海量日志的 web 应用异常入侵检测模型研究

龚锦红[1]　凌仕勇[2]

1.华东交通大学电气学院，江西 南昌 330013

2.华东交通大学网络信息中心，江西 南昌 330013

摘要：高校应用系统中的 web 日志数据是系统运维、安全分析的重要来源，文章基于 MapReduce 架构，结合属性长度、字符分布特征、属性域枚举的学习与检测模型，给出了一种海量数据入侵检测学习模型和检测算法。系统运行结果证明,该平台可以有效地发现校园网中的异常入侵,检索效率高,能有效提供运维效率和异常排查速度。

关键词：MapReduce；入侵检测；logstash

中图分类号：TP393.08　　**文献标识码：**A

Research on Web application anomaly intrusion detection model based on massive logs

Gong Jinhong[1]　Ling Shiyong[2]

1. School of electrical engineering, East China Jiaotong University, Jiangxi Nanchang 330013

2. Network information center of East China Jiaotong University, Jiangxi Nanchang 330013

Abstract: Web log data in university application system is an important source of system operation and maintenance and security analysis. Based on MapReduce architecture, combined with the learning and detection model of attribute length, character distribution characteristics and attribute domain enumeration, this paper presents a massive data intrusion detection learning model and detection algorithm. The system operation results show that the platform can effectively find abnormal intrusion in the campus network, has high retrieval efficiency, and can effectively provide operation and maintenance efficiency and abnormal troubleshooting speed.

Keywords: MapReduce; Intrusion detection; logstash

0 引言

随着高校教育信息化的发展，积累了大量的师生、教学、科研、管理方面的业务数据。而随着各业务系统的对外访问，网络安全问题日趋严重。目前，校园网安全运维主要是通过网络安全设备产品如防火墙、IDS、IPS 等设备来实现，总体效果不佳，一个重要的原因是忽视了日志在校园网管理中的作用。校园网中的网络产品、服务器、应用系统等软硬件运行过程中产生大量的日志，记录了系统运行、使用者、攻击者的访问行为，可以通过对这些日志的综合分析和处理，有效解决校园网运行中遇到的安全问题。

目前，校园网安全运维主要是通过网络安全产品如防火墙、IDS、IPS 等设备来实现，总体效果不佳，一个重要的原因是忽视了日志在校园网管理中的作用。校园网中的网络产品、服务器、应用系统等软硬件运行过程中产生大量的日志，记录了系统运行、使用者、攻击者的访问行为，可以通过对这些日志的综合分析和处理，有效解决校园网运行中遇到的安全问题。

Web 入侵检测是针对 web 应用的一种入侵检测技术，通过对 web 应用的请求分析，检测和识别 web 攻击行为。周勇禄[1]使用 web 日志中动态页面的参数值长度、字符分布等数据，建立了基于统一异常的检测模型。Estevez-Tapiador 等[2]对日志 URL 进行了划分，

对应到马尔科夫模型的不同状态，使用状态转移矩阵，根据模型达到终态的概率判断日志的合法性。

高校信息系统一般分散部署在各个服务器中，导致所产生的日志也比较分散，高凯[3]研究了大数据环境下，采用分布式数据流的四个子系统：数据采集子系统、消息处理子系统、流式计算子系统和数据存储子系统，用户大规模日志安全分析。陈付梅等[4-6]介绍了大规模系统的日志模式提炼算法的优化方法。文章主要针对数据中心产生的 Web 日志进行研究，基于 Hadoop 的 MapReduce 计算模型，开发出多种异常入侵学习模型和检测算法，对同一 Web 站点的不同日志事件进行联动挖掘分析，发现传统安全设备漏报或无法检测出的新型攻击事件，得出平台整体的安全态势，为数据中心正常运转提供安全保障。

1 异常检测模型

1.1 属性域枚举

属性域枚举指的是HTTP请求参数的值来源于一个特定的枚举集，如人员类型、职称、科研类型等等，当恶意用户试图使用这些参数传递非法参数值时就会被检测出异常。判断一个参数是枚举型还是随机型，只要判断参数的不同取值个数是不是在一定的数值范围内，但对于不同应用，这个枚举数量并不方便人为认定。故可以考虑，当某个参数的不同取值个数随着请求量的增加成比例增加时，此参数为随机型的，否则为枚举型的。

令 p 为请求中的一个参数，i 表示样本次数， 定义如下函数：

$$\begin{cases} f(i)=i & \\ g(i)=g(i)+1 & \text{如果p的第i次出的值是新的} \\ g(i)=g(i)-1 & \text{如果}p\text{的第}i\text{次出的值已存在} \\ g(i)=0 & i=0 \end{cases}$$

