供学有余力的读者扩展知识结构，提升开发技能。

源代码下载

本书所有程序代码均在 Python 3.6 下调试通过，源代码 Github 下载地址：

https://github.com/xyjw/python-Reptile

你也可以扫描下面的二维码下载。

如果你在下载过程中遇到问题，可发送邮件至 554301449@qq.com 获得帮助，邮件标题为“实

战 Python 网络爬虫下载资源”。

技术服务

读者在学习或者工作的过程中，如果遇到实际问题，可以加入 QQ 群 93314951 与笔者联系，笔者会在第一时间给予回复。

读者对象

本书主要适合以下读者阅读：

- Python 网络爬虫初学者及在校学生。
- Python 初级爬虫工程师。
- 从事数据抓取和分析的技术人员。
- 学习 Python 程序设计的开发人员。

虽然笔者力求本书更臻完美，但由于水平所限，难免会出现错误，特别是实例中爬取的网站可能随时更新，导致源码在运行过程中出现问题，欢迎广大读者和高手专家给予指正，笔者将十分感谢。

黄永祥

2019 年 1 月

目　录

第1章

理解网络爬虫

1.1 爬虫的定义

网络爬虫是一种按照一定的规则自动地抓取网络信息的程序或者脚本。简单来说，网络爬虫就是根据一定的算法实现编程开发，主要通过 URL 实现数据的抓取和发掘。

随着大数据时代的发展，数据规模越来越庞大，数据类型繁多，但是数据价值普遍较低。为了从庞大的数据体系里获取有价值的数据，从而延伸了网络爬虫、数据分析等多个职位。近几年，网络爬虫的需求更是井喷式地爆发，在招聘的供求市场上往往是供不应求，造成这个现状的主要原因就是求职者的专业水平低于需求企业的要求。

传统的爬虫有百度、Google、必应等搜索引擎，这类通用的搜索引擎都有自己的核心算法。但是，通用的搜索引擎存在着一定的局限性：

（1）不同的搜索引擎对于同一个搜索会有不同的结果，搜索出来的结果未必是用户需要的信息。

（2）通用的搜索引擎扩大了网络覆盖率，但有限的搜索引擎服务器资源与无限的网络数据资源之间的矛盾将进一步加深。

（3）随着网络上数据形式繁多和网络技术的不断发展，图片、数据库、音频、视频多媒体等不同数据大量出现，通用搜索引擎往往对这些信息含量密集且具有一定结构的数据无能为力，不能很好地发现和获取。

因此，为了得到准确的数据，定向抓取相关网页资源的聚焦爬虫应运而生。聚焦爬虫是一个自动下载网页的程序，可根据设定的抓取目标有目的性地访问互联网上的网页与相关的 URL，从而获取所需要的信息。与通用爬虫不同，聚焦爬虫并不追求全面的覆盖率，而是抓取与某一特定内容相关的网页，为面向特定的用户提供准备数据资源。

1.2 爬虫的类型

网络爬虫根据系统结构和开发技术大致可以分为 4 种类型：通用网络爬虫、聚焦网络爬虫、增量式网络爬虫和深层网络爬虫。

通用网络爬虫又称全网爬虫，常见的有百度、Google、必应等搜索引擎，爬行对象从一些初始 URL 扩充到整个网站，主要为门户站点搜索引擎和大型网站服务采集数据，具有以下特点：

（1）由于商业原因，引擎的算法是不会对外公布的。

（2）这类网络爬虫的爬取范围和数量巨大，对于爬取速度和存储空间要求较高，爬取页面的顺序要求相对较低。

（3）待刷新的页面太多，通常采用并行工作方式，但需要较长时间才能刷新一次页面。

（4）存在一定缺陷，通用网络爬虫适用于为搜索引擎搜索广泛的需求。

聚焦网络爬虫又称主题网络爬虫，是选择性地爬取根据需求的主题相关页面的网络爬虫。与通用网络爬虫相比，聚焦爬虫只需要爬取与主题相关的页面，不需要广泛地覆盖无关的网页，很好地满足一些特定人群对特定领域信息的需求。

增量式网络爬虫是指对已下载网页采取增量式更新和只爬取新产生或者已经发生变化的网页的爬虫，它能够在一定程度上保证所爬取的页面尽可能是新的页面。只会在需要的时候爬取新产生或发生更新的页面 ，并不重新下载没有发生变化的页面，可有效减少数据下载量，及时更新已爬取的网页，减小时间和空间上的耗费，但是增加了爬取算法的复杂度和实现难度，基本上这类爬虫在实际开发中不太普及。

