国外科技网站反爬虫研究及数据获取对策研究

张晔 孙光光 徐洪云 庞婷 曲潇洋

北方科技信息研究所，北京 100089

摘 要：当前，来自国外网站的互联网开源科技信息已经成为科技情报的重要表现形式和组成部分，利用垂直爬取技术抽取、集成、解析、跟踪、研究这些网页信息可帮助科研人员实时、全面、深入地了解领域内的研究现状。然而国内目前访问国外某些网站困难；且国外很多网站都加强了反爬虫技术策略与应用，爬虫技术总是不断被反爬虫技术超越，特定主题内容规模化信息获取尤为困难。采用简单的搜索方式难以获取，且有些信息具有很强的时效性，人工跟踪难度大、时间耗费多，不利于数据的长期积累。为此，我们重点针对开源信息获取的反爬虫技术开展了研究，提出针对性的解决方案，系统地介绍了反爬虫技术和爬虫技术的应用。

关键词：爬虫；反爬虫；信息采集；搜索引擎；python

DOI:10.19442/j.cnki.ci.2020.01.004

Research on Foreign Science Websites' Anti-crawling Technologies and Data Acquisition Strategies

ZHANG Ye, SUN Guangguang, XU Hongyun, PANG Ting, QU Xiaoyang

Northern Science and Technology Information Institute, Beijing 100089, China

Abstract: Currently, the Internet-based science information originating from foreign key websites has become an important form and an integral part of scientific intelligence. To extract, integrate and parse those web page information by using vertical crawling technology helps scientific researchers gain an overall but in-depth understanding of the up-to-date scientific achievements in various fields in real time. But it is difficult to have access to some of foreign websites as they have also increased the research and application of anti-crawling technology. With the crawling technology surpassed by anti-crawling technology, it becomes particularly difficult to obtain information on topic-specific contents in large scale. We analyze typical scientific websites based in foreign countries to give systematic introduction of crawling and anti-crawling technologies and corresponding solutions.

Keywords: crawling; anti-crawling; search engine; data acquisition; python

张 晔 女，北方科技信息研究所高级工程师，研究方向为信息资源建设。电子邮箱：48537796@qq.com。

孙光光 男，北方科技信息研究所工程师，研究方向为多媒体资源建设。

徐洪云 男，北方科技信息研究所工程师，研究方向为信息资源建设。

庞 婷 女，北方科技信息研究所工程师，研究方向为信息资源建设。

曲潇洋 女，北方科技信息研究所工程师，研究方向为信息资源建设。

0 引言

当前，互联网开源信息已经成为科技情报的重要表现形式和组成部分，特别是重点科技网站的信息，代表了全球最新的科技发展现状和趋势，但有些信息处于网站深层结构中，采用简单的搜索方式难以获取，且具有很强的时效性，人工跟踪难度大、时间耗费多，不易于长期积累。通过爬虫技术对这些开源数据自动抽取、集成、解析后得到的信息，可支撑科研人员在当前大数据背景下对情报作出快速反应，满足这种需求需要基于爬虫技术的数据采集与加工处理[3]。目前，国外很多重点科技网站都加强了反爬虫技术研究与应用，爬虫技术总是不断被反爬虫技术超越，特定主题内容规模化信息获取尤为困难，因此迫切需要系统地开展网络爬虫和反爬虫技术研究[1]。

1 网站反爬虫策略研究

爬虫和反爬虫技术是矛与盾之争，且换代周期越来越短，故需要长期化、系统化、平台化、标准化的研究，以避免每次遇到不同的爬虫和反爬虫问题都重新进行重复冗余的工作[2]。一是针对信息源网站开展反爬虫技术研究，进而针对性地提出解决方案，从而合法、高效、便捷地获取开源信息。二是针对恶意爬虫攻击，研究相应的反爬虫解决方案（例如，如何制定上下策略，如何匹配规则，如何更换惩罚等），限制网络爬虫大量无效的访问以及恶意爬取信息。只有加强网站反爬虫技术措施的应用，才能有效屏蔽爬虫工具恶意窃取数据。

在用爬虫工具爬取数据时，经常会遇到数据虽然在浏览器上显示但却抓取不到的情况，其原因也许是向服务器提交不恰当的表单被拒绝，也许是需要注册才能访问、IP地址已经被限制请求、复杂的验证码拦截等。我们共分析了50个国外科技门户网站，共133条信息源，其中网站栏目也叫“爬虫入口”（同一个网站包含多个信息源，但同一网站不同信息源的反爬虫策略可能不同，例如网站的文献类栏目跟视频类栏目反爬虫策略不同）。由于篇幅原因，表1列举了几个具有代表性反爬虫措施的典型网站栏目所应用的反爬虫策略。

