UCan下午茶

AI和大数据
在云安全中的应用

刘少东 qq: 389292268

Web应用防火墙的重生

UCLOUD

目录

- 01 WAF概念介绍
- 02 机器学习技术改进WAF质量
- 03 Web Bot检测与识别

WAF是什么？

- WAF是web应用防火墙（Web Application Firewall）的简称，对来自Web应用程序客户端的各类请求进行内容检测和验证，确保其安全性与合法性，对非法的请求予以实时阻断，为web应用提供防护，也称作应用防火墙，是网络安全纵深防御体系里重要的一环。

- WAF对请求的内容进行规则匹配、行为分析等识别出恶意行为，并执行相关动作，这些动作包括阻断、记录、告警等。

云WAF趋势

Waf 规则防护原理

目录

- 01 WAF概念介绍
- 02 机器学习技术改进WAF质量
- 03 Web Bot检测与识别

WAF误报与漏报

- 基于规则体系构建的WAF，一定会存在误报和漏报的问题。
- 如何评判规则WAF的质量？发现它的误报和漏报数据。
- 重新构建一套学习模型的检测引擎，进行对比改进。

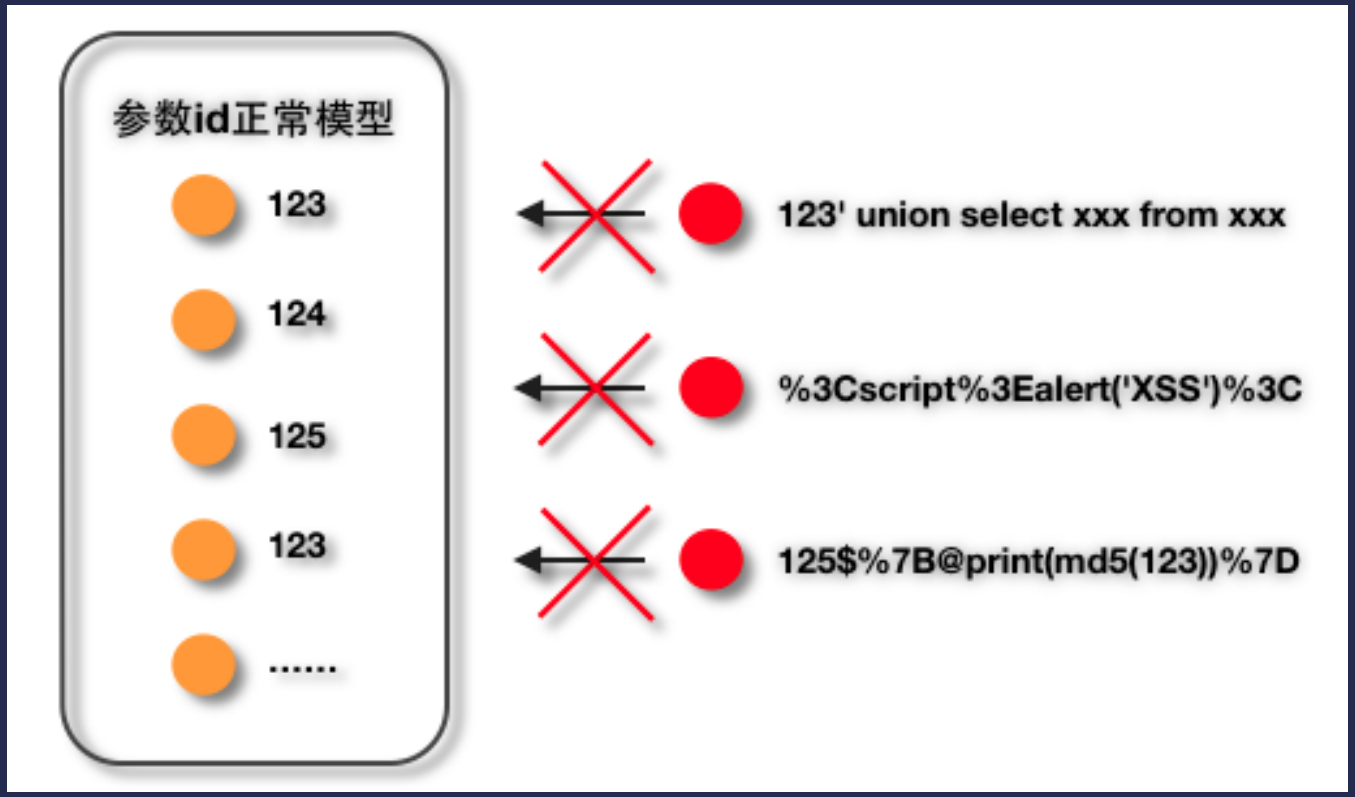
学习模型原理

如何发现异常载荷？

学习模型原理

如何发现异常载荷？

学习参数的值类型？长度概率？统计特征？ 这些方法太基础，效果不好



学习模型原理

先验原理：正常的载荷总是相似的，异常的各有各的不同。

学习模型原理

HMM 算法检测异常载荷

训练和检测性能消耗都比较低

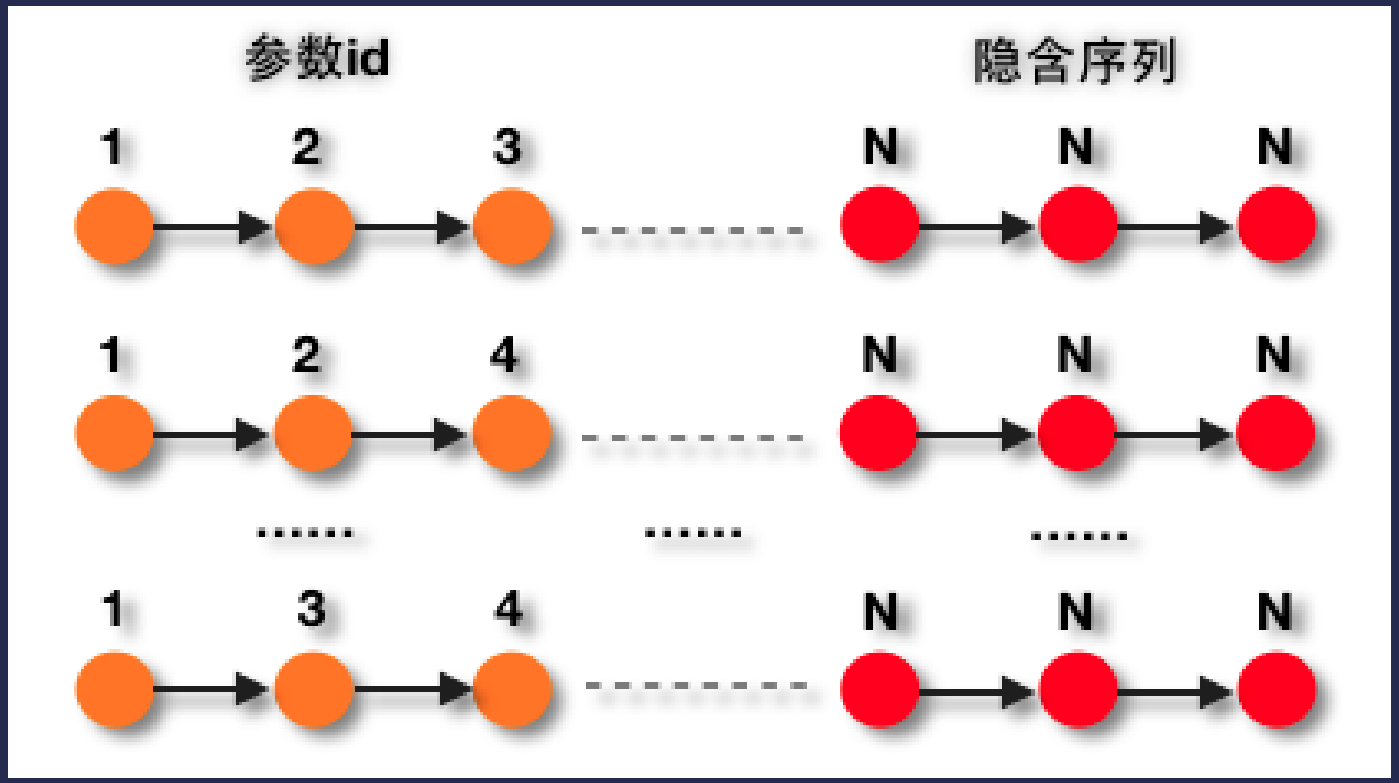
无监督训练模型

相比Kmeans和One-class SVM而言，HMM的性能消耗是很低的。

相比贝叶斯和决策树，HMM优势在于无监督学习。

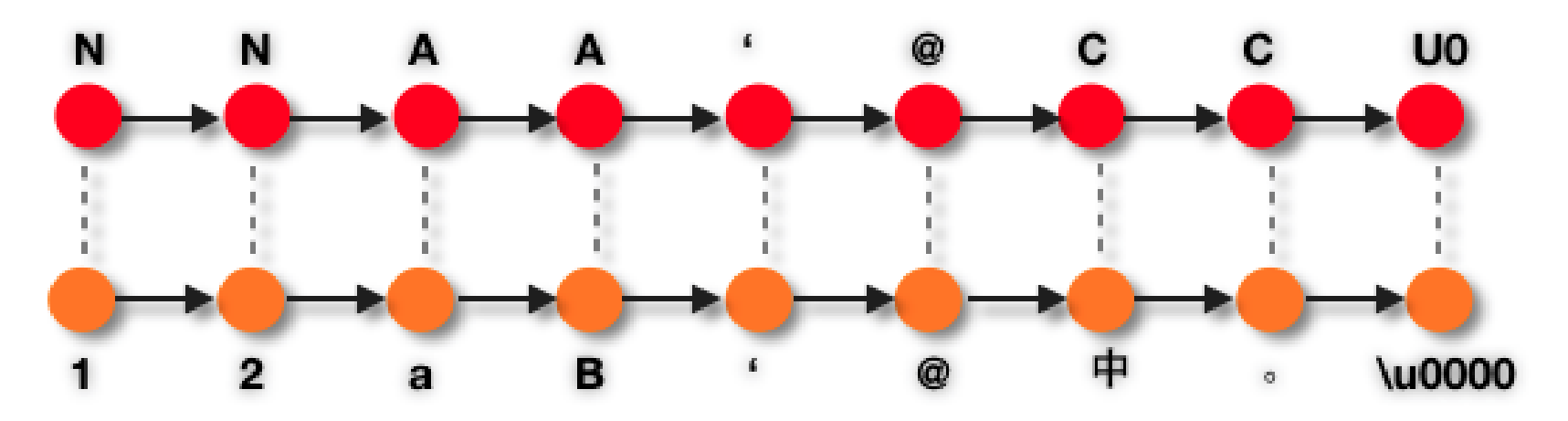
学习模型原理

HMM的隐藏状态和观测状态



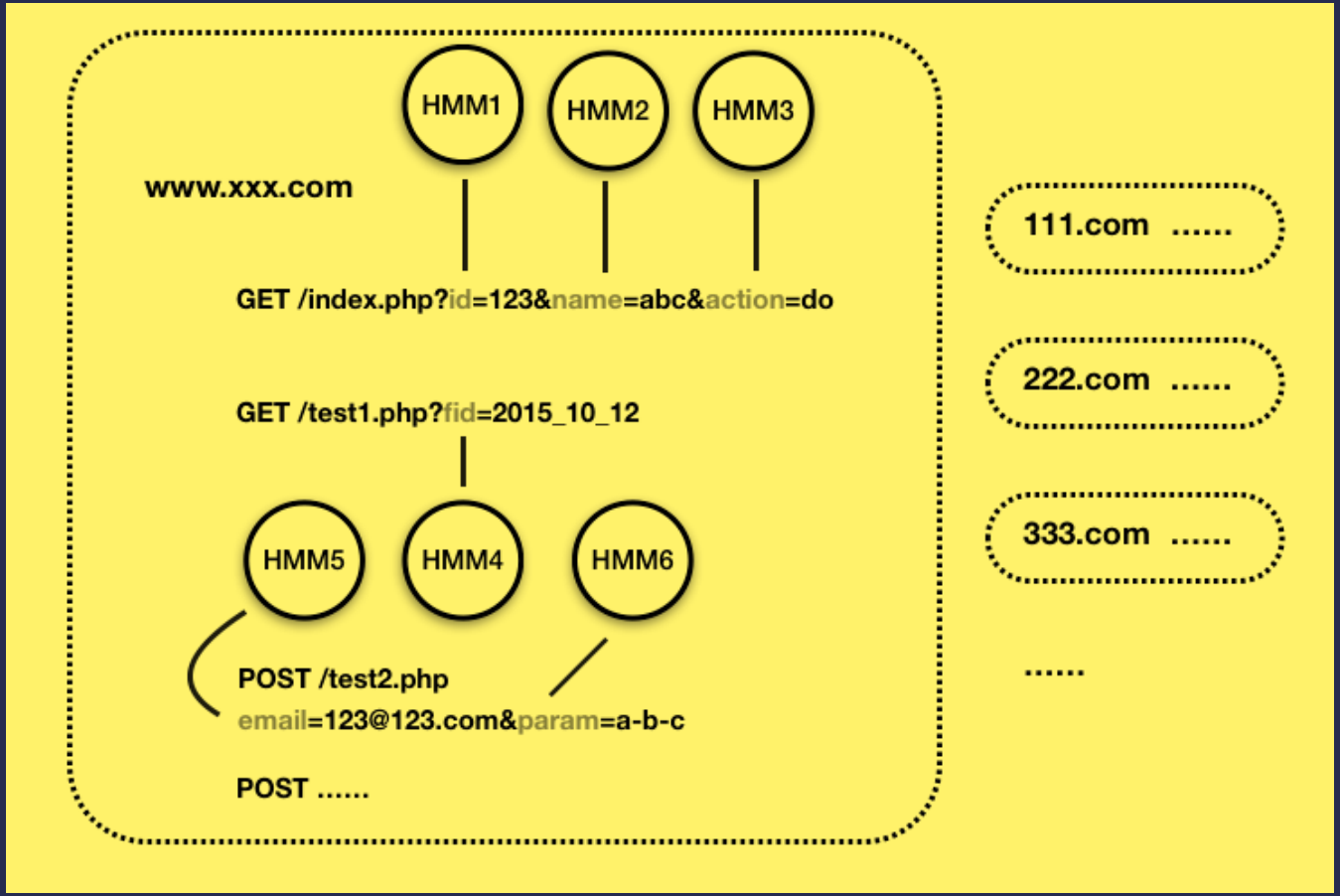
学习模型原理

泛化得到隐藏状态



学习模型原理

模型对参数进行学习，得到模型表示



学习模型原理

Trick

1.参数解码以及泛化的选定

2.隐藏状态数目的取值

3.Hmm学习的时候使用K-means算法效果更好

4.数据污染的问题需要一些工程方法保护（ip情报）

问题：检测的异常能直接用么？

NO，异常的大部分不是攻击行为。攻击属于异常，但是异常不一定是攻击。

如何处理？

学习模型原理

- 异常是不是攻击？ --分类问题。

深度学习算法，CNN RNN在分类问题上都表现出卓越的性能。但是性能需求目前阶段还是硬伤。

综合测试考虑，使用SVM，在分类问题上也表现出卓越准确率，与深度学习的差距可忽略不计。特别是2分类问题。

学习模型原理

- 载荷的分词和向量表达 Wordvec TF-IDF (样本量也有关系)
- Wordvec可以更加准确的表达words之间的关系。

- 给出一个word，它可以帮你找到和这个word存在紧密联系的其他words。 例如我给出sqli注入的常见key select，联系紧密的此应该有when from and 等等。当然这些依赖语料库的数据充分性。
