网络爬虫反爬策略研究

胡俊潇,陈国伟
(中国传媒大学,北京 100024)

摘 要:网络爬虫在工作时会对目标站点发送大量的请求,这样的爬虫工作方式决定了其会消耗不少目标站点的服务器资源,这对于一个服务器不大的中小型站点来说负载是巨大的,甚至会导致该站点直接崩溃。另外某些网站也不希望自己的内容被轻易的获取,如电商网站的交易额,这些数据是一个互联网产品的核心,因此采取一定的手段保护敏感的数据。因此很多网站都在站点中加入了反爬机制。例如 User-Agent+Referer 检测、账号登陆及 Cookie 验证等。文章讨论了几种主流的方法来避免爬虫被目标站点服务器封禁,从而保证爬虫的正常运行。

关键词:网络爬虫;反爬虫;抓取策略
中图分类号:TP393.092 文献标志码:A 文章编号:2095-2945(2019)15-0137-03

Abstract: Web crawlers send a large number of requests to the target site when they work. this way of crawler work determines that it will consume a lot of server resources of the target site, which is a huge load for a small and medium-sized site with small and medium-sized servers. It can even cause the site to crash directly. In addition, some websites do not want their content to be easily accessed, such as the transaction volume of e-commerce websites, these data is the core of an Internet product, so take certain means to protect sensitive data. As a result, many sites have added anti-crawling mechanisms to their sites. For example, User-Agent + Referer detection, account login and Cookie verification. In this paper, several mainstream methods are discussed to avoid the crawler being blocked by the target site server, so as to ensure the normal operation of the crawler.

Keywords: Web crawler; anti-crawler; crawling strategy

网络中承载的信息随着互联网的发展呈现出了爆炸式发展的趋势,如何从丰富庞大的信息中快速找到自己需要的信息变得越来越重要。越来越多的机构、个人使用网络爬虫技术去网站中获取信息,大量的爬虫访问会使该网站的访问速度减慢甚至无法访问[1],因此,很多网站引入了反爬机制来防止网络爬虫频繁的访问。下面列举了几种主流的反爬策略。

1 模拟登陆

模拟登陆指通过计算机程序代码模拟普通用户使用浏览器登陆某网站的过程。有的网站在进行内容展示的时候,需要用户完成登陆操作,如果需要爬虫此类型的页面,就要求爬虫能够模拟用户的正常登陆行为。其技术核心主要可以分为两个部分:(1)抓包分析;(2)程序实现。

互联网中的网页在进行通信时是遵循 HTTP 协议的,网站服务器向访问者返回的 response 是由访问时的参数决定的,抓包的目的就是通过捕捉人为操作中的一系列请求记录,从中找到用户操作的关键请求参数。

在请求访问某网页时,request 请求中携带了两部分参数:Request Method 与 Headers。

Request Method 即 HTTP 请求方法,目前较为主流的是 GET、POST 两种请求方式,GET 方法用来访问已知的页面,访问成功后指定资源的服务器返回响应内容。POST 向指定资源提交特定的数据进行请求,例如提交表单,这些特定的数据包含在请求体中。POST 请求可能会导致新的资源的建立和/或已有资源的修改。

Headers 中包含了 Accept、Accept-Encoding、Accept-Language、Connection、Cookie、Host、Referer、User-Agent 等信息。Accept 指的是可以接收的信息类型,例如仅接收视频文件。Accept-Encoding 是指浏览器能够进行解码的数据编码方式,比如 gzip。Accept-Language 指的是浏览器所希望收到的数据语言种类,有的服务器提供多种语言,但这种时候必须要用到这个参数。Connection 指的是是否需要保持连接。Host 指的是初始 URL 中的主机和端口。

Cookie 是 Headers 中最重要的请求头信息之一,网站的服务器通过 Cookie 来标识区别不同的用户。HTTP 是一种无状态的连接,当服务器同时收到多个请求时,服务器是无法判断该请求来自哪一个客户端,于是需要 Cookie 来标识用户的身份,储存用户的状态信息,如登陆状态。当用户第一次登陆网站时,服务器会给浏览器下发一个 Cookie,浏览器将 Cookie 保存在本地,当在该网站域名内跳转时,用户无需再次登陆该网站,同时服务器也能够识别该请求来自同一个用户。获取 Cookie 的方式有很多种,常用的就是通过抓包直接在浏览器中找到相应的 Cookies 信息,将该信息填入 Headers 对应的字段中,即可实现访问。还可以通过在 Python 使用 Selenium 库来调用浏览器,使用 Selenium 库中提供的 find_element()_by_xxx 的方法来找到目标站点中的账号密码输入框、登陆按钮等元素进行输入点击等操作,登陆后将 Cookies 保存下来。值得注意的是,Cookie 是有时效的,超过时效后在使用该 Cookie 无法实现访问了。具体的时效长短每个网站设定的时

