项目初步实施路径

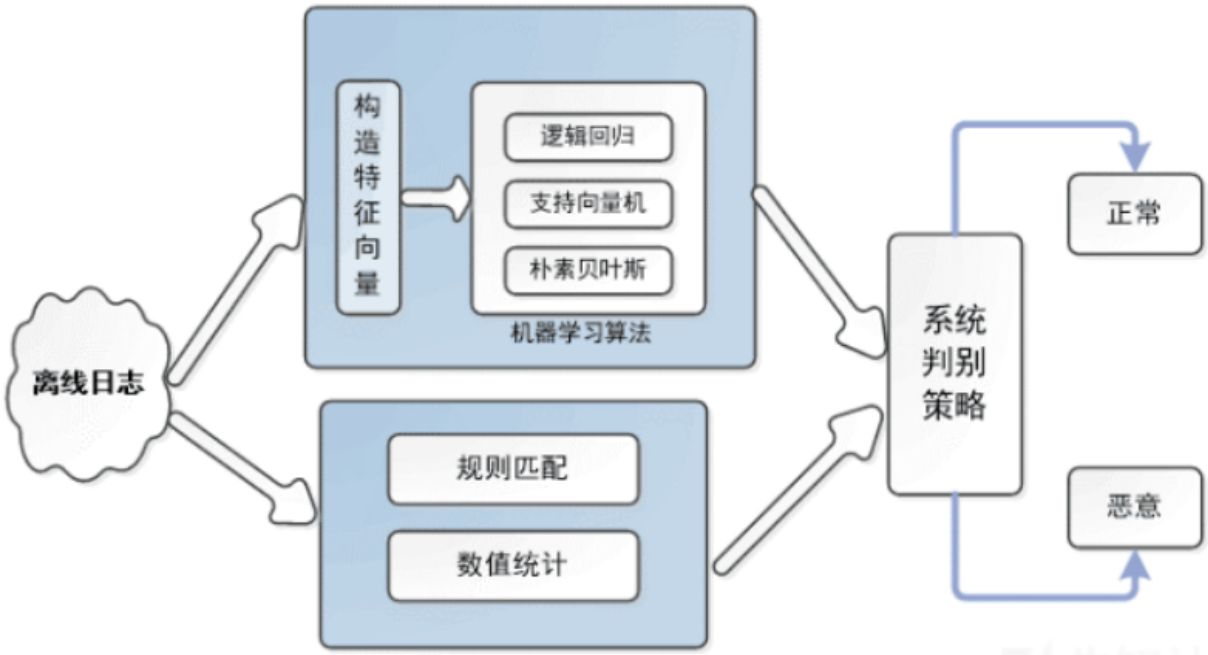
一、了解基础知识

第一周期项目所使用的方法也可作为此次项目的方法之一。

本次新增知识：统计学习之逻辑回归、支持向量机、朴素贝叶斯算法。

系统使用了三种机器学习算法进行恶意攻击的识别：逻辑回归、支持向量机和朴素贝叶斯，同时包含了传统的正则匹配算法，此外还应建立统计模块进行恶意IP访问的数值统计。

对**正则匹配、数值统计和机器学习**（三种机器学习算法两两取交集，即进行投票机制，三者中两者检测到异常则认为异常）进行处理，得出一条日志的识别结果，然后将结果存储到数据库中，结构图如图：



二、日志预处理及攻击分类

1、分析日志

利用正则表达式或常见安全产品（参考《Web日志安全分析浅谈》）提取日志内各字段信息，如请求资源、IP、时间、User-Agent等。

日志的识别主要是针对日志记录中的请求资源、referer、user-agent进行分析。

> 日志样本可在此寻找（http://www.secrepo.com/、https://github.com/foospidy/payloads）

2、攻击分类

(1)正则匹配

利用网络上的和一些CMS厂商的正则代码，对已知的Web攻击进行匹配。

> https://xz.aliyun.com/t/2136#toc-14

(2)数值统计

在海量的日志文本中通过一些潜在异常行为对目标进行统计，进而对可能存在攻击行为的IP进行相应处理。

> https://xz.aliyun.com/t/2136#toc-14

(3)机器学习

使用机器学习算法的前提是构造好特征向量，主要针对日志记录中的请求资源、referer和user-agent进行处理（原文作者举了这三部分，其实我认为还存在请求方法，因为有时对目标的bypass利用寻常的请求方式会被拦截，而使用畸形的请求比如LOL则可以绕过WAF的检测，这也可以作为训练之一，不过这一点应该补充在正则匹配中）。