- 30 Words closest to 'select': [case, when, concat, else, then, from, dual, end, cast, int, or, union, 001, and, elt, where, q706b6a71, extractvalue, convert, null, j, q, like, generate, series, q6a717871, k, count, as, p]

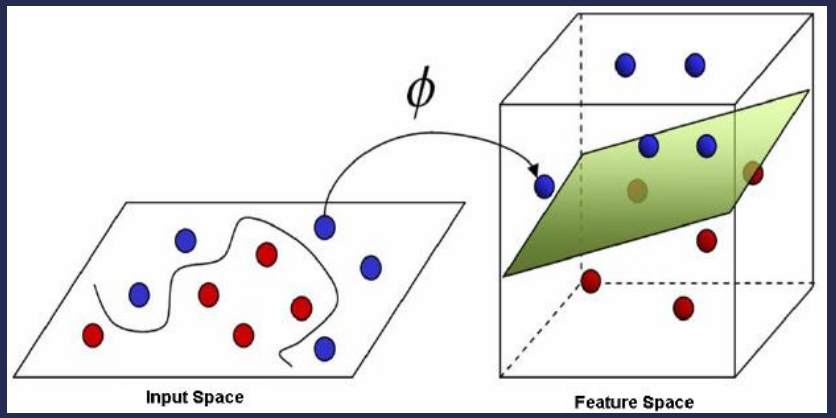
- 30 Words closest to 'script': [body, svg, javascript, alert, onload, document, src, style, div, >, iframe, onerror, object, source, xml, <, xss, input, cookie, eval, ”, location, prompt, write, img, frameset, red, a, execscript, test]

学习模型原理

向量机分类

准确率：99.3%

召回率：98.1%



Trick：

1.多分类的向量机没有2分类的性能好

2.多项式核就可以达到很高的准确率，更复杂的反而降低计算性能

3.编码的载荷需要解码，除非样本足够大

4.word2vec的模型参数超参数的选定

学习模型效果 - 漏报检测举例

- %22%2Bconvert(int,CHAR(55)%2BCHAR(68)%2BCHAR(80)%2BCHAR(54)%2BCHAR(121)%2BCHAR(106)%2BCHAR(120)%2BCHAR(119)%2BCHAR(108)%2BCHAR(65)%2BCHAR(101))%2B%22
- /m/1.2.3/hSea.min.js/(#_memberAccess=@ognl
- ?id=1' AND sleep 5&Submit=Submit'
- SELECT%25C0%25AAFIELD%25C0%25AAFROM%25C0%25AATABLE%25C0%25AAWHERE%25C0%25AA2%25C0%25BE1
- 1+AND+GREATEST%28A%2CB%2B1%29%3DA
- I%2F**%2FN%2F**%2FSERT
- 1%2523nVNaVoPYeva%250AAND%2523ngNvzqu%250A9227%3D9227
- 1%2523ngNvzqu%250AAND%2523nVNaVoPYeva%250A%2523lujYFWfv%250A9227%3D9227
- ;WAITFOR DELAY '0:0:5'--
- “2) ORDER BY 3146-- bFqx”