2 反爬虫解决方案

2.1 服务端限制

反爬虫技术通常先在服务器端进行请求限制，防止爬虫进行数据请求，从源头限制恶意数据爬取。通常有如下几种方式。

（1）“请求头设置” 反爬虫策略：HTTP的请求头是在每次向网络服务器发送请求时，传递的一组属性和配置信息。HTTP 定义了十几种请求头类型，如python-requests、User-Agent等，易被发现，网站运维如发现携带有这类请求头的数据包，拒绝访问，爬虫任务即刻失败，通常会返回403错误。目前几乎所有网站都模拟了请求头设置。例如www.northropgrumman.com-Annual Reports，www.afcea.org-Magazine等。

虽然目标网站可能会对HTTP请求头的每个属性进行“是否常规访问”的判断，但如果把User-Agent属性设置成其他无关参数，伪装成通用搜索引擎或者其他浏览器请求头，例如设置r=requests.get (url,headers={’User-Agent’: ’Baiduspider’}) 就可解决。

（2）“签名请求规则”反爬虫策略：签名请求指在请求url中增加一个sign字段，通常取值为自定义字段的md5校验码。对于每一次HTTP或者HTTPS协议请求，网站根据访问中的签名信息验证访问请求者身份，判断是否允许继续访问。例如www.militaryaerospace.com网站应用的就是此种反爬虫策略 。爬虫技术人员对待此类网站，通常会判断发起请求方，如果是JS发起的请求，签名规则可以在JS函数中寻找，再根据规则去构造签名；如果是App发起的请求，最大可能是由于前端调用原生封装，或者原生发起等多种原因。情况复杂的，需要反编译App包，但也不一定能成功，需要反复调试验证。

（3）“流量限制” 反爬虫策略：防护措施完备的网站会监控用户是否快速地提交表单，或者快速地与网站进行交互，从而限制速度异常、短时间大量下载信息的IP访问。但此种方法极其容易误伤其他正常浏览用户，因为同一区域内的其他用户可能有着相同的IP，所以一般运维人员很少采用此方法限制爬虫。而爬虫技术人员如果发现请求被限制，可尝试请求延

表1 典型网站的代表性反爬虫措施栏目的应用策略

序号	网站名称	网站栏目	反爬虫情况
1	www.ict.fraunhofer.de	Films	国内无法直接访问Youtube网站，同时Youtube网站还设置了链接或编码加密阻止视频直接下载
2	ieeexplore.ieee.org	Popular	访问频率高，IP会被封锁
3	www.cmtc.com	Magazine	访问URL参数是乱码
4	www.sto.nato.int	Tech Reports	字体样式反爬机制，即网站定义了字体文件
5	www.northropgrumman.com	Annual Reports	请求头限制
6	www.enisa.europa.eu	Press Releases	“元素隐藏式”反爬虫策略，元素的属性隐藏和显示，主要是通过type="hidden"和style="display:none;"属性隐藏
7	www.militaryaerospace.com	White Papers	“签名请求规则”反爬虫策略，签名请求指在请求url中增加一个sign字段，通常取值为自定义字段的md5校验码。对于每一次HTTP或者HTTPS协议请求，网站根据访问中的签名信息验证访问请求者身份，判断是否允许继续访问
8	www.militaryaerospace.com	Farnborough Report News and Updates	访问过多会出现“验证码限制”反爬虫策略
9	www.armymantech.com	AIR	“流量限制” 反爬虫策略，访问频率高，IP会被封锁
10	www.gao.gov	Bid Protest Regulations	没有cookie就无法访问，同一个用户的cookie有访问次数限制
11	www.enisa.europa.eu	Corporate Documents	“元素隐藏式”反爬虫策略，主要是通过type=＂hidden＂和style=＂display:none；＂属性隐藏
12	www.drdo.gov.in	Alphabatical List Of DRDO Labs & Establishments	“CSS或者HTML标签干扰”反爬虫策略，利用css来控制图片的偏移量，或把文字伪装成图片，干扰混淆关键数据

迟，通过AJAX延时加载、异步更新脚本技术延迟网页加载的速度，避免被目标网站查封，具体延迟时间应根据实际情况设定。如www.ict.fraunhofer.de、www.ieeexplore.ieee.org、莱茵金属防务公司网站等可采用此种方法。除此之外还可考虑使用分布式爬取或者购买代理IP设置代理池的方式解决，笔者就是采用直接购买专业代理的方式进行解决，实践证明应用效果很好。

目前有很多收费的代理IP服务平台，有各种服务方式，可满足各种应用需求。但需要注意，合理控制数据爬取速度是爬虫行业应该遵守的规则，恶意速度的访问爬取会消耗服务器资源，严重情况甚至会把目标网站拖垮。