由定义可知，函数 f 严格递增，函数 g 只有在样本未出现才增加。同时，定义函数 f 和 g 的相关系数 ρ ：$\rho=\frac{\mathrm{cov}(f,g)}{\sqrt{\mathrm{var}(f)\mathrm{var}(g)}}$，当 f 随着样本增加而增加时，出现过的样本会导致 g 减少，即 ρ <0 时，f 和 g 负相关，此样本值可认定属于枚举类型。

1.2 属性长度

通常同一 web 应用请求的某一属性字节数变化不大，异常事件的请求参数长度比一般请求字节总数大，例如缓冲区溢出攻击。为此，Vigna[7]提出了一种属性值长度模型，以长度期望和方差作为检测标准，利用切比雪夫不等式计算给定参数值字符串异常概率。

在学习阶段，计算某一属性域的 HTTP 请求串长度 l_1， l_2,…， l_n，并计算期望和方差 σ^2。

在检测阶段，对某一属性值长度进行模型偏离程度评估。切比雪夫不等式给出了在分布未知情况下对事件发生的概率进行估计的一种方法。根据切比雪夫不等式可计算出字符串长度的概率：$P(|x-\mu|\geq t)\leq\frac{\sigma^2}{t^2}$，可知长度为 l 的属性值出现的概率为：$P(|x-\mu|\geq|l-\mu|)\leq P(l)=\frac{\sigma^2}{(l-\mu)^2}$

1.3 属性域字符分布

属性域字符分布模型通过查看属性域字符分布规律来判断请求是否异常。对于正常的 web 请求，属性域属于可打印字符，如果实施缓冲区的二进制字节攻击、或者目录遍历漏洞的同一目录字符如“./”等，都会展现一种完全不同的字符分布。

字符分布是一个由 6 个概率值组成的序列，字符串 s 的字符分布用 $CD(s)$ 表示，$CD(s)$ 计算过程如下：

（1）ASCII 映射：将参数属性域的每个字符对应一个 ASCII 码。例如”abcda”对应的 ASCII 码序列为“97,98,99,100,97”。

（2）概率序列：将 256 个 ASCII 码在 ASCII 序列中出现的概率降序排列，例如上述概率序列为“0.4,0.2，0.2,0.2，0，…”，后面为 252 个 0。

（3）序列分组：将概率序列依据组号分成 6 组，组号 1 包含字符序号 1，组号 2 包含字符序号 2-3，组号 3 包含字符序号 5-7，组号 4 包含字符序号 8-12，组号 5 包含字符序号 13-16，组号 6 包含字符序号 17-256。

（4）组内求和：将序列分组后的各个小组内概率求和，得到 6 个概率值，即为字符串的字符分布。如 ”abcda” 的字符分布 $CD(s)$ 为概率序列“0.4,0.6,0,0,0,0”。

在模型学习阶段，计算正常样本的理想字符分布 ICD 。设训练集为 $\{s_1,s_2,...,s_n\}$，其字符分布为 $CD(s)$ 中的第 i 个概率值为 $CD(s)_i$， ICD_i 表示 ICD 中的第 i 个概率值，定义为：$ICD_i=\frac{1}{n}\sum_{k=1}^{n}CD(s_k)_i$，其中 i =1,2,…,6。即 ICD 中的第 i 个概率值是样本集中所有样本分布的第 i 个概率值的均值。

在模型检测阶段，给定任意 ICD 和一个字符串 s，

则 s 出现概率由下列式子计算：$p(s)=\chi^2(\alpha,5)$，$\alpha=\sum_{i=1}^{6}\frac{(CD(s)_i-ICD_i)^2}{ICD_i}$。其中 χ^2 检测称为卡方检测[8]，用来衡量任意观测样本属于某给定分布的可能性，在此 α 为符合自由度为 5（ICD 的变量个数 6 减 1 得到）的 χ^2 分布。

1.4 混合模型评估

采用多个分类器组成的混合分类器，将多个决策综合以获取更好的分类[9]。本文采用四种静态融合函数：最大值、最小值、算术均值、几何均值进行分类融合，在满足评估效果的基础上以提高计算性能。混合分类器的输出是有效载荷属于正常数据的概率，如果待检测的数据概率高于预定义的阀值，那么该条数据判断为正常，否则，判断为异常。

$$\begin{cases} \text{最大值} & s_i^*=\max\{s_{ij}\} \\ \text{最小值} & s_i^*=\min\{s_{ij}\} \\ \text{均值} & s_i^*=\frac{1}{K}\sum_{j=1}^{K}s_{ij} \\ \text{几何均值} & s_i^*=\left[\prod_{j=1}^{K}s_{ij}\right]^{\frac{1}{K}} \end{cases}$$