- ${new java.lang.String(new byte[]{105, 116, 108, 116, 119, 115, 113, 118, 106, 114, 122, 111, 118, 101, 104, 117, 109, 106, 112, 102, 97, 107, 98, 107, 97, 103, 100, 103, 115, 119, 122, 99, 117, 99, 120, 104, 112, 111, 113, 109, 120, 98, 114, 100, 105, 121, 102, 110, 121, 101, 113, 98, 102, 107, 111, 116, 120, 97, 114, 104, 119, 106, 112, 112, 97, 110, 117, 115, 120, 101, 106, 108, 118, 116, 114, 99, 105, 110, 111, 107, 121, 98, 103, 100, 108, 109, 102, 118, 122, 99, 109, 101, 122, 104, 105, 121, 103, 113, 117, 119})}\n"

学习模型效果 - 漏报检测举例

- encryption_data=Be%2FGuUNISh4%2BoR%2B5iPBKR0OtSVUc2XEJOpwyYyxTn%2FGsmwMpXB5KTg%3D%3D&uuid=6ca6a538c30a53bacb7d34ab9d07a98d
- en_num3=AiK%2BAnD1/1gzjwAqe%2BNzBsgQUbTa4NF7a0fAzbVXRLkp1ptHo5DBIKivJwEJ%2BQ1XGXzIUiM/1GAzPpbpMK4OGu0cL1xsu5R0hySE/HcE4x0z
- "iyzmobile=2&iyzversion=4&version=4&module=platformlogin&platform_id=5&utm=18&u=http%3A%2F%2Fbbs.qiuchenglicai.com%2Fforum.php%3Fmod%3Dguide%26view%3Dnewthread%26mobile%3D2&p=ezfsGJHcP3q1MVH8XSks2DZIlN4vBoU3/SMUFvwNffza1hcdyQZD3Ocwr5+3nVPFxkwT9vkVV6JebxWNBlbvrh3Ha0qC+XKHYBbIs7+R6BYnZ6+or0/78M0doyASbRaMUwLBlD7KRYtmtFphePnC23REuOooLjUR6jU2hZkBKQNSErWhLrB9N03LKfg4g/T6&tokenVerified=false",
- "referer" : "http://as-vip.xxxxxxx.cn/v1/red_packet/newUserGiftBag/open?formSource=ympyqh5-2&inviteCode=&platform=wechat&inviteId=&wx_aid=1608782867&wx_traceid=wx0joapeqfdm2ptu&state=mryx&comp_id=9817584&fromSource=ympyqh5-2&wxad_extend=%7B%22aid%22%3A%221608782867%22%2C%22comp_id%22%3A%229817584%22%2C%22pass_ticket%22%3A%22OR3%2F5jiI8hD7wQCb7ce41WaCp2TwflB%2FcutbTsGqT2mjJ3aD%2FBRJTzfttz1LQLas%22%2C%22traceid%22%3A%22wx0joapeqfdm2ptu%22%7D",

