



专题：网络空间安全

基于自更新威胁情报库的大数据安全分析方法

侯艳芳, 王锦华

(移动互联网系统与应用国家工程实验室, 上海 201315)

摘要: 大数据安全分析方法一般是基于已有的特征库对恶意行为进行匹配检测, 该方法只能检测到已知攻击, 不能对新型的未知攻击与复杂变种攻击进行分析判断。通过大数据分析技术, 对采集数据进行安全价值提炼, 实现从采集数据、威胁信息、威胁知识, 再到威胁情报的价值转换, 形成具有隐蔽性、价值选择性、对抗性、预测性、决策性、可举证性、时效性的高价值威胁情报。首先分析了已有安全分析方法的局限性, 然后详细介绍了威胁情报库的构建方法, 最后阐述了基于自更新威胁情报库的大数据安全分析方法。基于自更新的威胁情报库进行大数据安全分析, 可以有效地挖掘出互联网数据中潜藏的威胁, 强化互联网企业已有的安全防护能力, 以便在安全博弈过程中掌握事前防御的主动权。

关键词: 威胁情报库; 大数据分析; 威胁信息; 威胁知识; 自更新

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018127

Big data security analysis method based on self-update threat intelligence database

HOU Yanfang, WANG Jinhua

Mobile Internet System and Application Security National Engineering Laboratory, Shanghai 201315, China

Abstract: Big data security analysis method is generally based on the existing signature database to detect and match malicious behaviors. This method can only detect known attacks and can't analyze and judge new types of unknown attacks and complex variant attacks. Through the big data analysis technology, the collected data was extracted for safety value, and the value conversion from the collected data, threat information, threat knowledge to threat intelligence was realized, high-value threat intelligence of forming hidden, value selective, adversarial, predictive, and decision-making, evidence-based, time-sensitive was formed. Firstly, the limitations of the existing security analysis methods were analyzed. Then the construction method of threat intelligence database was introduced in detail. Finally, the security analysis method of big data based on self-renewing threat information database was described. The analysis of big data security based on the self-updated threat intelligence database could mine the hidden threats in internet data effectively and strengthen the existing security protection capabilities of internet companies so that the preemptive defense initiative could be mastered during the security game.

Key words: threat intelligence database, big data analysis, threat information, threat knowledge, self-update

1 引言

科技伴随着便利服务,同时也伴随着信息安全问题,从通信安全、计算机安全、信息系统安全、信息安全保障,到网络空间安全,信息安全问题越来越得到重视,甚至已经上升到国家法律法规层面。

近年来,信息安全事件频发,各互联网行业采取相应的安全防护措施进行安全防御已是刻不容缓。大数据时代的到来,使得人们摆脱了传统离散、碎片化、体量有限的数据分析的困扰,进入了海量的大数据时代,可以利用应运而生的大数据分析技术,对海量的基础信息及事件数据进行挖掘分析,得出有价值的安全信息,为安全防护提供有效的支撑。

大数据安全分析技术一般是对已知的攻击行为进行检测分析,以分布式的数据存储与处理架构为分析平台,再结合自有的一定规模的特征库,对上网流量数据进行简单的检测判断,而对于未知攻击,甚至由已知病毒衍生而来的复杂变种威胁,都无法进行有效的分析检测。面对越来越严峻的安全形势,作为一名安全从业人员,应该从不同的角度深入思考针对目前棘手安全问题的解决方法。基于目前安全形势的思考,本文提出了一种基于自更新威胁情报库的大数据安全分析方法,该方法的重点是构建具有自更新能力的威胁情报库,该威胁情报库是具有隐蔽性、价值选择性、对抗性、预测性、决策性、可举证性、时效性的一系列高价值数据的集合,将其应用在大数据安全分析中,可快速有效地对外部威胁进行分析检测,实现事前防御,对强化已有的安全防护能力具有指导意义。