作者简介:胡俊潇(1994,11-),男,彝族,云南砚山人,硕士研究生,研究方向:新媒体。

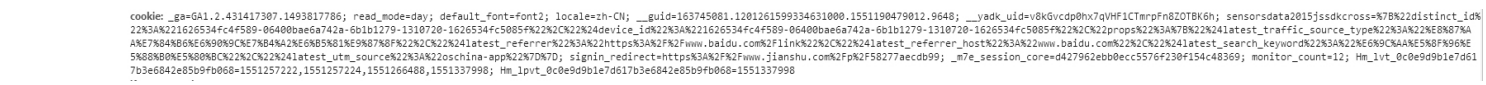

图 1 访问某网站时的 cookie 信息

间不一致，需根据具体网站设置更新 Cookie 的时间。

2 模拟 User-Agent

User-Agent 指的是用户在访问站点时所使用的客户端种类和版本，一般指的是所使用的浏览器。它是 HTTP 协议中的一个字段，在用户访问站点时，服务器通过该字段就可区分是什么浏览器访问的网站。目前主流的浏览器基本上都是 IE 和 Chrome 两种内核的，每个浏览器、知名的爬虫都有其固定的 User-Agent [2]，因此只要知道这些浏览器或者知名爬虫的 User-Agent 就可以伪装成该浏览器访问站点。但知名的爬虫一般都有固定的 IP，如搜索引擎的爬虫，不容易伪装。因此伪装浏览器的 User-Agent 是目前比较主流的伪装方法之一，因为浏览器下载后任何人都可以使用，没有固定的 IP 地址。

当访问一个服务器过于频繁时，单一的 User-Agent 会被服务器识别为异常访问，需要不断的更换不同浏览器的 User-Agent 来进行反爬，以下是两个不同内核的浏览器 User-Agent。版本不同也就意味着所使用的浏览器不同，因此除了切换不同内核之外，还可以选择不同版本的 User-Agent。

Chrome 内核：Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/63.0.3239.132 Safari/537.36

IE 内核：Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Win64; x64; Trident/5.0)

有的网站在进行反爬虫检测时，会将 Refer 与 User-Agent 一起检测。Refer 代表的意思是这个页面是从哪里跳转过来的，如从 www.baidu.com 跳转到 https://www.sina.com.cn/，Refer 就是 www.baidu.com。百度统计网站流量来源时就是依靠这个字段。有些站点在收到爬虫的访问请求时候会做识别，不是要求内的 Refer 则不能访问该页面，Refer 输入错误也会导致无法访问该页面。一般来说，使用目标站点的首页地址作为 Refer 就可以了。

3 设置爬虫爬取频率

爬虫在一定时间内的爬取量是检验一个爬虫效率的重要标准之一，因此为了加快爬虫的爬取效率，一般在使用爬虫时会采用多线程或多进程的方式来对爬虫节点的 CPU 内存带宽等资源进行充分利用，理论上在目标站点服务器没有做访问限制的前提下，只要爬虫节点的内存带宽资源足够，采用多进程或多进程的爬虫爬取效率将会直线上升。一般来说，普通用户访问一个网站时往往需要 10 秒以上，在没有做出限制时爬虫一秒能够爬取上百个页面。上文提过，这样的频繁访问会导致目标站点服务器崩溃，因此在实际使用中，网站的服务器往往会对爬虫一段时间内的访问请求次数做出限制，如果超过限制的阈值，则判定该访问是一个爬虫行为，对其封禁或采取其他限制。降低爬虫的爬取频率在一定程度上是尽可能的模拟出真人访问一个普通网站的行为[3]，但将爬虫的频率降得太低会影响爬虫的爬取效率，在爬取过程中，需要通过不断的测试找到网站的阈值限制，在对爬虫的程序进行修改。在 Python 中可以直接使用 tme.sleep 函数进行延时设置。本项目中通过设置间隔抓取时间从而达到降低爬取频率的目的。

