文献标识码：B 文章编号：1003-0492（2018）06-0076-05 中图分类号：TP274

基于网络日志进行大数据分析的安全感知

Network Log Analysis based Big Data Security Perception

★向永谦，李欣，满建文（62101部队，湖北 武汉 430010）

摘要：基于网络日志的大数据分析感知可以从技术上有效通过网络安全应用进行安全事件发生前的迹象捕捉，从而进行预防与设置陷阱，对安全进行有效防护。

关键词：大数据分析；网络日志；安全感知；陷阱捕捉；Hadoop；Spark

Abstract: Network log analysis based big data perception can be technically effective for capturing security incidents. Built upon network security, it is effective for safety protection and can be implemented to prevent and set the trap.

Key words: Big data analysis; Web logs; Safety awareness; Trap; Hadoop; Spark

1 引言

大数据时代，万物互联，网络的安全访问形势日趋严峻，根据2016年国际电信联盟（ITU）的数据统计，每天新增恶意软件20000个，网络犯罪的受害者人数达到5.56亿，直接净损失达1100亿美元。2017年5月12日的勒索病毒，据《华尔街日报》报道，硅谷网络风险建模公司Cyence的首席技术官George Ng称，此次网络攻击造成的全球电脑死机直接成本总计约80亿美元（约合人民币550亿元），这些安全事件的背后除了人为的恶意攻击外，也反应了网络安全应对策略及预测感知能力需要进一步提升。基于大数据分析的安全感知，尤其是基于网络设备日志方面的实时分析、进行安全感知、提前预测可能发生的安全问题，具有很高可行性。

而有网络就会有网络设备，有网络设备就会产生设备日志，这些日志正在被浩瀚的信息流所蒙蔽，不能很好地发挥作用。如此庞大的廉价海量数据源，急切需要大数据技术来进行挖掘，一旦得以应用将对网络安全态势的感知及有效防护产生极大的影响。本文正是着眼于大数据的数据挖掘、分析技术，利用网络日志进行安全模式识别，陷阱（Trap）捕捉，提出安全感知的一种新思路。

2 安全威胁的分析

当前的网络安全的形势严峻，威胁来源主要来自于通过网络设备的访问、通过存储设备进行的网络存储、基于运算设备的网络协同运算等，而这些网络访问的客户端多以软件应用方式，包括浏览器、传统的使用网络的软件、云虚拟化类软件、云端应用等。

这些威胁多是利用病毒感染、黑客攻击等呈现不同的形式，如非授权访问，信息泄露和丢失，网络基础设施传输过程中破坏的数据完整性，拒绝服务攻击，进一步网络病毒的传播。有的威胁在服务器不停机扩容时支

持热插拔而接入介质感染，有的在复杂多样的存储要求中而暴露漏洞，有的在前端程序的 SQL/NoSQL注入中产生，而常规扫描无法完成海量数据处理，实时性上也不能达到即时报警，这些安全威胁一旦变成了真正的攻击，则后面的任何侦测、采取的措施，都将成为被动的防御而已，如果能从威胁存在时进行感知，提前进行分析、模式识别、预警，则在安全防护上更有意义。

按照Gartner的定义——“大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产”，它具有海量数据规模、快速数据流转、多样数据类型以及价值密度低四大特点。对于网络访问过程中的日志，应用传统技术很难解决信息的多样化、海量、快速增长带来的实时分析难题，而大数据相关的Hadoop，Stream，Spark技术等正适合应用与日志分析的诸多方面。

3 安全威胁的大数据感知

基于网络日志，进行大数据分析感知，主要是基于网络日志进行大数据技术应用，并采取模式识别分析感知，总体思路基于大数据的Hadoop及Flume对日志类非结构化数据进行分类过滤，进行文本语义编码、融合全息数据进行模式识别，进行分析结果可视化输出，同时对于报警威胁类提前预测触发相关保护措施。

