阿里云 | 云盾先知 安全情报

WAF是时候跟正则表达式说再见

破见

先知白帽大会
开放创新·先知先行

议题内容

- Part 1 正则表达式不适合用于构建WAF
- Part 2 现有WAF的解决方案
- Part 3 如何构建未来的WAF

Part 1

正则表达式不适合用于构建WAF

感性认识—误报和漏报难以平衡

关键字【waf】的搜索结果共112记录

提交时间	标题
2016-05-17	中国石油某电商SQL注入(waf绕过)
2016-05-04	韩国本土最大电商linterpark全球站/主站存在sql注入/9库/大量表/双编码/有waf/可union
2016-04-29	中石化某业务SQL注入漏洞（绕过WAF）
2016-04-21	汽车安全之奔驰某站SQL注入/可影响大量客户信息(bypass waf)
2016-04-21	虎扑体育某站注入（绕waf）
2016-04-13	2345某主要站点SQL注入影响所有用户数据影响N个同服网站数据（注入需绕过WAF）
2016-04-09	迅雷官方APP存在SQL注入（跨70库/艰难绕WAF）
2016-04-05	绿盟WAF SQL注入检测bypass
2016-03-28	宁波某p2p平台存在SQL注入漏洞（可绕过WAF）
2016-03-28	申银万国证券mssql注射绕waf写shell
2016-03-19	车易拍某系统SQL注入40W用户数据（绕过WAF）
2016-03-05	海尔某站绕过waf注入至Getshell(附脚本)
2016-02-20	天融信数据安全管理系统存在SQL注入无需登陆（非注释绕waf）
2016-02-03	搜狐某站3处SQL注入漏洞（sqlmap绕过数据编解码流程与waf）

