•可信计算与信息安全•

DOI:10.15961/j.jsuese.201801053

基于告警属性聚类的攻击场景关联规则挖掘方法研究

陈兴蜀[1,2],何　涛[1],曾雪梅[2*],邵国林[2]

(1.四川大学 网络空间安全学院,四川 成都 610065;2.四川大学 网络空间安全研究院,四川 成都 610065)

摘　要:针对现有攻击场景重构方法中存在关联规则挖掘不充分、攻击场景链断裂的问题,以及安全设备的误告警影响攻击场景重构准确性的现状,提出一种基于告警属性聚类的攻击场景关联规则挖掘方法。该方法能够有效挖掘攻击场景关联规则,减少攻击链断裂,还原实际的多步攻击,更好地帮助安全管理员深入理解攻击者入侵行为并掌握攻击全貌。以真实网络中的安全设备的原始告警为数据源,首先,对原始告警数据进行预处理,实现告警数据的归一化。然后,通过构建告警时间序列,利用FFT和Pearson相关系数对误告警周期特性进行分析,生成误告警过滤规则。接着,提出一种基于动态时间阈值的告警属性聚类方法,通过告警属性相似性刻画告警间相似度,并根据告警发生的时间间隔结合动态时间阈值方法更新聚类时间,对属于同一攻击场景的告警进行聚类。最后,利用Apriori频繁项挖掘算法生成攻击场景序列模式,并对具有重复攻击步骤的攻击场景序列模式进行融合生成关联规则。在四川大学校园网真实环境中进行实验,结果表明所提方法可有效缓解攻击链断裂问题和误告警的影响,相较于对比方法可有效提升生成的攻击场景关联规则的完整性。

关键词:攻击场景重构;告警关联;属性相似度;误告警

中图分类号:TP393.0　　**文献标志码**:A　　**文章编号**:2096-3246(2019)03-0144-07

Research on Attack Scene Association Rule Mining Method Based on Alarm Attributes Clustering

CHEN Xingshu[1,2], *HE Tao*[1], *ZENG Xuemei*[2*], *SHAO Guolin*[2]

(1.College of Cybersecurity, Sichuan Univ., Chengdu 610065, China; 2.Cybersecurity Research Inst., Sichuan Univ., Chengdu 610065, China)

Abstract: In order to solve the problems that the association rules are not fully exploited, the attack scenario chain breaks in the existing attack scene reconstruction methods, and false alarms of security device affect the accuracy of attack scene reconstruction, an attack scenario association rule mining method based on alarm attributes similarity clustering was proposed in this paper. The method can effectively mine attack scene association rules, reduce attack chain breaks, restore actual multi-step attacks, and help the security administrator to deeply understand the attacker's intrusion behaviors and master the attack. First, the alarm data including the original alarms of security device in the real network and the data source was preprocessed and normalized. By constructing an alarm time series, the FFT and Pearson correlation coefficients were used to analyze the characteristics of the false alarm period to generate a false alarm filtering rule. Then, an alarm attributes clustering method based on dynamic time threshold was proposed. The similarity between alarms was characterized by the similarity of alarm attributes. The clustering time was updated according to the interval between alarms and the dynamic time threshold. Finally, the Apriori frequent item mining algorithm was used to generate the attack scene sequence pattern, and the attack sequences with repeated steps were merged to generate the association rules. The experiments results showed that the proposed method can effectively alleviate the impact of attack chain breaks and false alarms. Compared with the comparison methods, the integrity of the generated attack scene association rules can be effectively improved.

Key words: attack scenario reconstruction; alert correlation; attribute similarity; false alarms

收稿日期:2018 – 09 – 24

基金项目:国家自然科学基金项目(61802270);国家"双创"示范基地之变革性技术国际研发转化平台(C700011);四川省重点研发项目(2018G20100);四川省科技支撑计划项目(2016GZ0038);中央高校基本科研业务费专项资金(2017SCU11059;2017SCU11065;SCU2016D009)

作者简介:陈兴蜀(1968—),女,教授,博士. 研究方向:云计算;信息安全. E-mail:chenxsh@scu.edu.cn

* 通信联系人 E-mail:zengxm@scu.edu.cn

网络出版时间:2019 – 04 – 24 14 : 19 : 17　　**网络出版地址**:http://kns.cnki.net/kcms/detail/51.1773.TB.20190424.1018.002.html

随着网络技术的快速发展，攻击者水平不断提高，攻击技术和手段日趋复杂多样，网络安全问题日益突出。国家计算机网络应急技术处理协调中心（CNCERT/CC）在2019年1月《CNCERT互联网安全威胁报告》[1]中提到，近年来危害较大的攻击几乎都是复杂的多步攻击[2−3]，尤其是针对高价值目标的APT攻击，这类攻击往往目标性更强、危害性更大。传统的安全检测设备能检测大部分此类攻击中的单步攻击并形成告警，但这些告警往往是攻击的碎片化信息，缺乏对大量单个攻击告警之间的关联性分析。挖掘攻击者的多步攻击模式，重构攻击场景，可以帮助网络安全人员深入了解攻击者的攻击动机、攻击技巧、攻击方式，甚至预测攻击者的下一步攻击[4−5]，感知网络安全态势[6]，从而为构建一个更加稳健的安全防护体系提供支撑。

