

# 目录

前言	1.1
爬虫简介	1.2
爬虫的叫法	1.3
为何叫爬虫	1.4
为何又叫模拟登陆	1.4.1
爬虫的核心步骤	1.5
下载	1.5.1
提取	1.5.2
保存	1.5.3
爬虫框架	1.6
为何需要爬虫框架	1.6.1
常见爬虫框架	1.6.2
爬虫相关总结	1.7
不同方案和爬虫步骤的对应关系	1.7.1
用Python写爬虫	1.7.2
相关名词和概念	1.7.3
附录	1.8
参考资料	1.8.1

# 爬取你要的数据：爬虫技术

- 最新版本：[v1.0](#)
- 更新时间：[20190328](#)

## 简介

整理爬虫技术的各种叫法，解释为何叫做爬虫，为何又被叫做模拟登陆，总结爬虫的核心步骤和阶段，以及每一步的各种细节包括优缺点和其他涉及的内容，继续解释为何要用爬虫框架，总结常见语言的各种爬虫框架，总结不用爬虫框架和常见爬虫框架与爬虫的不同步骤和功能的对应关系，总结爬虫相关名词和概念。

## 源码+浏览+下载

本书的各种源码、在线浏览地址、多种格式文件下载如下：

### Gitbook源码

- [crifan/crawl\\_your\\_data\\_spider\\_technology: 爬取你要的数据：爬虫技术](#)

### 如何使用此Gitbook源码去生成发布为电子书

详见：[crifan/gitbook\\_template: demo how to use crifan gitbook template and demo](#)

### 在线浏览

- [爬取你要的数据：爬虫技术 book.crifan.com](#)
- [爬取你要的数据：爬虫技术 crifan.github.io](#)

### 离线下载阅读

- [爬取你要的数据：爬虫技术 PDF](#)
- [爬取你要的数据：爬虫技术 ePUB](#)
- [爬取你要的数据：爬虫技术 Mobi](#)

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间：2019-03-28 23:03:50

# 爬虫简介

爬虫，此处主要指的是，能够从网站的页面或app等数据源中爬取到你所需要的数据的代码程序。

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 22:05:36

# 爬虫的叫法

爬虫有很多种常见的叫法，整理如下：

- 爬虫
  - 常见英文说法：
    - `crawler` =爬取数据的工具
      - `crawl` 英文原意：爬，爬行
    - `spider` =蜘蛛 =像蜘蛛捕获昆虫一样你去捕获你要的数据
      - `spider` 英文原意：蜘蛛
        - 为何把爬取数据的工具叫做蜘蛛，后续有类比解释
    - `scraper` =刮取到你想要的数据的工具
      - `scrape` 英文原意：刮取
    - `grab` =抓取你要的数据的工具
      - `grab` 英文原意：攫取，夺取
- 爬取数据
  - 常见英文说法：
    - `crawl data` = `crawling data`
    - `scraping data`
    - `grabbing data`
- 爬取网站 = 爬取网页
  - 常见英文说法：
    - `crawl website`
- 模拟登录
  - 常见英文说法：
    - `emulate login`
      - `login emulation`
  - 至于为何也会被叫做 模拟登录，后面会解释

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间： 2019-03-28 22:08:45

# 为何叫爬虫

下面解释一下，为何被叫做爬虫：

- 现实世界的蜘蛛网
  - 蜘蛛 =
  - 织网
  - 捕获 自己要的 东西
    - 食物 = 昆虫
- 计算机 世界的 互联网
  - 你 自己
  - 写 爬虫 (代码)
    - crawler = spider
  - 爬取 自己想要的 数据
    - 并且保存下来
  - 说明：
    - 互联网：是一个包含众多资源的大网络
    - 狹义 上说，主要指的是：
      - 各种 网站 = 网页
        - 里面有各种（我们想要爬取的）数据
        - 比如想要爬取汽车的车型车系，可以从 汽车之家等网站爬取
      - 广义 上说包含：
        - （上面提到的）各种网站=网页
        - 各种 app
          - 包括各种 Android 和 ios 中的app软件
          - 比如想要爬取别人的app中的一些数据
            - 比如爬取大众点评app中的商家和用户评论数据
        - 各种 其他渠道、终端 的数据和资源
          - 微信公众号
            - 理论上也是属于 网页
          - 小程序
            - 微信小程序
            - 支付宝小程序
          - 等等