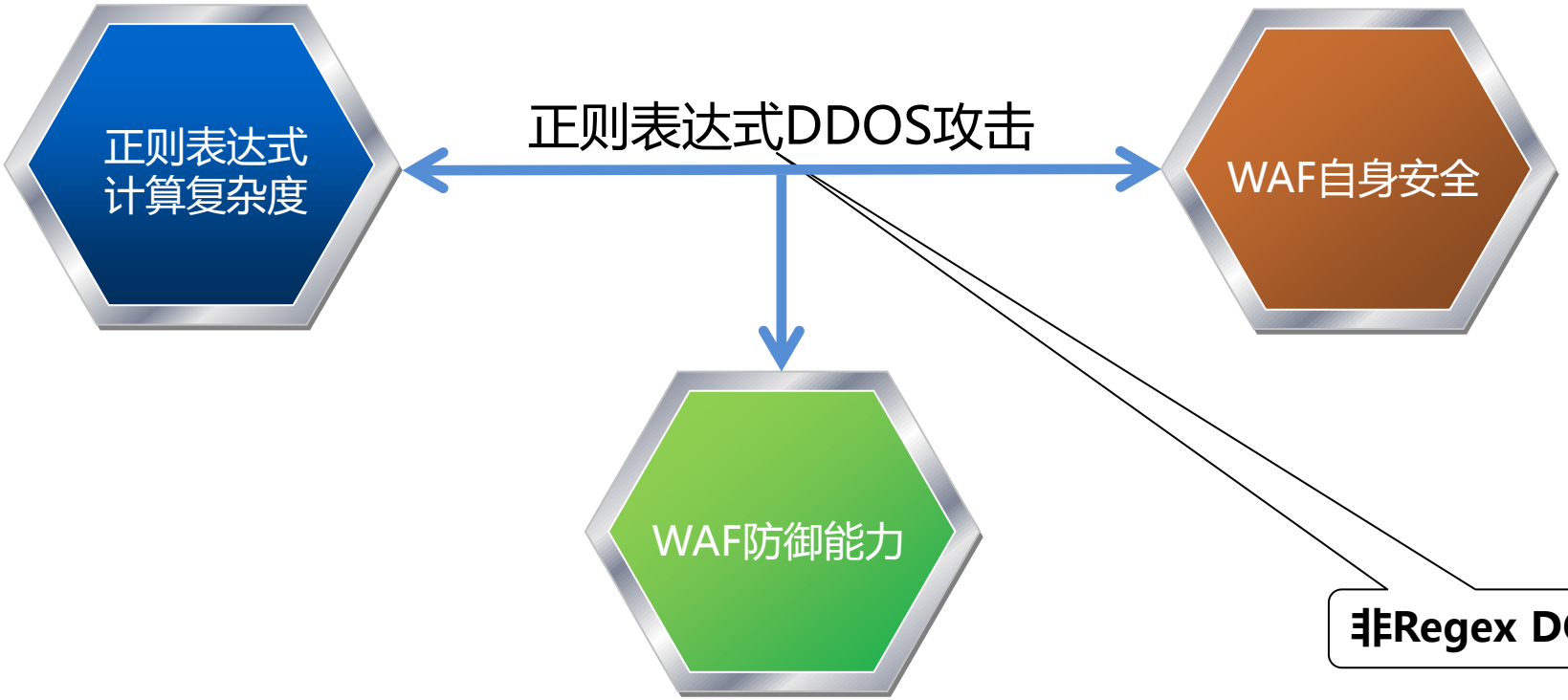
WAF攻防研究之四个层次Bypass WAF

破-见 | 2016-08-11 12:15:58 | 删除

见招拆招：绕过WAF继续SQL注入常用方法

mikey 2014-06-17 +10 共578132人围观，发现 34 个不明物体 WEB安全

尝试寻找有理证明



正则表达式DDOS攻击

提出一种正则表达式的DDOS攻击：

正则表达式的最坏时间复杂度大于等于$O(n^2)$，该正则表达式可被DDOS攻击

Regex DDOS与Regex DOS不同
- $O(n^2)$，$O(2^n)$
- Regex DDOS目前普遍存在
- Regex DOS很难找到了

输入长度(K)	PCRE/PHP(ms)	JAVA(ms)
1	0.5	32
2	23	53
4	111	142
8	458	500
10	720	786
20	2910	2941