多步攻击关联又称“攻击场景重构”，指由多个不同步骤构成的多步攻击产生的告警构建完整攻击场景过程。目前，对于攻击场景重构的研究，根据所采用方法的不同可以大致分为以下两类：

1）基于因果知识库的关联方法，通过专家经验知识定义攻击场景重构关联规则。Templeton等[7]提出一种灵活的可扩展的攻击模型，利用特定的攻击模型语言分析攻击间的因果关系识别出多步攻击。Ning等[8]在Templeton的基础上，利用攻击之间的依赖关系提出了具有代表性的基于前因后果的告警关联方法，通过专家知识定义攻击步骤之间先决条件集和产生的结果集进行告警关联。Morin等[9]通过融合安全设备告警、漏洞、主机信息及软件产品信息，提出一种基于安全系统的1阶逻辑数据模型，用于查询和断言有关安全事件及其发生的上下文信息，以实现告警关联。这类方法都是事先定义攻击的关联知识库，依赖于专家经验知识，需要人工配置大量参数。

2）基于数据挖掘的方式生成关联规则，基于数据挖掘算法从历史数据中挖掘攻击场景关联规则，从而关联攻击告警。樊迪等[10]提出一种基于因果知识发现的攻击场景重构方法，利用概率统计方法发现各告警类型间的关联关系。冯学伟等[11]先对告警进行聚类，再基于马尔可夫链的无后效性挖掘不同攻击类型间的转移概率矩阵，构建攻击场景，但是，该方法在聚类时只提取了IP相关性，未考虑其他属性对告警聚类的影响。Zhang等[12]提出采用时间窗口对告警序列进行分段，在每个时间窗口中发现攻击场景。Ramaki[13]、Kavousi[14]等提出使用贝叶斯网络构建贝叶斯攻击图（BAG），并根据生成的关联规则预测攻击者的后续步骤，但是，这种方法通过经验设置时间窗口，存在聚类时间窗口难以确定的问题。田志宏等[15]提出了一种告警关联模型A3PC，借助异常检测思路对误报警进行自动鉴别，同时采用模式挖掘和聚类分析算法相结合的方法重构攻击场景。Daneshgar等[16]提出一种基于模糊事件聚类的关联分析方法，先根据事件的相关性将模糊事件聚类，再通过挖掘频繁模式生成关联规则。这类方法虽不需要大量的先验知识，但由于未考虑误告警情况，会影响攻击场景重构的准确性，并且需在聚类时设置固定时间窗口，导致属于同一攻击场景告警发生分裂，造成多步攻击链断裂。

针对上述问题，作者提出一种基于告警属性聚类的攻击场景关联规则挖掘方法。该方法首先基于入侵检测消息交换格式（IDMEF）规范安全设备告警日志，并对告警日志中的不完整数据进行初步过滤；之后，对误告警数据的周期特性进行分析，利用FFT变换计算准确周期值以及相关系数验证周期值的准确性；接着，基于提出的告警属性相似度聚类方法聚合告警，分别对IP、端口、攻击类型相似度进行计算，并利用动态时间阈值对告警进行聚类；最后，通过频繁项挖掘算法Apriori生成攻击场景序列，并采用树形图形式对具有重复步骤的序列进行融合，生成攻击场景关联规则。

1　基于属性相似度聚类的关联规则挖掘方法

1.1　日志预处理

1.1.1　安全日志格式化

由于安全日志的格式缺乏规范标准，导致厂商不同、安全设备不同，产生的告警日志格式也不统一，因此在对告警日志进行分析前需要对其进行规范化。采用IDMEF格式对告警日志进行格式化处理，便于后续的聚合与关联分析。将日志规范化为如下格式：*raw_alert*=(*time*,*srcIP*,*srcPort*,*dstIP*,*dstPort*,*attackType*,*warningLevel*)。其中，*time*表示告警产生时间，*srcIP*表示源IP，*srcPort*表示源端口，*dstIP*表示目的IP，*dstPort*表示目的端口，*attackType*表示攻击类型，*warningLevel*表示告警等级。

1.1.2　日志过滤

由于大规模的网络具有复杂性，再加上网络安全攻击事件的不确定性，导致网络安全设备中产生的日志信息存在大量不完整、误告警的数据，而这些不完整、误告警数据会影响网络安全管理人员对安全事件分析的准确性。所以，需要对这些信息进行过滤操作。具体处理步骤如下：