学习模型效果

- 借助学习模型的检测引擎，配合平台上的大量流数据进入分析系统，使得平台可以及时的发现新的漏报载荷，同时及时的发现WAF的误报行为。

甲方为什么没有直接部署线上？

1.相比正则检查延迟要高，线上使用对性能需求高, storm延迟更低。

2.有的客户的数据请求量偏少，导致模型准确度不够

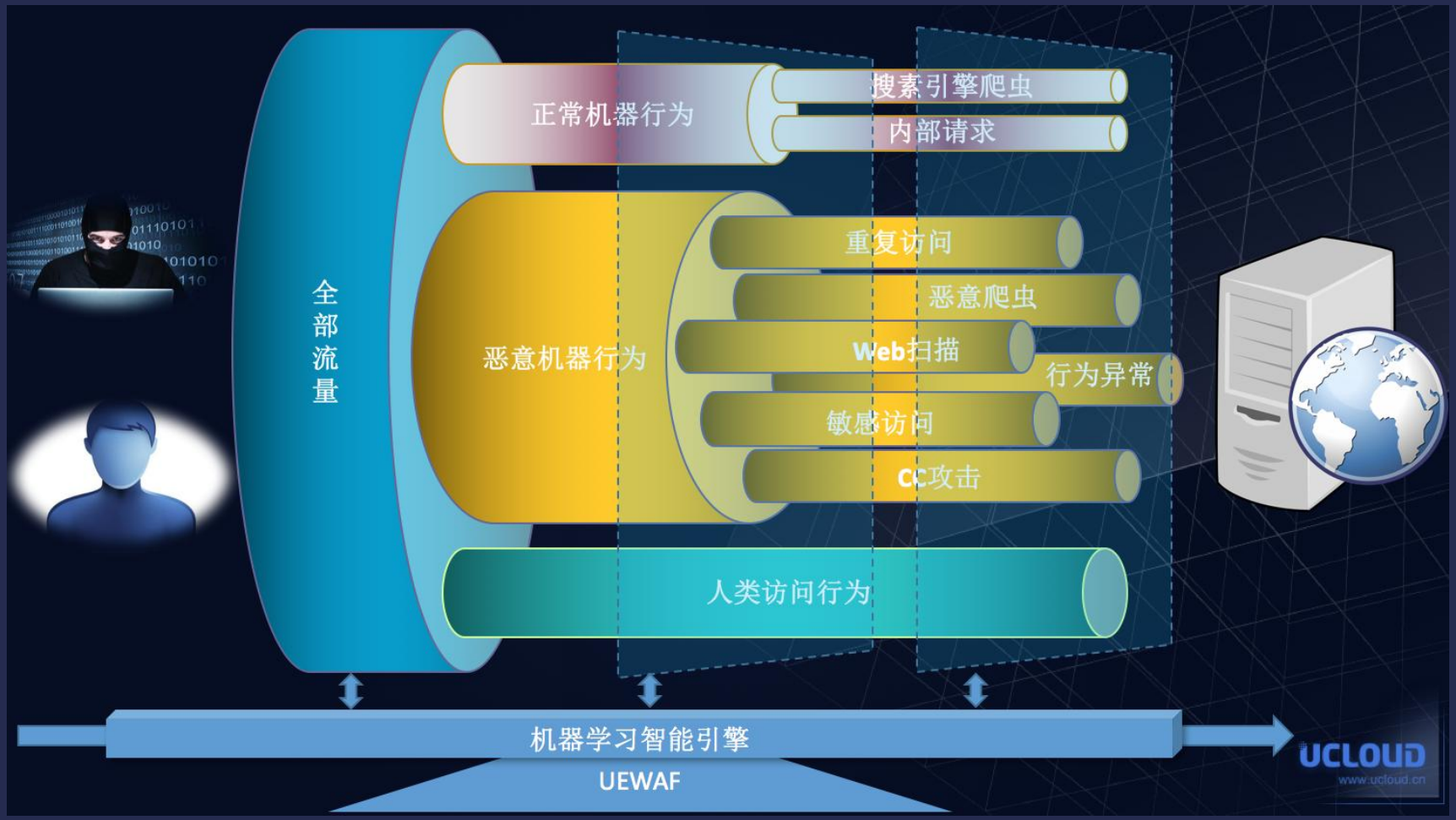
3.依赖运维持续的反馈学习达到很好的效果

乙方可以部署这套检测机制，属于白模型机制。

目录

- 01 WAF概念介绍
- 02 机器学习技术改进WAF质量
- 03 Web Bot检测与识别

Web Bot检测与识别



Web Bot检测与识别

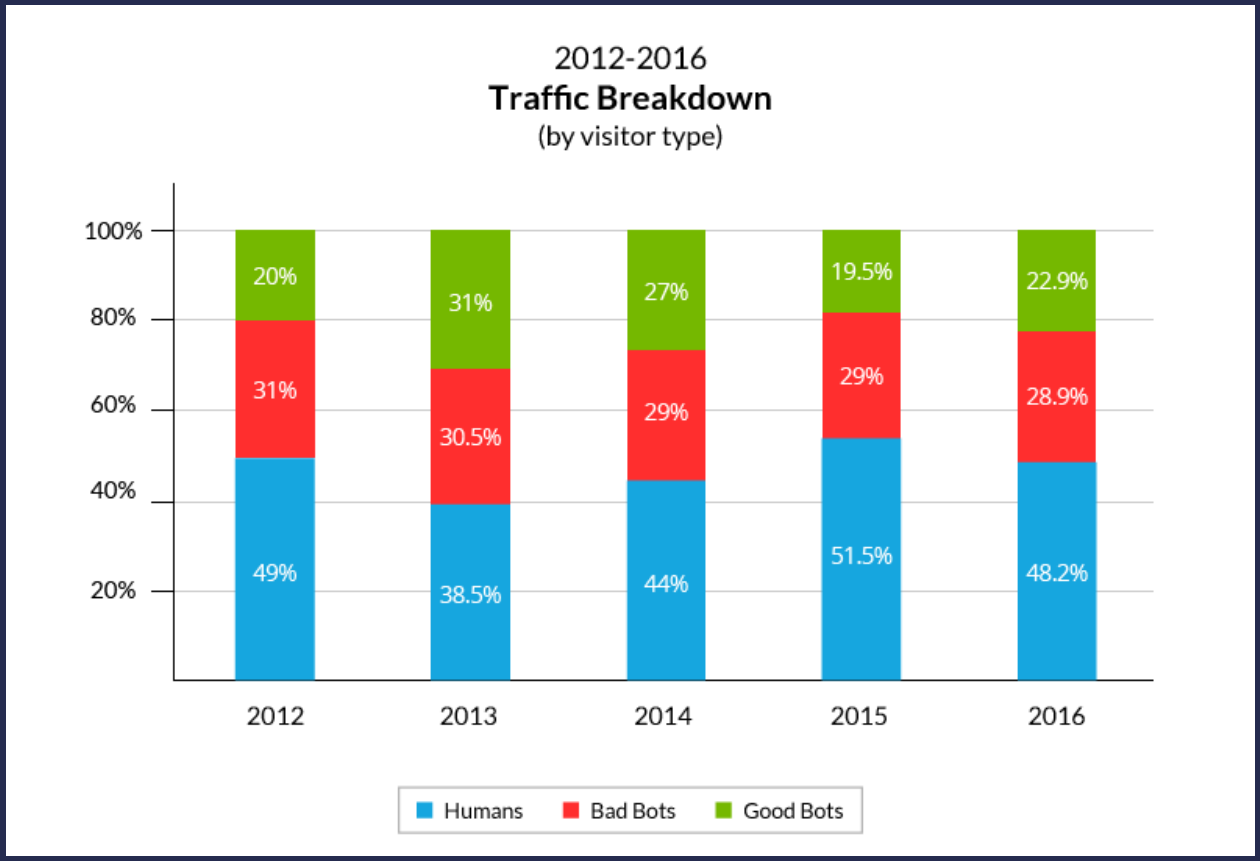
- **正常的bot举例**
- 搜索引擎的爬虫
- 自动化完成的业务逻辑（例如应用自动提取关联信息以及一些自动化工具的统计类工作）
- 监控机器人，主要用于网站可用性等指标的实时监控

Web Bot检测与识别

- **异常的bot举例**
- 僵尸网络发起的ddos以及cc攻击
- 接口烂刷：比如耍票，羊毛党，短信接口，垃圾注册等等
- 垃圾邮件机器人
- 自动化扫描工具，探测网站潜在漏洞，为后续攻击收集情报
- 不规范搜索引擎的爬虫，不遵循robots.txt规范