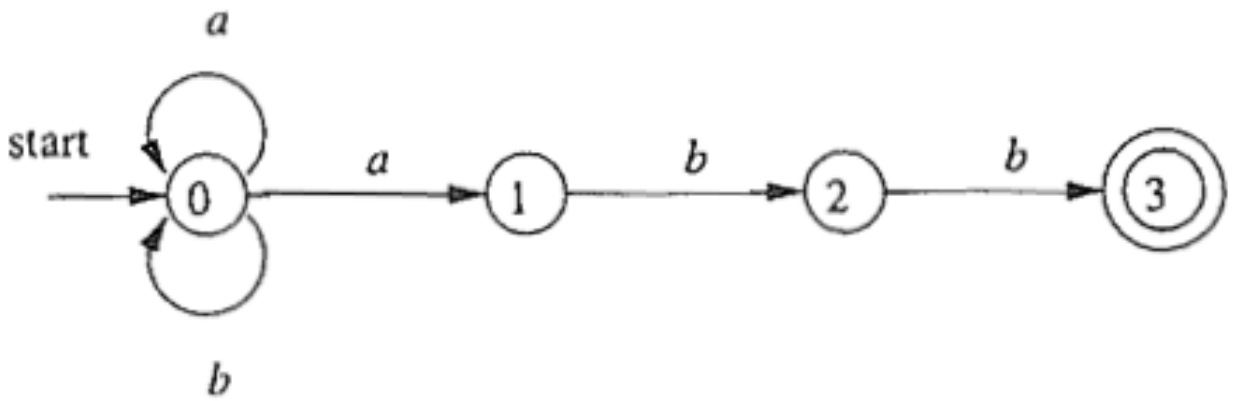
寻找能被DDOS的正则表达式

寻找最坏时间复杂度大于等于$O(n^2)$的正则表达式

利用正则表达式匹配的回溯

正则表达式匹配原理：NFA

正则表达式：(a|b)*abb 对应的NFA

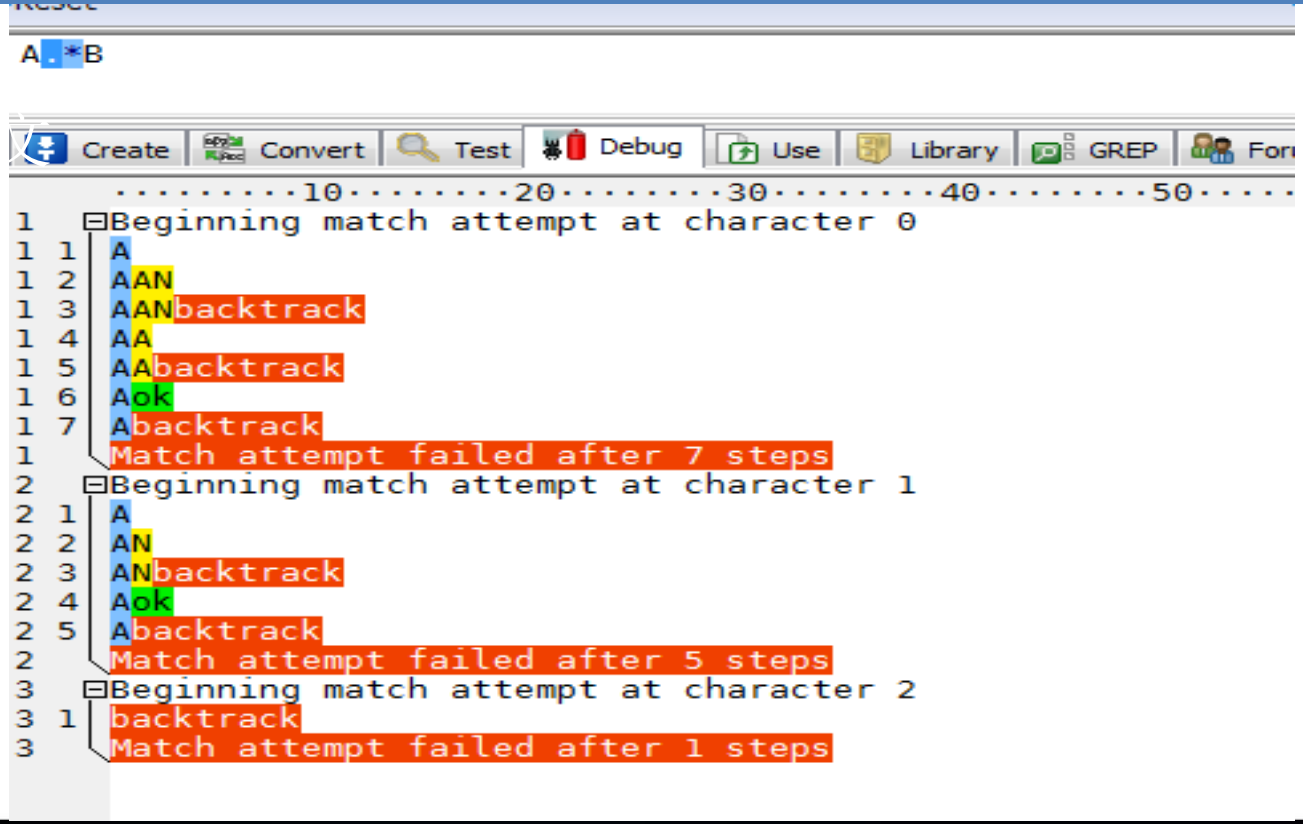


匹配算法需要尝试每一条路径，直到找到一条匹配路径。尝试所有路径失败则匹配失败。

- 尝试所有匹配路径
- 路径尝试失败，需要回溯

正则表达式DDOS原理

正则：A.*B
文本：AAN



可被DDOS的一种正则表达式模式

$$\boldsymbol{Pattern} = (\boldsymbol{SubA})(\boldsymbol{SubB})^{*}(\text{SubC})，\ 其中\boldsymbol{SubA} \in \boldsymbol{SubB}$$

例如

- select.*from
- <script>.*</script>
- <title>.*</title>

影响范围

owasp-modsecurity-crs

(?i:(?:(union(.*?)select(.*?)from)))
(?i:<META[\s/+].*?charset[\s/+]*=)

wordpress-4.7.1

class-wp-text-diff-renderer-table.php: Line 266:
(<ins>.*?</ins>|~~.*?~~)