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间： 2019-03-28 22:12:12

## 为何又叫模拟登陆

那 爬虫 为何也会被叫做 模拟登录 呢？

- 对于这种情况：想要爬取很多网站上的数据，需要用户（使用账号和密码等方式）去登录后才能获取到
- 所以要先去 模拟（用户）登录，然后才能 爬取数据
- 而模拟登录的过程，有时候或者经常，比后续的爬取数据更难，更复杂
- 所以此时的爬取全称是 先要模拟用户登录后再去爬取数据
- 也就常简称为 模拟登陆
  - 用 模拟登陆 指代 爬虫

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 22:14:13

## 爬虫的核心步骤

接下来介绍爬虫的原理、过程和步骤，以及相关涉及到的知识。

从原理上来说，写爬虫去爬数据的过程，最核心的就这3步：

- 下载 = `download`
- 提取 = `extract`
- 保存 = `save`

下面详细解释每一步的各种细节：

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间： 2019-03-28 22:16:07

# 下载

- 下载 = 下载网页
  - 做了什么：请求网址或api接口，去下载返回
  - 得到什么：html网页或json字符串
  - 实现方式：
    - 原始：自己写网络函数
      - 比如
        - Python 的 urllib
        - C# 的 HttpWebRequest + HttpWebResponse
        - [crifanLib.cs之Http](#)
      - 优点：更贴近和了解底层技术
      - 缺点：要求熟悉底层技术，相对用已有的库，写起来比较复杂
    - 使用已有第三方库
      - 优点：省心，高效
      - 缺点：
        - 要额外引入库，且要了解如何使用
        - 对于新手，往往是直接用了第三方库后，不了解内部机制
  - 涉及到
    - （尤其是新手需要学习）Http基本知识
      - Request
        - Method
          - GET
          - POST
          - 等
        - Header
          - User-Agent
          - Content-Type
          - Accept
          - Authorization
          - 等
        - Cookie
        - Body
          - data
            - json
      - Response
        - Status Code
        - Header
        - Cookie
        - Body
          - data
            - json
    - 主流数据格式：JSON
  - 教程：[HTTP知识总结](#)
  - 如果被爬方（网站，app等）
    - 需要用户登录后才能看到数据
      - 用技术绕过限制
        - 模拟登陆
        - 先要抓包分析出登录逻辑

- 再用代码模拟用户登录
- 做了一些 反爬 措施
  - 验证码
    - 用技术绕过限制
      - 验证码识别
      - (用第三方) 打码平台
  - IP限制 + 抓取频率 限制
    - 用技术绕过限制
      - IP代理池
      - 设置抓取的 间隔时间
  - 身份限制
    - Http的Headers
      - UA = User-Agent
- 被爬网站所含页面层级很多
  - 抓取策略
    - 深度优先遍历策略
    - 宽度优先遍历策略
    - 反向链接数策略
    - Partial PageRank策略
    - OPIC策略策略
    - 大站优先策略
  - 抓包
    - 什么是抓包
      - 对于下载来说，具体要请求网站 url 是什么，调用什么 api 接口，传递什么 参数，需要事先去分析和研究清楚，这个过程一般叫做： 抓包
    - 抓包的难度
      - 普通网页： 抓包分析，一般比较简单
      - 复杂网站： 对于需要登录才能获取到数据，且加了验证码等做了其他反爬措施和手段的网站和 app， 抓包分析起来，一般都很复杂
        - 复杂网站的抓包分析和破解，往往比（之后的，单纯的）写爬虫去 下载+提取+保存，要难多了
  - 常用辅助工具
    - 通用类
      - Wireshark
      - Postman
        - 用于对于api去设置参数并发送请求测试是否能获取数据
    - 针对网站网页类
      - Chrome 浏览器
        - 用于分析网络请求，页面元素等内容
    - 针对app (的api接口) 类
      - Charles
      - app抓包利器： Charles