目前在威胁情报方面的技术研究如下。

- 参考文献[1]在网络安全数据可视化分析现状基础上,深入讨论了当前网络威胁安全数据可视化研究存在的问题,并进一步

探讨了网络空间安全数据可视化的发展趋势。

- 参考文献[2]宏观论述了网络空间安全发展态势,并系统性分析了其核心要素及层次化模型。
- 参考文献[3]论述了结果与分析过程中其他元素之间的关系对于分析来源的重要性,并阐述了可视化系统与分析路径系统可以实现解析关系的自动化提取,结合已知经验的再提取过程,可实现对犯罪来源的精准分析。

2 已有安全分析方法的局限性

目前,市场上层出不穷的威胁情报产品,对于威胁情报的认知与定义也仅限于特征库层面,没有上升到对可能出现的针对资产的威胁或危险的趋势预测以及能够为主体响应相关威胁或危险提供决策信息的层面。因此,这些威胁情报产品也只能根据安全经验知识(特征库)对已知的攻击行为进行检测分析,而对于未知攻击与复杂变种攻击,则失去了检测分析能力,这使得该类产品的漏报率很高,面对日益严峻的安全威胁的实用价值则大打折扣。基于目前对威胁情报的认知及输出能力,进行的大数据安全分析具有很大的局限性,其主要体现在以下几个方面。

(1) 可依赖的特征库价值体现十分有限

已有的安全特征库是从历史安全事件中获取到的经验知识总结而来,是基于已发生的已知攻击的特征抽取,涵盖范围有限,对于日益复杂的网络结构来说,特征数据的价值也十分有限,仅适用于已知攻击。

(2) 漏报率高

由于特征库的局限性,使得对于可能发生的复杂攻击进行检测分析,会有很高的漏报率。

(3) 无法检测新型的攻击

已有的基于特征库匹配的大数据安全分析方法



法不能对新型的未知攻击与复杂变种攻击进行分析判断,这也成为了目前所使用的大数据安全分析方法的检测盲区。对于新型的攻击来说,企业即使采取了相应的安全防护措施对重要信息系统进行了防护,也依然是“裸”防护。

本文提出一种基于自更新威胁情报库的大数据安全分析方法,能够克服已有大数据安全分析方法的局限性,可以对恶意行为进行精准、全面的分析,从而有效地检测出新型的未知攻击与复杂变种攻击。

3 威胁情报库的构建方法

威胁情报是“基于证据的知识,包括场景、机制、指标、含义和可操作的建议。这些知识是关于现存的或者即将出现的针对资产的威胁或危险的,可为主体响应相关威胁或危险提供决策信息(Gartner对“威胁情报”的定义)”^[4]。可知,威胁情报是对具有隐蔽性、价值选择性、对抗性、预测性、决策性、可举证性、时效性等几大属性的高价值数据的统称,威胁情报库则是这一系列高价值数据的集合。威胁情报的生成过程是基于大数据分析技术对威胁数据进行价值提取的过程,其主要生成流程如图1所示。

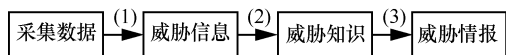


图1 威胁情报的生成流程

如图1所示,在威胁情报的生成过程中,所使用的大数据分析技术^[5]为图1中的(1)、(2)、(3),具体说明如下。

(1) 对采集到的用户上网流量数据与日志数据进行数据预处理,然后再根据经验知识(特征库)进行匹配判断,获取匹配成功的恶意行为数据,然后进行数据分析,获取威胁关键词,生成威胁信息。威胁信息主要包含以下7个指标:攻击者、被攻击者、攻击类型、攻击类别、攻击手段、攻击开始时间、攻击结束时间。

(2) 利用攻击树的构建方法与聚类、关联规则等数据挖掘算法,结合常用的统计分析方法,对生成的威胁信息进行数据分析,确定攻击工具、攻击方式、攻击路径、攻击影响面等攻击信息,并对攻击者进行画像,明确攻击者的意图,找出被攻击的原因,生成威胁知识。威胁知识主要包含以下7个指标:攻击工具、攻击方式、攻击路径、攻击影响面、攻击意图、攻击者画像、被攻击原因。