普遍存在

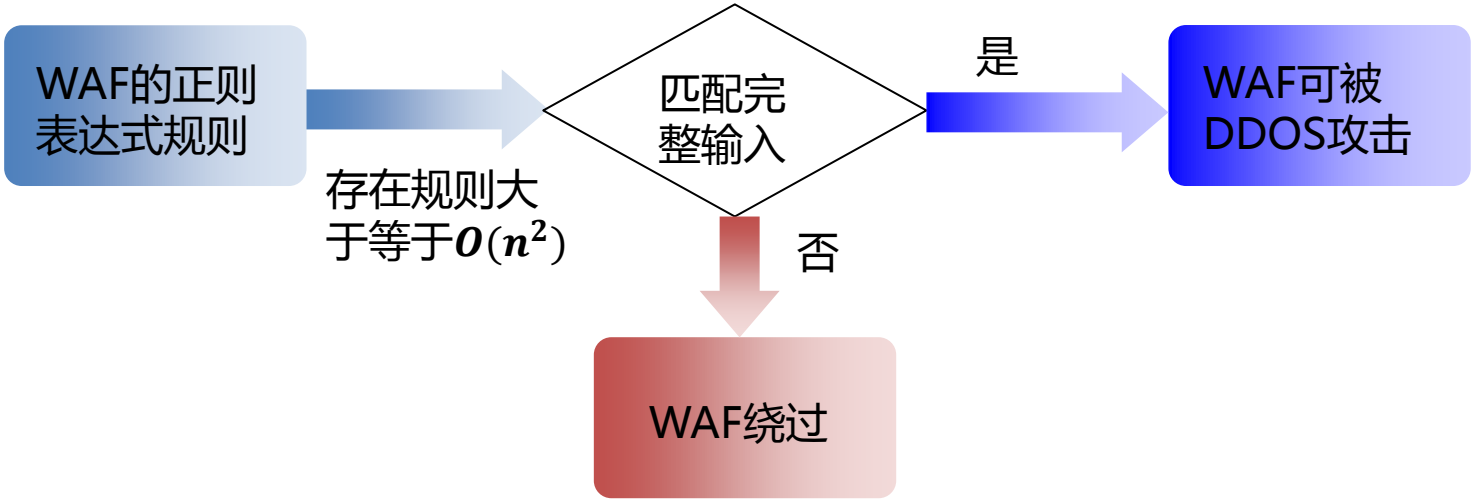
Discuz_X3.3_SC_UTF8

admincp_announce.php
- Line 136: '/<b>(.*?)<\/b>/i',
- Line 139: '/<i>(.*?)<\/i>/i',
- Line 142: '/<u>(.*?)<\/u>/i',

某云WAF/360_safe3.php

某云WAF: union\s+select.*from
360_safe3.php:
- \/*.+?*\/
- <\s*script\b
- UNION.+?SELECT

