ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.15, No.28,October. 2019

E-mail: jslt@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

网站反爬虫策略的分析与研究

伏康,杜振鹏

(山东财经大学 管理科学与工程学院,山东 济南 250014)

摘要:随着大数据时代的来临,大数据在日常生活中的应用显得尤为重要。如何便捷、快速地获取数据将成为提高竞争力的重要手段,而通过网络爬虫这一新兴技术能够非常高效的获取网络中的数据。但是如果不对爬虫进行控制,爬虫也会对网站造成巨大危害。为了减少网络爬虫对于网站的危害,详细阐述了网络爬虫的工作原理、种类以及URL的搜索策略,针对某些网站的"反爬"措施,提出相应的提出几种反爬策略。从而减轻了网站服务器压力,保护了数据,防止数据的大量流失。

关键词:大数据;网络爬虫;爬虫;反爬措施;反爬策略

中图分类号:TP393 文献标识码:A

文章编号:1009-3044(2019)28-0028-03

开放科学(资源服务)标识码(OSID):

DOI:10.14004/j.cnki.ckt.2019.3535

The Analysis and Research on Anti-crawler Strategy of Website

FU Kang,DU Zhen-peng

(The School of Management Science and Engineering, Shandong University of Finance and Economic, Jinan 250014, China)

Abstract: With the advent of the era of big data, big data is particularly important in daily life. How to get data conveniently and quickly becomes an important means to improve competitiveness, and the new technology of web crawler can obtain data in the network efficiently. If the crawler is not controlled, the crawler will also cause great harm to the website. In order to reduce the harm of web crawlers to websites, this paper elaborates the working principle, types and URL search strategies of web crawlers in detail, and proposes several anti-crawling strategies for some websites. Thus, the pressure of the website server is reduced, the data is protected, and the massive loss of data is prevented.

Key words: big data; cyber crawler; spider; anti-crawling strategies; anti-spider technology

随着大数据时代的来临,在信息与知识爆炸增长的时代,为了分析预测人或者其他事物的客观规律以及行为预测和分析,需要从网络中获取大量的数据进行后续分析。人们为了更快、更准确地获取网络中个各种各样的数据,会写各种各样的爬虫来直接爬取网站数据。大量的不加限制的网络爬虫会对网站服务器造成巨大压力,严重者会造成服务器宕机,从而降低其他正常用户的使用体验,并可能会造成其他大量用户的潜在流失,给网站经营者造成损失。为了缓解爬虫对网站造成的压力,本文在对网络爬虫原理分析的基础上,针对不同爬虫抓取策略提出一些有效的"反爬"策略。

1 爬虫总述

1.1 爬虫简介

网络爬虫是抓取网页信息的一种程序,现阶段多种编程语言都可实现,包括现在主流的搜索网站百度、Google等都属于爬虫范畴,只不过百度、Google等是对于整个Internet进行爬取,而我们普通的爬虫是对特定网站的特定信息进行爬取。

1.2 爬虫原理

爬虫可以通过构造Url以及后缀参数对网站发起Request请求,当网站接收到Request请求之后将Servlet处理之后结果嵌入HTML中,之后将HTML代码返回到爬虫端,将得到的结果进行解析、提取有效数据、将有效数据清洗、数据存储。

1.3 爬虫组成

基本爬虫由五部分组成:Url队列(UrlQueue)、调度器(Scheduler)、下载器(Downloader)、解析器(Parser)、存储器(Storage)。Url队列是通过研究网站结构,生成待爬取的Url队列。调度器(Scheduler)是将Url调度到下载器(Downloader)中进行下载,并且记录服务器返回的状态码(state_code),将下载成功的Url从队列中剔除掉,下载失败的Url重新调回Url队列中,等之后再次进行下载。下载器(Downloader)则是通过发起Request请求,将服务器返回的数据传到爬虫端。解析器(Parser)则是对数据进行后续解析,从HTML标签中抽取研究所需要的数据。存储器则是将处理完的数据进行清洗、存储。