(3) 利用神经网络、时间序列等预测模型,对生成的威胁信息与威胁知识进行数据分析,预测攻击趋势,并根据相关结果,给出指导性的安全防护策略,统计分析出资产的损失情况,生成威胁情报。威胁情报主要包含以下17个指标:攻击者、被攻击者、攻击类型、攻击类别、攻击手段、攻击开始时间、攻击结束时间、攻击工具、攻击方式、攻击路径、攻击影响面、攻击意图、攻击者画像、被攻击原因、攻击趋势预测、指导性决策、资产损失率。

综上所述,假设“ T_i ”代表威胁信息,“ T_k ”代表威胁知识,“ T_g ”代表威胁情报,则:

$$T_g = V(T_i, T_k) \quad (1)$$

式(1)表示威胁情报是以威胁信息与威胁知识为变量的价值函数的结果,该函数可以对攻击趋势进行预测,能够给出指导性的安全防护策略,并且可以统计分析出资产的损失情况。

威胁情报所包含的17个指标,能够充分体现出威胁情报的隐蔽性、价值选择性、对抗性、预测性、决策性、可举证性及时效性这几大特征属性。运用大数据分析技术对采集到的数据进行分析处理,可以形成威胁情报,从海量数据中挖掘到的一系列威胁情报就可以构成威胁情报库。因此,威胁情报库可以看作以威胁情报的17个指标为列向量、以一系列完整的威胁情报为行向量的数据矩阵。现分别对威胁情报所包含的17个指标表示的含义进行说明。

(1) 攻击者

任何能代表攻击者身份的信息,如物理 ID、IP 地址、MAC 地址、地理位置等。

(2) 被攻击者

任何能代表被攻击者身份的信息,如物理 ID、IP 地址、MAC 地址、地理位置等。

(3) 攻击类型

分为已知与未知两种类型。已知攻击是指具有历史指纹信息的攻击行为经验知识,未知攻击是指超出经验知识的复杂变种或新型攻击。

(4) 攻击类别

已知攻击中一般包含如下攻击类别:Web 入侵攻击、僵尸网络、DDoS 攻击、移动应用协作攻击、APT 攻击、网络钓鱼攻击。未知攻击中的攻击类别一般是指新型未命名的攻击。

(5) 攻击手段

在宏观上可以分为以下三大类:社会工程攻击、蛮力攻击和技术攻击。

社会工程学源于黑客凯文·米特尼克悔改后所著的《欺骗的艺术》^[6]一书,是一种通过对受害者心理弱点、本能反应、好奇心、信任、贪婪等心理陷阱进行诸如欺骗、伤害等危害的手段。

蛮力攻击主要针对目标网站,猜测并组合不同的用户名和密码,多次尝试登录以获取正确的登录信息,从而达到进入网站的目的。这是一个简单而有效的网站攻击方式,可用于任何网站,尤其是那些登录区域安全性较差的网站^[7]。

技术攻击主要是指攻击者利用一些计算机技术在技术层面发起的攻击。

在本文中,将攻击手段作为攻击类别的子类,具体划分如下。

Web 入侵攻击类:SQL 注入、XSS 攻击、CSRF 攻击、Web shell、远程命令执行、文件包含等。

僵尸网络类:IRC botnet、AOL botnet、P2P botnet 等。

DDoS 攻击类:基于 ARP、基于 ICMP、基于

IP、基于 UDP、基于 TCP、基于应用层等(按照 TCP/IP 的层次)。

移动应用协作攻击类:移动应用协作攻击。

APT 攻击:水坑式攻击等。

网络钓鱼攻击:普通钓鱼、鱼叉式钓鱼。

新型未命名的攻击:根据新发现攻击的分类情况而定。

(6) 攻击开始时间

指攻击动作发起的时间。

(7) 攻击结束时间

指攻击动作结束的时间。

(8) 攻击工具

主要包括目标扫描工具(SuperScan、X-Scan、流光等)、扩大入侵工具(Iris Network Traffic Analyzer、Sniffer pro 等)、嗅探器工具(TcpDump、影音神探等)。