- 6.恶意爬虫，抓取网站数据，尤其是来自竞争对手的抓取
- 7.恶意行为，多为模拟人操作特征，但是访问逻辑异常，多为网站存在漏洞被发现，攻击者专门开发的恶意程序的访问

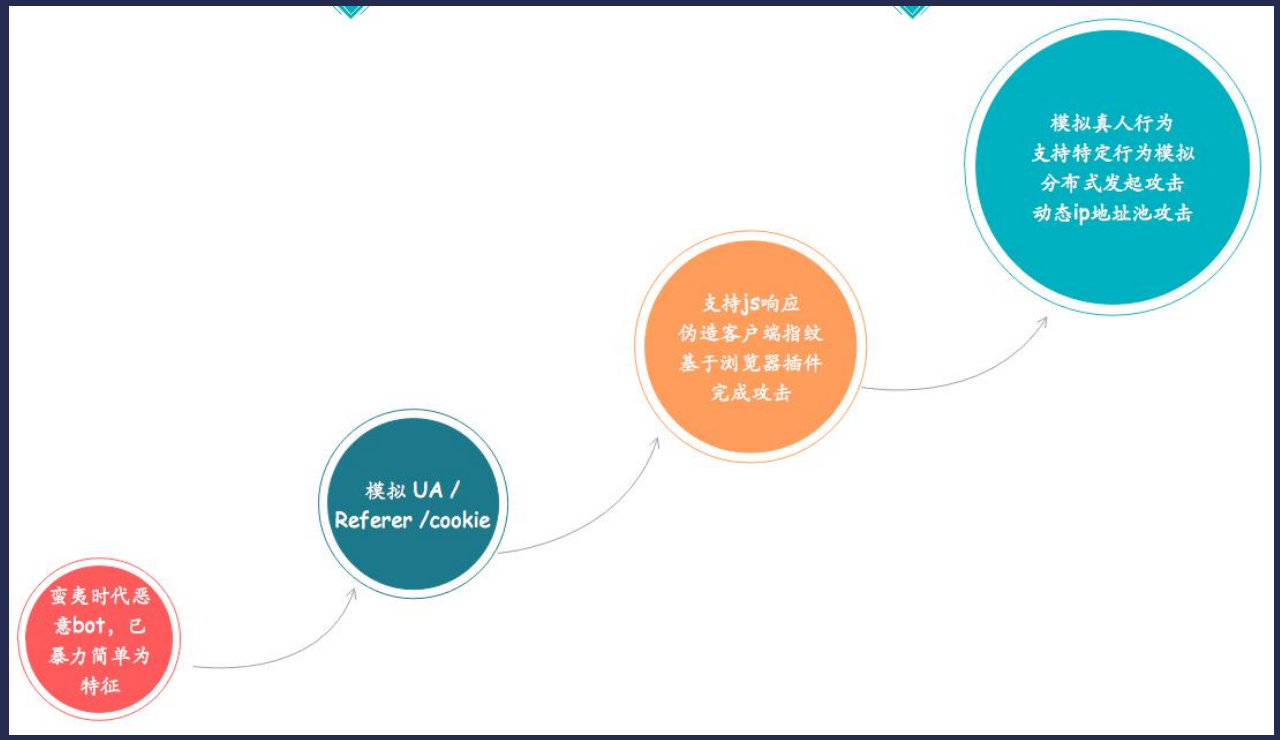
Web Bot检测与识别

- bot流量究竟如何？



Web Bot检测与识别

- bot发展进化--模拟人的行为



Web Bot检测与识别

- 常见防御策略--ip限速
- 存在的问题： 高误报率
- 一些秒杀、抢购等会导致瞬间请求激增的业务和该方案是矛盾的。
- 代理的问题也是很突出的，高校机构以及手机运营商的网关都是代理模式的多。
- 如果一个页面存在大量的资源文件，会导致请求改页面时的关联请求激增，此方案也是不友好的。
- 稍微复杂一点的业务，本身也会提供一些接口给其他服务使用，速率方面的浮动范围也是很大的。

Web Bot检测与识别

- 常见防御策略—钓鱼的方法发现恶意bot
- 原理
- 正常的bot都会请求robots.txt文件，然后遵循robots.txt描述进行后续的bot行为(robots.txt是一种君子协议，标记bot可以访问哪些内容，哪些内容不可以访问)。借用此机制，如果在robots.txt中将一个不存在的url标注为拒绝，然后在网页中，内嵌这个隐藏的url连接，隐藏意味着human不会点击到这个连接，但是恶意的bot有很大概率会访问。

