AI 技术应用

基于模糊关联规则的网络攻击溯源技术研究

于少中，于雷，张晨，赵蓓，刘胜兰
（中国移动通信集团设计院有限公司，北京 100080）

摘　要：随着计算机技术的发展以及应用的普及，互联网承载的服务越来越多，各类应用场景不断扩大，使得近年来网络攻击事件愈发频繁。黑客利用非法手段获取机密信息，破坏正常的网络环境，给人们的生产生活带来极大的危害。针对上述频繁的网络攻击行为，目前有多种安全防范措施，主要分为攻击前预防措施、攻击后检测与封堵措施以及攻击过程中的溯源反制措施。其中，攻击过程中的溯源反制措施相较于预防和检测封堵措施，能够更加有效地避免安全设备的漏洞，数据丢失风险更小，同时能够对攻击者起到威慑作用。但 IP 协议本身的缺陷，使攻击者可以通过地址伪造技术隐藏自己的真实身份，以往溯源措施大多利用路由器的入口调试功能来追踪攻击源，但一般情况下攻击者往往采用伪造地址或者间接攻击的方法来躲避追踪，这对溯源反制提出了较大的挑战。本文将采用模糊关联规则算法，借鉴主动防御网络攻击的思想实现对攻击行为的溯源分析。

关键词：网络攻击；模糊关联规则；溯源分析
中图分类号：TP393.0
文献标识码：A
doi：10.11959/j.issn.1000−0801.2021165

1　引言

随着信息化在各领域的深入应用，各行业与互联网的融合程度进一步加深，涉及敏感信息的海量数据分布在互联网中，这使得网络攻击事件愈发频繁。据统计，2020 年上半年信息诈骗、敏感信息泄露、病毒木马攻击、DoS 攻击等事件频发，有效的网络安全防范措施显得尤为重要。为了对网络攻击展开有效的防御，有关研究提出了对网络攻击进行追踪溯源的技术方法，用于对定位追踪攻击源头。

2　综述

2.1　相关技术背景

2.1.1　追踪溯源技术

目前，针对各类网络攻击行为，一些追踪溯源技术相继被提出。一般来说，按照溯源目标，可以将追踪溯源技术分为 4 类，分别是针对攻击主机的溯源、针对攻击控制主机的溯源、针对攻击者的溯源以及针对攻击组织的溯源。目前业界大多根据溯源目标来进行不同维度的研究工作。上述网络追踪溯源技术需要获取网络的拓扑结构作为支撑，但是由于现网环境中网络结构复杂多样，一般无法直接获取网络拓扑，因此网络追踪溯源存在较大的困难。目前较多研究为基于包标记的溯源方法，需要大量的数据包进行传送路径重构，通常只适用于 DoS 攻击。现网一般采用多跳中继代理服务器，这些代理服务器一般不在同一个区域，即使采用包标记方法来解决跨区域的问题，也会由于中继代理服务器的路由器路径过长而带来路径组合爆炸的问题。

2.1.2　模糊关联规则

通常的关联规则挖掘都是基于 Apriori 算法的

收稿日期：2021−05−01；修回日期：2021−07−15

思想[1]，模糊关联规则一般也是基于此类算法思想。模糊理论是将取值范围从{0,1}扩展到[0,1]概念的理论，假设域中存在元素 n 和集合 M，则在一般的集合理论中，n 与 M 的关系为：

$$L(M,n,N)=\begin{cases}1, n\in M\\ 0, n\notin M\end{cases} \tag{1}$$

在模糊理论中，假设 $f_x(n)$表示元素 n 对集合 M 的隶属度，其取值范围为[0,1]，则 n 与 M 的关系为：

$$L(M,n,N)=f_x(n), f_x(n)\in[0,1] \tag{2}$$

模糊集合通过将数据集合 $I=\{I_1,I_2,I_3,\cdots,I_n\}$ 划分成 K 类，并在每一类中设置一个聚类中心，设定 U 表示模糊隶属度矩阵，则模糊理论的目标函数可以表示为：

$$F_x(U,k_1,\cdots,k_n)=\sum_{i=1}^{n}\sum_{j=1}^{m}\left(u_{ij}\right)^w {d_{ij}}^2 \tag{3}$$