收稿日期:2019-06-26

作者简介:伏康(1997—),男,山东临沂人,本科,主要研究方向为大数据与网络爬虫;杜振鹏(1998—),男,山东济南人,本科,主要研究方向为java。

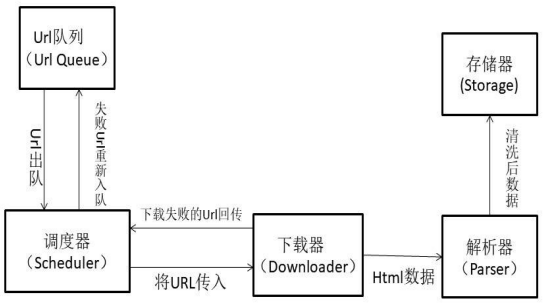

图1　爬虫工作原理

1.4 爬虫抓取策略

在爬虫爬取数据过程中，待抓取的URL队列也是非常重要的一部分，对于Url如何抓取也是可以深入研究的一部分。抓取策略即为一个网站内数据如何去抓取，即抓取的先后顺序，其中牵扯到程序的时间复杂度与空间复杂度，对于数据量比较小可以忽略时间复杂度与空间复杂度，但对于GB级数据或者TB级数据的抓取，必须去考虑算法的时间、空间复杂度问题。常见的爬取策略有：

深度优先策略。此策略主要运用的思想主要是，爬虫从起始页开始爬，根据网页对链接进行一个个的各种，处理完一整条线路之后再对下一个起始页进行爬取。

广度优先策略。将新下载网页中发现的连接置入队列的末尾，当横向链接全部爬取完毕之后，再对下一层次的连接进行爬取。

反向链接策略。该策略指的是一个网页被其他网页链接指向的数量，即其他网站链接对于该网站链接的推荐程度，但是往往由于网站内广告流量主会对爬虫产生错误引导，故该策略对实际参考意义上并不是很大。

大站优先策略。对网站页面进行分类，对于链接内可下载的链接数比较多时，提高该网页链接的优先值，对该网页进行优先爬取下载。

Partial PageRank策略。该策略借鉴了PageRank算法思想，对于已经下载的网页连同待抓取的Url队列中的Url形成网页集合，之后随每个页面计算PageRank值，之后按照PageRank值进行从大到小排列，并且按照顺序进行抓取页面。之后我们设置一个k值，每抓取队列中k个页面之后，对剩余页面重新计算PageRank值，重新排列，按照顺序抓取。

2 网络爬虫常用的伪装机制

了解爬虫常用的伪装机制之后，就能对一些基础爬虫进行识别，从而能够准确地过滤掉部分爬虫，减轻滥用网络爬虫对于服务器产生的负担。一般的网络爬虫使用的伪装机制有：

1）爬虫请求时设置Header：在Headers中存在的一些属性例如:host、Referer、user-agent、cookie等，一般的爬虫会设置一些假的user-agent来躲过检测。一般可以结合referer+user-agent来识别，Referer字段会指明这个页面是从哪里跳转过来。

2）设置定时休眠：一般爬虫在爬取的时候一般会采用设置爬虫休眠来模拟人为登录状态。可以用过客户端联入服务器的时间分析出是否爬虫，若访问时间非常规律，则证明其为网络爬虫。

3）使用普通代理服务器：爬虫为了躲避单一ip限制，会使用代理服务器进行访问，使用普通代理服务器比较容易被识别，可以通过用户端所发送的环境变量REMOTE_ADDR、HTTP_VIA、HTTP_X_FORWARDED_FOR三个字段检验出来。非爬虫用户后两个变量会没有数值，而普通代理服务器第三个变量会透露出爬虫真正IP。

4）使用高匿代理服务器：使用此类服务器和普通代理服务器功能相类似，但是三个变量值全为代理服务器IP。能隐藏爬虫真实IP，但会向服务器透露是使用代理服务器访问。

5）伪装网站Cookie：该类爬虫会获取网站发送给客户端的Cookie值，并将Cookie加入请求字段，模仿用户登录获取数据。用户初次访问网页时生成一个随机Cookie，如果用户之后多次访问都不携带Cookie，则可判定为爬虫。

3 “反爬”策略

3.1 妥协式策略

设置Robots.txt协议。在网站根目录下放入Robots.txt文件，来告诉一些规范的爬虫使用者哪些页面是可以爬取，哪些页面是不准使用爬虫爬取的，在Scrapy框架中默认是遵守Robots.txt协议。通过Robots协议的规范可以减轻一小部分爬虫对于服务器的负担。

设置Sitemap.xml静态文件。将整个网站的所有链接以及其他元数据(上次更新时间、更改频率以及相对于网站上其他网址的重要程度等)，通过该静态网页可以减少爬虫对于网站整体动态网页的爬取，从而能够减轻服务器压力，属于妥协式策略中比较有效的方法。

3.2 非妥协式策略

1）限制IP单位时间内访问次数。通过分析系统日志，对于一段时间内单一IP突然产生大量请求基本可以判别为爬虫。将其IP放入封禁池中，封禁一段时间之后再给予解封，使其能再次合理访问服务器。