1）采用IDMEF格式对安全设备告警日志格式化。

2）对于每一条告警日志，先判断五元组（源/目

的IP, 源/目的端口, 协议)是否完整, 当上述属性缺少两个及以上时将其过滤掉。否则步骤转3)。

3)根据误告警特性, 对误告警数据进行过滤, 详情见第1.1.3节。

1.1.3 误告警特性研究

安全设备往往会产生大量的误告警数据, 而误告警混杂在真实告警中, 会影响最终攻击场景重构的正确性。通过对告警数据进行统计分析, 发现存在大量的误告警, 并且误告警数据具有周期特性。虽然部分网络攻击也会具有周期性, 但是具体体现在安全设备告警上可能并不存在该特性。为了对误报数据进行过滤, 作者基于李冬等[17]提出的方法对日志中周期性误告警数据进行处理。具体过程为: 对于预处理后的告警日志, 以小时为单位统计$\langle srcIP, dstIP, attackType\rangle$每个三元组产生的告警数, 由此得到该三元组的时间序列$\{ats(i), i=0,1,2,\cdots,N\}$, 其中, $ats(i)$表示第i小时的告警数, N表示时间序列长度(即小时数), 图1为某三元组的告警时间序列。

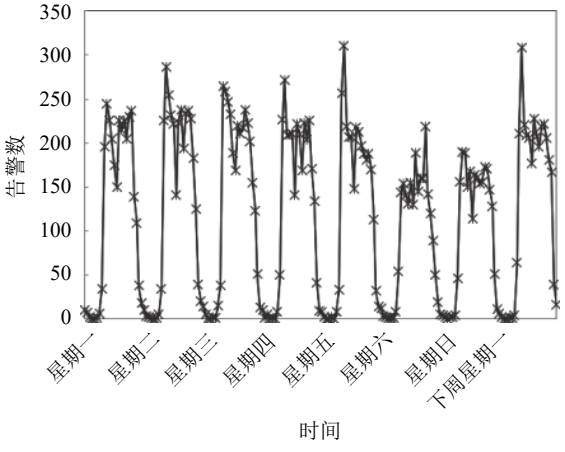


图 1 告警数时间序列

Fig. 1 Alarm number time series

从图1可以看出每天产生的告警数量具有明显的周期性。为了获得告警时间序列的周期值, 对告警时间序列通过快速傅里叶变换(FFT)计算准确周期值。傅里叶变换函数定义如下:

$$DFT(n)=\sum_{i=0}^{N-1} ats(i)\mathrm{e}^{\frac{\mathrm{i}2\pi kn}{N}}, n=0,1,\cdots,N-1 \qquad (1)$$

式中, $\mathrm{e}^{2\pi k}=1$, $ats(i)$表示第i小时的告警数。采用FFT将时序图转换为频率图, 图2为某三元组的频谱图, 通过找到图2中的峰值对应的频率点f, 根据$p=1/f$计算出告警周期p。

告警时间序列经过FFT后能产生一个周期值, 而并非所有的告警时间序列都具有周期性, 所以, 针对此问题采用Pearson相关系数对FFT后计算出的周期值的正确性进行检验。对于时间序列$\{ats(i), i=0,1,\cdots,N\}$, 假设存在周期$p$, 对该序列进行分组$\{ats(i), ats(p+i),\cdots, ats((m-1)p+i)\}, m=N/p$, 等价于$\{ats_{i1}, ats_{i2},\cdots, ats_{im}\}$, m为分组数。相似度计算公式为:

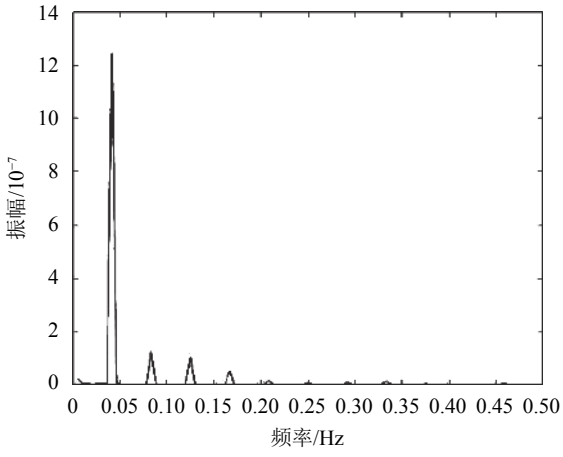


图 2 FFT谱频

Fig. 2 FFT frequency

$$r=\frac{\sum_{j=0}^{p}\left(ats_{ij}-\overline{ats_i}\right)\left(ats_{i(j+1)}-\overline{ats_i}\right)}{\left(\sum_{j=1}^{p}\left(ats_{ij}-\overline{ats_i}\right)^2\right)^{\frac{1}{2}}\times\left(\sum_{j=1}^{p}\left(ats_{i(j+1)}-\overline{ats_i}\right)^2\right)^{\frac{1}{2}}} \qquad (2)$$