其中，d_{ij}表示第 i 个聚类区间的聚类中心与第 j 个数据之间的聚类，w 表示海量数据增量参数，u_{ij} 表示第 j 个数据与第 i 个聚类中心隶属度，需要满足：

$$\sum_{i=1}^{k}u_{ij}=1, u_{ij}\in[0,1], \forall j=1,2,\cdots,n \tag{4}$$

根据建立的目标函数，要保证目标函数取得最小值，从而完成模糊集合。

在实际的网络安全事件追踪溯源的过程中，一般会面临大规模数据处理以及数据分析的挑战。通常情况下，经典逻辑归纳假设法[2]依据的是数据的有效性和精确性，而现实情况中，网络攻击事件中的各类数据维度往往存在较高的非线性、复杂性以及模糊性，这时需要将模糊理论应用到大规模网络数据分析中才能有效降低数据维度，解决“硬区间”问题[3]。

2.1.3　蜜罐网络

蜜罐[4]是一种借鉴主动防御思想的网络资源，目的是利用自身已有的脆弱性来诱骗攻击者进行探测、攻击甚至破坏。蜜罐网络[5]一般是由大量蜜罐主机、路由器、防火墙、WAF 等组成的网络系统，目的是为攻击者提供一个接近真实的网络环境，诱骗攻击者对其发动攻击。

蜜罐网络结构如图 1 所示，出入陷阱网络的数据包都经过防火墙和路由器，防火墙的功能是控制内外网之间对陷阱网络的访问。路由器安放在防火墙和陷阱网络之间，路由器可以隐藏防火墙，即使攻击者控制了陷阱网络中的蜜罐主机，发现路由器与外部网连接，也能被防火墙发现。

2.2　研究现状与意义

国外针对网络取证[6-7]的研究起步较早，目前

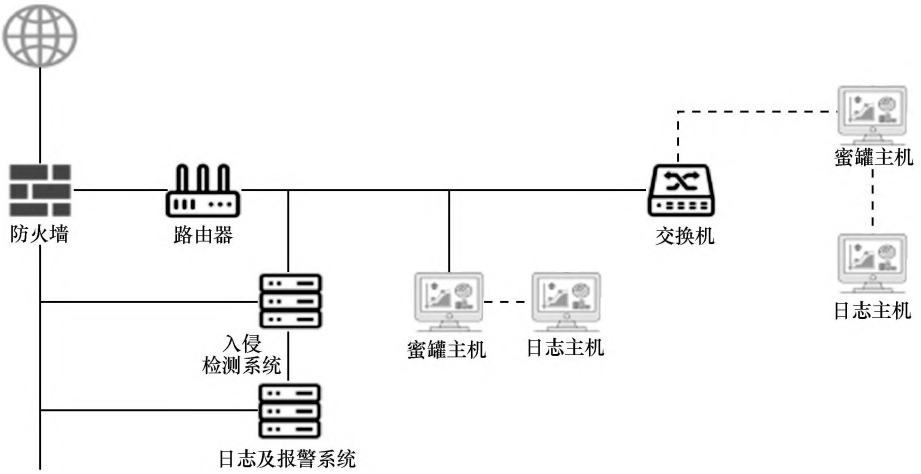

图 1　蜜罐网络结构

已经取得了很大的发展。网络攻击行为分析一直以来都是业内研究的重点，目前较为成熟的攻击行为分析技术就是网络入侵检测系统（NIDS）[8]，传统的网络入侵检测系统大多基于签名的检测。如今，对网络攻击行为的分析正在尝试检测常规行为定义、触发器偏离警告等信息。采用基于统计的方法，将待检测的数据与常规记录数据进行对比，通过偏离值打分的方法得到加权均值，当超过某个阈值时，判定为异常。此外，还可以采用机器学习的方法为各类数据打上标签，以此来进行分类。文献[9]提出利用时间序列来进行异常信息判定也是一种较为主流的分析技术，目前时间序列分析技术已经被应用于系统故障检测、垃圾邮件分类等领域。但是由于追踪溯源本身的复杂性，目前各类分析技术还不能高效精准地定位攻击源，尤其是面对网络环境中多维度的数据信息，往往会导致溯源过程的低效性。本文利用蜜罐网络采集网络攻击数据，采用模糊关联规则的方法，将网络攻击行为中大量信息进行降维分析，进而提高网络攻击溯源的效率。

3 基于模糊关联规则的溯源技术

3.1 技术原理

网络攻击溯源的根本目的是通过对网络中的攻击行为进行解析，找到各类攻击行为之间的逻辑关系，从而复盘攻击过程。利用模糊关联规则，进行攻击溯源的一般过程如图 2 所示。

