HZ BOOKS
华章IT

以实战为导向，讲透Python网络爬虫各项核心技术和主流框架，帮助读者快速、深度掌握网络爬虫的爬取技术与反爬攻关技巧

韦玮◎著

Deep in Python Web Crawler
Core Technology、Frame and Practices

精通Python网络爬虫

核心技术、框架与项目实战

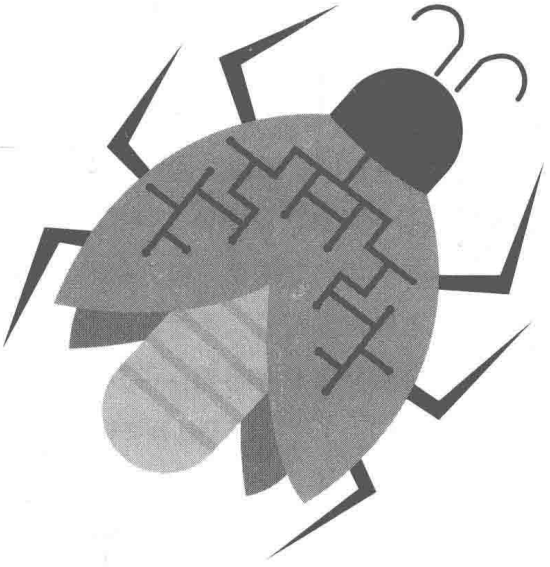
机械工业出版社
China Machine Press

Deep in Python Web Crawler
Core Technology、Frame and Practices

精通Python网络爬虫

核心技术、框架与项目实战

韦玮◎著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

精通 Python 网络爬虫：核心技术、框架与项目实战 / 韦玮著 . —北京：机械工业出版社，2017.2（2017.5 重印）

ISBN 978-7-111-56208-5

I. 精… II. 韦… III. 软件工具 – 程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2017）第 040626 号

精通 Python 网络爬虫
核心技术、框架与项目实战

出版发行：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：何欣阳　　责任校对：李秋荣

印　　刷：中国电影出版社印刷厂　　版　　次：2017 年 5 月第 1 版第 2 次印刷

开　　本：186mm × 240mm　1/16　　印　　张：19

书　　号：ISBN 978-7-111-56208-5　　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）88379426　88361066　　投稿热线：（010）88379604

购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

Preface 前　言

为什么写这本书

网络爬虫其实很早就出现了，最开始网络爬虫主要应用在各种搜索引擎中。在搜索引擎中，主要使用通用网络爬虫对网页进行爬取及存储。

随着大数据时代的到来，我们经常需要在海量数据的互联网环境中搜集一些特定的数据并对其进行分析，我们可以使用网络爬虫对这些特定的数据进行爬取，并对一些无关的数据进行过滤，将目标数据筛选出来。对特定的数据进行爬取的爬虫，我们将其称为聚焦网络爬虫。在大数据时代，聚焦网络爬虫的应用需求越来越大。

目前在国内 Python 网络爬虫的书籍基本上都是从国外引进翻译的，国内的本版书籍屈指可数，故而我跟华章的副总编杨福川策划了这本书。本书的撰写过程中各方面的参考资料非常少，因此完成本书所花费的精力相对来说是非常大的。

本书从系统化的视角，为那些想学习 Python 网络爬虫或者正在研究 Python 网络爬虫的朋友们提供了一个全面的参考，让读者可以系统地学习 Python 网络爬虫的方方面面，在理解并掌握了本书的实例之后，能够独立编写出自己的 Python 网络爬虫项目，并且能够胜任 Python 网络爬虫工程师相关岗位的工作。

同时，本书的另一个目的是，希望可以给大数据或者数据挖掘方向的从业者一定的参考，以帮助这些读者从海量的互联网信息中爬取需要的数据。所谓巧妇难为无米之炊，有了这些数据之后，从事大数据或者数据挖掘方向工作的读者就可以进行后续的分析处理了。

本书的主要内容和特色

本书是一本系统介绍 Python 网络爬虫的书籍，全书注重实战，涵盖网络爬虫原理、如何手写 Python 网络爬虫、如何使用 Scrapy 框架编写网络爬虫项目等关于 Python 网络爬虫的方方面面。