（4）“cookie/cookies限制”反爬虫策略：“cookie/cookies限制”指服务器对每一个访问网页的用户都设置cookie/cookies，给其一个cookie/cookies字段。网站为了辨别用户身份、进行session跟踪，当该cookies访问超过某一个阈值时就禁止掉该cookie/cookies，导致数据爬取失败。如https://www.militaryaerospace.com/sea-technology/sea-technology-articles.html就是此类情况。网络爬虫想要模拟真实用户请求发送给目标站点，就需要拟造匿名身份，然后填入cookie/cookies中，在每一次访问时带上cookie/cookies，如果登录用

户cookie/cookies信息在固定周期内失效，那就要找到登录接口，重新模拟登录，存储cookie/cookies，再重新发起数据请求，不断循环此步骤。

（5）“验证码限制”反爬虫策略：验证码是基于人能从图片中识别出文字和数字而机器却不能的原理产生的，是网站最常用来验证是爬取机器人还是普通用户在浏览的方式之一。但由于近几年机器学习和人工智能技术的飞速发展，机器和人之间的差距越来越小，验证码技术的发展已经迭代了多次。从最初的数字字母验证码到中文验证码、再到图像验证码，网络安全技术人员不断地与爬虫技术作斗争，验证码技术的发展史就是爬虫技术和反爬虫技术的博弈史。目前滑动拼图验证则是验证码的升级版，要求必须滑动拼图到指定位置才能通过验证进行下一步操作。爬虫工具可建立简单的验证码库，如对图片里的字母或者数字进行识别读取，可使用识图的模块包或一些验证码识别第三方库（pytesser，PIL）来破解。但复杂验证码，无法通过识图识别，可以考虑使用第三方收费服务或通过机器学习让爬虫自动识别复杂验证码，识别后程序自动输入验证码继续数据爬取。

（6）“数据加密”反爬虫策略：有些网站把ajax请求的所有参数全部加密，根本没办法构造所需要的数据请求，如美国复合材料世界网站Magazine栏目，全文文件加密内嵌在flash插件中，无法爬取。有的网站反爬虫策略更复杂，还把一些基本的功能都封装了，全部都是在调用网站自己的接口，且接口参数也是加密的，如www.compositesworld.com。遇到这样的网站，爬虫可以考虑用selenium+phantomJS框架，调用浏览器内核，并利用phantomJS执行js模拟人为操作，触发页面中的js脚本。从填写表单到点击按钮再到滚动页面，不考虑具体的请求和响应过程，全程模拟人浏览页面获取数据的过程。用这套框架几乎能绕过大多数的反爬虫，因为它不是伪装成浏览器来获取数据，它本身就是浏览器。

（7）“Youtube链接”反爬虫策略：很多国外科技公司网站都以Youtube为平台设有专门视频频道，介绍其最新产品、技术路线、技术原理等。针对Youtube平台市场上有较为成熟的开源工具，爬取解决方案主要是解决代理访问和开源工具的有机结合。先爬取采集任务入口下所有列表页地址，根据地址调用国外代理，再利用Youtube-DL开源工具进行二次封装爬取视频，通过技术手段判定爬取任务是否完成。代理负责避开流量监控，开源工具负责解决Youtube加密防爬。

2.2 前端限制

前端通常利用“CSS或HTML标签”“自定义字体”“元素错位”等干扰混淆关键数据的反爬策略，保护数据安全。

（1）“CSS或HTML标签”干扰反爬虫策略：前端通过CSS或者HTML标签控制一些关键信息安全，例如利用CSS来控制图片的偏移量显示出来，或把文字伪装成图片，干扰混淆关键数据，如https://www.defensemedianetwork.com/sections/photos-videos/等网站就是如此。针对此类反爬虫机制没有通用手段，需要对网页抽样分析，反复测试，寻找其规则，然后替换成正确的数据。例如需要先请求初始网页得到CSS文件和相应数据的span标签的CSS属性，再从CSS文件中提取出svg文件和一些CSS属性的偏移量。

（2）“自定义字体”反爬虫策略：某些网站在源码上的字体不是正常字体编码，而是自定义的一种字体，调用自定义的TTF文件来渲染网页中的文字，真实内容通过一种对应关系最终在页面上展示，而不在网页源代码中展示，通过复制或者简单的采集无法爬取到真实的数据，例如www.sto.nato.int等网站就是这种情况。虽然反爬虫在源代码中隐藏了真正的字体，但最终如果要在页面上展示还是需要导入字体包，找到字体文件，下载后使用font解析模块包对TTF文件进行解析，解析出一个字体编码集合，与模块包里的文字编码进行映射，再反推转换对应关系即可获得真实正确的内容。