# 提取

- 提取数据：
  - 做了什么：从（返回的）网页（的 `html`，`js` 等）或 `json` 中提取
  - 得到什么：自己需要的内容
  - 提取数据的方式：
    - 从 `json` 中提取想要的内容
      - 用 `json` 库，把 `json` 字符串转换为 `json` 对象（`dict`, 字典）即可
        - 无需（`html`）解析相关的库
      - 常见的库
        - Python
          - `json`
        - C#
          - `Newtonsoft.Json`
          - `JavaScriptSerializer`
      - 从 `html`, `js` 等内容中提取想要的内容
        - 原始方式=自己（用内置库）裸写代码
          - 正则
            - 【整理Book】应用广泛的超强搜索：正则表达式
          - Python
            - `re` 模块
              - 【整理Book】Python中的正则表达式：`re`模块详解
        - XPath
          - 【整理Book】XPath 知识总结
      - 用第三方库
        - Python
          - `lxml`
            - 【记录】Python中尝试用 `lxml` 去解析 `html` – 在路上
          - `BeautifulSoup`
            - 【整理Book】内容网页提取利器：`BeautifulSoup`
          - `PyQuery`
            - 【整理Book】Python心得：HTML 解析 `PyQuery`
          - `python-goose`
          - 等
        - C#
          - HTML 解析
            - `HtmlAgilityPack`
            - `sgml`
    - 涉及到
      - 字符编码 的问题
        - 否则编码搞不清，就会出现各种乱码问题
        - 需要学习相关 编码 知识
          - 【整理Book】字符编码详解与应用
          - 【整理Book】Python 心得：字符串和字符编码
        - `html` 的 `meta charset`
        - 编码检测
          - Python
            - `chardet`
      - 如果被爬方做了反爬

- 数据加密
- 用技术绕过
  - 找到解密的逻辑和方法
  - 【整理Book】安卓应用的安全和破解
  - 【已解决】尝试破解小花生app安卓apk希望看到api返回的json中的J的解密算法得到明文

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 22:36:59

# 保存

- 保存数据：
  - 做了什么：把数据保存到对应的地方
  - 得到什么：包含了我们要的特定格式的数据的文件或数据库
  - 保存成不同格式：
    - 文件
      - `txt`
      - `csv` / `excel`
        - 【整理Book】Python心得：操作CSV和Excel
    - 数据库
      - `mysql`
        - 【整理Book】主流关系数据库：MySQL
      - `mongodb`
        - 【整理Book】主流文档型数据库：MongoDB
      - `sqlite`
      - 等等

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间：2019-03-28 22:37:44

## 爬虫框架

而上面的三个步骤：下载+提取+保存，其中包含很多通用的，重复的逻辑和操作，所以有些人开发出来，独立的爬虫框架，方便我们去实现爬虫。

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 22:38:08

# 为何需要爬虫框架

接着来解释，为何要用爬虫框架：

- 框架帮你把大部分重复的工作都实现了
  - 做了哪些通用的事情
    - 下载
      - 网络异常时 自动重试 retry
      - 还可以设置
        - 最大 重试次数：最多重试几次
        - 如果还是不行，才视为下载失败
        - 重试间隔：两次重试之间的间隔时间
      - 好处：不用担心偶尔某次网络有问题，就导致下载失败了，因为还可以自动重试
      - 对比：自己裸写代码，就要考虑这种异常情况，导致自己爬虫代码臃肿和逻辑复杂
        - 花了太多精力在和爬取数据关系不大的方面，不值得，效率低
    - 下载 进程的管理
      - 同时发出多个url请求去下载内容
      - 有专门的进程管理和调度策略
      - 好处：能同时并发多个请求
      - 对比：自己裸写代码去下载，往往同一时刻只能有一个请求
        - 否则就要花很多精力去实现并发
    - url去重
      - 前后（不同页面，不同场景下）发出的多个url中是否有重复的
      - 如果有，则自动忽略掉，去掉，去除重复=去重
    - 提取
      - 做了啥
        - 内置常用的内容提取的库
          - PySpider 集成 PyQuery
          - Scrapy 集成选择器，支持： xpath、css、re
        - 同时支持可选的第三方的库
          - Scrapy 也支持用 BeautifulSoup 提取内容
      - 好处：不用额外安装和使用这些库
      - 对比：自己裸写代码就要考虑选用哪些合适的库去提取内容
    - 保存
      - 做了啥
        - 集成各种保存数据的接口和框架
          - PySpider
            - 自带默认保存为 sqlite 中
              - 可以从界面中导出 csv 或 json
            - 其他数据库接口
              - mysql
              - mongodb
              - 等等
          - 好处：可以方便的选择保存数据的方式，无需过多操心细节
          - 对比：自己裸写代码，还要安装不同的数据库的库，再手动写（sql等）代码去保存数据
    - 还带了很多额外的好用功能
      - PySpider
        - 带 UI界面， 调试非常方便
        - 支持网页内容是 执行js 后才生成的