式中, r为相关系数, $\overline{ats_i}$表示第i个分组的均值。r的大小决定了变量的相关程度, 其值越大相关性越高。若相似度大于0.8, 则具有强相关性, 认为周期值正确。基于Pearson相关系数对周期值进行检验, 能够避免过滤真实的网络告警, 保证误告警的准确性。误告警过滤算法的描述如算法1所示。

算法1 误告警过滤算法

输入: 原始告警$RA=\{ra_1, ra_2,\cdots, ra_n\}$;

输出: 误告警过滤规则R。

```
{
1. RA′ = preprocess(RA);
2. for i = 1:n
3.    ats = getAlertTimeSequence(RA′_i);
4.    f = FFT(ats); //FFT转换时间序列
5.    find max(f);
6.    p = 1/f;  //p为告警序列周期值
7.    //计算告警序列的相关系数
8.    r = calculateCorrelation(ats);
9.    if(r>0.8)
10.       R.add(i);
11.   end if
12. end for
}
```

1.2 告警日志相似度聚类研究

误告警过滤后的告警日志仍然有信息分散、质量低等特点。通过告警聚类，将大量分散、有关联的告警进行整合，使得聚类后的每条超级告警都能代表这一类告警。超级告警（SuperAlert）定义如下：$SA=(srcIP,srcPort,dstIP,dstPort,attackTypeList,warningLevel,count,startTime,endTime)$。其中，$attackTypeList$表示聚合后的攻击类型序列，$count$表示聚合的告警数量，$startTime$表示这类告警的开始时间，$endTime$表示这类告警结束时间。马琳茹等[18]通过将属性分层，对每一层赋予不同的阈值，通过简单比较属性是否相等，确定最终的相似度。该方法虽然将属性分层，但是属性计算方法单一，难以刻画多种属性的相似度。因此，作者提出一种基于动态时间阈值的告警属性相似度计算方法。

1.2.1 告警属性相似度计算

常用的属性相似度计算方法采用距离度量和余弦相似度计算，适合于数值类型的相似度计算，而告警属性大多数为非数值型。针对告警属性相似度的计算，Valdes等[19]提出一种告警相似度计算方式。所有原始告警用$\{ra_i,i=0,1,2,\cdots,n\}$表示，$ra_i$为第$i$条告警，$n$为告警总数。任意两条告警相似度计算公式为：

$$S(ra_i,ra_j)=\frac{\sum_k w_k s(ra_{ik},ra_{jk})}{\sum_k w_k} \tag{3}$$

式中，S为两条告警总体相似度，k为告警属性索引，w_k为告警的第k属性的权重，ra_{ik}、ra_{jk}为告警第k个属性值。这里，主要提取$\langle time,srcIP,srcPort,dstIP,dstPort,attackType\rangle$共6种属性实现对告警相似度的刻画。

1）IP地址相关性

由同一攻击活动触发的告警在IP地址分布上具有相关性，前一攻击步骤的目标节点可能就是下一攻击步骤的源节点。如果两条告警ra_i、ra_j的IP地址具有相关性，那么，ra_i的源IP或目的IP中总有一个与ra_j的源或目的IP相同。IP相关性计算公式如下：

$$s_{\mathrm{IP}}=\begin{cases}1, sip_{ra_i}=sip_{ra_j}\|sip_{ra_i}=dip_{ra_j};\\ 1, dip_{ra_i}=sip_{ra_j}\|dip_{ra_i}=dip_{ra_j};\\ 0, \text{other}\end{cases} \tag{4}$$

式中，sip表示源IP，dip表示目的IP。

2）端口相似度

对于端口相似度的计算，采用概念层次树的归纳方法将端口分类，如图3所示。

相似度计算公式如下：

$$s_{\mathrm{port}}=\begin{cases}1, p_{ra_i}=p_{ra_j};\\ 1-l(p_{ra_i},p_{ra_j})/3, p_{ra_i}\neq p_{ra_j}\end{cases} \tag{5}$$