网络日志对于安全中正常的信息流来讲属于非结构化的杂乱数据，数据受限于设备本身的能力、应用的设置、存储的限制等，基于前端计算在前端服务器进行实时分析同时进行日志的压缩上传，利用后端云平台，进行文本语义编码校对、安全模式查询与分析，尤为适用于连续性的跟踪或同类模式主题的安全信息流追踪（数据量大、语义同质化比较高）。通过动态地调用不同的日志语义分类算法对非结构化数据进行分类，提高安全预测态势的性能和效率，从而大大提高了安全模式判别的稳定性和可靠性。

考虑现有人工编码分析技术分析速度慢，很难支持数据量较大的信息挖掘，现有的统计分析技术，对数据形式要求高，对日志文本类行文数据分析效果差，无法结合语境对语义进行分析与快速查询，且无法应对非结构化的日志数据。当网络日志数据量每天以10G、100G增长的时候，单机处理能力已经不能满足需求。我们就需要增加多台服务器，用计算机集群、大数据Hadoop技术来解决。Hadoop的出现，大幅度降低了海量数据处理的门槛， Hadoop非常适用于日志分析。有了大数据环境后，经过的分类语义过滤后，结合MapReduce技术进行自动编码，进行大数据Hadoop的Mapreduce变量定义示例如下：

Map过程{key:$request,value:1000} ——表示需要跟踪的自动编码开始

Reduce过程{key:$request,value:特征值求和(sum)}——自动编码的各项每10分钟、30分钟（可自定义）总和

Map：{key:$request,value:$remote_addr}——独立IP的访问量数量

Reduce：{key:$request,value:去重再求和(sum(unique))}

Map：{key:$time_local,value:1} ——Time：用户每小时访问量的数量

Reduce：{key:$time_local,value:求和(sum)}

Map：{key:$http_referer,value:1} ——Source：用户来源地址的挖掘

Reduce：{key:$http_referer,value:求和(sum)}

Map：{key:$http_user_agent,value:1}——agent：用户的访问设备或代理信息挖掘

Reduce：{key:$http_user_agent,value:求和(sum)}

基于以上但不局限于以上变量的定义，可以得出每个用户在访问期间的各个行为、特征的不同自定义周期的总和，这块总和可以基于通常网络访问的情况设立规则、权重，并进行机器自学习，发现特征异常值，立即触发模式识别进行陷阱捕捉后，给予相关的系列反应。

