人工智能在WEB安全中的实践

冯景辉
百度安全技术总监

北京　伦敦　纽约　旧金山　圣保罗　上海　东京

QCon 全球软件开发大会

[上海站]

信息安全　机器学习　人工智能　黑产　互联网金融（FinTech）　团队管理　基础设施　云计算　软件性能　硅谷　微服务　互联网架构

主办方 Geekbang 极客邦科技　InfoQ

2017年10月17-19日
上海·宝华万豪酒店

扫描二维码
开启软件开发新思路

EGO聚合

TOGETHER CHANGE

2017

Geekbang 极客邦科技 | EGO EXTRA GEEKS' ORGANIZATION NETWORKS

EGO会员招募季

EGO旨在组建全球最具影响力的技术领导者社交网络，联结杰出的技术领导者学习和成长。

2017年6月30-7月10

扫码报名

China · Shenzhen

SPEAKER

INTRODUCE

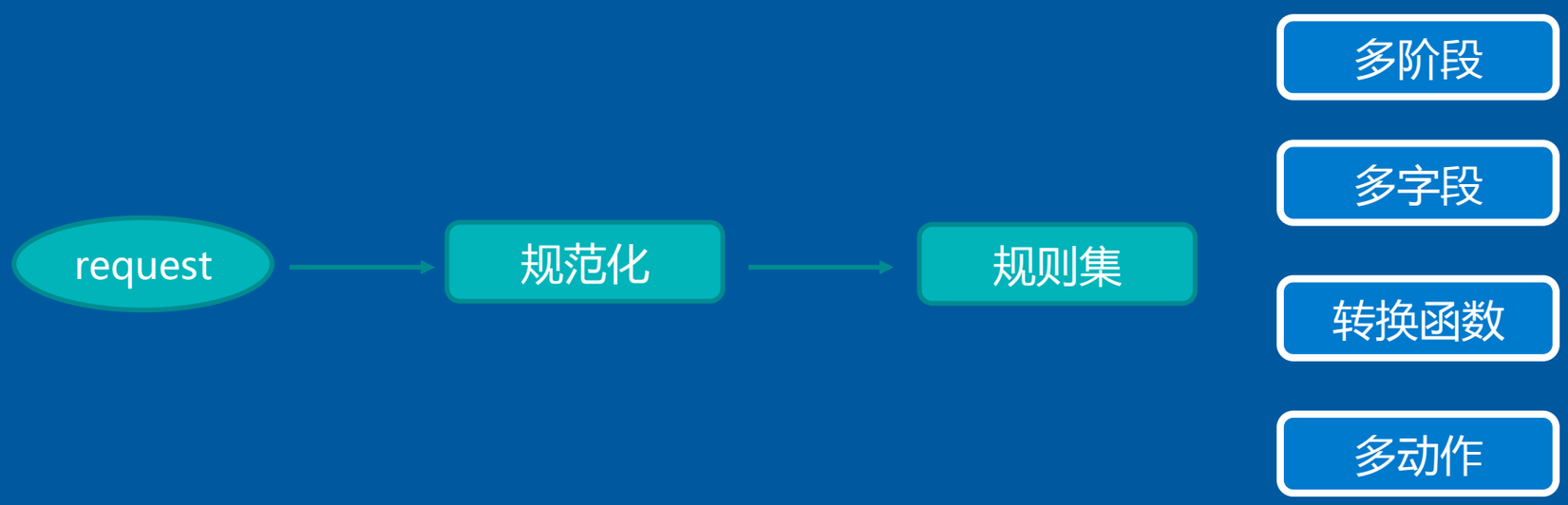
冯景辉 百度安全

- 百度安全技术总监，百度商业安全产品总经理。安全行业11年老兵，国内第一家完全基于SaaS的云安全服务厂商安全宝的联合创始人兼研发副总裁，安全宝系统架构总设计师。