本书的主要特色如下：

- 系统讲解 Python 网络爬虫的编写方法，体系清晰。
- 结合实战，让读者能够从零开始掌握网络爬虫的基本原理，学会编写 Python 网络爬虫以及 Scrapy 爬虫项目，从而编写出通用爬虫及聚焦爬虫，并掌握常见网站的爬虫反屏蔽手段。
- 有配套免费视频，对于书中的难点，读者可以直接观看作者录制的对应视频，加深理解。
- 拥有多个爬虫项目编写案例，比如博客类爬虫项目案例、图片类爬虫项目案例、模拟登录爬虫项目等。除此之外，还有很多不同种类的爬虫案例，可以让大家在理解这些案例之后学会各种类型爬虫的编写方法。

总之，在理解本书内容并掌握书中实例之后，读者将能胜任 Python 网络爬虫工程师方向的工作并学会各种类型网络爬虫项目的编写。此外，本书对于大数据或数据挖掘方向的从业者也非常有帮助，比如可以利用 Python 网络爬虫轻松获取所需的数据信息等。

本书面向的读者

- Python 网络爬虫初学者
- 网络爬虫工程师
- 大数据及数据挖掘工程师
- 高校计算机专业的学生
- 其他对 Python 或网络爬虫感兴趣的人员

如何阅读本书

本书分为四篇，共计 20 章。

第一篇为理论基础篇（第 1 ~ 2 章），主要介绍了网络爬虫的基础知识，让大家从零开始对网络爬虫有一个比较清晰的认识。

第二篇为核心技术篇（第 3 ~ 9 章），详细介绍了网络爬虫实现的核心技术，包括网络爬虫的工作原理、如何用 Urllib 库编写网络爬虫、爬虫的异常处理、正则表达式、爬虫中 Cookie 的使用、手写糗事百科爬虫、手写链接爬虫、手写微信爬虫、手写多线程爬虫、浏览器伪装技术、Python 网络爬虫的定向爬取技术及实例等。学完这一部分内容，读者就可以写出自己的爬虫了。这部分的爬虫编写采用的是一步步纯手写的方式进行的，没有采用框架。

第三篇为框架实现篇（第 10 ~ 17 章），主要详细介绍了如何用框架实现 Python 网络爬虫项目。使用框架实现 Python 网络爬虫项目相较于手写方式更加便捷，主要包括 Python 爬虫框架分类、Scrapy 框架在各系统中的安装以及如何避免各种“坑”、如何用 Scrapy 框架编写爬虫项目、Scrapy 框架架构详解、Scrapy 的中文输出与存储、在 Scrapy 中如何使用 for 循环实现自动网页爬虫、如何通过 CrawlSpider 实现自动网页爬虫、如何将爬取的内容写进数据库等。其中第 12 章为基础部分，读者需要着重掌握。

第四篇为项目实战篇（第 18 ~ 20 章），分别讲述了博客类爬虫项目、图片类爬虫项目、模拟登录爬虫项目的编程及实现。其中，也会涉及验证码处理等方面的难点知识，帮助读者通过实际的项目掌握网络爬虫项目的编写。

勘误和支持

由于作者的水平有限，书中难免有一些错误或不准确的地方，恳请各位读者不吝指正。

相关建议各位可以通过微博 @ 韦玮 pig 或通过 QQ 公众号 a67899 或微信公众平台 weijc7789（可以直接扫描下方二维码添加）进行反馈，也可以直接向邮箱 ceo@iqianyue.com 发送邮件，期待能够收到各位读者的意见和建议，欢迎来信。

扫描关注 QQ 公众号

扫描关注微信公众号

致谢

感谢机械工业出版社华章公司的副总编杨福川老师与编辑李艺老师，在近一年的时间里，是你们一次次在我遇到困难的时候，给予我鼓励，让我可以坚持写下去。创作一本图书是非常艰苦的，除了技术知识等因素之外，还需要非常大的毅力。特别感谢杨福川在写作过程中对我各方面的支持，尤其是对我毅力的培养。