(9) 攻击方式

一般包含以下几种攻击方式:获取口令、放置特洛伊木马程序、Web 欺骗、电子邮件攻击、通过一个节点来攻击其他节点、网络监听、寻找系统漏洞、利用账号进行攻击、偷取特权。

(10) 攻击路径

是指基于 Intrusion Kill Chain 模型,从每个攻击阶段获取相关信息构成的一条攻击路径。

Intrusion Kill Chain 模型(或称为 Cyber Kill Chain 模型)由 Lock Martin 公司 Hutchins E M 等 3 位安全研究员在 2011 年 3 月举行的 ICIW 大会上公布,精髓在于明确提出网络攻防过程中攻防双方互有优势,防守方若能阻断/瓦解攻击方的进攻组织环节,就能成功地挫败对手的攻击企图。模型将攻击者的攻击过程分解为如下 7 个步骤:reconnaissance(踩点)、weaponization(组装)、delivery(投送)、exploitation(攻击)、installation(植入)、C2(控制)、actions on objectives(收割)。

(11) 攻击影响面

是指目标系统被攻击的漏洞的集合。



(12) 攻击意图

攻击者实施攻击一般包含以下两种目的：获取金钱和财物、满足个人心理需求。具体表现形式是：窃取资产数据、实施资金诈骗、公布未披露的企业漏洞及进行恶作剧等。

(13) 攻击者画像

利用攻击树的构建方法，结合聚类、关联规则等数据挖掘算法，对从上述组成威胁情报的(1)~(12)的指标中获取的威胁信息进行分析，描述攻击者在实施攻击过程中的一种行为状态，对攻击身份、攻击工具、攻击手法及攻击目标等攻击要素进行关系图的梳理，实现对整个安全事件前因后果的推理。攻击者画像过程示例如图2所示。

(14) 被攻击原因

主要说明资产所存在的脆弱性，确定企业遭受攻击的根本原因，如桌面漏洞、服务器漏洞、Web 服务器虚拟托管、显性/开放式代理、内部人员安全意识薄弱、采用不安全的通信协议进行互联网访问、在邮件中采用嵌入式HTML等。

(15) 攻击趋势预测

利用历史数据与从上述组成威胁情报的(1)~

(14)的指标中获取的威胁信息，基于数据挖掘中的预测模型，对整个攻击趋势进行可视化展现，从而掌控该攻击的整个态势情况。

(16) 指导性决策

根据上述威胁信息，形成具有指导性的安全防护策略，如打补丁、内部人员的安全意识培训、网络关键节点中的防护措施等。

(17) 资产损失率

资产损失率主要说明该攻击对企业所造成的伤害程度。具体计算式如下：

$$\text{资产损失率} = \frac{\text{资产损失金额}}{\text{资产投资成本}} \quad (2)$$

4 基于自更新威胁情报库的大数据安全分析方法

目前所使用的大数据安全分析方法一般是基于分布式的架构对海量数据进行存储与处理，然后通过基于已知经验知识所形成的特征库进行规则匹配，实现对已知攻击的检测分析及预警。基于自更新威胁情报库的大数据安全分析方法的核心理念是通过具有隐蔽性、价值选择性、对抗性、预测性、决策性、可举证性及时效性等特征属性的威胁情报所提供的高价值信息，弥补传统特征库