4 设置代理 IP 池

与降低爬取频率相比，设置代理 IP 池是以效率为先。在爬虫高频的爬取时，网站服务器会封禁其主机 IP，而为了解决这个问题，人们往往使用与模拟 User-Agent 相似的方法，设置代理服务器，不断的更换主机的 IP 及访问端口，以解除网站服务器对爬虫的限制。设置代理服务器的方式有两种：(1) 购买可以拨号更改 IP 的 VPN 服务。(2) 使用代理池中 IP 不断更换 IP。第二种方式是目前比较主流的。IP 代理池可从网上找到开源的免费代理 IP 池或自建代理 IP 池，但这类型的 IP 往往稳定性较差，不可用的几率很高，如果只是进行简单的数据爬取，自建 IP 池能够满足需要，当爬取量巨大且爬取周期较长时，自建 IP 和开源免费的 IP 往往是不够用的。所以需要购买服务商提供 IP 池服务，在爬取时提取 IP 池中大量的 IP 进行更换使用。Python 在 urllib.request 库中封装了 ProxyHandler[4]函数来设置使用代理服务器。

5 其他反爬策略

目前进行反爬的主流方式还有通过官方 API 获取及混合反爬两种方式。(1) API 获取。数据目前已经成为了一种资产，各行业各公司在对数据进行保护时，也有很多公司选择性的对用户开放，以供数据相关从业者或者专家学习研究，这些数据往往会封装成统一的对外 API 接口，数据从业者可从该接口获取那些不容易提取的网站数据。目前大型的网站基本都提供了 API 接口服务，如国外的 Facebook，Instagram，国内的微博、腾讯、百度等网站。当有从 API 中获取数据的需求时，可以在官方网站上进行 API 接口的申请，在通过相关的资质认证后，就能通过调用相关 API 接口来完成数据获取的操作。因为接口是官方提供的，使用 API 来进行目标数据获取的好处显而易见：避开了网站的各种反爬限制，直接进行数据获取，减少项目中在反爬部分精力投入。但使用 API 进行获取时，缺点也十分突出，只能从接口获取官方规定好的字段，未加入数据获取的无法从接口处获取。如 Facebook 在获取数据时，API 仅提供了文章的基本数据，而用户的信息是禁止获取的。(2) 混合反爬。混合反爬指的是在同一项目中使用了多种反爬方法进行数据获取，一个项目可能会需要两种以上的方法共同完成爬取任务，同时过程中需要不断的测试才能得出适合的反爬策略。在实际的项目爬取过程中，因为每个网站所采用的反爬策略并不是单一的，因此会遇到各种混合搭配的反爬问题，如有的网站需要模拟登

(下转 140 页)

```
class Handler(BaseHandler):
    headers = {"Accept":"text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8",
           "Accept-Encoding":"gzip, deflate, sdch",
           "Accept-Language":"zh-CN,zh;q=0.8",
           "Cache-Control":"max-age=0",
           "Connection":"keep-alive",
           "User-Agent":"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.101 Safari/537.36"}
```

图 2 模拟 User-Agent 代码

发区健康产业企业将农作物种植、饲养以及半成品加工过程的环境友好性和资源使用效率作为原材料选择的主要依据。同时，要加强农副产品有机废物的综合利用，如粮油生产的菜籽粕可以进行再利用。

2.2 资源集约和污染减排

在加强节能降耗、污染减排、提高环境质量等方面加大工作力度，开发区通过双管齐下，一方面推进完善集中供热和推广试点光伏发电示范工程，另一方面对企业进行大规模技术改造，优化生产工艺，逐步降低开发区产业资源消耗和污染排放，进一步强化开发区产业低碳发展，以期实现资源集约利用水平较大提升，污染排放水平总体较低，产业生态效率全面优化。

能源集约利用方面，完善集中供热体系，推进 LED 路灯改造，试点合同能源管理和能源审计，提高工业能源利用效率，推广绿色建筑，推广清洁能源示范，加强节能宣传。水资源集约利用方面，加强用水定额管理，试点非常规水资源利用，推进建筑楼宇节水，全面推进海绵城市建设，制定节水激励政策。土地资源集约利用方面，完善土地管理制度，充分挖掘土地集约潜力，建立统一的土地管理标准。