大数据日志分析系统架构如图1所示。

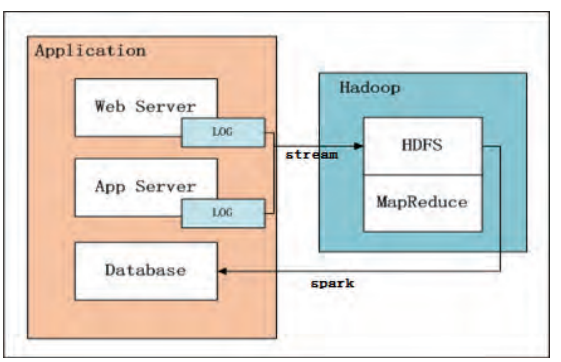

图1 大数据日志分析系统架构

图1中，左边Application是日志分析系统，右边Hadoop的HDFS用来做日志分析的文件存储，MapReduce进行上面的特征数值计算。主要思路如下：

日志是由设备系统、网络系统或业务系统产生的，我们可以设置网络服务器每天产生一个新的目录，目录下面会产生多个日志文件，每个日志文件64M。

设置系统定时CRON或者对于实时性高的访问日志使用Stream技术，向HDFS导入的日志文件。

完成增量导入后，启动MapReduce程序，提取并计算统计特征指标。

完成MR计算或Spark实时计算后，从HDFS导出统计指标结果数据到数据库的同时进行相应的模式识别如果发现威胁存在，则触发报警。

其中基于大数据的网络日志的模式分类分析过程的主要示意如图2所示。

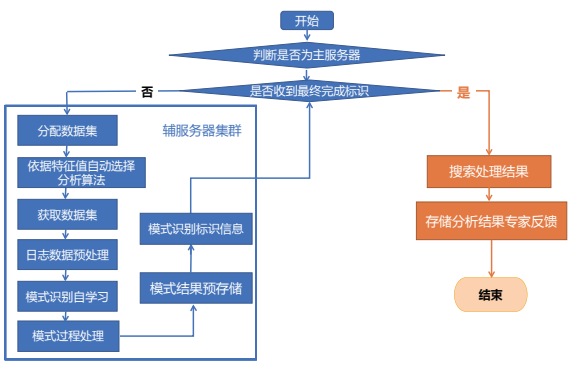

图2 模式分类分析过程

（1）主服务器对训练或者需要模式识别分类的日志集进行划分，分类包括安全等级、设备属性、IP地址、频度等类别，比如同一IP，在某个标准下快速连续访问将为模式分类提供很好的组合识别。

（2）主服务器选择本次分类所采用的具体日志分析算法，把该算法命令传递给抽象算法程序，此算法主要是下面提到的路径分析、协同过滤分析等算法，具体见下文。

（3）主服务器等待接收各个处理器发送模式分类完成标识符，将分类识别后的运算结果最终通过模糊算法进行标识。

（4）完成标识后，各单独服务器获取来自主服务器分配的日志数据集，进行进一步计算识别。

（5）对获取的日志数据集进行预处理，把数据集划分成Mapreduce形式如上面的< key ,value> 的格式。

（6）创建具体算法的一个实例，调用该算法实现的map函数、reduce函数对数据集进行处理。对于需要调用实时计算的，利用Hadoop stream及Spark进行实时数据传输及计算。

（7）处理完成后对处理后的数据进行模式的预判、概率值统计等并结合机器自学习机制进行自学习训练，最后把处理结果存储在本地文件系统。

（8）发送执行完成标识符到主服务器。

（9）主服务器接收各个服务器的完成标识符，如果所有的服务器都发送了标识符，则进行下一步处理，若还有未完成的处理器，则主服务器继续等待。

（10）主服务器到各个处理器搜集处理后的结果集，然后再在本地系统对结果集进行一次类别预判，得到最终的模式识别结果，把识别结果存储到数据库服务器的同时提供专家干预的机制，专家可以通过设置权重、打分来提升机器学习中各个参数的精准性。

上面谈到的过程中，基于网络日志进行大数据分析具体的涉及的分析算法，主要应用如下：

3.1 路径分析

日志分析中，我们将应用路径分析的算法来跟踪用户的访问行为，这个路径可以被用于判定在一个网络访问中最频繁访问的路径，还有一些其它的有关路径的信息通过关键路径分析中可以得出。路径分析可以用来确定该用户的频繁访问路径，从而调整和优化安全访问的策略，使得用户访问在策略下更加规范，还可以根据用户典型的操作访问模式用于陷阱捕捉和有针对性的安全报警，同时融入群体特征下的路径分析，对于同一类或同一工作职责的相关人，访问路径的差异性也体现了不同的安全动机。例如：80%的SQL注入都是通过网页的脚本进行的，这些脚本会在网络访问日志中去直接访问数据库，这种跳过应用层面直接访问的方式经过的必然路径就是我们提前在访问数据库之间进行捕捉的关键路径，在实践中很有效果。而将海量路径混合在一起进行大数据分析，可以发现趋势的共性和特殊性，为共性路径及特殊路径提供了很强的预测能力。

3.2 关联规则分析

使用关联规则的发现方法，可以从网络日志的访问事务中找到的相关性。关联规则是寻找在同一个事件中出现的不同项的相关性，用数学模型来描述关联规则发现的问题。在日志分析中，利用的关联规则初步思路