Web Bot检测与识别

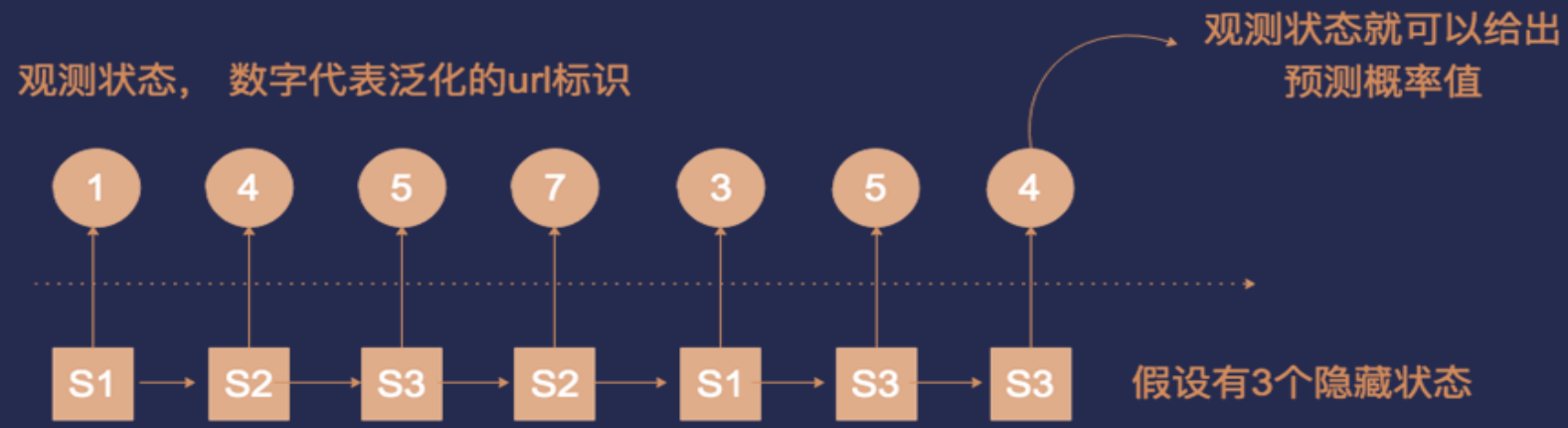
- 常见防御策略—常规技术
- cookie支持
- 验证码（损害正常客户体验）
- javaScript支持（收集客户端动作）
- 设备指纹技术（浏览器指纹）

Web Bot检测与识别

- 会话追踪与行为分析技术
- “行为”是有时间维度上属性的，发生的行为是一系列动作在时间维度上的偏序关系，动作是客户端发起的请求的抽象。行为分析模型首先会动态追踪活动的会话，模型会动态的选择合适的检测时机触发评估逻辑。
- 例如一个爬虫bot，不论它做深度优先还是广度优先的遍历，或者改进的针对特定模式的url的爬虫时，不论它访问频率是高是低，请求资源表现出的偏序关系是异常的。这种偏序关系里即包含了”异常”的来源。

Web Bot检测与识别

- 会话追踪与行为分析技术



隐藏状态之间箭头表示
转换概率，学习得到

Web Bot检测与识别

- **人机识别—创新的信息熵检测技术**
- 信息熵可以用来衡量离散随机事件的出现概率。对于url资源的访问，这里被当成一个离散事件。网络bot请求很人请求资源时，时间间隔上存在不同：人是依据主观需求对目标url进行点击触发，而bot是程序设定好的，例如间隔多久，或者伪造随机时间触发等。

