



面向大数据的异构网络安全监控及关联算法研究^{*}

刘 兰¹, 林 军², 蔡 君¹

(1. 广东技术师范学院 广州 510655; 2. 工业和信息化部电子第五研究所 广州 510610)

摘 要: 我国信息安全战略在大数据时代需考虑大规模、异构的网络安全行为的复杂性和时效性问题。针对大数据具有的数据量巨大、查询分析复杂的特点, 分析面向大数据的异构网络安全监控的相关技术, 提出对各类异构数据源进行清洗整合, 通过安全事件的关联和分布式序列图模式等方式进行网络安全监控的整体态势分析, 为大数据环境下的异构网络安全监控提供了一些思路, 也为信息安全企业提供了一个分析大数据下隐含规律的参考模型。

关键词: 大数据; 关联算法; 数据模型; 信息安全

doi: 10.3969/j.issn.1000-0801.2014.07.013

Research on Network Security Monitoring Model and Associated Algorithm in the Age of Big Data

Liu Lan¹, Lin Jun², Cai Jun¹

(1. Guangdong Polytechnic Normal University, Guangzhou 510655, China;

2. China Electronic Product Reliability and Environmental Testing Research Institute, MIIT, Guangzhou 510610, China)

Abstract: In the age of big data, large-scale, complexity of heterogeneous network security behavior should be considered. According to the features of huge amount and complex, big data analysis technologies for network security monitoring were proposed. Various types of heterogeneous data sources by data cleaning, and the key data through security event correlation, traffic rules based on wavelets and distributed sequence diagram model and other methods, were analyzed. Some ideas of network security monitoring system for big data environment were provided.

Key words: big data, correlation algorithm, data model, information security

1 引言

网络安全技术经过多年的发展, 已相对成熟, 然而目前处在大数据时代, 要从一个大型网络中扫描并且发现真正的安全问题可能需要很长的时间。大数据时代, 数据以几何速度增长, 现有的安全措施不能满足信息化发展的需要。

大数据已成为网络信息安全领域必须面对的事实^[1-3], 怎样处理异构的网络安全设备所产生的数据量庞大的安全数据, 如何分析及处理大数据量的网络事件, 这些都是信息网络安全需要思考的问题。构建一个面向大数据环境的网络安全监控体系, 分析和协调异构、大数据量的各类安全事件, 根据结果调整相应的安全策略, 使整体网络安全适应大数据网络环境的动态性和全局性要求。

^{*} 广东高校优秀青年创新人才培养项目(No.2012lym_0088), 广东省自然科学基金资助项目(No.S2013040014339), 广东省教育厅科研基金资助项目(No.2013KJ CX0120)

2 面向大数据的异构网络安全监控新需求

2.1 大数据的特点

大数据有4个特点,被总结为4V特性:volume、velocity、variety、veracity。volume指大数据的数据量巨大,从TB级别跃升到PB级别;velocity指大数据要求对数据的处理速度要快,能及时反映数据的真实意义,这一点是大数据与传统的数据挖掘技术最不同的地方;variety指大数据的数据来源种类繁多,包括日志、图片、音频、视频及地理信息等;veracity指大数据的数据中可能有用的实际数据所占源数据的比重比较少,也就是需要从价值密度很低的大量数据中挖掘出有价值的信息。

大数据作为时下热门的IT词汇,许多围绕大数据商业价值的技术(如数据仓库、数据安全、数据分析及大数据的挖掘等)被科学研究者深入探讨^[2,3]。在RSA2012会议上,有很多信息安全公司提出以大数据分析为基础的理论及技术,这些研究中,有的是在大数据分析的技术层面做出改进,有的是从大数据分析的应用层面提出解决方案。对于信息安全业界,其安全数据量随着安全事件的发生越来越多,呈几何级数增长,这也体现了大数据的4V特点。

国内外对于大数据处理技术均非常重视。美国在2012年3月底就发布了大数据研发专项研究计划,投入2亿多美元研究开发科学探索、生物医学及国家信息安全等重大领域所需的大数据处理技术和相关工具,将大数据研究提升到国家发展战略的高度。IBM公司投入大量资金用于大数据的开发利用;EMC则提出了数据科学的概念。2011年底,在我国工业和信息化部发布的物联网“十二五”规划中,信息处理技术作为关键技术创新工程被提出,其包括海量数据的存储、数据挖掘和视频智能分析,这些均为大数据的重要组成部分,其提出的信息感知、信息安全和信息传输技术均与大数据密切相关,可见我国对于大数据的重视程度。

