项目初步实施路径

1、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。

2、日志预处理及攻击分类

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》）提取日志内所需信息，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。

攻击分类

1、确认日志结构，在数据库中建立表来存储日志，并建立相应的字段，如请求时间、客户端IP、请求方法、请求资源等。

2、给Web攻击进行分类

3、建立攻击规则表对应不同的攻击类型

3、数据统计及模型搭建

构造向量与模型搭建

使用机器学习算法的前提是构造好的特征向量，日志的识别主要是针对日志记录中的request、referer和user-agent。这三部分是用户可控且可能注入payload的地方。

（1）基于统计学习模型

基于统计学习模型的方法，首先要对数据建立特征集，然后对每个特征进行统计建模。对于测试样本，首先计算每个特征的异常程度，再通过模型对异常值进行融合打分，作为最终异常检测判断依据。

（2）基于文本分析的机器学习模型

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。

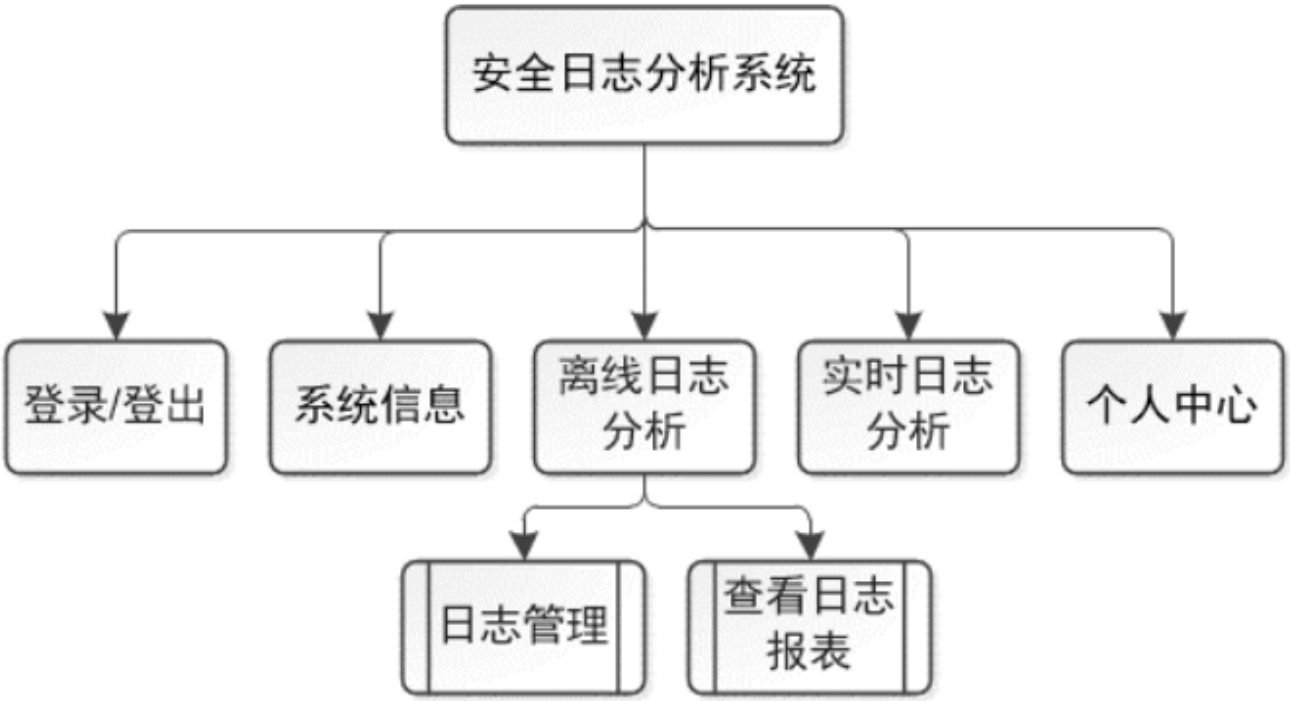
（3）基于单分类模型

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)。

5、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