如下：

设x=>y的蕴含式，其中x,y为属性——值对集（或称为项目集），且X∩Y空集。在数据库中若S%的包含属性——值对集X的事务也包含属性——值集Y，则关联规则X=>Y的置信度为C%，则C%在一定安全区间时，则说明安全访问是在正常范围内，越出此空间则触发模式识别的相关流程。在关联规则时，通过海量访问信息的日志进行无序规则关联，最终识别出安全区间的大数据推荐区间，同时采用打分、权重的专家评估共同干预的方式，并加以机器学习，通过大数据对模型训练逐步提升关联规则的准确性。

3.3 序列模式

在有时间戳日志的有序事务集中，序列模式的发现那些如“一些项跟随另一个项”这样的内部事务模式，能结合应用发现安全访问数据中如“在某一段时间内，用户导出数据A，接着导出数据B，尔后又导入数据C，即序列A→B→C出现的连续性”之类的信息。序列模式可以描述在给定的日志访问序列数据库中，每个序列按照访问日志的时间排列的一组数据集，通过挖掘序列函数，返回该数据库中高频率出现的序列进行安全模式识别，这个模式基于时间戳，也可以基于设定的某种特殊序列，如访问设备的先后顺序，如果某个用户跳过了某个设备直接访问，则出现了安全问题，经试验，这种模式在大数据安全分析中很有价值。

3.4 分类分析

日志中利用大数据的分类规则可以给出识别一个特殊网络群体的公共属性的描述，这种描述可以用于分类访问者的属性。分类包含的挖掘技术将找出一个项或事件是否属于安全数据中某特定子集或类的规则。分类算法可以采用决策树方法、神经元网络、Bayesian分类等，最终分析出同一类群体、或者某个固定团体在网络访问中的不同公共属性。此分类规则是分层次的，不是同一层次的，试验中，在不同层级运算不同的分类，最终基于大类、中类、小类得出分类的特殊标识符。

3.5 聚类分析

可以从网络日志访问信息数据中聚类出具有相似特性的访问者。在网络日志事务日志中，聚类访问者信息或数据项能够将群体与其职责进行匹配，是基于大数据挖掘出安全事件中的内鬼模式。

日志聚类分析将日志数据集划分为多个类，使得在同一类中的数据之间有较高的相似度，而在不同类中的数据差别尽可能大。在日志聚类技术中，没有预先定义好的类别和训练样本存在，所有日志记录都根据彼此相似程度来加以归类。主要应用算法k—means、DBSCAN，通过把具有相似特征的访问用户或数据项归类,在网络日志管理中通过聚类具有相似操作行为的用户，分析有2种：基于模糊理论的网络日志页面聚类算法分析或群体聚类算法的模糊聚类分析。比如客户访问情况可用access(Uj)表示（U代表用户）。对于聚合分析的用户访问j项结果Suj，有Suj={(Ci，fSuj(Ci))|Ci∈C}，其中fSuj(Ci)→[0，1]是客户Ci和URL(Uj)间的关联度，C为当前访问客户，i为客户的数量，hits(Ci)表示客户Ci访问access(Uj)的次数。利用Suj和模糊理论中的相似度度量Sfij定义建立模糊相似矩阵，再根据相似类[Xi]R的定义构造相似类，合并相似类中的公共元素得到的等价类即得出相关网络日志的聚类情况，此种分析需要建立的模型稍有复杂，在此论文中因篇幅及题目所限不能详述。

3.6 基于大数据统计挖掘

基于大数据统计挖掘方法是从网络日志中抽取知识，通过分析会话文件，对浏览时间、浏览路径等进行频度、平均值等统计挖掘分析的同时，结果应用到机器自学习的相关参数中，可用于改进网络日志的结构配置，增强系统安全性，提高网络日志访问的侦查性等。