2.2 面向大数据的网络安全监测新需求

预测是大数据面对的最核心的科学问题。目前学术界关心的预测问题包括两类:一类是趋势预测,指通过事物的基本属性和前期的态势分析,预测出事物的发展轨迹和可能的最终发展态势;另一类是缺失信息预测,指观察到的信息只是全部真实信息中很小的一部分,研究在已知信息的基础上,如何利用当前信息预测未观察

到的信息。

大数据时代,异构网络安全监控系统也应从这两种预测角度进行研究,考虑怎样通过已发生事件的种种迹象,预测真正安全事件和安全态势,并考虑通过已发生的安全事件全局分析事件对整个网络安全环境的影响。基于目前已经相对成熟的各种安全技术(如IDS/IPS、防火墙及防病毒等),面向大数据的异构网络安全监控系统应满足如下需求:第一,应具有快速反应能力,能够迅速对异常的安全事件报警采取相应措施,减少事件对网络服务的影响,这实际上就是趋势预测的一种应用^[4,5];第二,应具有从大量告警数据中提取核心、非冗余数据的能力,能够及时找出问题所在并进行解决。大数据环境下的异构网络安全监控体系,其误报和漏报是一对难以解决的矛盾。在获得的警报信息中,难免会有一些缺失,这里实际上可以应用大数据的缺失信息预测,获得真正的网络攻击或网络异常的全貌。

大数据要求数据能够充分发挥其外部性,并通过与某些相关数据进行交叉合并分析,产生远大于简单叠加的巨大价值。对于大数据环境下的异构网络安全监控,最重要的环节是事件的关联。网络环境中数量巨大的网络设备和安全设备产生了大量安全事件,国内外对于传统事件关联有很多科研成果,如贝叶斯网络推理理论、基于规则的关联规则、人工神经网络等^[6]。这些分析方法是面向异常检测系统提出的,没有针对整个网络安全的范围,因此缺乏完整性和全局观,对于整个信息网络的监控缺少各个设备数据之间的有效联系,并不能联合进行数据分析和反应。在大数据环境下,网络安全监控的发展需要适应这种变化,新的研究应该从架构、算法等方面进行考虑,在尽量不损失数据价值的条件下,清洗数据,缩减核心数据的规模,并在此基础上进行数据关联,发现网络安全信息域中的异常行为并进行预警。

3 面向大数据的异构网络安全监控体系模型

3.1 新环境下网络安全监控体系的方向

大数据环境下,安全事件数据量迅速增加,这些事件数据充分体现了大数据的“4V”特性,在面向大数据环境的异构网络安全监控模型的研究中,应涉及以下几个方面。

从海量数据源中收集数据,数据来源应考虑整个网络环境的网络设备、安全设备、主机和服务器的日志数



据、事件信息和告警信息,应考虑提供 SNMP、OPSEC、ODBC、Log Files、Socket、Serials Port、Cisco、SDK、NAP 等接口能力,将不同厂商、不同产品产生的海量日志信息进行统一收集。

大数据时代,数据是随处可见的,所缺少的是从大量纷繁复杂的数据中提取知识的能力。通过对安全事件数据的去冗和整合,清理数据格式,去除过时、冗余的数据,实现对数据的进一步清洗。此时,通过设置用户对特殊安全事件属性的关注度调整,丢弃对自定义特别属性不重要的安全事件,对特别关注的安全事件进行特别标记。

大数据收集是希望根据需求从海量数据中提取出有用的知识,并将其应用到实际领域中。经过前期的处理,大数据的关联成为面向大数据异构网络安全监控系统的重要环节。而大数据环境下的异构网络安全监控系统的最终目标是实现对全网安全事件的处理,实现实时信息安全行为的可视和量化。

3.2 体系模型

现有网络安全监控的事件分析能力不够,事件的监测数据主要是较原始的数据,无法进行知识化处理。大数据环境下各种安全技术的发展,导致整个网络安全环境数据异构,对整体网络的安全态势没有准确、及时的数据分析。数量巨大的安全事件中包含着大量不可靠信息,这使得大数据环境下信息安全的原始事件数据价值很低。因此,以往的一些算法存在适应性的问题,算法的时间复杂度偏高、准确率偏低,不能适应大数据的环境,有相关研究表明,一些简单的算法对于大数据的处理反而很有效^[6]。