三、机器学习模型搭建

构造向量与模型搭建

（1）基于统计学习模型（逻辑回归）

拿到正常请求和恶意请求的数据集，对其进行处理得到特征矩阵，使用逻辑回归方式对特征矩阵建立训练模型，最后计算模型的精准度，使用检测模型判断未知请求是恶意还是正常。

> https://www.freebuf.com/articles/network/131279.html

（2）基于文本分析的机器学习模型（HMM）

对文本序列模式的建模，相比较数值特征而言，更加准确可靠。其中，比较成功的应用是基于隐马尔科夫模型(HMM)的序列建模。

也就是在上一周期使用的方法，对数值进行泛化，对每个状态统计之后一个状态的概率分布。

利用模型判断输入序列是否符合样本模式，通过合适的阈值来进行异常识别。

（3）基于单分类模型（SVM）

在二分类问题中，由于我们只有大量白样本，可以考虑通过单分类模型，学习单类样本的最小边界，边界之外的则识别为异常。

这类方法中，比较成功的应用是单类支持向量机(one-class SVM)。

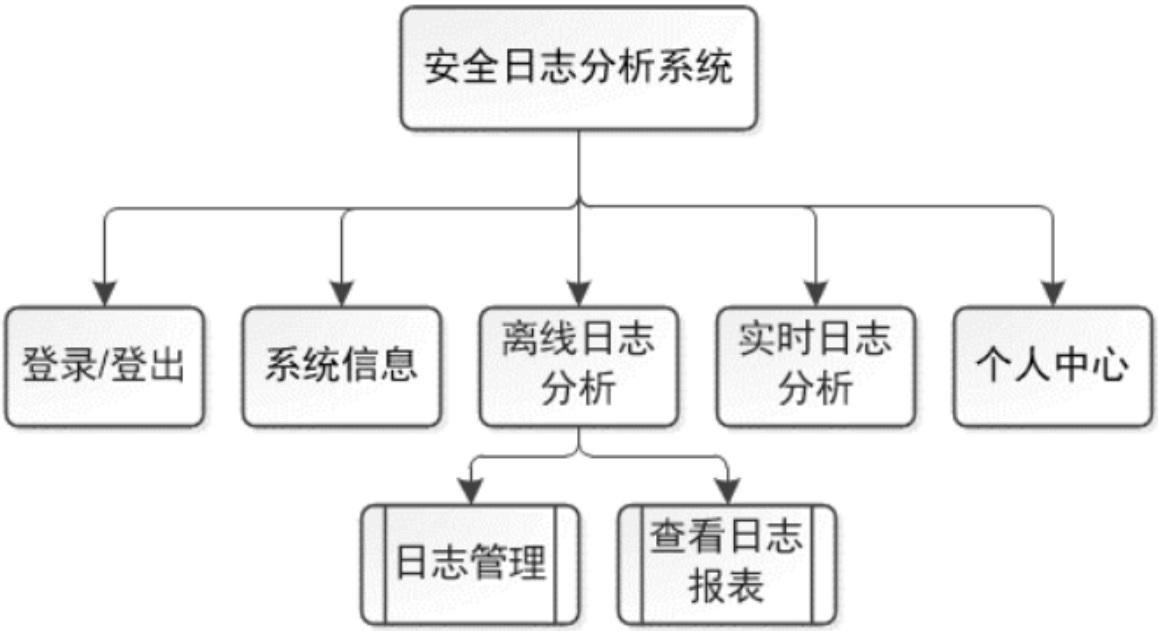
（4）基于朴素贝叶斯分类器模型

通过朴素贝叶斯分类器对数据进行训练和测试，依此对未知的Web请求进行判断来检测其Web攻击类型。检测主要包括：特征选择与提取，朴素贝叶斯训练阶段以及分类阶段.

> http://www.doc88.com/p-9945245389424.html

四、系统展示（可选）

系统应包括系统监控、用户管理（系统使用人员）、日志管理、实时分析、离线分析等功能，并为用户提供可视化的操作、分析与结果展示界面。



参考文章

Web日志安全分析浅谈：https://xz.aliyun.com/t/1121?accounttraceid=9ef7efd4-0316-406a-9129-88852da08abc

Web日志安全分析系统实践：https://xz.aliyun.com/t/2136#toc-2

用机器学习玩转恶意URL检测：https://www.freebuf.com/articles/network/131279.html

基于机器学习的Web异常检测：https://www.freebuf.com/articles/web/126543.html

基于大数据和机器学习的Web异常参数检测系统Demo实现：https://www.freebuf.com/articles/web/134334.html

我的日志分析之道：简单的Web日志分析脚本：https://www.freebuf.com/sectool/126698.html

基于朴素贝叶斯的Web攻击检测：http://www.doc88.com/p-9945245389424.html