正则表达式不适合用于构建WAF



维护几十条到几百条正则表达式规则，保证拦截率，误报率前提下，所有规则最坏时间复杂度小于$O(n^2)$，是一件很难事情。

正则表达式不适合用于构建WAF

Part 2

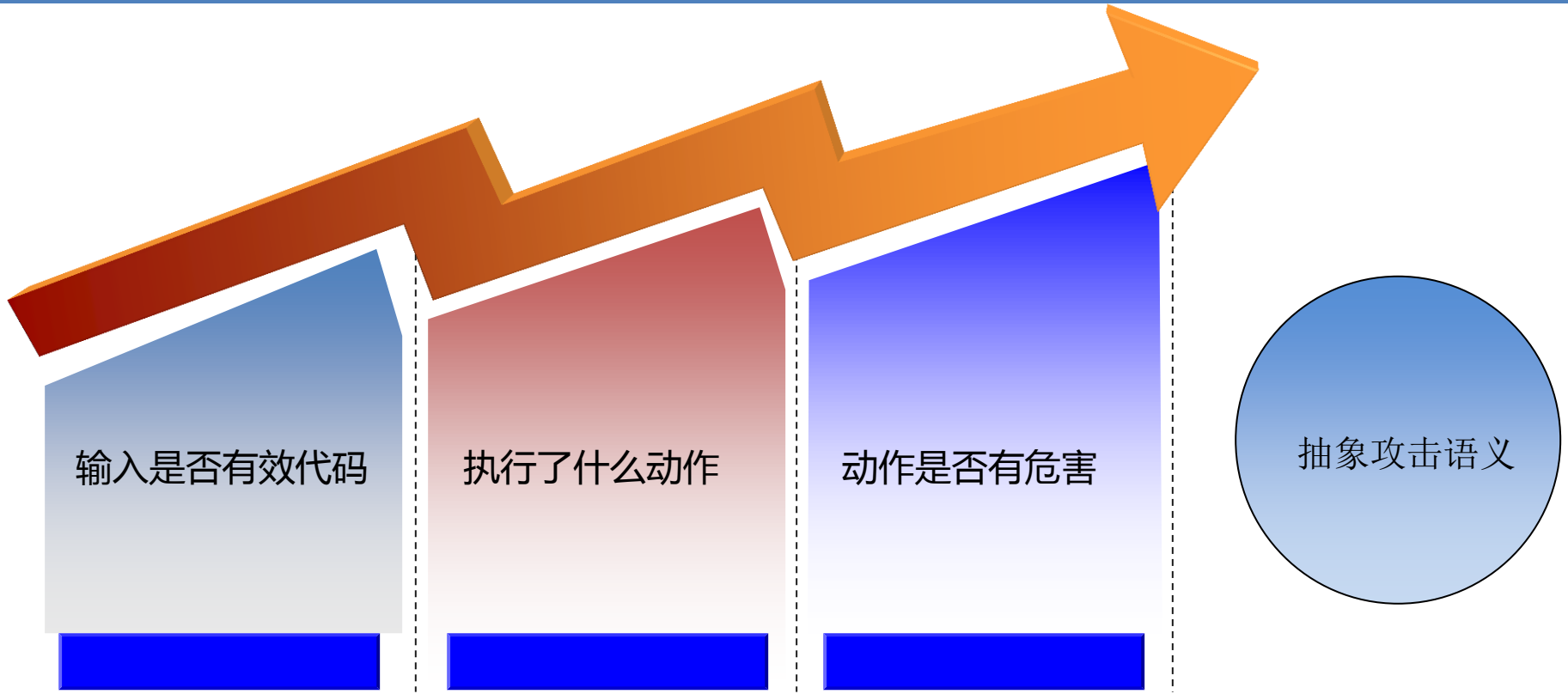
现有WAF的解决方案

基于语义检测的WAF

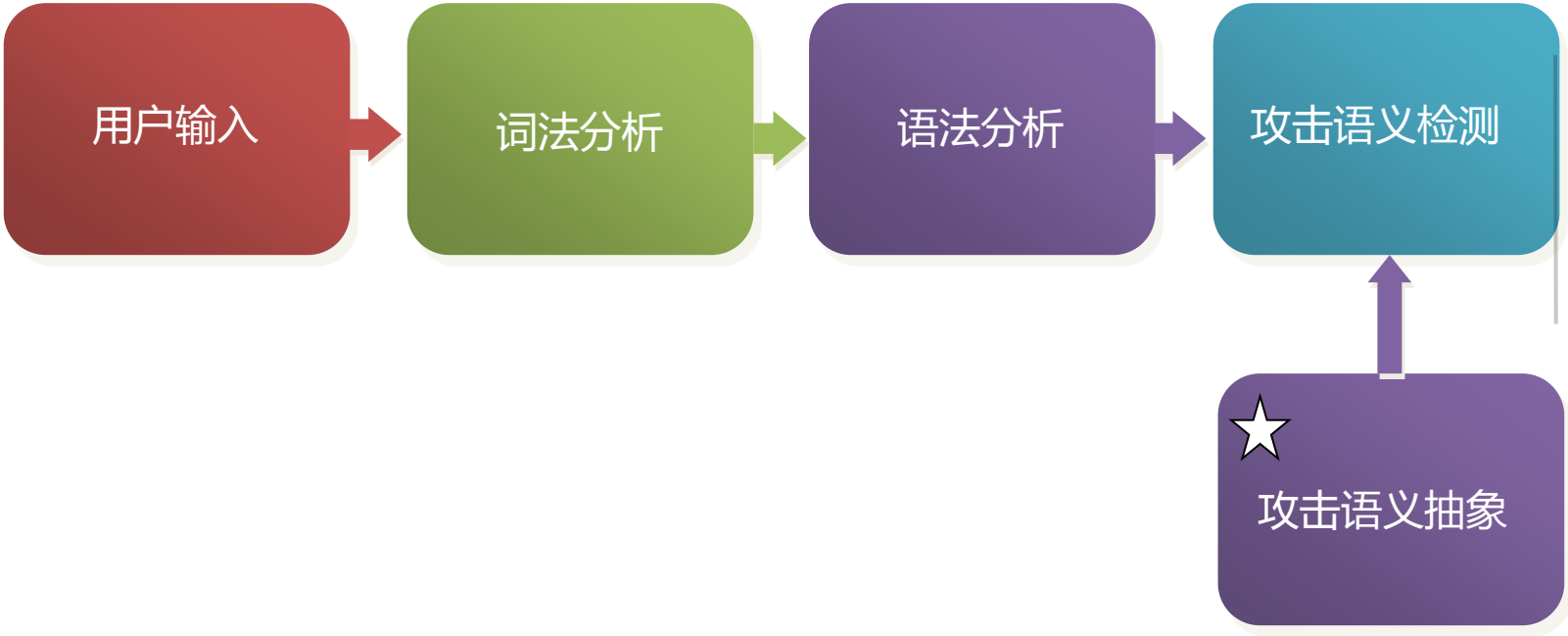
文本：What’s problem about ‘Select id,name from’, give me a hand。