面向大数据环境的安全态势发展需要,在以往研究的基础上,对网络安全监控体系进行分析,提出一种高效、标准的体系结构^[5-7],系统模型由4个模块组成,如图1所示。

(1) 大数据收集模块

此模块对当前信息网络安全环境中的各种数据源进行收集,采集源数据,原始数据包括信息网络系统中的各种网络设备、安全设备、主机和服务器的日志数据、事件信息和告警信息;提供对 SNMP、OPSEC、ODBC、Log Files、Socket、Serials Port、Cisco、SDK、NAP 等的接口能力。相关研究表明,大数据集上的简单算法比小数据集上的复杂算法能产生更好更高效的结果。

(2) 大数据整合模块

在大数据收集模块中得到的数据大部分是凌乱的,各

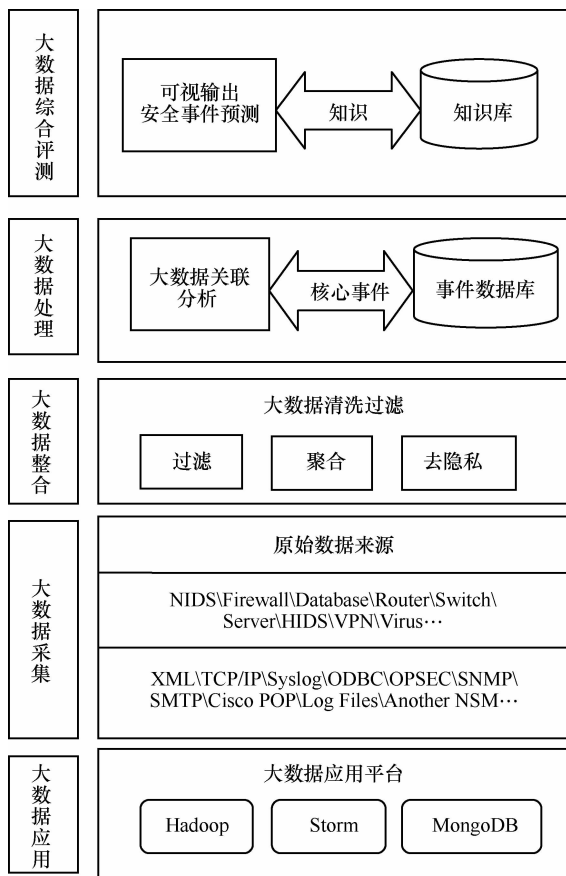


图1 体系模型结构

种各样的信息源产生的数据格式不一,这给监测系统处理带来很大的不便,用大数据整合模块来处理这一问题,对源数据进行过滤、去冗处理,根据用户设定的相应分类规则进行数据分类。在这个过程中也考虑大数据中隐私信息的去隐私化处理,在由源数据获得的隐私数据中提取索引字段,根据去隐私映射关系表中的映射规则实现对隐私数据的去隐私化。由于大数据的非结构化特性,其工作任务任重道远。这里需要研究大数据的清洗过滤和质量管理技术。

(3) 大数据关联分析模块

经过前面两个步骤处理的数据虽然去除了冗余,但对于整个网络安全监控系统来说仍然只能称为初始数据。真正需要找出的核心数据可以根据整个网络拓扑环境、事件发生的时间、频度等关联分析得出。关联分析模块的工作包括资产信息关联、网络拓扑关联、攻击模式关联、事件和已有安全漏洞关联、事件和规则库中行为模式的关联等。关联信息包括对实时事件与历史趋势的对比分析,从而发现真实的攻击行为。大数据关联分析模块的分析能力取决

于其事件规则库,事件库中存放相应的安全漏洞、异常行为模式和攻击场景等,当网络安全事件发生时,大数据关联分析模块通过对核心数据的分析细化,把事件映射到对应的逻辑拓扑中,对事件进行追踪。通过解析安全事件的行为流程,鉴别真正的攻击行为并定位安全事件发生的位置。大数据关联分析模块中的最重点和最难点在于算法和事件规则库的表示,就此提出一些想法,从并行关联规则和流量规则等方面着手解决大数据集的挖掘问题。

(4) 大数据综合评测模块

此模块将分析出来的结果进行综合处理,给出可视化的输出显示,通过大数据分析预测当前的网络安全状况、发展趋势等,根据分析结果采取相应的措施,制定相应流程,反馈给特定的管理员。在此模块可方便地关联查询相关资产的漏洞和风险库,为事件管理员提供帮助和指导信息。