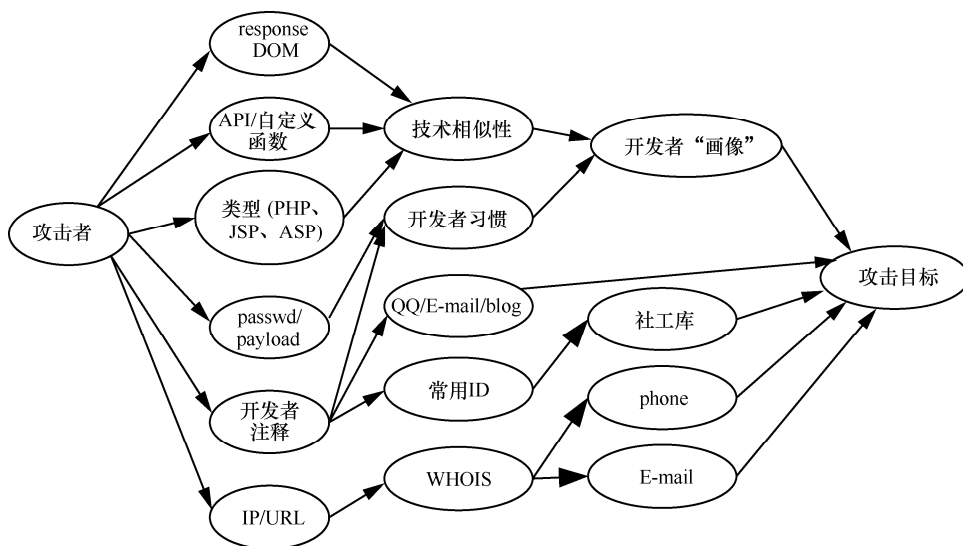


图2 Web入侵(Web shell)过程的攻击者画像示例

的知识局限性,对可疑攻击行为进行精准预判,实现事前防御的安全防护能力,并通过自动化读写能力,进行威胁情报库的自更新操作。该方法的实现同一般的大数据安全分析方法一样,均是基于分布式的架构对采集到的数据进行存储与处理,所有的数据分析过程与数据库的构建存储过程均是在分布式环境下进行的,其实现逻辑框架如图3所示。

