Atitit 基于**scrapy的 爬虫 架构原理**

**[1.1. 有移动版的网站优先爬移动版 站点选取](#_Toc6104)** [1](#_Toc6104)

[1.2. 网页爬虫组成： 1](#_Toc1014)

**[1.3. 模拟登录 与](#_Toc21663)**[cookie模拟登录 3](#_Toc21663)

**[2. 参考](#_Toc15881)** [3](#_Toc15881)

## ****有移动版的网站优先爬移动版** 站点选取**

现在的大网站基本除了pc端都会有移动端，所以需要先确定爬哪个。

比如爬新浪微博，有以下几个选择：

1. www.weibo.com，主站
2. www.weibo.cn，简化版
3. m.weibo.cn，移动版

上面三个中，主站的微博数据是动态加载的，意味着光看源码是看不到数据的，想爬的话要么搞清楚其api访问情况，要么模拟js，那样的话花的力气可能就有点多了。weibo.cn是一个简化版，数据能直接从网页源代码中解析出来，但使用正则或xpath取网页中的元素仍然是无聊且费时的，更不用说偶尔的页面结构错误更让人抓狂。 相比之下，移动版的爬虫最好写，因为移动版能直接拿到json格式的数据。

**一般来说，****有移动版的网站优先爬移动版，会节省很多力气。**

## 网页爬虫组成：

1. **一个url队列。**我们的爬虫从这个队列中读取url，并将新的url放入这个队列。这里最重要的是判重。简单的哈希就能达到判重的目的，但是为了节约空间（url的数量往往很多），一般使用bloomfilter的思想。bloomfilter与普通的哈希算法最大的不同就是bloomfilter只需要一个bit来表示某个元素是否存在，所以能节约空间。bloomfilter有一个小缺点，即准确率并不是百分百：判断一个元素是不是已经存在时，已有的有很小的可能会判断为不存在，但是没有的元素一定会判断为不存在。
2. **网页爬取模块。**需要能模拟浏览器发送请求。
3. **网页分析模块。**爬下来的是网页源码，可以用正则或者其他方法提取我们需要的信息。
4. **新的url生成模块。**生成新的url，放入队列。

从这张图上看，scrapy包含了以下模块：

1. scrapy engine，主引擎。在处理数据流时负责管理整个系统，同时各种事件的触发也由其负责。
2. spider，即我们的爬虫。主要的爬虫代码都在这部分中，包括发起请求，处理返回的网页等等。
3. spider middleware，spider中间件。中间件的一种，主要工作对spider发送的request做一些处理。
4. scheduler，调度器。上面说到的url队列就是调度器在管理，一方面接收spider发送的request请求，放入队列中；另一方面会从队首取出request交由downloader去下载网页。
5. downloader，下载器。将网页的html源码下载下来供之后的网页分析和信息提取。
6. downloader middleware，下载器中间件。中间件的一种，在下在网页之前和之后都会运行，可以用来设置发送请求的header，cookies，代理ip以及处理一些错误返回的情况。
7. item pipeline， 管道。一个网页被爬取下来并解析之后，其后续的信息存储之类的工作在pipeline中进行。当然也可以直接在spider中完成这些工作，但在pipeline中完成则显得整个项目结构清晰。

上面列出的里面spider，pipeline需要自己写，两种middleware需要的话可以自己添加自己写的。

## **模拟登录 与**cookie模拟登录

上面只是一些简单的登录情况，如果验证码很变态（比如需要鼠标滑动）或者登录过程很复杂，需要各种加密（比如新浪微博pc端的登陆）的话，模拟登录确实是个很让人头大的问题。这时候有另一个通用办法，那就是cookie模拟登录。网站想知道我们的登录状态，都是通过cookie来确认的，所以我们只需要在每次request的时候都附带上cookie即可实现已登录的效果。

那么，如何获得cookie呢？有chrome的可以F12打开Network界面，这时候人工网页登录，即可在headers中看到cookie。得到cookie后，只需要在request中加入自己的cookie即可。

# **参考**

**[网页](http://www.cnblogs.com/rubinorth/p/5963066.html)****[爬虫--](http://www.cnblogs.com/rubinorth/p/5963066.html)****[scrapy进阶](http://www.cnblogs.com/rubinorth/p/5963066.html)**

Atitit.网页爬虫的架构总结.doc nd spider api design manau

Atitit 网络爬虫与数据采集器的原理与实践attilax著 v2

@depre Atitit 网络爬虫的原理与实践attilax著.docx

Atitit.数据检索与网络爬虫与数据采集的原理概论 （book

网页爬虫--scrapy入门 - 刀刀流 - 博客园.html