2）设置较为复杂的验证码图片、滑动拼图。对于突然产生大量请求的客户端进行人机验证，采用隔时验证的手段。并且防止爬虫使用ORC自主识别平台，要采用较为复杂，但用户能够准确识别的验证码，例如：物品识别、问题选取、计算结果等方法。也可以使用滑块拼图的手段来进行人机验证。

3）通过JS脚本防爬虫。网站大量使用Ajax技术对网站数据进行异步加载，爬虫直接访问链接时异步加载的数据并不会返回到HTML中，使爬虫无法爬取到重要信息。

4）Css反爬虫机制。该反爬机制是比较先进的反爬技术，解密起来比较复杂。原理是通过加载SVG资源的方法，对页面的部分文字和数字进行加密。运用起来主要是通过css资源文件定义样式，通过样式的坐标值去SVG文件中定位最终的文字内容。例如使用该项技术的大众点评公司。

5）字体反爬。此项技术也属于比较常见的反爬技术，网站采用了自定义的文体文件(即新式的.ttf字体文件)，在浏览器上正常显示，但是如果使用爬虫爬取下来之后，数据会成为乱码或者被替换为其他字符

6）部分重要数据使用图片代替。例如一些数字数据比较重要，可以使用相应的数字图片去代替相应的文本。

4 研究结论

现阶段比较流行的爬虫无非普通的发起Request请求的简单爬虫、多线程处理数据结构化的多线程爬虫，以及类似基于

自动化测试框架的Selenium、PhantonJS爬虫等。以上都可以通过"反爬"措施来对爬虫进行封禁,唯一识别有些难度的就是基于WebDriver的自动化测试框架的Selenium、PhantonJS爬虫,但也并不是完全没有办法去识别以上爬虫。在使用Selenium时,会暴露出webdriver属性,可以通过HTML中JS来识别webdriver属性,从而封禁相应的爬虫。而对于比较简单的爬虫即使经过伪装还是比较容易被识别,从来网站服务器端可以采用相应的技术来限制此类爬虫对于资源的消耗。爬虫"突破"反爬措施的能力在不断地提升,这也迫使我们不断研究最新的爬虫,反向研究相应的"反爬"措施,从而减少网站珍贵数据的流失。并且与传统的"反爬"技术相比较,本研究所发现的"反爬"技术更加完善,对于爬虫更加有针对性,能够更加有效地防止爬虫的滥用对服务器资源的消耗。

5 结语

本文深入探究网络爬虫基本原理,从原理出发对网络爬虫进行封禁,并且提出比较主流的"反爬"措施,能够减少不规范爬虫对网站的爬取,减轻服务器压力,减少了宝贵数据的流失。但是随着互联网的迅速发展,各种爬虫性能以及伪装能力也会不断地提高,作为网站运营方也需要不断学习研究,针对层出不穷的爬虫新技术提出更加先进完美的反爬措施。

参考文献:

[1] 胡俊萧,陈国伟.网络爬虫反爬策略研究[J].科技创新与应用,2019(5):137-139.
[2] 陈利婷.大数据时代的反爬技术[J].电脑与信息技术,2016(24)60-62.
[3] 刘清.网络爬虫针对"反爬"网站的爬取策略分析[J].信息与电脑,2019(3):23-25.
[4] 邹科文,李达,邓婷敏,等.网络爬虫针对"反爬"网站的爬取策略研究[J].电脑知识与技术,2016(12):61-64.

【通联编辑:代影】

(上接第26页)

4 结束语

云计算目前已经渗透到社会的方方面面,学校、机关单位等都将云计算引入日常工作,很多数据直接在云端完成存储,云计算给网络传输带来的便利无疑是巨大的,但同时也带来巨大风险,因此合理使用云平台,降低平台风险是目前网络安全研究的重点。相信在研究者的不断努力下,云计算环境下的网络平台,其安全程度会越来越高。

参考文献:

[1] 黄海军.基于云计算的网络安全评估[J].电子设计工程,2016(12):115-117,120.
[2] 杨志杰.云计算技术下的网络安全防御技术[J].科技与创新,2018(6):69-70.
[3] 邱慕涛.基于云计算的计算机实验室网络安全技术应用探讨[J].中国管理信息化,2017(18):148-149.
[4] 盛丹丹.基于云计算环境下计算机网络安全问题的思考[J].电脑知识与技术,2018(14):25-26.
[5] 魏斯超,张永萍.基于云计算技术的网络安全评估技术研究及应用[J].数字技术与应用,2016(5):211.

【通联编辑:代影】