Web Bot检测与识别

技术原理

人在请求资源的时候，下一个请求和上一个请求是存在关联的。因此这也会导致时间间隔随机变量和上一个值也是存在关联性。本质上，至少是一阶马尔科夫过程 。

Ucloud使用一种创新机制的信息熵检测算法，能准确的在请求时间维度上检测出bot，即使bot使用随机时间来请求，同样能被检测出。

Web Bot检测与识别

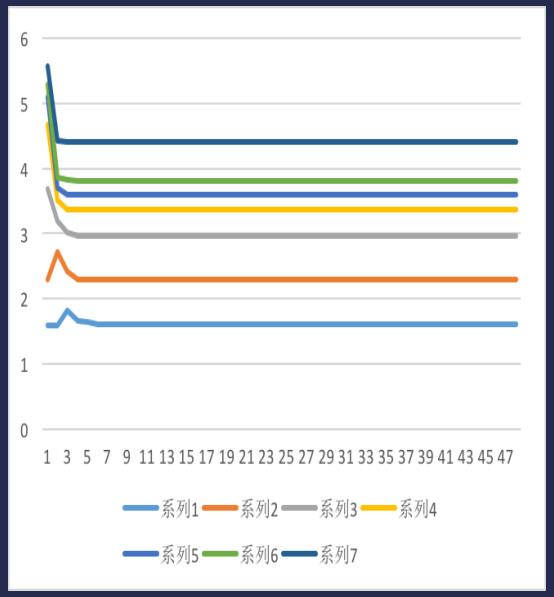


图1：随机请求CCE值

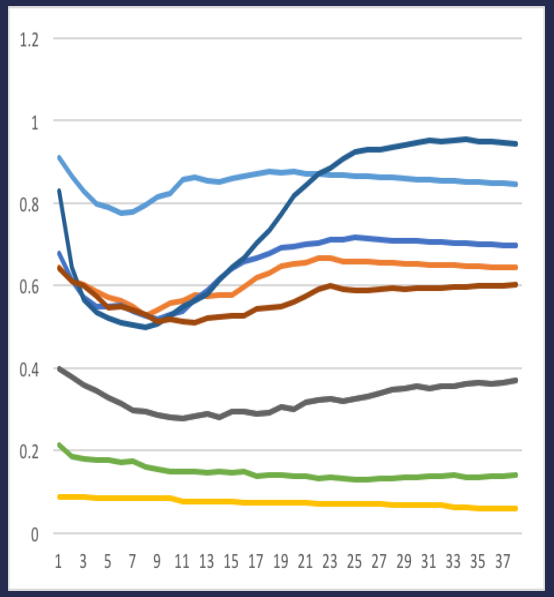


图2 ：恶意Bot LCCE值

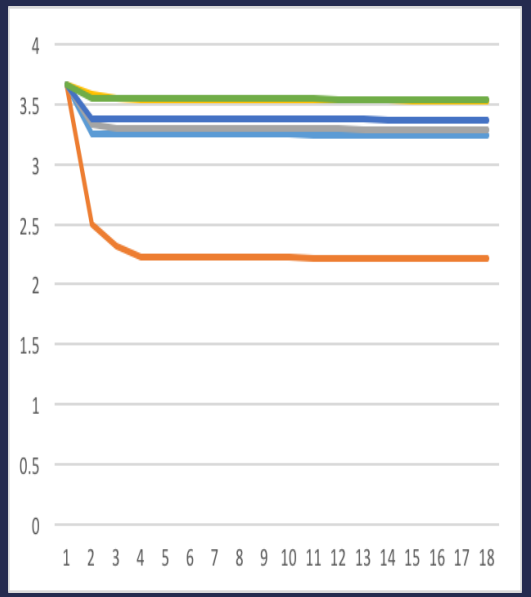


图3： 正常访问LCCE值

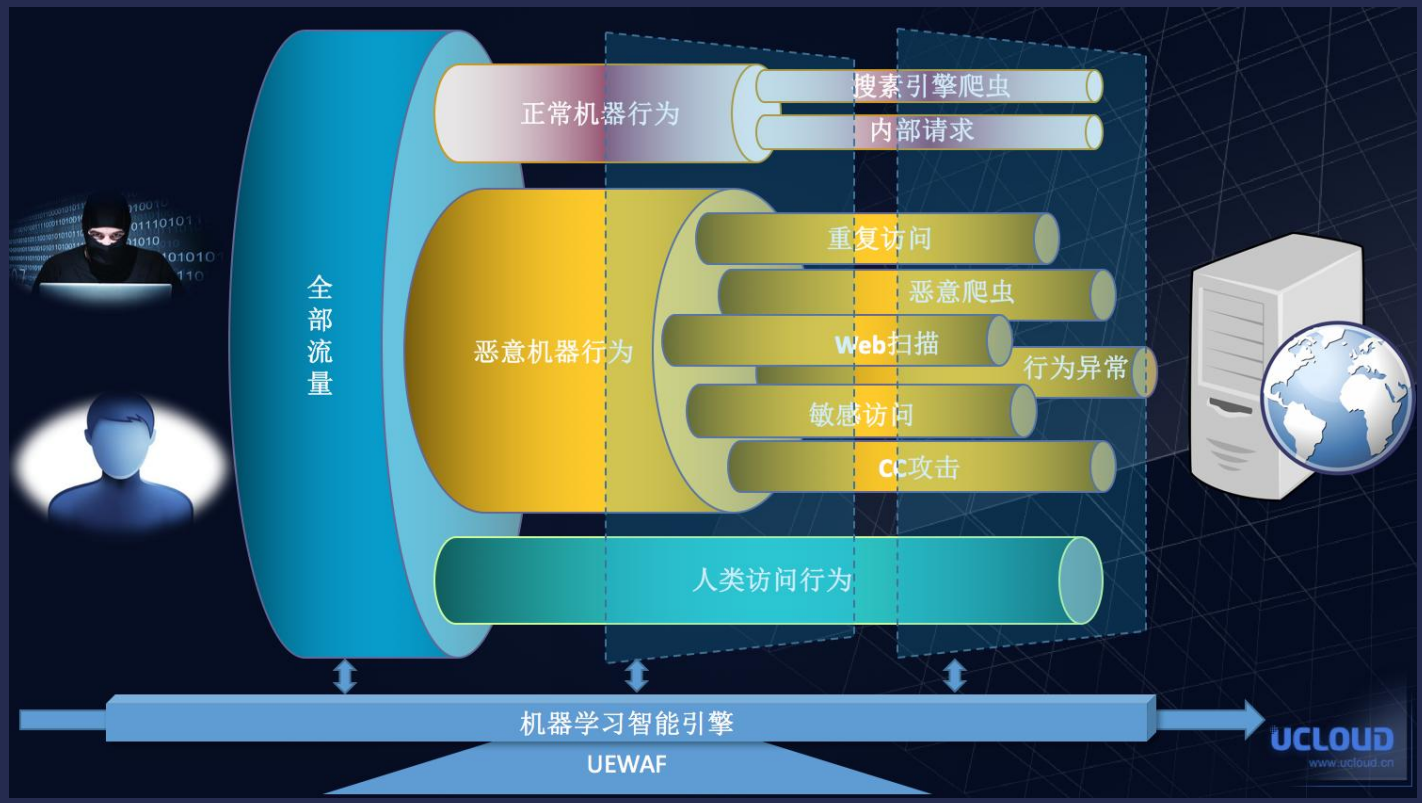
图的横坐标是m值。图1的系列1到系列7分别对应的随机范围是5，10，20，30，40，50，100.图2中最底下的3条(从下到上)分别对应着图1的系列7，系列6和系列5.

Web Bot检测与识别

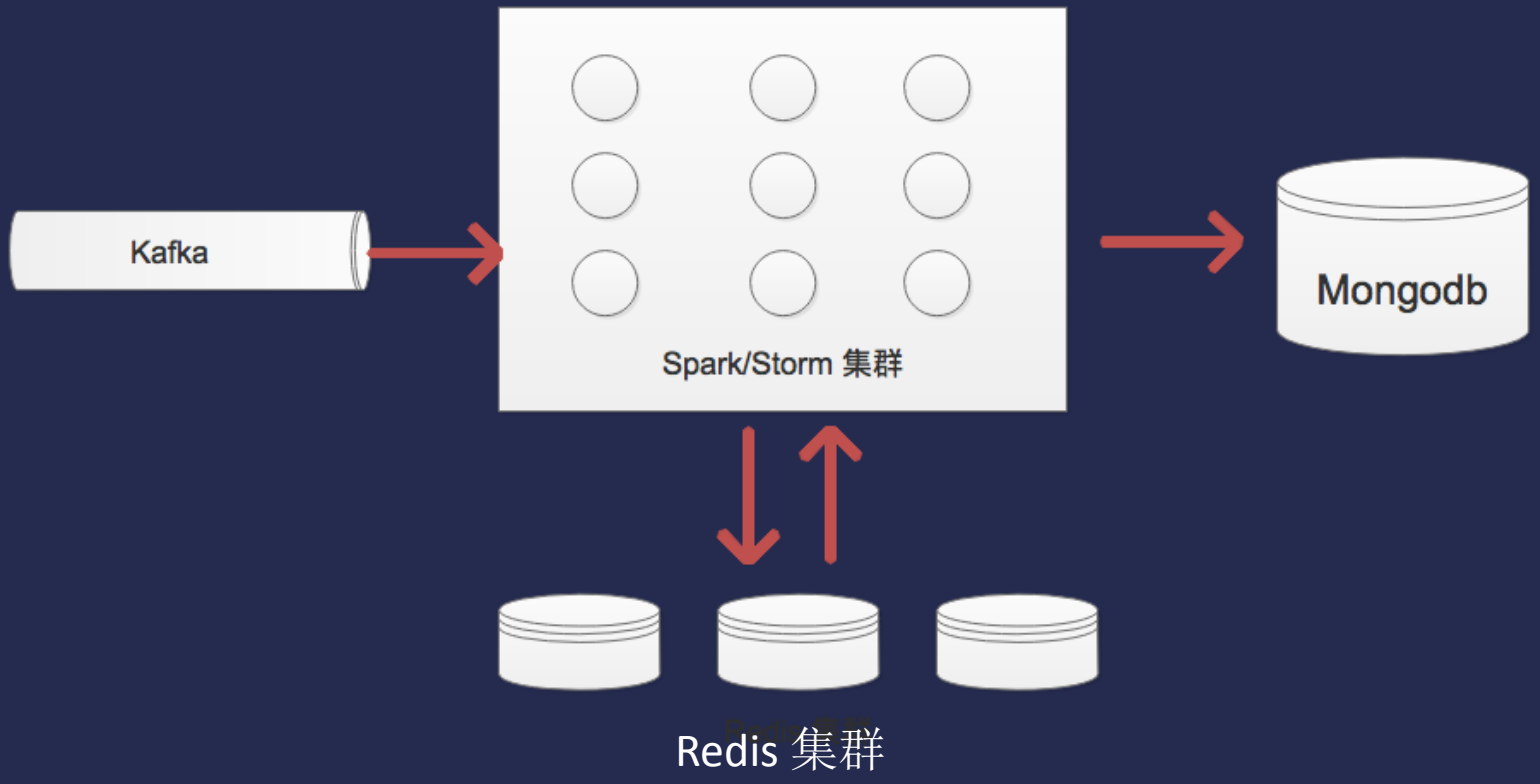
- IP情报技术
- 合作伙伴微步在线
- UCloud自己收集维护有常见搜索引擎IP