水污染控制方案方面，实行严格的水污染物总量控制，推进重点污染源在线监控系统，改进企业生产工艺，提高企业水循环利用率，开展地表水环境整治。大气污染控制方面，加大治理设施监管力度，控制挥发性有机物排放，加强扬尘污染控制，防治汽车尾气污染，建立重污染天气预警和应急工作机制。固体废物减量及资源化染方案，推进生活垃圾资源化、工业废物资源化利用、危险废物安全处置，加强废物综合管理，适当发展静脉产业。此外，开发区临近自然保护区，需做好生态监测和生态防护林建设。

2.3 生态宜居环境建设

生态工业园区应以生态为准绳，本着人与自然协调发展、资源开发与生态保护统一的原则，打造生态宜居环境，全面提升区域生态服务能力。围绕全面提升开发区生态服务能力，做好生态景观规划，塑造美学景观美化开发区环境，增加开敞空间提升开发区精神品味。开展低碳社区建设和绿色学校，绿色生活培育，建设低碳人居环境。同时引导生态文明行为，加大创建绿色细胞工程，结合开发区特色，加大力度建设区内绿色社区、绿色宾馆等系列工程。在区内大型道路电子宣传屏幕及各企业内宣传生态工业园建设理念，加大财政对生态工业园建设的投入力度，如环境友好型企业可采取减免部分税收等经济措施鼓励及补助。

2.4 强化环境管理

生态工业园区建设需要社会、经济、环境各个部门进行观念的创新、制度的创新、体制的创新，用全新的视角和手段对其进行综合的、全方位的管理，从而为规划实施创造良好环境和支撑保障。规划的实施需要相应的内部和外部的条件作为保障，这些保障措施包括规划的组织机构、政策体系、管理制度、技术保证和实施手段等方面。其中，重点做好环境管理和风险防控工作，强化环境管理方面，严格执行环境保护各项法规制度，实施清洁生产审核，开展循环经济试点，推进企业 ISO 14001 体系认证，建设生态工业信息平台，定期开展环境质量监测；强化风险防控方面，开展风险评估和应急预案编制，建设开发区环境风险防范预警系统，严格安全生产管理，企业环境安全达标建设。

3 结束语

新建开发区尚处于工业发展初始阶段，企业相对较少，纵向分工和横向合作的产业链较短，各企业相互之间没有形成良好的上下游依托关系，互补性不强，集聚效应较弱，因此重点是在顶层设计阶段做好工业链建设、产业布局及环境基础设施建设等，同时在开发区开发建设过程中做好规划的贯彻落实，从而保障开发区绿色发展，实现区域可持续发展。

参考文献：

[1]刘丹丹，康婷婷，林立清.传统开发区转型升级过程中的生态工业园区建设路径——以苏北某开发区为例 [J]. 再生资源与循环经济，2018，11(11)：15-17.
[2]包惠玲.中国生态工业园发展现状研究[J].特区经济，2019(01)：59-61.
[3]赵满华，田越.贵港国家生态工业（制糖）示范园区发展经验与启示[J].经济研究参考，2017(69)：42-50.
[4]郭辰，黄付平，李泽鹏，等.广西生态工业园区发展对策研究[J].生产力研究，2017(5)：85-88.
[5]田金平，刘巍，臧娜，等.中国生态工业园区发展现状与展望[J].生态学报，2016，36(22)：7324-7334.
[6]腊孟坷，刘会成，林立清，等.生产性服务业主导的生态园区建设模式及实例研究[J].江苏科技信息，2016(28)：3-10.

（上接 138 页）

陆的同时还需要进行网页重定向处理，有的网站需要降低爬取频率，同时还需要不断更换代理，headers 等。

6 结束语

本文主要研究了目前主流的几种反爬策略，在进行实际的项目爬取的时候，根据网站的不同，我们先通过观察测试找出该网站的爬虫限制策略，然后使用相应的反爬策略进行数据获取。

参考文献：

[1]逄菲.基于 Python 的分布式网络爬虫系统的设计与实现[J].电子技术与软件工程，2018(23)：6.
[2]陈利婷.大数据时代的反爬虫技术[J].电脑与信息技术，2016，24(6)：60-61.
[3] 刘石磊. 对反爬虫网站的应对策略 [J]. 电脑知识与技术，2017，13(15)：19-21，23.
[4]王星，刘李敦.基于移动代理（Agent）的智能爬虫系统的设计和实现[J].科技资讯，2007.