基于模糊关联规则的攻击溯源步骤如下。

（1）首先选定网络流量中的关键信息。

（2）判断当前处理的流量所含关键信息是否满足最小置信度，如果不满足则不处理当前流量，如果是有效流量，则进行持久化。

（3）在持久化的流量中进行源 IP 或域名存活判定，如果不满足，则将该持久化记录进行休眠处理。

（4）确定用于最小置信度判定的关键信息，并进行最小置信度判定。

（5）挖掘网络数据中关键信息是否具有关联的属性信息。

（6）对这些属性信息进行模糊处理，将模糊处理得到的样本变量通过隶属度函数，完成模糊

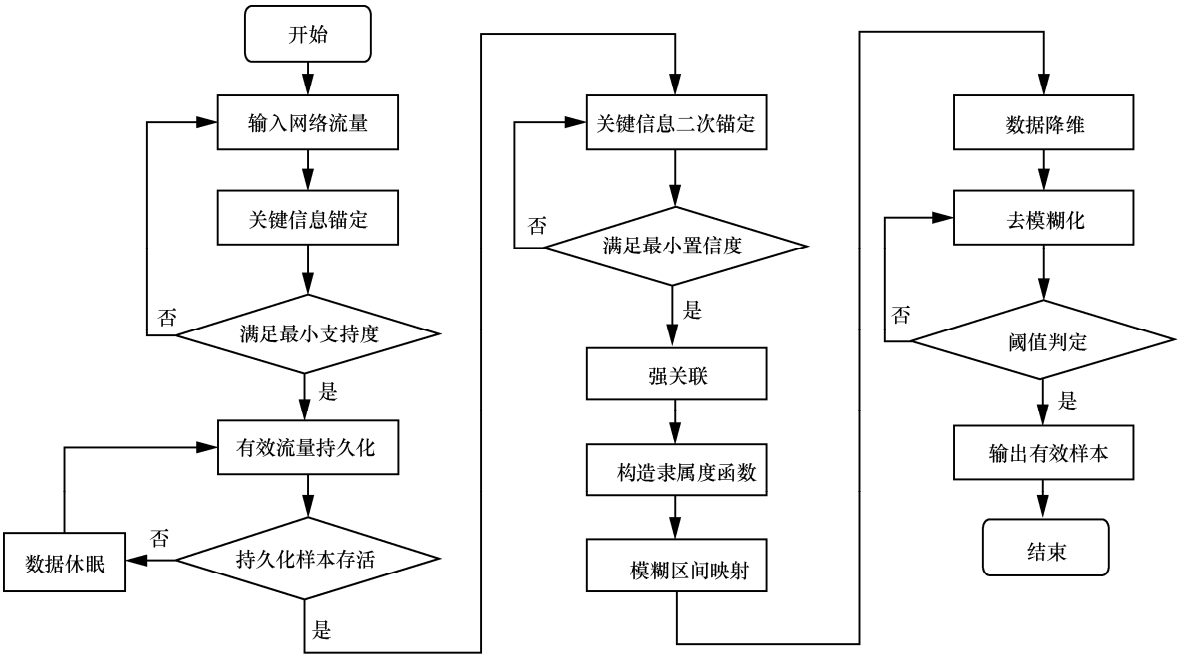

图 2 模糊关联规则溯源技术原理

样本与模糊值的映射关系。

（7）进行离散化处理，将连续的模糊值映射到模糊集合空间，确定网络攻击行为中各属性值的对应关系。

（8）尝试复盘整个网络攻击过程。

3.2　模糊关联规则算法流程

模糊关联规则就是将模糊理论引入关联规则中，通过模糊关联规则实现海量数据的快速分析与挖掘。模糊关联规则的应用一般包括原始数据模糊化、数据分类聚类、数据离散化、关联规则分析 4 个阶段。模糊关联规则算法流程如图 3 所示。

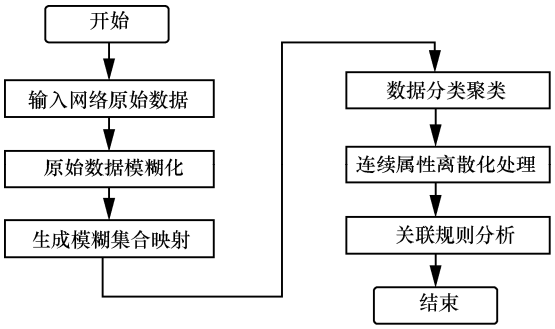

图 3　模糊关联规则算法流程

原始数据模糊化。首先要建立模糊集合与模糊隶属函数，这个过程其实是将量化属性映射到模糊集合，即将数据集中的属性都看作一个模糊变量，每一个变量都对应多个模糊值，这个模糊值与模糊变量的关系由一个模糊隶属函数来描述。