正则：select.*from	Improved
正则只关注‘Select id,name from’，忽略了上下文的信息。在做注入判断时，对输入进行片面的理解，导致误报。	整个输入作为一个整体，尝试理解意图。 How?

基于语义检测的WAF



基于语义检测的WAF—实现

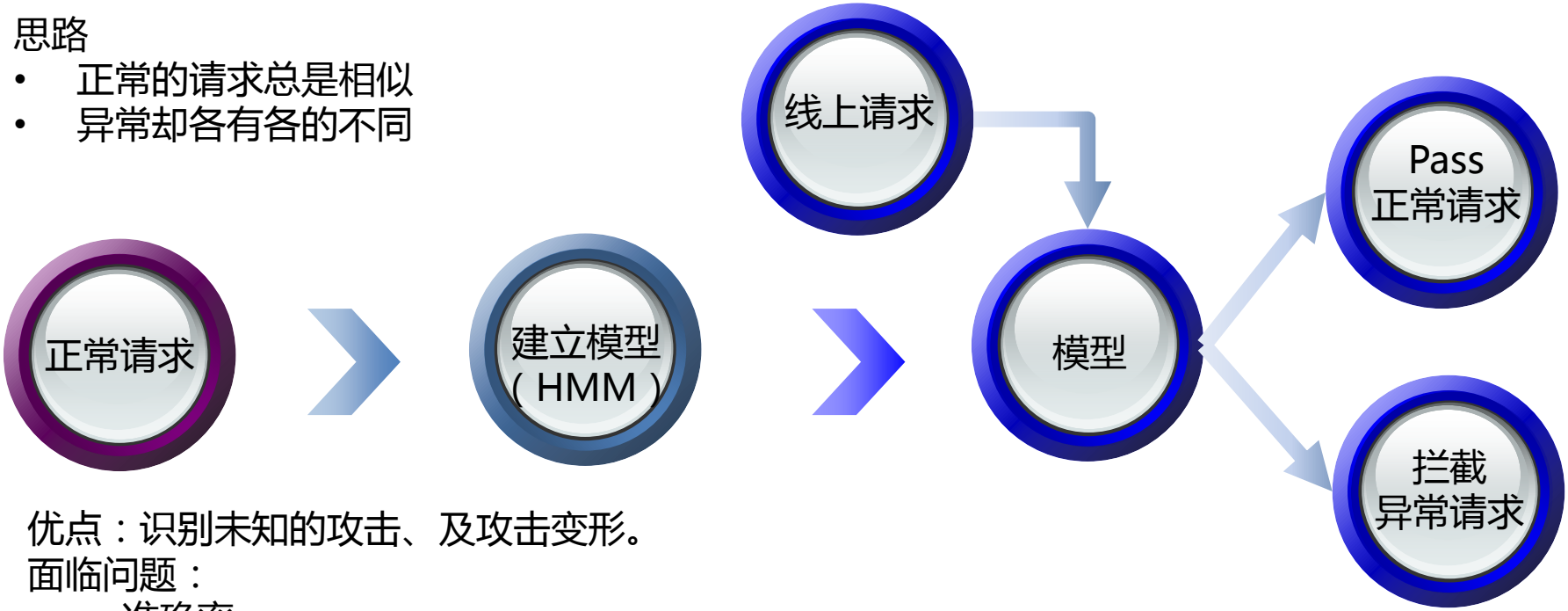


基于语义检测的WAF—优缺点

基于统计的机器学习WAF—异常模型

思路

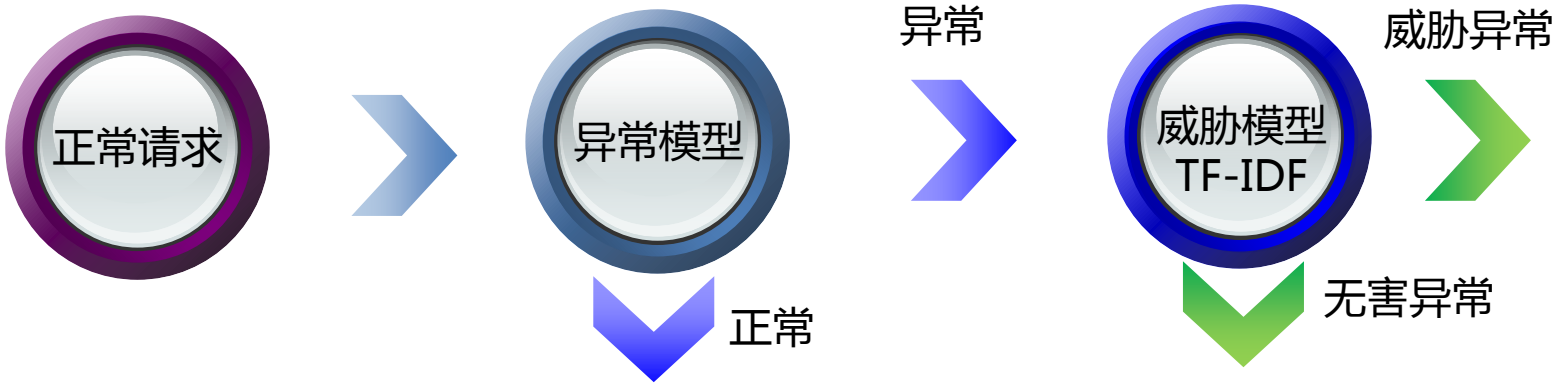
- 正常的请求总是相似