4 关联算法及知识表示

大数据的数据量巨大,但希望能提炼出核心的数据,这要求数据组织好、数据内在的各种内容和关联清晰,并且容易获取和调用,研究人员和开发人员可以自如地在处理后的数据中获取数据间的逻辑片段。对于面向大数据的异构网络安全监控系统而言,重点和难点在于数据的关联分析。

面向大数据环境的异构网络安全监控系统的关联算法需要从以下几个方面进行考虑。

首先,需要考虑各种处理后的核心数据中不同数据项之间的关联关系,将之称为规则关联,研究中使用安全事件模糊等量约束的因果关联算法对其进行处理;其次,由于面向大数据环境的异构网络安全监控系统研究的对象是整个网络,网络中各种相关设备的流量信息也需要重点分析,发现其中的规律并提取流量规则;最后,异构的各种设备核心数据和设备流量数据之间的相关性亦需要协调分析,这种规则间关联使用分布式序列图模式挖掘。

4.1 事件模糊等量约束因果关联

面向大数据的异构网络安全监控系统中,通过关联分析、相关的模式挖掘方式可以发现清洗过的网络数据的系统属性之间的联系^[8,9]。在以往研究的基础上,提出大数据通过去除冗余和规范化处理,并聚集数据导出新属性,用四元组 (A, P, C, O) 描述和分析安全事件之间的因果关系,其

中 A 表示安全事件名称, P 表示安全事件所需的前提条件集合, C 设定为安全事件以后可能产生的结果集合,记做 $C(A)$, O 为与安全事件相关的一组属性的集合,比如 $O = \{SIP, SPORT, DIP, DPORT\}$, 属性的不同取值代表不同事件类别具有不同的实例。

研究如何对处理后的大数据分析所用的约束进行分类,提出模糊等量约束的因果关联算法,在安全事件集不完备和安全事件多样化的情况下,根据约束的强制程度改进安全事件因果关联计算的性能,发现事件间潜在的因果关联关系,从而预测事件的发生。

根据模糊理论,对任意安全事件 e_1 和 e_2 , 引入 $C(e_1) \times P(e_2)$ 上的二元模糊因果关系 $\tilde{R}$, 其定义了事件 e_1 与 e_2 之间是否具有因果关联关系: $\tilde{R}$ 是 $C(e_1) \times P(e_2)$ 上的模糊集合, 定义隶属度函数 $\mu_{\tilde{R}}(c, p)$, 其取值范围为实轴闭区间 $[0, 1]$, 它的大小表示序偶 (c, p) 在模糊集合 $\tilde{R}$ 上的隶属度, 也就是 c 和 p 具有模糊因果关系的程度大小。

设对于处理后的高级安全事件 $e_1, e_2, A_1 = \{u_1, u_2, \dots, u_n\}$ 为事件 e_1 的属性集, $A_2 = \{v_1, v_2, \dots, v_m\}$ 为事件 e_2 的属性集, $B_1 = \{u_1, u_2, \dots, u_q\}$ 和 $B_2 = \{v_1, v_2, \dots, v_q\}$ 分别为 e_1 与 e_2 的基本属性集, 其中 $q \leq n, q \leq m, B_1 \subseteq A_1, B_2 \subseteq A_2$, 假设 $e_1.end_detectTime < e_2.begin_detectTime$, 并且存在 $e_1.u_1 = e_2.v_1 \wedge \dots \wedge e_1.u_k = e_2.v_k, k \leq n, k \leq m$, 其中 $B_1 \subseteq \{u_1, u_2, \dots, u_k\}, B_2 \subseteq \{v_1, v_2, \dots, v_k\}$, 称事件 e_1 与 e_2 满足模糊等量约束; 即存在谓词 $pred$, 对于 $c \in C(e_1), p \in P(e_2), c = pred(u_1, u_2, \dots, u_n), p = pred(v_1, v_2, \dots, v_m)$, 使得 $(c, p) \subseteq \tilde{C}(e_1) \times \tilde{P}(e_2)$, 则称事件 e_1 和 e_2 具有模糊等量因果关系, 集合 $C(e_1) \times P(e_2)$ 上具有模糊二元因果关系 $\tilde{R}$ 。

定义隶属度函数 $\mu_{\tilde{R}}(c, p)$:

$$\mu_{\tilde{R}}(c, p) = \frac{\prod_{i=1}^q W(u_i, v_i) \times \sum_{j=1}^k W(u_j, v_j)}{\sum_{i=1}^{Mat(c, p)} W_i} \quad (1)$$