数据分类聚类。通过模糊隶属函数计算后得到大量的模糊数据，将这些模糊数据分为多个类别，根据每个类别的相似度对这些模糊函数进行加权，然后利用分类器对大量的数据样本进行预测，确定分类器的准确性，如果有错分的样本，则给予更高的权重。通过上述方法完成模糊数据分类器的构建，并在此基础上进行基于模型的聚类。

数据离散化。为保证数据分析过程中的高效性和准确性，需要对模糊数据进行离散化处理。这个过程将特定的连续属性值划分成多个模糊离散化区间，然后通过隶属度函数对模糊值进行离散空间归域，进而实现模糊数据离散化处理。

关联规则分析。对数据离散化处理后得到模糊后的数据，通常这类数据也被称为降维后的数据，利用这些数据可以有效提高数据挖掘以及统计分析的效率。

3.3　网络攻击溯源过程

溯源过程一般需要在一个网络拓扑中的任意两个路由之间部署追踪溯源系统[10]，该系统会记录流经的数据流信息，并将数据存储在系统的存储空间[11]中。当提交溯源请求时，溯源系统将在存储空间中对提交的溯源信息进行检查，如果发现匹配数据，则判定该溯源信息曾流经本系统，并将匹配数据添加到溯源路径中，继续进行下一级查询。

如图 4 所示，H 为攻击者，R 为攻击数据流经的路由器，攻击者 H 发出的攻击报文流经 R10、R6、R3、R2 之后到达受害者 C。在图 4 中，每两个相邻的路由器之间都会部署一个溯源系统 S，这些系统时刻都在记录着流经自身的数据包信息。当

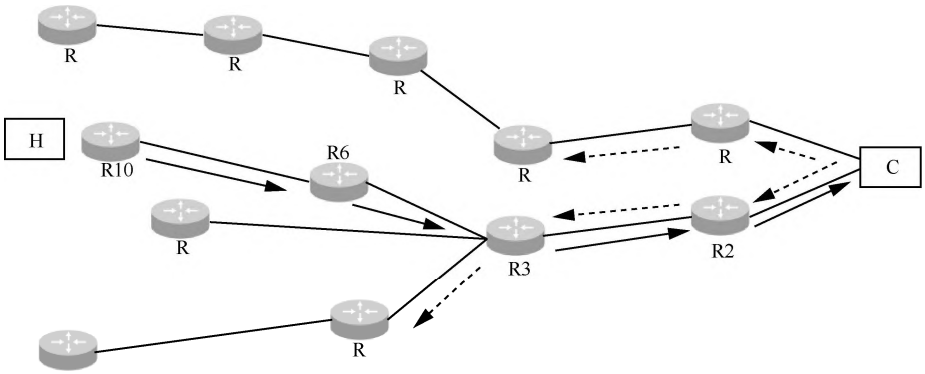

图 4　蜜罐网络各类业务部署效果

受害者 C 检测到攻击时，其将向相邻的溯源系统发出查询请求，当溯源系统 S 收到查询请求后，将通过系统的存储结构逐一判定查询单位是否流经当前系统，然后将结果添加到溯源路径中，最终完成攻击路径的重构。

4 仿真分析

为了验证模糊关联规则算法在网络攻击溯源中的可行性，本节将进行基于模糊关联规则的攻击溯源实验，实现对网络攻击过程的复盘。

4.1 实验环境

本实验采用基于Golang开发的跨平台蜜罐平台，将其分别部署在 5 台 Centos 服务器上，每台服务器开放了多种服务环境，包括 SSH、SFTP、Redis、MySQL、FTP、Telnet 等，此外还模拟真实业务系统部署了大量的 Web 系统，如图 5 所示。