深层网络爬虫是大部分内容不能通过静态 URL 获取的、隐藏在搜索表单后的、只有用户提交一些关键词才能获得的网络页面。例如某些网站需要用户登录或者通过提交表单实现提交数据。这类爬虫也是本书讲述的重点之一。

实际上，聚焦网络爬虫、增量式网络爬虫和深层网络爬虫可以通俗地归纳为一类，因为这类爬虫都是定向爬取数据。相比于通用爬虫，这类爬虫比较有目的性，也就是网络上经常说的网络爬虫，而通用爬虫在网络上通常称为搜索引擎。

1.3 爬虫的原理

通用网络爬虫的实现原理及过程如图 1-1 所示。

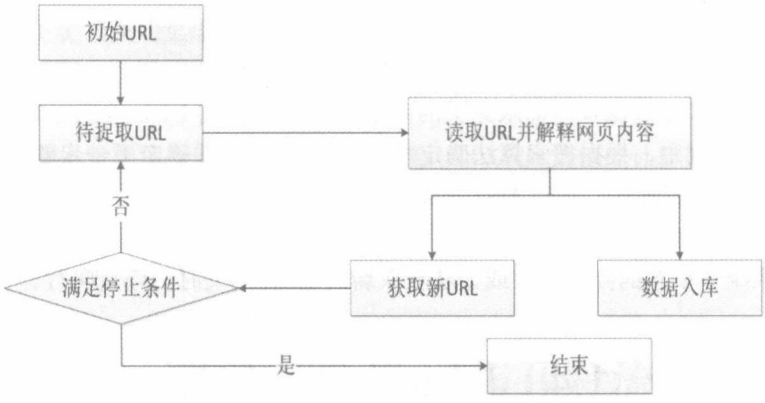

图 1-1　通用爬虫实现的原理及过程

通用网络爬虫的实现原理：

（1）获取初始的 URL。初始的 URL 地址可以人为地指定，也可以由用户指定的某个或某几个初始爬取网页决定。

（2）根据初始的 URL 爬取页面并获得新的 URL。获得初始的 URL 地址之后，先爬取当前 URL 地址中的网页信息，然后解析网页信息内容，将网页存储到原始数据库中，并且在当前获得的网页信息里发现新的 URL 地址，存放于一个 URL 队列里面。

（3）从 URL 队列中读取新的 URL，从而获得新的网页信息，同时在新网页中获取新 URL，并重复上述的爬取过程。

（4）满足爬虫系统设置的停止条件时，停止爬取。在编写爬虫的时候，一般会设置相应的停止条件，爬虫则会在停止条件满足时停止爬取。如果没有设置停止条件，爬虫就会一直爬取下去，一直到无法获取新的 URL 地址为止。

聚焦网络爬虫的执行原理和过程与通用爬虫大致相同，在通用爬虫的基础上增加两个步骤：定义爬取目标和筛选过滤 URL，原理如图 1-2 所示。

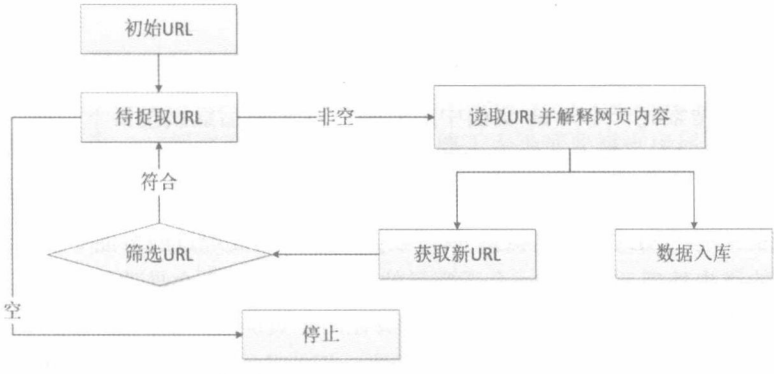

图 1-2　聚焦网络爬虫的原理

聚焦网络爬虫的实现原理：

（1）制定爬取方案。在聚焦网络爬虫中，首先要依据需求定义聚焦网络爬虫爬取的目标以及整体的爬取方案。