2 异常入侵学习与检测模型的实现

整个系统分为模型学习阶段和异常检测阶段，均采用 Hadoop MapReduce 架构。在日志采集模块，logstash 将 nginx 获取到的日志，通过 webhdfs 插件抽取到 HDFS，并做到与 elasticsearch 的集成，后续的模型学习与检测均从 HDFS 获取数据；在模型学习阶段，根据本文的学习与检测模型，进行模型学习与增量更新，将模型全部持久化到数据库。在入侵检测阶段，根据检测模型，对输入数据进行分析，将分析结果储存，供 web 系统进行前端数据展示。

2.1 日志采集

校园网应用服务主要包括 web 服务、FTP 服务、域名服务、数据库服务等，每种应用在运行中产生各种重要事件的记录。根据不同日志源环境不同，需要采用不同的日志协议，传统方式需要根据不同场景采用 syslog 协议、SNMP 协议、文本方式采集等。Logstash 基于 C/S 架构，支持多种输入选择，其客户端部署到日志源设备中，监听日志数据增量发送到指定目标，如 redis、kafka、es 等。

对于 web 服务，校园网的日志主要有 IIS、apache、nginx。以 nginx 为例，在 nginx 所在的服务器上部署 logstash 服务，配置 logstash shipper，将 nginx 日志输出到 kafka。

为了更好的对日志进行分析，需要将原有 web 日志消息进行解析，将消息内容解析为各个独立指标，记录内容包括：远端 IP、请求时间、请求地址、协议、请求返回的状态、访问时间、访问时长、访问文档的大小等。为此，需要自定义正则表达式植入 logstash 的 grok-patterns 中，以支持消息的在线解析[10]。如针对 nginx 的访问日志，定义 gork 如下：NGINXACCESS %{IP:remote_ip} \- \[%{HTTPDATE:timestamp}\] ″%{WORD:method} % {WZ:request} HTTP/ %{NUMBER:httpversion}″ % {NUMBER:status} % {NUMBER:bytes} % {NUMBER:bytes} % {NUMBER:bytes} % {QS:request_body} % {QS:agent} %{QS:referer} %{QS:xforward}

在日志采集模块，采用 logstash 从多个数据源获取数据，并对这些数据进行转换和过滤， logstash shipper 支持各种数据源的 web 日志，多个不同的 shipper 抽取不同数据源的各种数据，缓存到统一的 redis 模块。logstash indexer 从 redis 模块获取数据，结合 webhdfs 插件，将数据存储到 HDFS，同时推送到 elasticsearch。

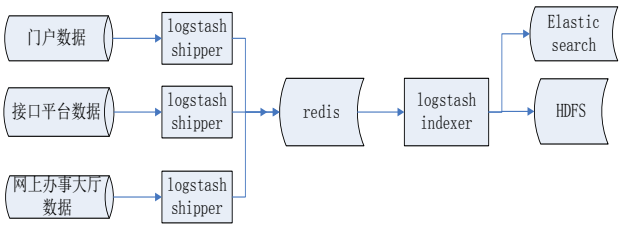

图 1 日志采集流程

2.2 模型学习

模型学习采用 Hadoop MapReduce[10-11]架构，含一组 map()和 reduce()。

map()阶段对每一条 HTTP 请求日志进行数据解析，解析出 GET 和 POST 请求的参数和参数值。首先，根据请求 url 解析请求函数，如果请求 url 串被加密，需要对 url 串进行解密。对于 GET 请求，根据请求 url 解析出所有的请求参数和参数值。对于 POST 请求，构造通用的 jsonToMap 函数，解析请求体中所有的请求参数和参数值。对每个请求参数和参数值，输出“systemLabel|请求函数|参数”为 key，参数值为 value。

reduce()阶段对 map()输出的参数和参数值，构建

请求参数枚举型模型、请求参数值的长度分布模型、请求参数值的字符分布模型。在模型学习的过程中，系统根据检测模型设定的阀值进行异常判断，若为正常请求则进行模型更新，将学习到的模型全部持久化到数据库。