其中，每台服务器之间存在业务连通性，若有黑客访问某些业务系统，相关记录将会共享在其他的服务器存储空间中。实验环境服务器分布效果如图 6 所示。

4.2 攻击溯源与结果分析

4.2.1 实验数据

实验数据采用自行部署的蜜罐网络采集到的网路流量数据，收集到包含各类网络操作行为的数据。由于蜜罐网络中存在业务交互系统，实验人员会模拟正常操作流程对业务系统进行访问，同时由于蜜罐网络部署在公网，每天将获取海量的访问请求，这些访问请求大部分是具有攻击性的。实验数据共有 20 个维度特征，见表 1。实验数据效果如图 7 所示。

图 5 蜜罐网络各类业务部署效果

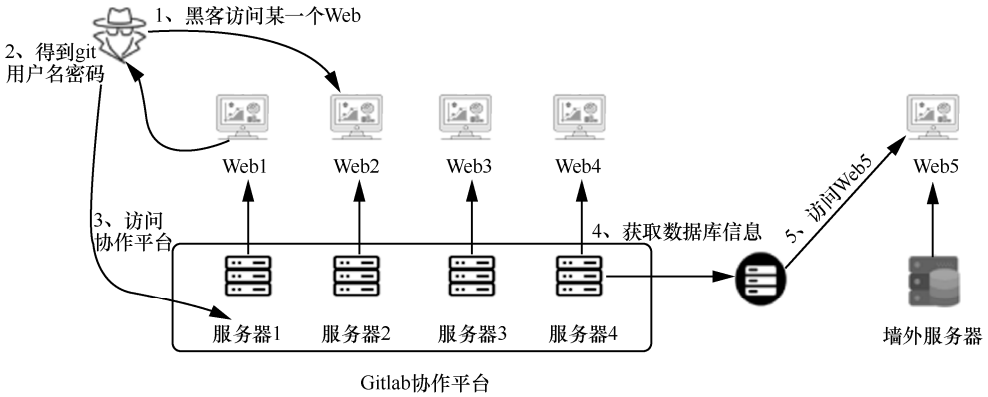

图 6 实验环境服务器分布效果

表 1　实验数据特征

编号	描述	编号	描述	编号	描述	编号	描述
n1	连接持续时间	n2	协议类型	n3	目标主机服务类型	n4	连接状态
n5	传输字节数	n6	是否同一端口	n7	敏感文件访问次数	n8	尝试登录次数
n9	是否登录成功	n10	是否获得 root	n11	是否出现 sudo su	n12	文件创建次数
n13	该问控制文件	n14	FTP 连接次数	n15	是否 guest 目录	n16	错误 SYN 占比
n17	错误 REJ 占比	n18	2 s 内相同服务数	n19	2 s 内不同服务数	n20	相同服务总数

图 7 实验数据效果

4.2.2　攻击溯源实验与结果分析

选取 20 000 条业务访问记录用作测试，利用 Python 作为工具语言进行仿真，使用基于模糊关联规则的网络攻击溯源技术对攻击过程进行还原。

```
conn = sqlite3.connect("X.db")
cursor = conn.cursor()
hosts = {}
for row in cursor.execute('SELECT * FROM events'):
    host = row[2].split(':')[0]
    if host in hosts:
        hosts[host] += 1
    else:
        hosts[host] = 1
sorted_dict = sorted(hosts.items(), key=lambda item: item[1], reverse=True)
for host in sorted_dict:
    print(host)
```

利用上述模糊关联算法对样本数据进行分析，在实验结果中发现，已将 20 个样本维度的弱相关规则进行了剥离，并分别进行了强相关匹配，利用上述规则可将维度有效降低。强相关匹配规则示例如图 8 所示。实验数据特征见表 2。数据降维攻击类别判定示例如图 9 所示。

```
5 ('84.75.207.23', 173) n3,n4,n5,n9,n10,n11,n14,n16
6 ('223.90.121.53', 106) n1,n2,n3,n4,n8,n14,n17
7 ('121.37.231.58', 74) n2,n3,n5,n6,n7,n13
8 ('185.234.218.42', 57) n1,n2,n4,n9,n19
9 ('188.217.244.81', 44) n4,n5,n6,n7,n8,n10,n20
11 ('2.37.182.228', 34) n1,n2,n3,n4,n5,n6,n14,n15,n16,n17,n20
12 ('66.102.6.186', 25) n1,n2,n4,n10,n12,n13,n14,n17
13 ('172.104.108.109', 24) n1,n2,n4,n9,n19
14 ('66.102.6.182', 23) n2,n3,n6,n7,n8,n9,n10,n14,n15,n17,n18
15 ('66.102.6.184', 22) n1,n2,n3,n4,n8,n14,n17
16 ('116.109.194.44', 21) n1,n2,n4,n5,n7,n8,n9,n14,n17,n19
17 ('66.102.6.42', 13) n1,n2,n4,n9,n19
18 ('139.162.106.181', 13) n1,n2,n4,n9,n19
19 ('66.102.6.46', 12) n1,n2,n5,n6,n7,n9,n10,n13,n15
20 ('128.14.134.170', 12) n3,n4,n5,n9,n10,n11,n14,n16
21 ('193.118.53.194', 11) n1,n2,n6,n10,n12,n13,n14
22 ('134.209.185.206', 11) n1,n2,n4,n8,n9,n10,n14,n15
23 ('94.177.214.123', 11) n1,n2,n4,n9,n11,n13,n14,n15,n16,n17
24 ('134.122.57.107', 11) n1,n3,n4,n8,n10,n11,n13,n14,n15,n19
25 ('66.102.6.44', 10) n2,n3,n4,n6,n8,n9,n10,n11,n18,n19,n20
```