如图3所示,可以将基于自更新威胁情报库的大数据安全分析方法的实现划分为以下7层。

(1) 数据源层

根据不同的网络环境,可以有3种数据来源:3G/4G数据、IDC DPI数据与移动上网日志数据。

志数据。

(2) 数据采集层

可通过主动爬取、旁路分光数据获取、增量数据同步获取这3种途径进行数据采集。

(3) 数据预处理层

主要通过数据清洗、数据集成、数据变换、数据规约及数据离散化的数据处理方式,对数据进行预处理,达到数据格式标准化与异常数据清除的目的。

(4) 数据仓库层

使用分布式架构,基于稳定的文件存储系统HDFS,部署HBase、MongoDB、MySQL等数据库,实现海量数据的存储。

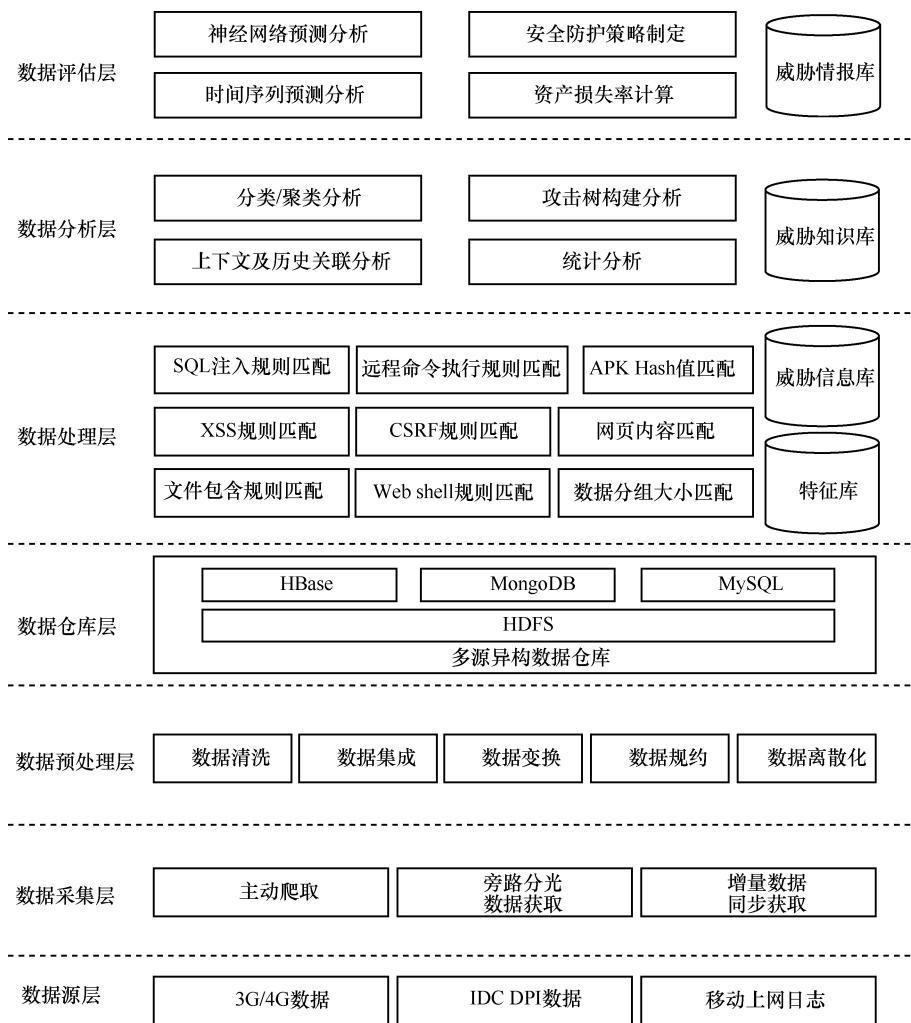


图3 基于自更新威胁情报库的大数据安全分析方法实现



(5) 数据处理层

基于已有的特征库，对基础信息数据与网络访问日志数据进行规则匹配：SQL 注入规则匹配、XSS 规则匹配、文件包含规则匹配、远程命令执行规则匹配、CSRF 规则匹配、Web shell 规则匹配、APK Hash 值匹配、网页内容匹配、数据分组大小匹配。根据匹配结果，生成威胁信息，进行入库操作。