式中，$l(p_{ra_i},p_{ra_j})$为两端口到达树中公共双亲的最大长度，p表示端口。

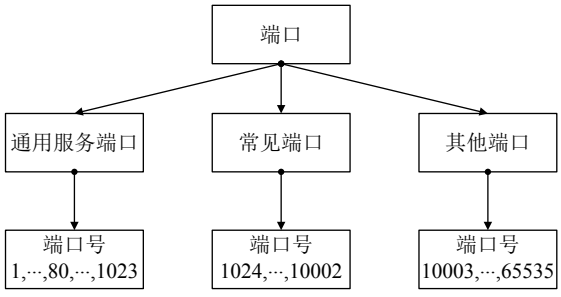


图 3　端口层次树

Fig. 3　Port hierarchy tree

3）攻击类型相似度

攻击类型属于非数值型属性，直接判断两种攻击类型是否相同。如果相同，相似度为1；如果不相同，相似度为0。

4）动态时间阈值

不同的网络攻击，它们的攻击时间存在差异，聚合时间过短会导致“欠聚合”，聚合时间过长会导致“过聚合”。李洪成等[20]提出自扩展时间窗的聚合方法，但是，由于需要通过专家经验设置聚合时间窗口，该方法可能会导致原本属于同一攻击场景的告警聚类为多个子场景，分裂多步攻击链。作者采用动态时间更新方法更新聚合时间，当有新告警满足聚合条件时，则聚合告警并更新动态时间阈值，减少攻击链断裂。$t_i(i=1,2,\cdots,n)$为原始告警ra_i产生告警日志的告警时间，$\tau_i=t_{i+1}-t_i(i=1,2,\cdots,n-1)$表示$ra_i$和$ra_{i+1}$的告警日志的时间差，则日志间的时间差集合为$\{\tau_1,\tau_2,\cdots,\tau_{n-1}\}$。时间阈值公式如下：

$$T=\tau_{\mathrm{avg}}+\tau_{\mathrm{avg}}\times\sigma^*(\tau) \tag{6}$$

式中，T为根据新接收的告警计算出的动态时间阈值，τ_{avg}为所有告警日志时间差的均值，$\sigma^*(\tau)$为计算出的动态时间阈值更新系数。τ_{avg}和$\sigma^*(\tau)$的表达式为：

$$\tau_{\mathrm{avg}}=\frac{\sum_i \tau_i}{n-1}(i=1,2,\cdots,n-1);$$

$$\sigma^*(\tau)=\frac{\sigma(\tau)}{\tau_{\mathrm{avg}}},\sigma(\tau)=\sqrt{\frac{\sum_i(\tau_i-\tau_{\mathrm{avg}})^2}{n-1}}(i=1,2,\cdots,n-1)。$$

动态时间阈值可以根据每一次的新告警动态更新，告警日志聚合越多，动态时间阈值越趋近于真实值。

1.2.2 告警聚类

告警聚类分析算法描述如算法2所示。

算法2　告警聚类算法

输入：误告警过滤后日志$RA=\{ra_1,ra_2,\cdots,ra_n\}$；

输出：超级告警$SA=\{sa_1,sa_2,\cdots,sa_m\}$。

```
{
1. for i=1:n
2.     for j=1:m
3.         ipSim =s_IP(i,j); //计算IP相似度
4.         portSim =s_port(i,j); //计算端口相似度
5.         if(i.attackType==j.attackType)
6.             s_attack = 1;
7.         else s_attack = 0;
8.         //计算两条告警整体相似度
9.         attributeSim = w×s(i,j)(v,:);
10.    end for
11.    //判断告警时间间隔是否小于阈值
12.    if(timeInterval<T)
13.        SA.modify(j);
14.    else
15.        SA.add(i);
16. end for
}
```

1.3 攻击场景关联规则挖掘

1.3.1 攻击场景序列模式挖掘方法

攻击场景序列指攻击者完成一次多步攻击活动产生的具有时序关系的攻击类型序列，记为$ASS=\langle a_1,a_2,\cdots,a_n\rangle$，其中$a_i(1\le i\le n)$为第$i$个攻击步骤。攻击序列模式挖掘的目标是通过数据挖掘算法发现入侵事件不同攻击步骤间的关联关系。采用Apriori算法[21]挖掘攻击类型间的先后顺序。攻击场景序列模式挖掘过程为：

1）利用告警聚类算法，形成超级告警，提取所有超级告警的*attackTypeList*（即超级告警的攻击类型序列），得到攻击序列集R。设置Apriori算法最小支持度MS与最小置信度MC。

2）扫描R中所有攻击类型序列，得到长度为1的序列模式L_1，作为初始的种子集。

3）根据长度为i的种子集L_i，通过连接操作和剪切操作生成长度为$i+1$的候选攻击序列C_{i+1}，计算C_{i+1}的支持度$Support$，判断是否大于MS，找出频繁候选攻击序列。

4）重复步骤2）、3），直至无频繁候选攻击序列产生。并判断所有频繁候选集是否满足MC，生成攻击场景序列模式。

1.3.2 攻击场景关联规则

生成的攻击场景序列模式有可能是局部的，而不同的攻击场景序列模式可能会有交叉重叠的攻击步骤，因此，需要对挖掘出的攻击场景序列模式进行融合，形成更完整的攻击场景序列模式，并将攻击场景序列模式转换为关联规则，为后续关联分析提供依据。攻击场景关联规则形成过程为：

1）按照时序关系将每个攻击场景序列关系转化为树形图表示。

2）对每个树形图进行匹配融合。若重复节点均为叶节点，则不融合；若重复节点一个为叶节点，一个为非叶节点，则将非叶节点及子节点连接到树形图的相同叶节点上；若重复节点都为非叶节点，则根据重复非叶节点将其中一棵树连接到另一棵树。