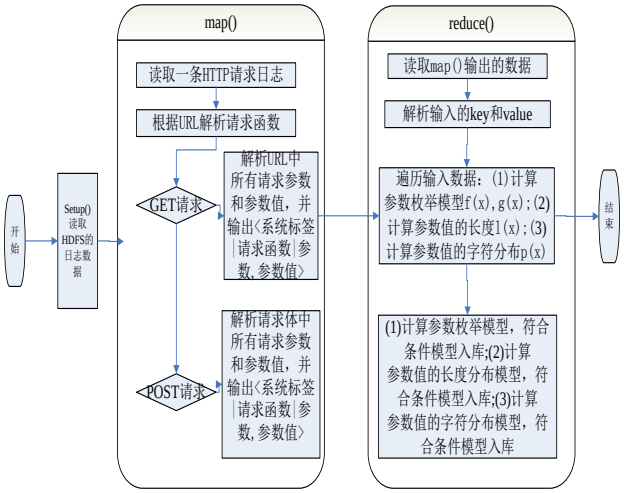

图 2　模型学习算法

2.3 入侵检测

入侵检测同样采用 Hadoop MapReduce 架构，含一组 map()和 reduce()。

map()阶段对每一条 HTTP 请求日志进行数据解析，解析出系统标签 systemLabel、请求所在的日期 day、以及客户 ip、组合构成输出的 key，请求串为输出的 value，供 reduce()进行详细数据分析用。

reduce()阶段对 map()输出的数据，经定义在 logstash shipper 端的，基于插件开发的安全分析专家规则[12]，以及检测模型进行检测分析，经混合模型评估后，给出分析结果，存储到数据库中，供 web 前端进行统计分析与查询。

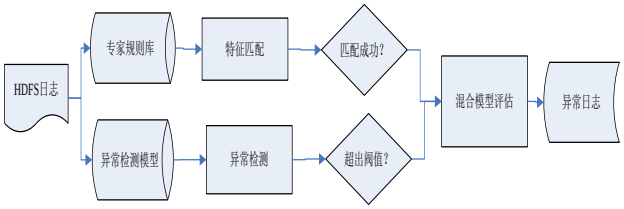

图 3　异常检测流程

3 实验与结论

本文数据全部来源于本校校园网络真实的数据，为获取更多的异常数据，特选取省网安中心组织的全省教育系统网站安全评估时间段的访问数据。系统接收门户、一站式网上办事大厅、接口服务平台等系统采集到的 nginx 日志数据，通过 logstash 将数据存储到 HDFS。系统从 HDFS 读取数据，进行模型的增量学习和异常检测。同时在模型学习阶段，通过部署在核心网络端的态势感知安全系统，通过接口方式获取入侵二次确认，提高了模型学习的准确性。

在模型学习阶段，系统目前共采集三个应用系统的 4105931 条 HTTP 请求，学习到 76 个请求函数、221 个请求参数、5929 个参数值的属性域枚举模型。以及均为 382 个请求函数，2453 条记录的属性长度模型和属性域字符分布模型中。

为了对模型进行验证，从态势感知安全系统中抽取部分已经确认的不同攻击类型的数据。图 4 显示了属性长度模型在阀值为 3 倍标准方差，字符分布模型的卡方置信度设为 90%，卡方检验阀值为 9.236 时，两者检测结果的 ROC 曲线。误报率约为 0.02，检测率达到 0.9 以上。

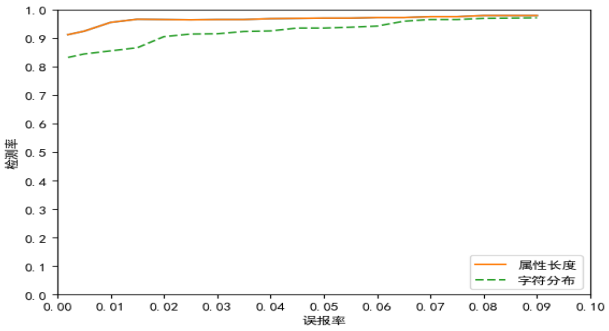

图 4　ROC 曲线（属性长度、字符分布）

表 1 显示了不同攻击的检测率。结果显示，因缓冲区溢出的字符输入长度一般明显异常于正常请求，属性长度模型对此基本可以做到完全识别。路径遍历攻击的字符分布明显呈现异常的差异，故字符分布识别对此比较敏感。而 SQL 注入在属性长度和字符分布的两个模型中均有所体现。而对于 XSS 和命令识别，本文主要以专家静态规则为主，故识别率有待提高。对于参数篡改，基于参数值的属性域枚举模型在这方面的识别还不能做到完全识别，需要借助于其它的一些模型如基于用户行为的马尔科夫模型做进一步的分析。

