带你进入网络爬虫与反爬虫的世界

文/鲁萍

为什么有众多的网络安全防御手段，还会存在爬虫和反爬虫的问题呢？

置身于互联网大数据时代，每个人、每个企业都是数据的生产者，同时又是信息的消费者。越来越多的企业开始关注数据的价值，挖掘数据的价值。爬虫作为一项获取数据的工具被广泛使用，40%～60%的网络流量来自爬虫。爬虫遍布各类网站，政府信息公示类网站、电商类网站、票务类网站，等等。爬虫爬得不亦乐乎，被爬的网站不堪其扰。

爬虫与反爬虫

互联网带来了海量数据，数据获取也变得更加便利，数据获取的渠道也多种多样。数据需求方可通过授权合规渠道获取数据，根据数据的价值，往往需要付出一定成本；有些情况下，比如同行业竞争企业之间，希望获得对方的一些数据信息，又不希望透露自己的身份，其授权方式也是行不通的；再有一些情况，发布方希望信息能被最终用户使用，但不希望其他人或者企业利用这些信息做商业用途，也可能不提供授权数据获取的方式，比如法院的执行公示信息。

通过入侵对方网络、系统是获取数据的手段。有别于一般的黑客攻击，APT（高级持续性威胁）出于经济或者政治目的，通过长时间的布局，各种手段综合使用，一层层突破对方的防御，攻入内网，在对方网络中搜寻，直到获取目标信息。通过入侵获取数据，通常具备非常强的技术能力，并且违法成本极大。对于一般的网络信息，如果可通过网站直接访问获得，入侵方式通常不会被选择。网络爬虫则是一个被广泛试用的工具，技术门槛远没有入侵那么高，风险也低得多。

为什么有众多的网络安全防御手段，还会存在爬虫和反爬虫的问题呢？从DDOS防御、防火墙防御到业务反欺诈，各种防御手段都有其针对性，却没有一劳永逸的手段。网络反爬虫聚焦OSI七层的应用层，充分利用HTTP报文，基于爬虫行为分析，识破伪装，准确定位爬虫并进行控制。

图 互联网时代数据获取渠道

	合法手段	非法手段	灰色手段
获取方式	获得数据发布方授权访问	通过入侵获取数据：黑客入侵；APT等	数据抓爬，模拟正常网站用户访问获取数据
获取障碍	需要经济成本付出； 需要隐藏身份； 数据发布方未提供面向非终端用户的授权访问方式	承担违法成本； 需要突破各层技术防御：基础防火墙、IDS、IPS、WAF等	数据发布方可能采取网络反爬虫措施

图 网络安全防御手段

各类网络安全防御	特点
DDOS防御	防御各种类拒绝服务攻击，PingofDeath、TearDrop、UDPflood、SYNflood、LandAttack、IPSpoofingDoS等，攻击者IP可伪造
基础防火墙	作用于OSI 七层网络模型的第二到四层，主要功能是限制对IP/port的访问，对应用层内容解析能力非常有限
IDS（入侵检测系统）	补充防火墙的不足，可以检测到应用层的深层攻击行为，侧重风险管理
IPS（入侵防御系统）	防御防火墙所不能防御的应用层的深层入侵行为，对恶意行为进行检测和防御，侧重风险控制
WAF（Web应用防火墙）	作用于第七层处理HTTP服务，防止网页篡改、木马植入等
网络反爬虫	作用于OSI 七层的应用层，充分利用HTTP报文信息，并对爬虫行为进行分析，识破伪装，识别并控制网络数据抓爬
反欺诈	业务层防护，业务耦合性强

图 反爬虫的常见方式

	云服务方式	本地化方式	本地化+云服务
方式	通过接口调用云服务；客户方主动集成	本地化产品，基于网络全流量，基于规则识别爬虫	本地化产品部署，基于网络全流量、云端反爬虫规则推送识别爬虫，一体化爬虫实时控制
特点	出于性能考虑，不能基于网络全流量判定爬虫，只能针对部分场景化接口，网络延时不能满足业务系统要求	产品作为反爬虫工具提供给客户，客户需要团队运维反爬虫规则	客户数据本地化，无数据外泄顾虑； 网络入口流量控制，无业务系统对接烦恼； 实时爬虫识别、控制一体化，性能更优； 服务方专业团队运维反爬虫规则，客户方更省心。

网站饱受爬虫困扰，该怎么办呢？

各家饱受爬虫困扰的网站也在积极应对，通常网站自己的手段可以是基于网络防火墙依靠IP识别做阻断，误伤概率较高；或者基于业务关键节点做控制，与业务层的耦合性比较高，维护成本高。有没有供应商方案可供选择呢?

市面上提到反爬虫的供应商方案也分不同的方式。纯粹的云方式反爬虫不能应付相对大流量场景，反应性能也是问题。纯粹的本地化方案往往将产品作为工具提供，需要客户方培养专业的反爬虫规则运维人员，反爬虫的效果往往依赖运维力量的投入。

产品本地化结合反爬虫规则运维服务模式另辟蹊径，是一个性价比相对高的方式。即使在传统的安全领域，安全服务也是越来越被推崇。把安全问题交给专业的服务团队，业务力量则更能集中拓展核心业务。

杭州邦睿科技有限公司作为杭州邦盛金融信息技术有限公司的全资子公司，邦睿网络反爬虫基于网络入口全流量，具备高性能毫秒级的爬虫识别控制能力，产品本地化部署，无信息外泄的顾虑，与网站系统轻耦合，对接无需业务系统改造，爬虫识别控制一体化也是一大特色。此外，专业的团队为运维反爬虫规则，进一步提升产品试用的便利性。

鲁萍
杭州邦睿科技有限公司CEO