Web Bot检测与识别

- 对于识别出的异常访问，结合安全领域知识，进一步进行分类，例如是否是cc攻击，web扫描，恶意爬虫等等

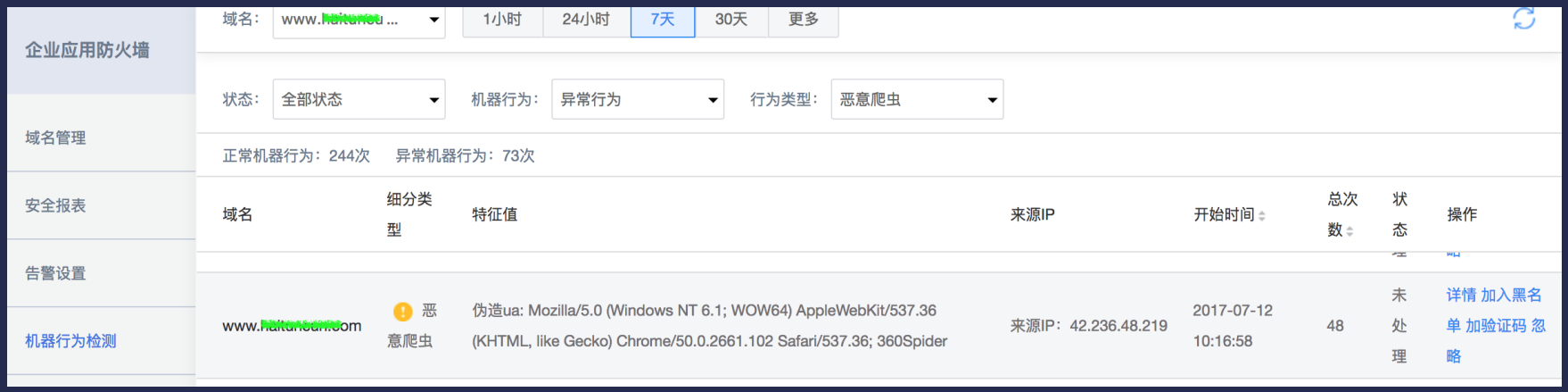


Web Bot检测与识别



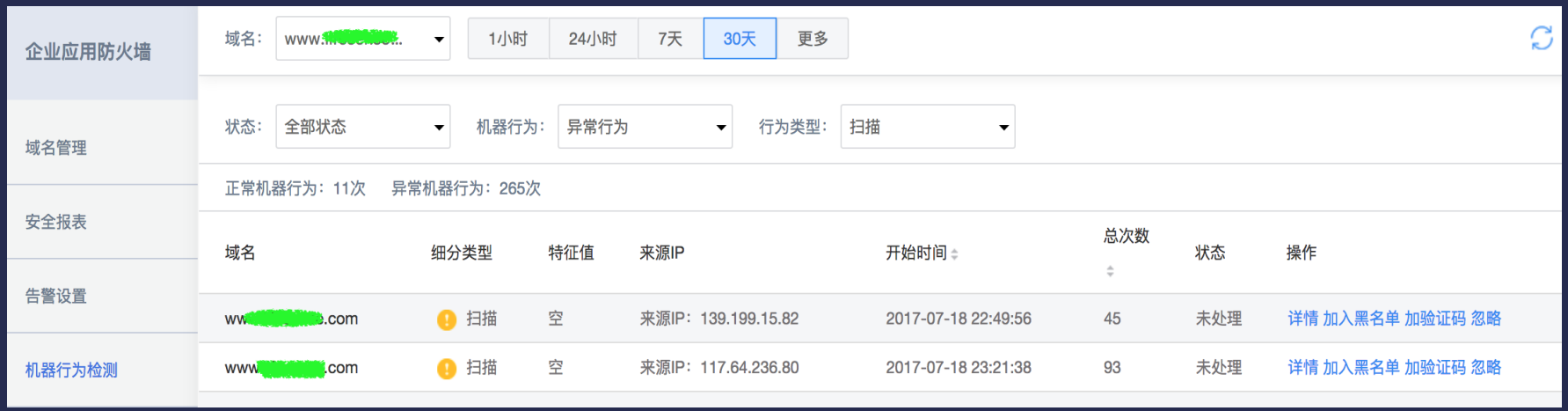
Web Bot检测与识别

- 结果举例



Web Bot检测与识别

- 结果举例



企业应用防火墙

域名管理

安全报表

告警设置

机器行为检测

域名：www.[redacted]..

1小时 | 24小时 | 7天 | 30天 | 更多

状态：全部状态　机器行为：异常行为　行为类型：扫描

正常机器行为：11次　异常机器行为：265次

域名	细分类型	特征值	来源IP	开始时间	总次数	状态	操作
ww[redacted].com	扫描	空	来源IP：139.199.15.82	2017-07-18 22:49:56	45	未处理	详情 加入黑名单 加验证码 忽略
www[redacted].com	扫描	空	来源IP：117.64.236.80	2017-07-18 23:21:38	93	未处理	详情 加入黑名单 加验证码 忽略

企业应用防火墙

域名管理

安全报表

告警设置

机器行为检测

域名：www.[redacted]...

1小时 | 24小时 | 7天 | 30天 | 更多

状态：全部状态　机器行为：异常行为　行为类型：行为异常

正常机器行为：244次　异常机器行为：73次

域名	细分类型	特征值	来源IP	开始时间	总次数	状态	操作
www.[redacted].com	行为异常	伪造refer: http://www.[redacted].com	来源IP：121.196.222.160	2017-07-14 10:53:32	108	未处理	详情 加入黑名单 加验证码 忽略
www.[redacted].com	行为异常	伪造refer: http://www.baidu.com	来源IP：107.189.143.194	2017-07-18 18:51:39	54	未处理	详情 加入黑名单 加验证码 忽略
www.[redacted].com	行为异常	伪造refer: http://www.baidu.com	来源IP：104.171.180.242	2017-07-18 19:50:58	147	未处理	详情 加入黑名单 加验证码 忽略
www.[redacted].com	行为异常	伪造refer: http://www.baidu.com	来源IP：23.19.26.146	2017-07-18 16:37:48	55	未处理	详情 加入黑名单 加验证码 忽略

Web Bot检测与识别

某电商客户检测到的撞库攻击实例