图 8　强相关匹配规则示例

表 2　实验数据特征

编号	描述	编号	描述	编号	描述
n1	连接持续时间	n2	协议类型	n6	是否同一端口
n7	敏感文件访问次数	n8	尝试登录次数	n9	是否登录成功
n10	是否获得 root	n13	该问控制文件	n14	FTP 连接次数
n16	错误 SYN 占比	n18	2 s 内相同服务数		

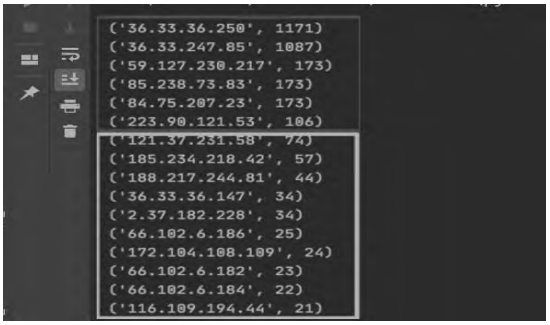

图 9　数据降维攻击类别判定示例

通过上述规则，可以看出在某段时间内虽然有大量来自不同域的主机对蜜罐网络发起了攻击，但是通过强相关匹配，可以将许多“肉鸡”排除。此外，根据特征维度可以判断某些 IP 地址的真实意图，例如 DoS 攻击中是通过 ICMP 数据包广播攻击还是通过 SMTP 邮件攻击。攻击溯源判定示例、攻击溯源态势感知分别如图 10、图 11 所示。

项目	集群名称	来源 IP	地理信息	信息	长度	上钩时间
TELNET Telnet蜜罐	本机	187.162.244.7	墨西哥	点击查看	24	2020-07-30 14:46:22
TELNET Telnet蜜罐	本机	150.116.32.153	中国 台湾 台北市	点击查看	25	2020-07-30 14:42:25
VNC VNC蜜罐	本机	153.101.29.58	中国 江苏 常州	点击查看	8	2020-07-30 14:28:26
TELNET Telnet蜜罐	本机	211.217.33.236	韩国	点击查看	24	2020-07-30 14:08:42
VNC VNC蜜罐	本机	153.101.29.58	中国 江苏 常州	点击查看	8	2020-07-30 14:04:53
REDIS Redis蜜罐	本机	106.52.221.116	中国 广东 广州	点击查看	31	2020-07-30 13:59:59
TELNET Telnet蜜罐	本机	14.33.59.147	韩国	点击查看	23	2020-07-30 13:53:52
TELNET Telnet蜜罐	本机	196.251.49.4	南非	点击查看	23	2020-07-30 13:46:43
MYSQL Mysql蜜罐	本机	47.100.64.86	中国 上海 上海	点击查看	23	2020-07-30 13:40:16
TELNET Telnet蜜罐	本机	196.251.49.61	南非	点击查看	24	2020-07-30 13:32:10

图 10　攻击溯源判定示例

图 11　攻击溯源态势感知

攻击溯源过程的重点就是查找攻击事件中的关联性，利用关联性来分析攻击行为的特点。本实验应用模糊关联规则，利用蜜罐采集攻击行为数据的时间关联性、特征行为关联性以及各类因果关联性，复盘网络攻击的整个过程。