3）循环，直到树形图无重复节点为止。

2 实验及结果分析

为验证本方法在真实网络环境数据上的有效性，采用四川大学一周的入侵防御系统（IPS）和Web应用防护系统（WAF）的真实告警作为数据集。其中，共有2 697 234条原始告警，误告警过滤的日志共有1 228 912条告警。误告警过滤日志数与原始日志数对比如图4所示，计算出平均误告警去除率为45%左右。对误告警过滤结果进行分析，在实验中检测到目的IP x.x.32.208与很多主机之间每天稳定地产生1 500条左右的告警日志，告警序列也具有明显的24 h周期变化。通过提取原始数据包，发现这些告警都是由某搜索引擎定期进行网页爬取产生的，属于正常行为。

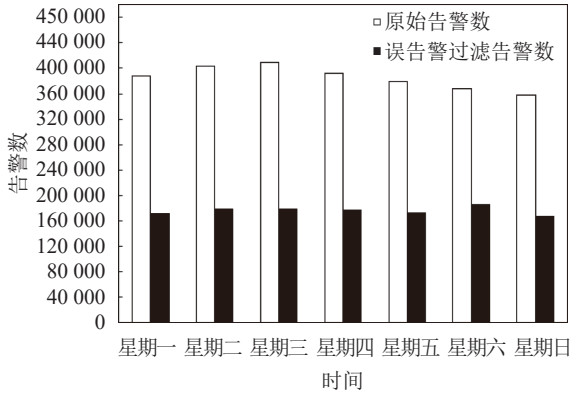


图 4 误告警过滤日志与原始日志对比

Fig. 4 Comparison of the number of false alarm filtering logs and the original number of logs

在误告警过滤后的日志中，攻击类型不为空的告警有291 459条，共45种攻击类型，表1为位于前5的攻击类型。

表 1 前5的攻击类型分布

Tab. 1 Top 5 attack type distribution

序号	攻击类型	数量	比例/%
1	僵尸网络	122 598	42.0
2	信息泄漏攻击	63 322	21.7
3	口令暴力破解	36 281	12.4
4	端口扫描	17 171	5.9
5	网站扫描	16 457	5.6

从上述告警数据中共得到25 353条超级告警，通过提取超级告警中的攻击类型序列，得到479条攻击场景序列。利用Apriori算法挖掘频繁攻击类型关联顺序，由于改变支持度和置信度值对产生的攻击场景序列数影响较大，分别测试了在置信度为70%、80%、90%时，支持度不同情况下挖掘的攻击场景序列模式数，如图5所示，从图5的结果得出，取置信度为80%、支持度为4%时攻击场景序列模式数较为稳定，产生的攻击场景序列模式为20条左右。

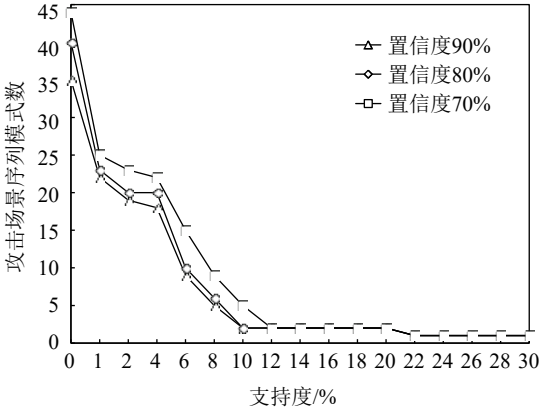


图 5　支持度–攻击场景序列模式数

Fig. 5　Support degree–attack scene sequence pattern number

图6为不同攻击场景序列模式进行融合后形成的攻击场景关联规则。每一个节点表示一种攻击类型，所有的攻击类型都从告警中提取出；每一条边为有向边，表示攻击类型间关联关系，每条边上标志数字表示来自不同的攻击场景序列模式，如“攻击序列模式1”表示攻击者实施了一系列对目的主机的攻击。攻击者在进行“网站扫描”后直接尝试利用“IIS畸形目录安全绕过漏洞”实施攻击，最后实现