3.7 协同过滤

利用大数据采用最近邻技术，利用访问用户的历史、常用的访问路径计算用户不同访问模式之间的距离，目标用户对Trap的喜好程度也将形成黑客特点识别的一个特征。

3.8 安全感知模式分析并进行Trap校验

基于以上的分类模式挖掘、聚类模式挖掘、时间序列模式挖掘、序列模式挖掘、关联规则等，对原始日志数据进行进一步分析，找出用户的网络访问规律，即用户的通常访问模式及其他用户的模式，并做可视化安全感知画像，为安全的策略规划及日志进一步分析的决策提供具体依据。主要方法不仅要使用大数据Hadoop的HDFS对原始日志进行存储，Flume技术进行日志的导入，Spark进行日志的实时分析，还要结合传统的基于SQL查询分析，因为以前历史数据的分析结果将在传统数据库mysql中保存，这样画像较为快速，或者用

OLAP工具进行分析并给出可视化的结果输出。对于威胁程度高的直接触发报警模块进行报警，并直接阻挡继续访问。

结合网络日志，基于大数据的安全感知，可以在网络监听程序中直接应用此感知结果作为监听安全的必要手段，并由此设立陷阱（Trap），当发现用户有黑客倾向时，引导用户到Trap中，给予虚假文件的诱导，对其之前模式识别的结果进行校验，一旦校验成功，则说明此用户是在进行网络安全的攻击或者破坏，立即可以锁定该用户，同时为犯罪留下了证据！

结合网络日志，基于大数据的安全感知，可以基于网关、防火墙等硬件设备进行感知，主要方法有2种，一种是基于设备的系统进行烧入感知程序，一种是在设备旁边放置前端感测的防火墙服务器，在此服务器中进行安全感知，只有通过感知校验后，才能进行后续访问。一旦通过Trap校验安全威胁，就会采取相关措施。

结合网络日志，基于大数据的安全感知，可以基于主机审计代理程序进行应用结合，审计代理结合安全感知发现的用户访问长期的模式，更有助于对用户的安全行为作出更进一步的审计。

4 基于大数据安全感知的有效保护

有了结合网络日志的大数据的安全感知，如何做好有效保护，在此论文中本人也做一点展望：

在隐私保护意识日益增强的时代，除了对关键数据进行存储安全和标准化外，可以基于大数据安全感知把用户数据采集到信任区域，对用户数据进行预处理和整合并对用户隐私进行保护转换如加入随机化算法、添加噪点，这样只有经过正常的程序访问才能读取正常的数据并正常显示出来，如果用户企图窃取数据，通过日志就可发现其路径及关联规则的异常的同时，对于其取得的噪点数据及随机化数据直接Trap捕捉，用户即使获得所谓“数据”，也是一些虚假数据。

结合网络日志的大数据的安全感知，对基于安全审计进行有效保护给予了支撑，可以通过大数据安全感知建立统一审计分析中心用于分析用户群体安全模式审计、预警分析中心用于安全审计策略的预警、基于安全感知调整策略管理中心，分析结果的数据审计等，将极大保护网络访问的安全。

5 结语

基于以上大数据分析的安全感知的思路及部分在实践中的应用，基于网络日志进行大数据分析的安全感知具有先进性，从感知过程到处理可操作性强，基于陷阱捕捉Trap的模式进行反模式校验及有效预防保护，在系统上可以上下连贯，形成大数据安全感知的天罗地网。AP

作者简介

向永谦（1962－），男，湖北武汉人，高级工程师，硕士，现任62101部队信息中心网络室主任，研究方向是网络与网络安全。

李　欣（1983－），男，湖北武汉人，工程师，硕士，现任62101部队信息中心参谋，研究方向是信息系统与安全。

满建文（1986－），男，山东枣庄人，高级技师，硕士，现任62101部队信息中心管理员，研究方向是网络与网络安全。

参考文献：

[1] 鲍永伟, 方兴东. 云计算蓝皮书2015-2016[M]. 北京：电子工业出版社, 2017.