（3）“元素错位”反爬虫策略：不管是爬虫还是自动化测试，元素定位是爬虫最基本而且必需的一个步骤，如用BeautifulSoup find定位，BeautifulSoup css定位、selenium定位等。“元素错位”反爬虫策略是指网站维护人员利用伪装或错位一些关键信息的定

位，让爬虫爬不到真实正确的内容。如设置一个合同数据相关网页内容中的价格显示，先用backgroud-image标签渲染，再用标签设置偏移量，展示错误的标签，形成视觉上正确的价格。本次研究的国外网站暂未遇到此情况，这种反爬虫策略在国内应用较多。通常先用上述各种方法找到样式文件，根据backgroud-postion值和图片数字进行映射，然后根据HTML标签里class名称，匹配出CSS里对应class中content的内容进行替换。

（4）“隐藏元素”反爬虫策略：用隐含字段阻止网络数据采集的方式主要有两种。第一种是表单页面上的一个字段可以用服务器生成的随机变量表示。如果提交时这个值不在表单处理页面上，服务器就认为这个提交不是从原始表单页面上提交的，而是由一个网络机器人提交。另一种是通过隐藏伪装元素保护重要数据，在重要数据的标签里加入一些干扰性标签，干扰数据的获取。元素的属性隐藏和显示，主要是通过type= " hidden " 和style= " display: none；" 属性控制，在元素属性里面让它隐藏，如www.enisa.europa.eu-Corporate Documents、Facebook等就是如此。绕开第一种表单交验的方式最佳方法为先采集表单所在页面上生成的随机变量，然后再提交到表单，处理页面第二种情况则需要过滤掉干扰混淆的HTML标签，或者只读取有效数据的HTML标签的内容。

3 结论与建议

除了掌握以上各种针对服务器端和前端的不同解决方案外，在策略管理上我们还需要遵循以下4个基本原则。

（1）遵守Robots协议

网络爬虫技术逐渐从仅作为搜索引擎的工具，到成为互联网公司数据竞争的标配装备，已被广泛地应用于各个行业，但由此引发的企业之间关于数据权利的争议层出不穷。Robots协议全称Robots Exclusion Protocol，也称为爬虫协议，该协议是网络爬虫行业需要广泛遵守的协议。网站通过Robots协议告诉搜索引擎哪些页面可以抓取、哪些页面不能抓取。爬取数据的前提是遵守Robots协议，在合法的情况下获得数据[3]。

（2）采用适用的代理以保障研究爬虫技术渠道畅通

任何项目想要大规模爬取数据或者解决反爬虫技术问题，基本需求就是选用合适的代理，没有代理IP，爬虫工作将寸步难行。从数据爬取规模、速度等需求方面考虑购买代理后，还需要优化方案，合理分配资源，才能更高效更快速更稳定地进行爬虫工作。有时还需要实现必要的IP轮转、请求限制、会话管理以及黑名单逻辑来预防代理被屏蔽。

（3）开发专用监控系统、及时发现网站变化

目标网站发生的结构性改变是爬虫失效的主要原因，靠人工检查采集任务发生变化与否是不现实的，如通过人工比对目标网站信息的标题、发布时间、摘要、内容、作者等，一旦采集任务超过10个，每天的工作量就会大大增加。因此，需要开发专用监控系统，这种系统会对爬取任务进行频繁的轮巡检查，一旦发现任何变化第一时间会发出通知。

（4）搭建专业的反爬虫平台

通过本项目研究发现，一般大型门户网站都会用到反爬虫策略，且其采用的反爬虫策略复杂多样而且更新迅速，一般规模化的爬取数据需要搭建一个专业的反爬虫和爬虫技术实验平台，来管理反爬虫策略并提供相应的解决方案，节省反复开发反爬虫策略的成本，提高爬虫技术的快速反应能力，减少反爬虫策略的失效时间。没有一个好的平台支撑，策略很难在最短的时间内生效。

参考文献：

[1] 潘晓英, 陈柳, 余慧敏, 等. 主题爬虫技术研究综述[J/OL]. 计算机应用研究: 1-6.[2019-10-21]. https://doi.org/10.19734/j.issn.1001-3695.2018.11.0790.

[2] 潘洪敏. 反爬虫探索 (1) ——爬虫[EB/OL]. [2018-07-25]. https://mp.weixin.qq.com/s/_yjAr_7nKu-IiShBmicyxQ.

[3] 张嘉琳. 由Robots协议引发的不正当竞争问题思考——以3百大战为视角[J]. 法制与社会, 2013 (23) : 96-97.