- 创立安全宝之前，冯景辉曾在当时中国最大的反病毒企业瑞星公司担任高级软件工程师、研发经理等职务，带领团队在企业级安全产品线上先后开发了9200、9300等多款安全防护系统

从ModSecurity开始说起

SecRule REQUEST_COOKIES|!REQUEST_COOKIES:/__utm/|REQUEST_COOKIES_NAMES|
ARGS_NAMES|ARGS|XML:/* "(?i:(?:(union(.*?)select(.*?)from)))"



绕过仍然无法避免

REVERSE(noinu)+REVERSE(tceles)

un?+un/**/ion+se/**/lect+

SQL Tokenizer Parser Analyzer

语法解析

- 关键词解析
- 语法规则
- 基本函数

语义分析

- SQL补全
- 环境感知
- 注入检测
- 语义行为

libinjection

兼容性

除了MySQL，其他SQL

误报

本质上，系统将尽量补全SQL，而SQL一旦通过语法分析，只要存在Token，误报就容易出现

机器学习初探

典型的机器学习场景

有监督学习 VS 无监督学习

图像识别

NLP

关联新闻

机器学习初探

特征选取

01

算法选择

02

样本训练

03

日志审计

04

特征选取

基于Payload的特征选择，需要结合安全特性，比如关键字、字符特征、甚至请求长度，同时避免过拟合

样本训练

选取大量黑白标注样本，同时要控制样本类型的分布

算法选择

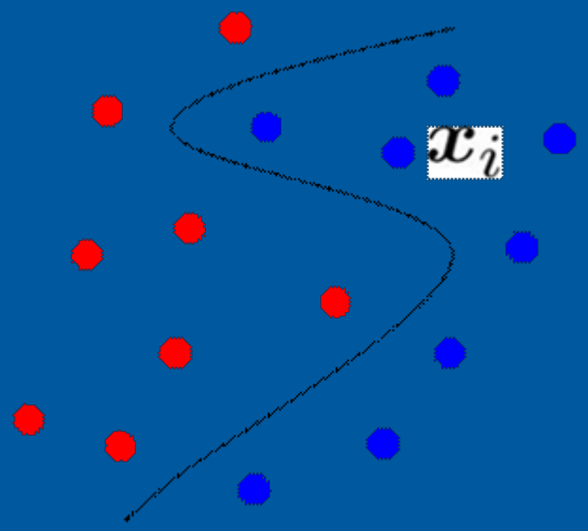
有监督学习有诸多常用算法、SVM、HMM、贝叶斯等等