- 异常却各有各的不同



- 优点：识别未知的攻击、及攻击变形。
- 面临问题：
  - 准确率
  - 应用变更

基于统计的机器学习WAF—异常模型&威胁模型

思路：在异常数据的基础上，注入一些领域知识，从而构成一个分类器，从异常中剥离出攻击



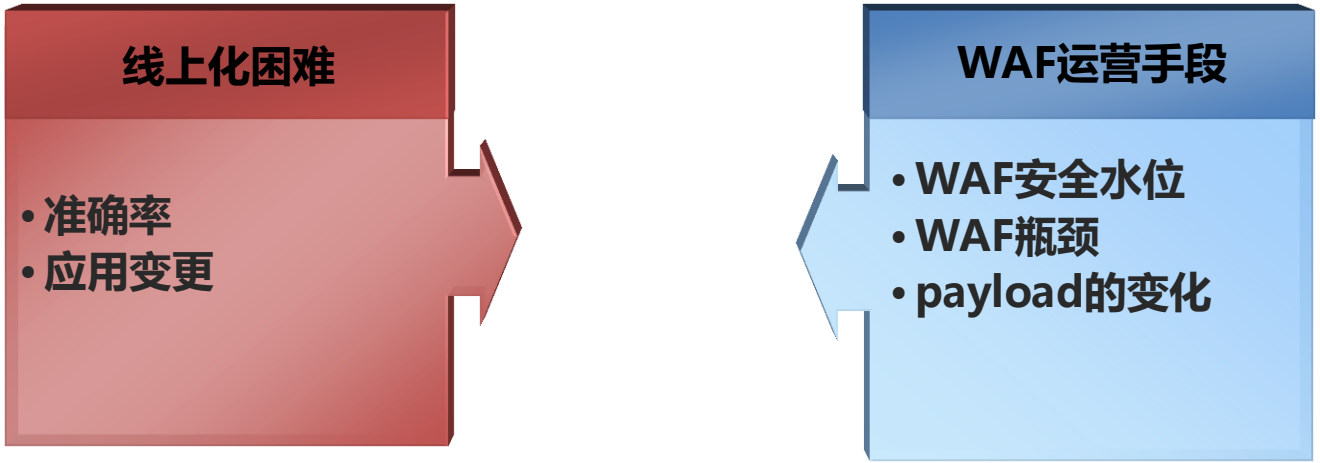
优点：

- 准确率相对单独异常模型，提升了许多。

问题：

- 模型滞后性：领域知识注入导致。
- 修复响应：出现漏报，在线上如何修复。

基于统计的机器学习WAF—个人总结

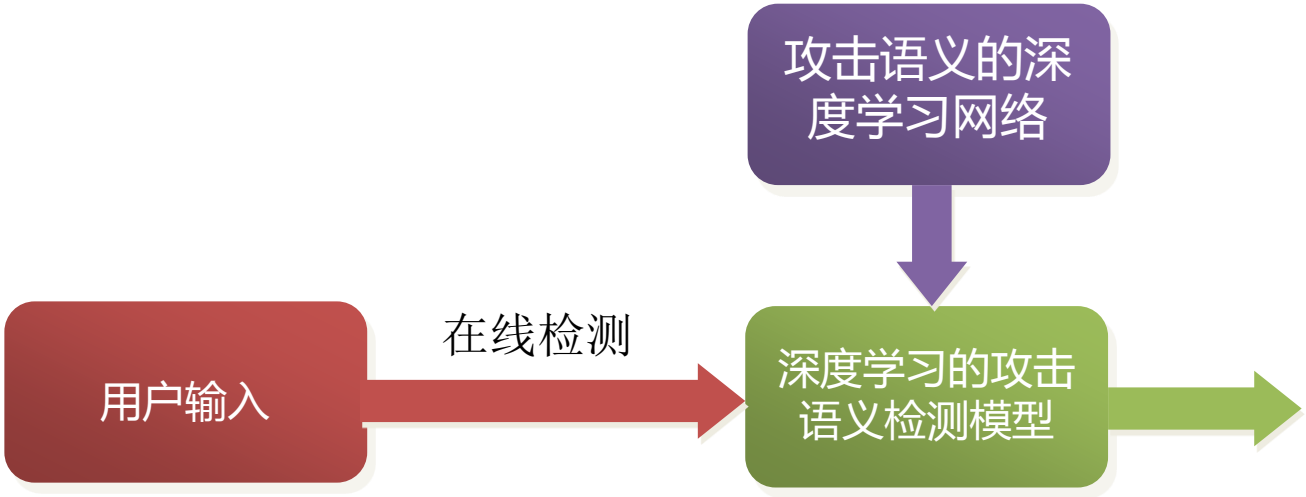