表 1　不同攻击的检测率

攻击类型	正常请求	异常请求	检测率
XSS	152	8455	0.9820
SQL 注入	0	3325	1.0
命令执行	8	346	0.9768
路径遍历	0	220	1.0
缓冲区溢出	0	77	1.0
参数篡改	42	1342	0.9687
总计	202	13765	0.9853

图 5 为系统给出了一周内异常入侵检测的攻击数

据在线统计。

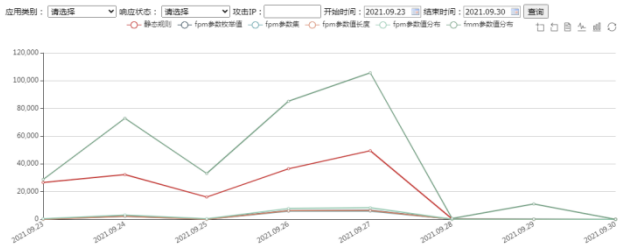

图 5　异常入侵统计

4 结语

校园网 web 日志数据对业务系统安全运维至关重要，文章基于 MapReduce 架构，结合属性长度、字符分布特征、属性域枚举的异常入侵学习与检测模型，给出了一种海量数据入侵学习与检测算法，构建了一整套数据采集、数据分发与传递、安全分析、数据存储与可视化呈现的校园网 web 安全分析平台，能较好地检测到系统访问状态、发现异常访问行为，并通过对访问数据的分析，找出系统可能存在的漏洞，为系统安全正常运维提供良好的基础运维。

本系统目前仅采集部分 web 日志数据，下一步将着力集成更多的业务系统数据，并将服务器系统日志、交换机和安全设备日志进行数据集成，结合多个数据源的数据进行联动分析，通过一系列攻击模式挖掘算法，做好校园网安全实时监控和风险预警。

参考文献

[1]周勇禄,吴海燕,蒋东兴.基于统一异常的 Web 应用入侵检测模型研究[J]. 计算机安全,2012,12(5):8-12.

[2]Estevez-Tapiador J M,Garcia-Teodoro P,Diaz-Verdejo J E.Detection of web-based attacks through markovian protocol parsing[C]//Proceedings.10th IEEE Symposium on Computers and Communications,2005(ISCC 2005).Los Alamitos:IEEE,2005:457-462.

[3]高凯.大数据索引与日志挖掘及可视化方案-ELK Stack:Elasticsearch、Logstash、Kibana[M].第 2 版.北京:清华大学出版社,2016.

[4]陈付梅,韩德志,毕坤,等.大数据环境下的分布式数据量处理关键技术探讨[J]. 计算机应用,2017,37(3): 620-627.

[5]赵一宁,肖海力.对于大规模系统日志的日志模式提炼算法的优化[J].计算机工程与科学,2017,39(5):821-828.

[6]姚攀,马玉鹏,徐春香.基于 ELK 的日志分析系统研究及应用[J]. 计算机技术与发展,2018,39(7):2090-2095.

[7]Mahoney M V and Chan P K.Learning nonstationary models of normal network traffic for detecting novel attacks.In Porc. of the 8th ACM SIGKDD international Conference[J]. on KNowledge Discovery and Data Mining,2002(1):376-385.

[8]Ester M,Kriegel H,Sander J and Xu X,A density-based algorithm for discovering clusters in large spatical databases with noise. In proc[J].of the 2nd International Conference on KNowledge Discovery and Data Mining,1996(1):226-231.

[9]I.Corona,G.Giacinto,C.Mazzarielo,F.Roli,C.Sansone,Information fusion for computer security:State[J].of the art and open issues,Information Fusion,2009(10):274-284.

[10]马文,朱志祥,吴晨,万一.基于 FP-Growth 算法的安全日志分析系统[J].电子科技,2016,29(9):94-97.

[11]舒远仲,戴海辉,吴小玲. FP-Growth 算法在 MapReduce 框架下的实现[J]. 软件导刊,2017,16(8):25-28.

[12] Jinhong Gong,Shiyong Ling. Security analysis of University web log based on elk[J].8th Annual International Conference-Information Technology & Applications,2021(1):491-498.

作者简介：龚锦红（1985—），女，汉族，江西南昌人，硕士研究生学历，讲师，华东交通大学电气学院。
通讯作者：凌仕勇。