日志审计与回归

当前有监督学习主要应用在离线日志分析中，快速发现未知攻击样本

支持向量机-XSS检测应用

SVM的典型问题



结构风险最小，
而非经验风险最小

特征选取

- URL长度
- 第三方域名个数
- 敏感字符
- JS关键字

召回率

93%

准确率

90%

支持向量机-不足

不适合大规模数据集训练
广泛采用的LibSVM，在最坏情况下复杂度为O(n^2)（训练样本数平方）

本质上与规则无异
可以对抗基本变形，只是对原有规则系统提供一定的宽容度

准确度无法满足需呀
对原有系统提供一个离线检查机制

是否能够结合更多的识别方法

隐马尔可夫

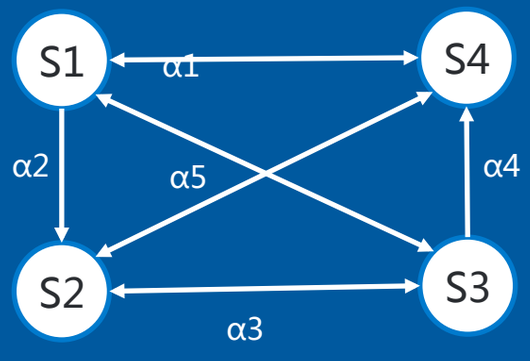
最大熵模型

<script>alert(0)</script>

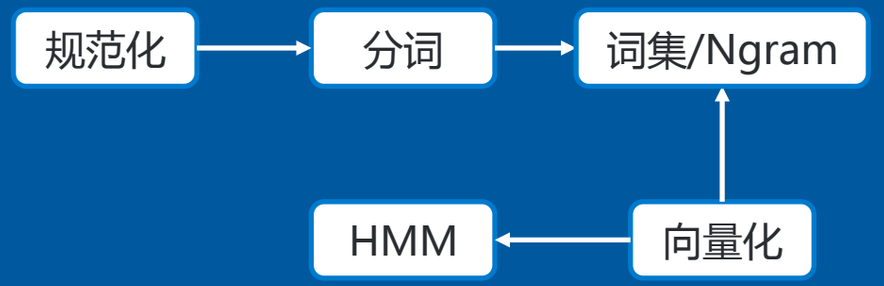
S1:符号
S2:字符
S3:数字
S4:分割符号

观察序列

隐含序列



加入词法之后

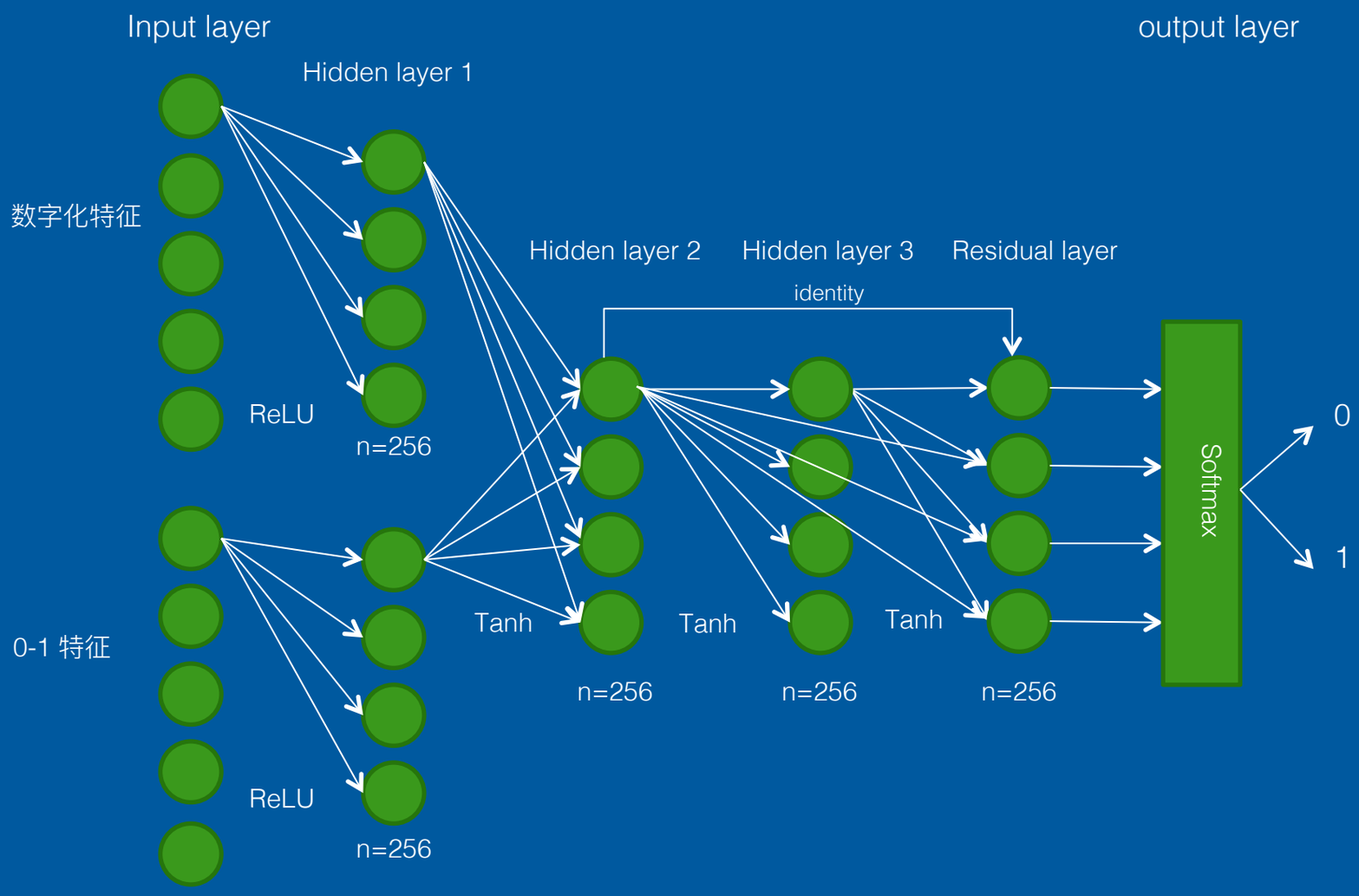


200维特征

召回率80%

准确率90%

从浅层学习走向深度神经网络



从浅层学习走向深度神经网络

特征提取

请求

结构特征

- URL长度
- 特殊字符数量
- JS关键字数量
- SQL关键字数量
- UA

经验特征

- “(“数量
- Union
- 参数个数
- 单参数section

数字化特征

205	3	34.5	143234
285	68	296	7
13850	157	11218	847
1.23e+9	422	1004	177
0	398	13.333	125