Part 3

如何构建未来的WAF

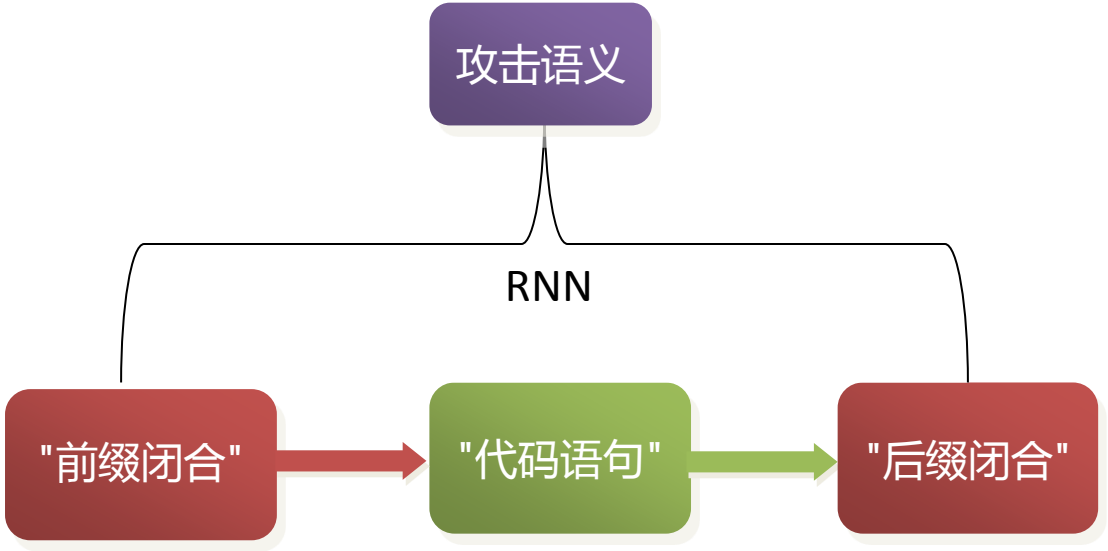
基于深度学习构建WAF

思路：用深度学习的模型代替语义检测中的词法分析、语法分析。



攻击语义的深度学习网络

一般攻击语义



为什么深度学习、RNN：

- 深度学习可通过学习一种深层非线性网络结构，实现攻击语义的逼近。
- RNN能够使信息持续保存，根据已有知识进行思考，更容易学习到攻击语义。


阿里云
云盾先知
安全情报
宁静致远
weibo.com/u/5261507198