- 通过集成第三方 phantomjs

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 22:43:38

# 常见爬虫框架

- Python
  - [PySpider](#)
    - Python爬虫框架：PySpider
  - [Scrapy](#)
    - 主流Python框架：Scrapy
  - 其他
    - [Grab](#)
    - [Portia](#)
    - [newspaper](#)
    - [ruia](#)
    - [Cola](#)
    - [Sasila](#)
- Java
  - [Nutch](#)
    - Nutch是一个基于Apache的Lucene，类似Google的完整网络搜索引擎解决方案，基于Hadoop的分布式处理模型保证了系统的性能，类似Eclipse的插件机制保证了系统的可客户化，而且很容易集成到自己的应用之中。
  - [Heritrix](#)
    - Heritrix是一个开源，可扩展的web爬虫项目。Heritrix设计成严格按照robots.txt文件的排除指示和META robots标签。
  - [crawler4j](#)
    - crawler4j is an open source web crawler for Java which provides a simple interface for crawling the Web. Using it, you can setup a multi-threaded web crawler in few minutes
  - [WebMagic](#)
    - 国人黄亿华先生的良心大作。无须配置、便于二次开发的爬虫框架，它提供简单灵活的API，只需少量代码即可实现一个爬虫
- Golang
  - [Pholcus](#)
  - [Colly](#)
- NodeJS
  - [headless-chrome-crawler](#)
- .NET
  - [abot](#)

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 22:50:22

# 爬虫相关总结

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 21:55:17

## 不同方案和爬虫步骤的对应关系

不用爬虫框架和常见爬虫框架和爬虫的不同步骤和功能的对应关系

	下载（网页）	提取（内容）	保存（数据）
自己裸写Python代码	<code>urllib</code>	<code>re</code>	<code>txt / csv</code>
用各种Python库组合	<code>requests</code>	<code>BeautifulSoup / lxml</code>	<code>csv / pymysql</code>
用框架 PySpider	<code>requests ( PySpider 的 self.crawl )</code>	<code>( PySpider 内置 的) PyQuery</code>	<code>( PySpider 内置) 各种 数据 库 (接口)</code>

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 23:02:15

# 用Python写爬虫

而不同语言去实现爬虫，其方便程度、难度、逻辑，又有一些区别，所以值得详细解释：

【整理Book】 Python心得：爬虫

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间： 2019-03-28 23:00:42

## 相关名词和概念

### 页面爬取策略 爬虫策略

在爬虫系统中,等待抓取URL队列是很重要的组成部分,等待抓取URL队列中的URL的顺序排列方式也是一个很重要的问题,因为这会决定到先抓取哪个页面,后抓取哪个页面.而决定这些URL排列顺序的方法,叫做抓取策略.下面主要介绍几种常见的抓取策略:

- 深度优先遍历策略
  - 深度优先遍历策略是指网络爬虫会从起始页开始,一个链接一个链接跟踪下去,直到处理完这条线路之后才会转入下一个起始页,继续跟踪链接.遍历的路径为: A-F-G ,E-H-I ,B ,C, D
- 宽度优先遍历策略
  - 宽度优先遍历策略的基本思路就是,将新下载网页中发现的链接直接放入待抓取URL队列的末尾.也就是说网络爬虫会优先抓取起始网页中链接的所有网页,所有网页都抓取完之后,再选择其中的一个链接网页,继续抓取在此网页中链接的所有网页.它的路径可以这样写:A-B-C-D-E-F ,G ,H, I
- 反向链接数策略
  - 反向链接数是指一个网页被其他网页链接指向的数量,同时反向链接数也是表示一个网页的内容受到其他人的推荐的程度.因此,很多时候搜索引擎的抓取系统会使用这个指标来评价网页的重要程度,从而决定不同网页的抓取先后顺序.
  - 而然在真实的网络环境中,由于许多广告链接、作弊链接等等的存在,反向链接数不能完全等同于重要程度.因此,许多的搜索引擎往往考虑一些可靠的反向链接数.
- OPIC策略策略
  - 这种算法实际上也是对网络页面进行一个重要性的打分.在算法开始前,会给所有页面一个相同的初始现金(cash).当下载了某个页面P之后,将P的现金分摊给所有从P中分析出的链接,并且将P的现金清空.对于待抓取URL队列中的所有页面按照现金数进行排序
- 大站优先策略
  - 对于待抓取URL队列中的所有网页,根据所属的网站进行分类.对于待下载页面数多的网站,优先下载.这个策略也因此叫做大站优先策略

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 23:01:32

## 附录

下面列出相关参考资料。

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 21:29:41

## 参考资料

- crifanLib.cs之Http
- HTTP知识总结
- app抓包利器：Charles
- JSON详解
- 主流数据格式：JSON
- 【已解决】C#中解析Json字符串 – 在路上
- 【记录】Python中尝试用lxml去解析html – 在路上
- python-goose
- PySpider
- Python爬虫框架：PySpider
- Scrapy
- 主流Python框架：Scrapy
- Grab
- Portia
- newspaper
- ruia
- Cola
- Sasila
- Nutch
- Heritrix
- crawler4j
- WebMagic
- Colly
- Pholcus
- headless-chrome-crawler
- scrapy中的提取正文的方法-python,爬虫,scrapy研究-51CTO博客
- 基于 Python 的 Scrapy 爬虫入门：页面提取 - 大虫 - SegmentFault 思否
- Scrapy定向爬虫教程(二)——提取网页内容 - 春华秋实 - CSDN博客
- Scrapy笔记04- Selector详解 | 飞污熊
- Scrapy爬虫抓取网站数据 | ShinChan's Blog
- Scrapy爬虫入门教程十二 Link Extractors (链接提取器) - inke的博客 - CSDN博客
- 基于WebMagic的CSDN博客爬虫 - zhuqiuuhui的专栏 - CSDN博客
- Heritrix与Nutch对比 - 爱专集
- Nutch、heritrix、crawler4j优缺点 - CSDN博客
- 爬虫用哪个好？ - 知乎
- 作为基础服务的数据采集，发展到哪个阶段了？搜狐科技搜狐网
- Python3网络爬虫(四)：使用User Agent和代理IP隐藏身份 - Jack-Cui - CSDN博客
- Python 爬虫一些常用的UA(user-agent) - abe\_abd的博客 - CSDN博客
- 如何评价可以自动更换 User-Agent 的爬虫设计？ - 知乎
- DarkSand/Sasila: 一个灵活、友好的爬虫框架
- Python有哪些常见的、好用的爬虫框架？ - 知乎
- 8个最高效的Python爬虫框架，你用过几个？ - 个人文章 - SegmentFault 思否
- 爬虫的几种抓取策略 | 阿布云 - 因为专业·所以简单
- 【爬虫工程师招聘】智慧芽爬虫工程师招聘-BOSS直聘
- 【数据采集招聘】智慧芽数据采集招聘-BOSS直聘
- 【高级爬虫工程师招聘】智慧芽高级爬虫工程师招聘-BOSS直聘
- 【记录】C#中的HTML解析 – 在路上

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2019-03-28 22:55:29