感谢 CSDN、51CTO 与极客学院，因为你们，让我在这个领域获得了更多的学员与支持。

感谢恩师何云景教授对我创业方面的帮助，因为有您，我才拥有了一个更好的创业开端及工作环境。

特别致谢

最后，需要特别感谢的是我的女友，因为编写这本书，少了很多陪你的时间，感谢你的不离不弃与理解包容。希望未来可以加倍弥补你那些错过吃的美食和那些错过逛的街道。

同时，也要感谢你帮我完成书稿的校对工作，谢谢你的付出与支持。因为有了你默默的付出，我才能坚定地走下去；因为有了你不断的支持，我才可以安心地往前冲。

感谢爷爷从小对我人生观、价值观的培养，您是一个非常有思想的人。

感谢远方的父母、叔叔、姐姐，那些亲情的陪伴是我最珍贵的财富。

谨以此书献给热爱 Python 的朋友们！

Contents 目　录

第一篇 Part 1

理论基础篇

网络爬虫也叫做网络机器人，可以代替人们自动地在互联网中进行数据信息的采集与整理。在大数据时代，信息的采集是一项重要的工作，如果单纯靠人力进行信息采集，不仅低效繁琐，搜集的成本也会提高。此时，我们可以使用网络爬虫对数据信息进行自动采集，比如应用于搜索引擎中对站点进行爬取收录，应用于数据分析与挖掘中对数据进行采集，应用于金融分析中对金融数据进行采集，除此之外，还可以将网络爬虫应用于舆情监测与分析、目标客户数据的收集等各个领域。当然，要学习网络爬虫开发，首先需要认识网络爬虫，在本篇中，我们将带领大家一起认识几种典型的网络爬虫，并了解网络爬虫的各项常见功能。

第 1 章 Chapter 1

什么是网络爬虫

随着大数据时代的来临，网络爬虫在互联网中的地位将越来越重要。互联网中的数据是海量的，如何自动高效地获取互联网中我们感兴趣的信息并为我们所用是一个重要的问题，而爬虫技术就是为了解决这些问题而生的。我们感兴趣的信息分为不同的类型：如果只是做搜索引擎，那么感兴趣的信息就是互联网中尽可能多的高质量网页；如果要获取某一垂直领域的数据或者有明确的检索需求，那么感兴趣的信息就是根据我们的检索和需求所定位的这些信息，此时，需要过滤掉一些无用信息。前者我们称为通用网络爬虫，后者我们称为聚焦网络爬虫。

1.1 初识网络爬虫

网络爬虫又称网络蜘蛛、网络蚂蚁、网络机器人等，可以自动化浏览网络中的信息，当然浏览信息的时候需要按照我们制定的规则进行，这些规则我们称之为网络爬虫算法。使用 Python 可以很方便地编写出爬虫程序，进行互联网信息的自动化检索。

搜索引擎离不开爬虫，比如百度搜索引擎的爬虫叫作百度蜘蛛（Baiduspider）。百度蜘蛛每天会在海量的互联网信息中进行爬取，爬取优质信息并收录，当用户在百度搜索引擎上检索对应关键词时，百度将对关键词进行分析处理，从收录的网页中找出相关网页，按照一定的排名规则进行排序并将结果展现给用户。在这个过程中，百度蜘蛛起到了至关重要的作用。那么，如何覆盖互联网中更多的优质网页？又如何筛选这些重复的页面？这些都是由百度蜘蛛爬虫的算法决定的。采用不同的算法，爬虫的运行效率会不同，爬取结果也会有所差异。所以，我们在研究爬虫的时候，不仅要了解爬虫如何实现，还需要知道一些常见爬虫的

算法，如果有必要，我们还需要自己去制定相应的算法，这些在后面都会为大家详细地讲解，在此，我们仅需要对爬虫的概念有一个基本的了解。

除了百度搜索引擎离不开爬虫以外，其他搜索引擎也离不开爬虫，它们也拥有自己的爬虫。比如 360 的爬虫叫 360Spider，搜狗的爬虫叫 Sogouspider，必应的爬虫叫 Bingbot。

如果想自己实现一款小型的搜索引擎，我们也可以编写出自己的爬虫去实现，当然，虽然可能在性能或者算法上比不上主流的搜索引擎，但是个性化的程度会非常高，并且也有利于我们更深层次地理解搜索引擎内部的工作原理。