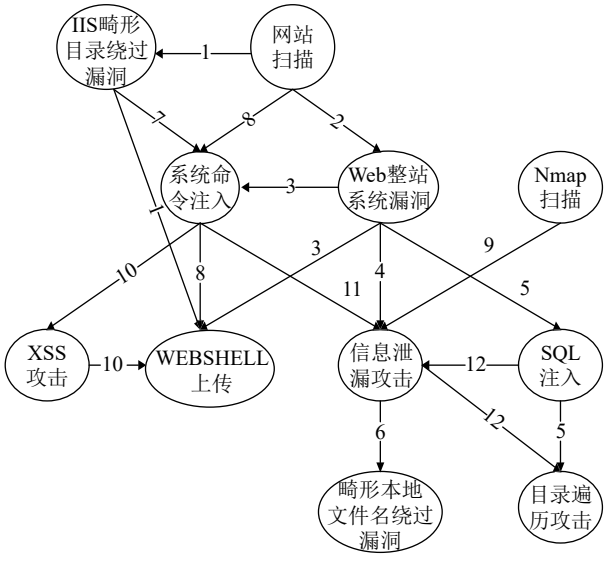


图 6　攻击场景关联规则

Fig. 6　Attack scene association rule

“WEBSHELL上传”。

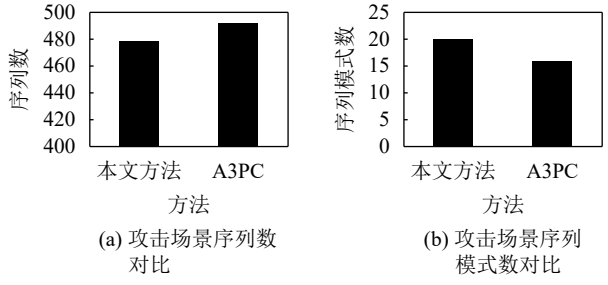
对所提出方法的有效性进行验证。采用同一周的IPS和WAF告警数据对比了本文方法和RTECA告警聚类方法[2]，误报过滤与告警聚类后告警数对比结果如表2所示。由表2可以得出，本文方法能够过滤部分的误告警，而RTECA并没有做相关分析。虽然，本文方法在告警聚合后告警数高于RTECA方法，但是，这是由于RTECA方法采用固定聚合时间阈值，阈值设置越高，聚合后告警越少，而本文聚合方法更接近真实攻击场景。

表 2　本文方法与RTECA方法的误报过滤与告警聚类后告警数对比

Tab. 2　Comparison of false alarm filtering and alarm clustering after the number of alarms by the proposed method and RTECA

时间	原始告警数	本文方法		RTECA	
		误报过滤后告警数	聚类后告警数	误报过滤后告警数	聚类后告警数
星期一	340 999	171 842	62 621	—	10 074
星期二	353 179	177 821	59 893	—	10 651
星期三	360 375	178 369	62 229	—	10 762
星期四	346 094	176 887	62 190	—	10 745
星期五	334 166	172 209	62 092	—	10 349
星期六	323 748	185 441	69 308	—	10 722
星期天	368 824	185 267	68 195	—	10 721
平均	346 769	178 262	63 789	—	10 574

考虑到本文方法与A3PC方法[16]都是采用Apriori算法挖掘攻击场景关联规则，进行对比实验，结果如图7所示。分析实验结果可知，A3PC方法提取的攻击场景序列数较多，这是由于采用固定时间窗口聚合，将原本属于同一攻击场景告警分裂为多个攻击场景序列。而本文方法挖掘出的频繁攻击场景序列模式比A3PC多，能够有效组合分裂的攻击链，将不同攻击序列进行融合，以形成更完整的攻击场景关联规则。因此本文方法更有优势。



(a) 攻击场景序列数对比　　(b) 攻击场景序列模式数对比

图 7　本文方法与A3PC方法的攻击场景关联规则挖掘对比

Fig. 7　Comparison of attack scene association rules mining by the proposed method and A3PC

3　结束语

在规范化安全设备告警数据的基础上提出一种

基于告警属性相似度聚类的攻击场景重构关联规则挖掘方法，采用数据挖掘的方法生成攻击场景序列模式，并将攻击场景序列模式进行融合形成攻击场景关联规则。最后通过在四川大学真实安全设备数据集中进行实验，验证了方法的有效性。与目前其他方法相比，本文方法能够形成更完整的攻击场景。在下一步研究中，将分析更多误告警特性，去除误告警；并基于攻击场景关联规则构建攻击场景，在实时环境中对多步攻击后续步骤进行预测研究。

参考文献:

[1] 国家计算机网络应急技术处理协调中心.CNCERT互联网安全威胁报告[EB/OL].[2019–01–20].http://www.cert.org.cn/publish/main/upload/File/CNCERT201901.pdf.

[2] Ramaki A A,Amini M,Atani R E,et al.RTECA:Real time episode correlation algorithm for multi-step attack scenarios detection[J].Computers & Security,2015,49:206–219.

[3] Navarro J,Deruyver A,Parrend P,et al.A systematic survey on multi-step attack detection[J].Computers & Security,2018,76:214–249.

[4] Wang Shuo,Tang Guangming,Kou Guang,et al.Attack path prediction method based on causal knowledge net[J].Journal on Communications,2016,37(10):188–198.[王硕,汤光明,寇广,等.基于因果知识网络的攻击路径预测方法[J].通信学报,2016,37(10):188–198.]

[5] Hu Hao,Liu Yuling,Zhang Hongqi,et al.Route prediction method for network intrusion using absorbing Markov chain[J].Journal of Computer Research and Development,2018,55(4):831–845.[胡浩,刘玉岭,张红旗,等.基于吸收Markov链的网络入侵路径预测方法[J].计算机研究与发展,2018,55(4):831–845.]