(6) 数据分析层

根据数据处理层得到的威胁信息，通过分类/聚类分析、上下文及历史关联分析、攻击树构建分析与统计分析，生成威胁知识，进行入库操作。

(7) 数据评估层

根据数据处理层与数据分析层得到的威胁信

息与威胁知识，通过神经网络预测分析、时间序列预测分析、安全防护策略制定与资产损失率计算，生成威胁情报，进行入库操作。

本文提出的基于自更新威胁情报库的大数据安全分析方法的主要分析过程如图 4 所示。

如图 4 所示，基于自更新威胁情报库进行大数据安全分析的实现过程可分为以下 6 个步骤。

步骤 1 通过数据采集设备采集用户上网流量数据与日志数据。

步骤 2 经过清洗、去重、标准化等数据预处理，得到基础信息数据与网络访问日志数据。

步骤 3 基于已有的特征库进行简单的规则匹配操作，匹配成功则判定是恶意行为数据，进行威胁信息的入库操作；匹配失败则基于威胁情

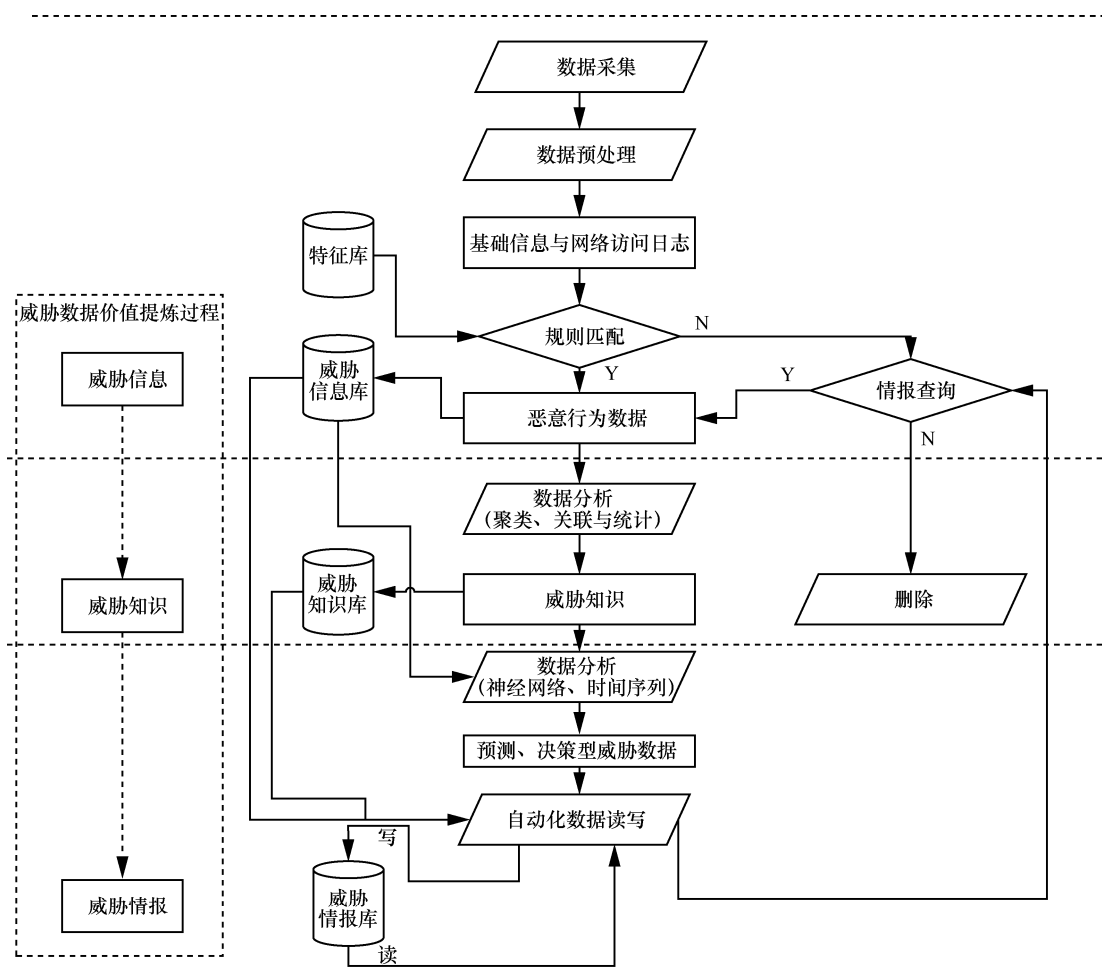


图 4 基于自更新威胁情报库的大数据安全分析过程

报库进行情报查询操作, 查询结果有反馈, 则判定是恶意行为数据, 进行威胁信息的入库操作, 查询结果无反馈, 则进行数据删除操作。

步骤4 利用聚类、关联规则等数据挖掘算法与攻击树的构建方法, 再结合常用的统计分析方法, 对判定的恶意行为数据进行数据挖掘分析, 确定攻击工具、攻击方式、攻击路径、攻击影响面等攻击信息, 并对攻击者进行画像, 寻找被攻击的原因, 生成威胁知识, 进行威胁知识的入库操作。

步骤5 利用神经网络、时间序列等数据挖掘中的预测模型, 对步骤4得到的威胁知识进行预测分析, 实现对攻击趋势的预测, 并结合上述步骤的相关信息, 给出指导性的安全防护策略, 并且根据式(2)对资产损失情况进行计算分析, 生成预测、决策型威胁数据。

步骤6 对步骤3~步骤5得到的威胁信息、威胁知识与预测、决策型威胁数据, 进行自动化读写处理, 形成威胁情报, 实现威胁情报库的自更新。自动化读写处理是通过图4中的自动化数据读写功能模块实现的, 具体过程如下。

(1) 写操作

将从威胁信息库、威胁知识库以及预测、决策型威胁数据模块中获取到的信息, 根据威胁情报所包含的17个指标分别进行信息抽取, 然后依次写入威胁情报库, 进行威胁情报的入库操作。