大数据时代也离不开爬虫，比如在进行大数据分析或数据挖掘时，我们可以去一些比较大型的官方站点下载数据源。但这些数据源比较有限，那么如何才能获取更多更高质量的数据源呢？此时，我们可以编写自己的爬虫程序，从互联网中进行数据信息的获取。所以在未来，爬虫的地位会越来越重要。

1.2 为什么要学网络爬虫

在上一节中，我们初步认识了网络爬虫，但是为什么要学习网络爬虫呢？要知道，只有清晰地知道我们的学习目的，才能够更好地学习这一项知识，所以在这一节中，我们将会为大家分析一下学习网络爬虫的原因。

当然，不同的人学习爬虫，可能目的有所不同，在此，我们总结了 4 种常见的学习爬虫的原因。

1）学习爬虫，可以私人订制一个搜索引擎，并且可以对搜索引擎的数据采集工作原理进行更深层次地理解。

有的朋友希望能够深层次地了解搜索引擎的爬虫工作原理，或者希望自己能够开发出一款私人搜索引擎，那么此时，学习爬虫是非常有必要的。简单来说，我们学会了爬虫编写之后，就可以利用爬虫自动地采集互联网中的信息，采集回来后进行相应的存储或处理，在需要检索某些信息的时候，只需在采集回来的信息中进行检索，即实现了私人的搜索引擎。当然，信息怎么爬取、怎么存储、怎么进行分词、怎么进行相关性计算等，都是需要我们进行设计的，爬虫技术主要解决信息爬取的问题。

2）大数据时代，要进行数据分析，首先要有数据源，而学习爬虫，可以让我们获取更多的数据源，并且这些数据源可以按我们的目的进行采集，去掉很多无关数据。

在进行大数据分析或者进行数据挖掘的时候，数据源可以从某些提供数据统计的网站获得，也可以从某些文献或内部资料中获得，但是这些获得数据的方式，有时很难满足我们对数据的需求，而手动从互联网中去寻找这些数据，则耗费的精力过大。此时就可以利用爬虫技术，自动地从互联网中获取我们感兴趣的数据内容，并将这些数据内容爬取回来，作为我们的数据源，从而进行更深层次的数据分析，并获得更多有价值的信息。

3）对于很多 SEO 从业者来说，学习爬虫，可以更深层次地理解搜索引擎爬虫的工作原

理，从而可以更好地进行搜索引擎优化。

既然是搜索引擎优化，那么就必须要对搜索引擎的工作原理非常清楚，同时也需要掌握搜索引擎爬虫的工作原理，这样在进行搜索引擎优化时，才能知己知彼，百战不殆。

4）从就业的角度来说，爬虫工程师目前来说属于紧缺人才，并且薪资待遇普遍较高，所以，深层次地掌握这门技术，对于就业来说，是非常有利的。

有些朋友学习爬虫可能为了就业或者跳槽。从这个角度来说，爬虫工程师方向是不错的选择之一，因为目前爬虫工程师的需求越来越大，而能够胜任这方面岗位的人员较少，所以属于一个比较紧缺的职业方向，并且随着大数据时代的来临，爬虫技术的应用将越来越广泛，在未来会拥有很好的发展空间。

除了以上为大家总结的 4 种常见的学习爬虫的原因外，可能你还有一些其他学习爬虫的原因，总之，不管是什么原因，理清自己学习的目的，就可以更好地去研究一门知识技术，并坚持下来。