[6] Gong Jian,Zang Xiaodong,Su Qi,et al.Survey of network security situation awareness[J].Journal of Software,2017,28(4):1010–1026.[龚俭,臧小东,苏琪,等.网络安全态势感知综述[J].软件学报,2017,28(4):1010–1026.]

[7] Templeton S J,Levitt K.A requires/provides model for computer attacks[C]//Proceedings of the 2000 Workshop on New Security Paradigms.New York:ACM,2001:31–38.

[8] Ning P,Cui Y,Reeves D S.Constructing attack scenarios through correlation of intrusion alerts[C]//Proceedings of the 9th ACM Conference on Computer and Communications Security.New York:ACM,2002:245–254.

[9] Morin B,Mé L,Debar H,et al.A logic-based model to support alert correlation in intrusion detection[J].Information Fusion,2009,10(4):285–299.

[10] Fan Di,Liu Jing,Zhuang Junxi,et al.Research on attack scenario reconstruction method based on causal knowledge discovery[J].Chinese Journal of Network and Information Security,2017,3(4):58–68.[樊迪,刘静,庄俊玺,等.基于因果知识发现的攻击场景重构研究[J].网络与信息安全学报,2017,3(4):58–68.]

[11] Feng Xuewei,Wang Dongxia,Huang Minhuan,et al.A mining approach for causal knowledge in alert correlating based on the Markov property[J].Journal of Computer Research and Development,2014,51(11):2493–2504.[冯学伟,王东霞,黄敏桓,等.一种基于马尔可夫性质的因果知识挖掘方法[J].计算机研究与发展,2014,51(11):2493–2504.]

[12] Zhang Aifang,Li Zhitang,Li Dong,et al.Discovering novel multistage attack patterns in alert streams[C]//Proceedings of the 2007 International Conference on Networking,Architecture,and Storage (NAS 2007).Guilin:IEEE,2007:115–121.

[13] Ramaki A A,Khosravi-Farmad M,Bafghi A G,et al.Real time alert correlation and prediction using Bayesian networks[C]//Proceedings of the 2015 12th International Iranian Society of Cryptology Conference on Information Security and Cryptology (ISCISC).Rasht:IEEE,2015:98–103.

[14] Kavousi F,Akbari B.A Bayesian network-based approach for learning attack strategies from intrusion alerts[J].Security and Communication Networks,2014,7(5):833–853.

[15] Tian Zhihong,Zhang Yongzheng,Zhang Weizhe,et al.An adaptive alert correlation method based on pattern mining and clustering analysis[J].Journal of Computer Research and Development,2009,46(8):1304–1315.[田志宏,张永铮,张伟哲,等.基于模式挖掘和聚类分析的自适应告警关联[J].计算机研究与发展,2009,46(8):1304–1315.]

[16] Daneshgar F F,Abbaspour M.Extracting fuzzy attack patterns using an online fuzzy adaptive alert correlation framework[J].Security and Communication Networks,2016,9(14):2245–2260.

[17] Li Dong,Li Zhitang,Lei Jie.Research on the method of reducing false positives with periodicity[J].Journal of Chinese Computer Systems,2009,30(7):1336–1340.[李冬,李之棠,雷杰.周期性误告警去除方法研究[J].小型微型计算机系统,2009,30(7):1336–1340.]

[18] Ma Linru,Yang Lin,Wang Jianxin,et al.Using fuzzy clustering to reconstruct alert correlation graph of intrusion detection[J].Journal on Communications,2006,27(9):47–52.[马琳茹,杨林,王建新,等.利用模糊聚类实现入侵检测告警关联图的重构[J].通信学报,2006,27(9):47–52.]

[19] Valdes A,Skinner K.Probabilistic alert correlation[C]//Proceedings of the 4th International Workshop on Recent Advances in Intrusion Detection.London:Springer-Verlag,2001:54–68.

[20] Li Hongcheng,Wu Xiaoping.Multistage aggregation and correlation for network alerts based on self-extending time windows[J].Advanced Engineering Sciences,2017,49(1):206–212.[李洪成,吴晓平.基于自扩展时间窗的告警多级聚合与关联方法[J].工程科学与技术,2017,49(1):206–212.]

[21] Li Ning,Zeng Li,He Qing,et al.Parallel implementation of Apriori algorithm based on MapReduce[C]//Proceedings of the 2012 13th ACIS International Conference on Software Engineering,Artificial Intelligence,Networking and Parallel/Distributed Computing.Kyoto:IEEE,2012:236–241.

（编辑 赵 婧）

引用格式: Chen Xingshu,He Tao,Zeng Xuemei,et al.Research on attack scene association rule mining method based on alarm attributes clustering[J].Advanced Engineering Sciences,2019,51(3):144–150.[陈兴蜀,何涛,曾雪梅,等.基于告警属性聚类的攻击场景关联规则挖掘方法研究[J].工程科学与技术,2019,51(3):144–150.]