数值型特征

0	0	0	0
0	0	0	1
0	1	0	0
0	1	1	0
0	1	0	0

布尔特征

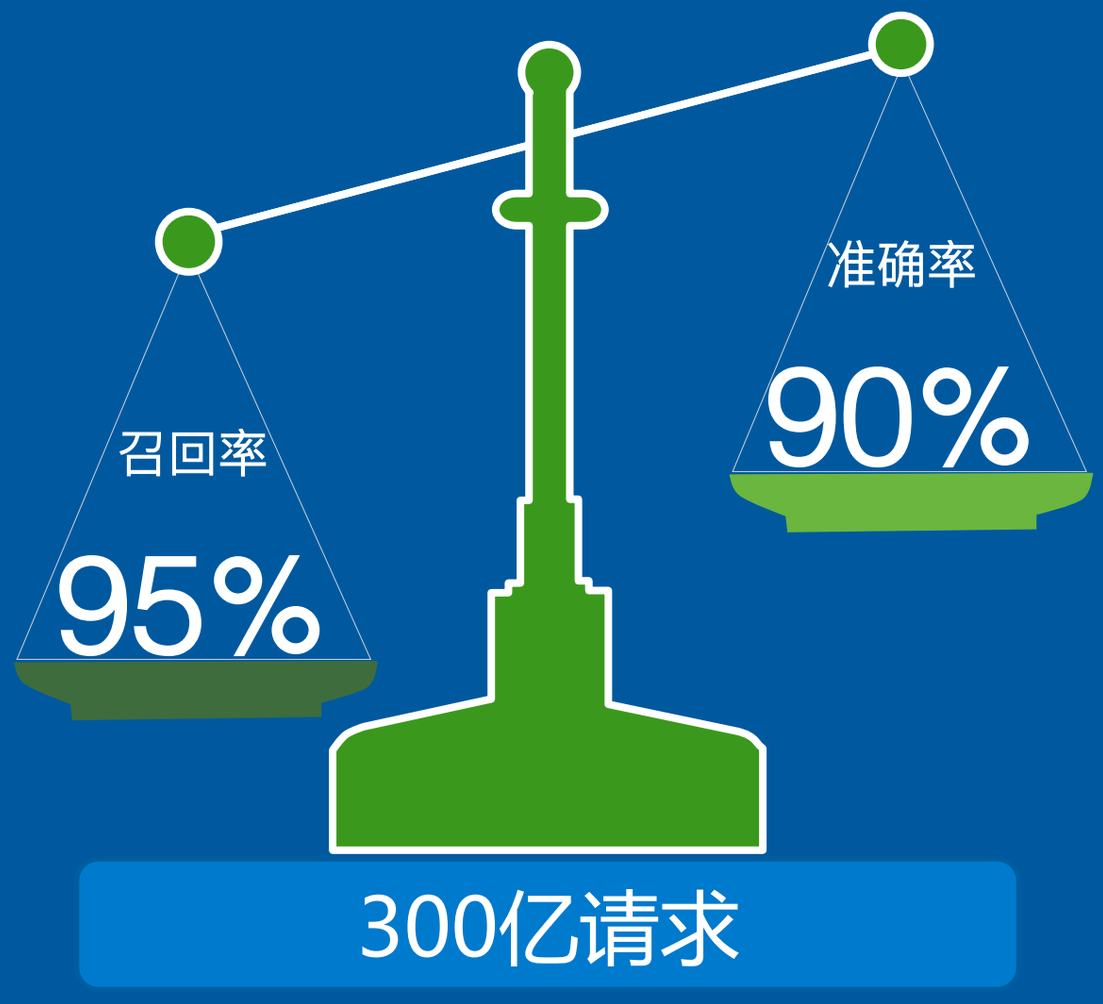
见证奇迹的时刻

一些奇怪的发现

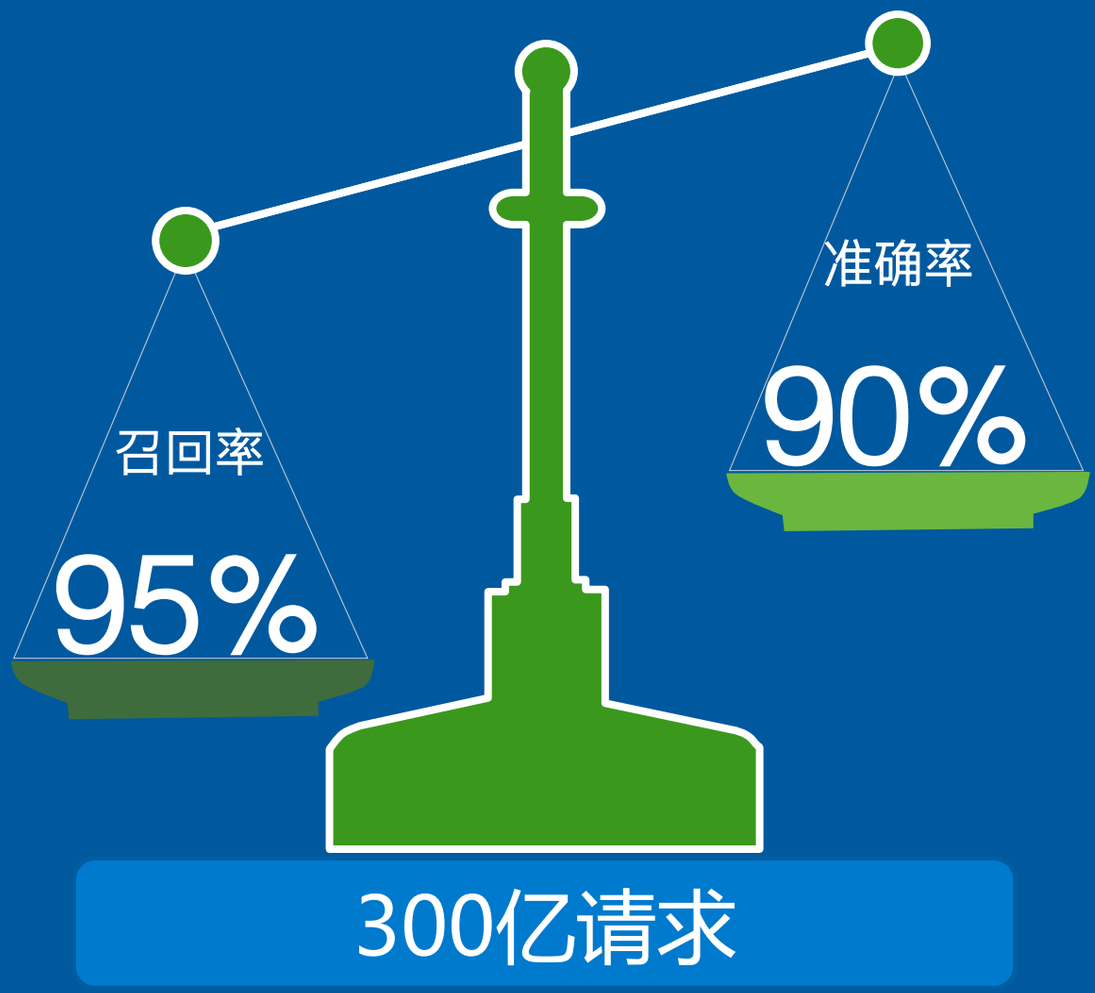
```
POST  /index/index.php?_c=zip://d://KAS/WebSource/ueditor/php/upload/file
/20170531/1496216087803962.zip#xxx&_m=captcha
cmd=echo "\n\n\n", system("dir C:");exit;

%2527!=(hex(user())>0x23)%2523
```