(2) 读操作

以规则匹配模块中匹配失败的数据为查询对象, 选取关键词对威胁情报库进行遍历读取操作, 若有反馈结果, 则查询停止, 并对反馈结果进行判定记录; 若无反馈结果, 则遍历完成后退出查询操作, 并对查询对象进行删除操作。

上述分析过程也是威胁数据价值的一种提炼过程, 从采集到的数据, 到威胁信息, 到威胁知识, 再到威胁情报, 数据的价值不断提升, 实用性也越来越高。

由上可知, 本文提出的大数据安全分析方法

是一种基于动态更新的威胁情报库的深度安全分析方法, 通过威胁情报库的构建过程可以看到, 对于网络中潜藏的未知威胁, 通过威胁情报中所提供的用户画像、攻击趋势预测与指导性决策指标, 可以有效地进行分析判断与阻断预警, 做到事前防御的安全防护效果, 加强了企业纵深防御的力度, 为未来复杂网络的安全防护提供了强有力的支撑。

5 结束语

大数据时代的到来为安全领域引入了强有力的大数据分析技术, 常用的基于特征库的匹配分析方法已经无法满足市场的需求, 随着攻击者攻击能力的不断提升, 安全防护工作对安全分析技术的要求也越来越高, 已有的安全经验知识已经无法解决越来越复杂的安全问题。即将进入的物联网时代, 网络结构会更加复杂, 网络节点的安全防护盲区也越来越大, 安全形势也更加不容乐观。

本文提出的基于自更新威胁情报库的大数据安全分析方法, 能够在具有实用性的高价值威胁情报库的协同分析下, 形成事前防御能力, 更精准、更全面地对已知及未知攻击行为进行分析检测, 能够强化各企业现有的安全防护体系, 降低企业遭受攻击的风险, 助力企业未来智能化的安全建设工作。

参考文献:

- [1] 张繁, 谢凡, 江颖. 网络威胁安全数据可视化综述[J]. 网络与信息安全学报, 2018, 4(2): 34-39.
ZHANG F, XIE F, JIANG X. Survey on the visualization technologies of threatening security data in cyber space[J]. Chinese Journal of Network and Information Security, 2018, 4(2): 34-39.
- [2] 方滨兴. 定义网络空间安全[J]. 网络与信息安全学报, 2018, 4(1): 1-5.
FANG B X. Define cyberspace security[J]. Chinese Journal of Network and Information Security, 2018, 4(1): 1-5.
- [3] ISLAM J, XU K, WONG W. Analytic provenance for criminal intelligence analysis[J]. Chinese Journal of Network and Information Security, 2018, 4(2): 18-33.



- ISLAM J, XU K, WONG W. Analytic provenance for criminal intelligence analysis[J]. Chinese Journal of Network and Information Security, 2018, 4(2): 18-33.
- [4] MITNICK K D, SIMON W L. 欺骗的艺术[M]. 潘爱民, 译. 北京: 清华大学出版社, 2014.
- MITNICK K D, SIMON W L. The art of deception[M]. Translated by PAN A M. Beijing: Tsinghua University Press, 2014.
- [5] 张云勇. 电信运营商大数据发展建议[J]. 电信科学, 2018, 34(1): 103-108.
- ZHANG Y Y. Suggestions for telecom operators' big data development[J]. Telecommunications Science, 2018, 34(1): 103-108.
- [6] Gartner. 威胁情报的定义[EB/OL]. (2013-05-16)[2018-01-08]. <https://www.gartner.com/doc/2487216/definition-threat-intelligence>.
- Gartner. Definition: threat intelligence[EB/OL]. (2013-05-16)[2018-01-08]. <https://www.gartner.com/doc/2487216/definition-threat-intelligence>.

[作者简介]



侯艳芳（1990-），女，移动互联网系统与应用安全国家工程实验室安全研究员，主要研究方向为大数据安全相关技术。



王锦华（1982-），男，现就职于移动互联网系统与应用安全国家工程实验室，主要研究方向为云计算、大数据安全相关技术。