5　结束语

本文提出了基于模糊关联规则的网络攻击溯源技术，首先分析了现阶段网络攻击溯源技术的研究现状，指出了目前溯源技术的不足，然后通过分析模糊关联规则的应用场景，验证了本文技术应用于网络攻击溯源的可行性。此外，本文还通过自建蜜罐网络，采集网络访问流量，应用模糊关联规则将样本数据进行处理，通过构建隶属度函数获取样本变量与模糊区间的映射关系，达到降维目的。实验结果表明，模糊关联规则可以有效实现网络攻击溯源，确定网络攻击源头。

下一步的研究重点将尝试将蜜罐网络应用于工业控制系统，通过本文的方法对工业控制系统的攻击行为进行溯源研究。

参考文献：

[1] 张继荣, 王向阳. 基于 XML 数据挖掘的 Apriori 算法的研究与改进[J]. 计算机测量与控制, 2016, 24(6): 178-180, 188.
ZHANG J R, WANG X Y. Research and improvement of apriori algorithm for XML data mining[J]. Computer Measurement & Control, 2016, 24(6): 178-180, 188.

[2] MARCHESE M, SURLINELLI R, ZAPPATORE S. Monitoring unauthorized Internet accesses through a ‘honeypot’ system[J]. International Journal of Communication Systems, 2011, 24(1): 75-93.

[3] PRASAD K M, KARTHIK M G, KRISHNA E S P. An efficient flash crowd attack detection to Internet threat monitors (ITM) using honeypots[C]//Advances in Computing and Information Technology, [S.l.:s.n.]. 2013: 177: 595-610.

[4] 杨德全, 刘卫民, 俞宙. 基于蜜罐的主动防御应用研究[J]. 网络与信息安全学报, 2018, 4(1): 57-62, 78.
YANG D Q, LIU W M, YU Z. Research on active defense application based on honeypot[J]. Chinese Journal of Network and Information Security, 2018, 4(1): 57-62, 78.

[5] 贾召鹏, 方滨兴, 刘潮歌, 等. 网络欺骗技术综述[J]. 通信学报, 2017, 38(12): 128-143.
JIA Z P, FANG B X, LIU C G, et al. Survey on cyber deception[J]. Journal on Communications, 2017, 38(12): 128-143.

[6] GUPTA S, GUPTA B B. Cross-Site Scripting (XSS) attacks and defense mechanisms: classification and state-of-the-art[J]. International Journal of System Assurance Engineering and Management, 2017, 8(1): 512-530.

[7] WANG W, LIU J Q, PITSILIS G, et al. Abstracting massive data for lightweight intrusion detection in computer networks[J]. Information Sciences, 2018, 433/434: 417-430.

[8] 尚进, 谢军, 蒋东毅, 等. 现代网络安全架构异常行为分析模型研究[J]. 信息网络安全, 2015(9): 15-19.
SHANG J, XIE J, JIANG D Y, et al. Research on abnormal behavior analysis of modern networking security architecture[J]. Netinfo Security, 2015(9): 15-19.

[9] NEUNER S, SCHMIEDECKER M, WEIPPL E R. PeekaTorrent: leveraging P2P hash values for digital forensics[J]. Digital Investigation, 2016, 18: S149-S156.

[10] DE LIMA I V M, DEGASPARI J A, SOBRAL J B M. Intrusion

detection through artificial neural networks[C]//Network Operations and Management Symposium. Piscataway: IEEE Press, 2008: 867-870.

[11] 郭征, 吴向前, 刘胜全. 针对校园网 ARP 攻击的主动防护方案[J]. 计算机工程, 2011, 37(5): 181-183.
GUO Z, WU X Q, LIU S Q. Active protection scheme against ARP attack in campus network[J]. Computer Engineering, 2011, 37(5): 181-183.

[作者简介]

于少中（1992− ），男，中国移动通信集团设计院有限公司咨询设计师，主要研究方向为信息与网络安全。

于雷（1991− ），男，中国移动通信集团设计院有限公司工程师，主要研究方向为信息与网络安全。

张晨（1980− ），男，中国移动通信集团设计院有限公司高级工程师，主要研究方向为信息安全、网络安全、内容安全、大数据分析。

赵蓓（1974− ），女，中国移动通信集团设计院有限公司教授级高级工程师，主要研究方向为网络安全、业务安全、数据安全及新技术安全。

刘胜兰（1994− ），中国移动通信集团设计院有限公司咨询设计师，主要研究方向为信息与网络安全。