通过不断调整特征，对于变形与绕过有了神奇的抵抗能力，但是准确率却无法提升

如果我们在结合Response呢？



见证奇迹的时刻



威力不止如此

如果机器学习只做文本特征检测，
不能称之为人工智能

威胁特征全貌

文本特征

用户身份特征报

访问行为特征の人机识别

业务行为特征

IP信息

代理

IDC

设备指纹

虚拟

伪造

黑白名单

肉鸡

历史

威力不止如此

如果机器学习只做文本特征检测，
不能称之为人工智能

威胁特征全貌

文本特征

用户身份特征报

访问行为特征の人机识别

业务行为特征

请求频率

停留

间隔

访问目的

质量

威力不止如此

如果机器学习只做文本特征检测，
不能称之为人工智能

威胁特征全貌

文本特征

用户身份特征报

访问行为特征の人机识别

业务行为特征

访问时序

关联

依赖

业务习惯

热点

异常

用户行为分析-电商案例

	A0	A1	A2	A3	A4	An
API	登录	浏览	搜索	购物车	抽奖	结算
Session	{A2,A1,A4,A8}	{A2,A1,A4,A4}	{A1,A1,A4,A8}	{A0,A1,A4,A8}		

无监督学习
K-means

产出	盗号	僵尸	恶意爬虫	未知威胁

用户行为分析-难点

业务抽象

通过n-gram算法，产生业务pattern,分析URL，将请求归类，实现业务抽象

不判别好坏，只寻找少数派，相信大多数用户都是正常业务

去噪

去除网络、浏览器等干扰，将Session中所有业务向量化

因为无法识别异常类型，还需要人工介入和辅助模型识别

关系向量化

每Session的API集合，交集

异常识别的准确率高达95%

算法

如何选择K值，还要考虑到的向量集合的方差

总结

- 有监督学习，有效降低规则维护工作量，但对于准召相比语法引擎没有突破
- 在样本空间扩大之后，DNN相比SVM能有效提高召回率，但更多的应用在离线场景
- UBA可以解决当前技术在高维空间上的不足，是安全的对抗的下一个风口
- 无监督学习是未来，能突破样本空间限制

THANKS!

让创新技术推动社会进步

HELP TO BUILD A BETTER SOCIETY WITH
INNOVATIVE TECHNOLOGIES

Geekbang.
极客邦科技

InfoQ

专注中高端技术人员的技术媒体

EGO EXTRA GEEKS' ORGANIZATION NETWORKS

高端技术人员学习型社交平台

StuQ 斯达克学院

实践驱动的 IT 教育平台

地址：北京市朝阳区洛娃大厦C座8层1801室　　网址：www.geekbang.org