1.3　网络爬虫的组成

接下来，我们将介绍网络爬虫的组成。网络爬虫由控制节点、爬虫节点、资源库构成。

图 1-1 所示是网络爬虫的控制节点和爬虫节点的结构关系。

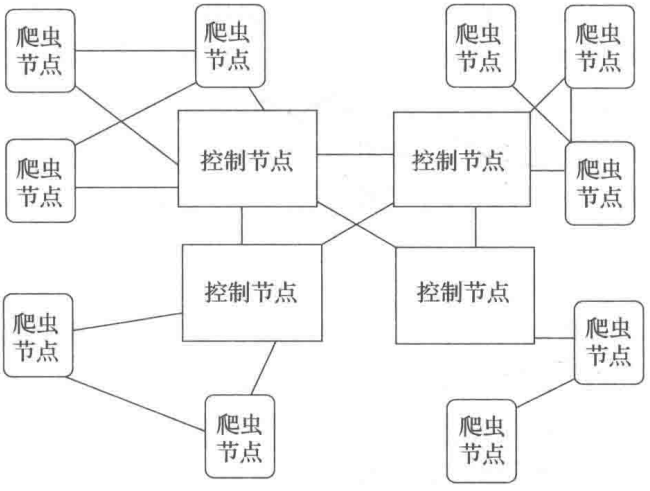


图 1-1　网络爬虫的控制节点和爬虫节点的结构关系

可以看到，网络爬虫中可以有多个控制节点，每个控制节点下可以有多个爬虫节点，控制节点之间可以互相通信，同时，控制节点和其下的各爬虫节点之间也可以进行互相通信，属于同一个控制节点下的各爬虫节点间，亦可以互相通信。

控制节点，也叫作爬虫的中央控制器，主要负责根据 URL 地址分配线程，并调用爬虫

节点进行具体的爬行。

爬虫节点会按照相关的算法，对网页进行具体的爬行，主要包括下载网页以及对网页的文本进行处理，爬行后，会将对应的爬行结果存储到对应的资源库中。

1.4 网络爬虫的类型

现在我们已经基本了解了网络爬虫的组成，那么网络爬虫具体有哪些类型呢？

网络爬虫按照实现的技术和结构可以分为通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深层网络爬虫等类型。在实际的网络爬虫中，通常是这几类爬虫的组合体。

首先我们为大家介绍**通用网络爬虫**（General Purpose Web Crawler）。通用网络爬虫又叫作全网爬虫，顾名思义，通用网络爬虫爬取的目标资源在全互联网中。通用网络爬虫所爬取的目标数据是巨大的，并且爬行的范围也是非常大的，正是由于其爬取的数据是海量数据，故而对于这类爬虫来说，其爬取的性能要求是非常高的。这种网络爬虫主要应用于大型搜索引擎中，有非常高的应用价值。

通用网络爬虫主要由初始 URL 集合、URL 队列、页面爬行模块、页面分析模块、页面数据库、链接过滤模块等构成。通用网络爬虫在爬行的时候会采取一定的爬行策略，主要有深度优先爬行策略和广度优先爬行策略。具体的爬行策略，我们将在第 3 章讲解，在此，我们只需要知道通用网络爬虫的基本构成和主要的爬行策略。

聚焦网络爬虫（Focused Crawler）也叫主题网络爬虫，顾名思义，聚焦网络爬虫是按照预先定义好的主题有选择地进行网页爬取的一种爬虫，聚焦网络爬虫不像通用网络爬虫一样将目标资源定位在全互联网中，而是将爬取的目标网页定位在与主题相关的页面中，此时，可以大大节省爬虫爬取时所需的带宽资源和服务器资源。聚焦网络爬虫主要应用在对特定信息的爬取中，主要为某一类特定的人群提供服务。

聚焦网络爬虫主要由初始 URL 集合、URL 队列、页面爬行模块、页面分析模块、页面数据库、链接过滤模块、内容评价模块、链接评价模块等构成。内容评价模块可以评价内容的重要性，同理，链接评价模块也可以评价出链接的重要性，然后根据链接和内容的重要性，可以确定哪些页面优先访问。聚焦网络爬虫的爬行策略主要有 4 种，即基于内容评价的爬行策略、基于链接评价的爬行策略、基于增强学习的爬行策略和基于语境图的爬行策略。关于聚焦网络爬虫具体的爬行策略，我们将在 1.5 节进行详细分析。

增量式网络爬虫（Incremental Web Crawler），所谓增量式，对应着增量式更新。增量式更新指的是在更新的时候只更新改变的地方，而未改变的地方则不更新，所以增量式网络爬虫，在爬取网页的时候，只爬取内容发生变化的网页或者新产生的网页，对于未发生内容变化的网页，则不会爬取。增量式网络爬虫在一定程度上能够保证所爬取的页面，尽可能是新页面。

深层网络爬虫（Deep Web Crawler），可以爬取互联网中的深层页面，在此我们首先需要