其中, $0 < q \leq k \leq \min(n, m)$ 。函数定义中, u_i, v_i 是 $c = pred(u_1, u_2, \dots, u_k), p = pred(v_1, v_2, \dots, v_k)$ 所对应的相关属性。 $Mat(c, p)$ 是 c 和 p 相匹配的属性个数。 $W(u_i, v_i)$ 表示属性 u_i, v_i 的权值, $W(u_i, v_i) \in [0, 1]$ 表示为:

$$W(u_i, v_i) = \begin{cases} 1, & u_i = v_i, u_i \in B_1, v_i \in B_2 \\ w_i, & u_i = v_i, u_i \notin B_1, v_i \notin B_2 \\ 0, & u_i \neq v_i \end{cases} \quad (2)$$

其中, 对于基本属性 $u_i, v_i, W(u_i, v_i) \in \{0, 1\}$; 而对于非基本属性 $u_i, v_i, W(u_i, v_i)$ 的取值由领域知识确定。在这样的前



提条件下,定义集合 $C(e_1) \times P(e_2)$ 的二元模糊关系的支持度函数 $Sup_{\tilde{R}}(c,p)$ 为:

$$Sup_{\tilde{R}}(C,P) = \frac{\sum_{k=1}^{\min(n,m)} \mu_{\tilde{R}}(c_k, p_k)}{Mat(C,P)} \quad (3)$$

其中, $p_i \in P(B)$, $c_j \in C(A)$, $i \in [1, m]$, $j \in [1, n]$, 支持度函数 $Sup_{\tilde{R}}(C,P)$ 取值范围为实轴开区间 $(0,1)$; $Mat(C,P)$ 表示集合 C 和 P 中相匹配的元素个数。当 $Sup_{\tilde{R}}(C,P)$ 的值大于支持度阈值 $T_{sup} \in (0,1)$ 时, 集合 $C(e_1) \times P(e_2)$ 存在二元模糊关系 $\tilde{R}$, 也就是说安全事件可能存在关联。

4.2 分布式序列图模式挖掘规则间关联

由于攻击发生的瞬时性和动态性,在大数据环境下分布式的挖掘事件序列是非常必要的。一个事件自身可能是正常的,但如果看成事件序列的一部分,就有可能是恶意的。因此,需要研究什么样的事件序列频繁遇到,发现序列模式并识别离群点。

攻击也可能是从多个不同位置发动并指向不同的目标。可以使用分布式序列图模式进行数据挖掘的方法,从多个网络位置分析网络数据,发现规则间的关联,检测出这类攻击。

分布式图模式挖掘是指通过挖掘不同的日志及记录之间的相关性发现异常。当异常事件发生时,在整个网络域中均会留下或多或少的痕迹,痕迹将体现在不同网络设备的日志记录和流量信息中,因此进行分布式图模式挖掘的目标是挖掘出海量实际事件的相关异常事件频繁图,分布式挖掘图模式的方法可以分成基于 Apriori 和基于模式增长的方法。在这个过程中,同样需要设定频繁子模式的最小支持度和最小置信度。

异常事件的频繁图在发生的时间序列上一般会呈现相关性。因此,把异常事件的分布式记录抽象为一个序列图事件,每个图事件在频繁子图集上也有相应的发生时间作为图边的序列值。因此,对于面向大数据的异构网络安全监控系统而言,形成有效的分布式图模式攻击时序知识库尤为重要。

这种分布式图模式的规则间关联分析用已知的异常事件模式序列图来预测正在发生的可能攻击事件。这属于大数据趋势预测很好的体现,通过分布式图序列模式找出异常事件在时间和空间上的相关性,以时间和空间关系作为关键属性进行关联分析。

本文通过分布式图模式挖掘的相应算法来挖掘异常

事件的频繁子模式,如果正在发生的事件与频繁子模式有超过支持度计数的倾向,则预测为网络异常行为。在这个过程中,用户指定的约束可以推进到分布式序列图模式挖掘过程中,提高效率。

5 实验平台

为了对提出的大数据环境下异构网络安全监控平台的分析算法进行测试和验证,搭建了一个测试环境并模拟攻击行为,对真实数据经过清洗、化简及规范化处理后,得到核心的分析数据,搭建了一个相对简单的 Hadoop MapReduce 集群实验环境,对核心数据进行并行规则处理。实验集群由 1 个主控节点(JobTracker)和 6 个计算节点(TaskTracker)组成。计算节点的配置见表 1。

表 1 计算节点配置情况

计算节点	配置情况
CPU	2 个 4 核 CPU
内存	8 GB
磁盘	1 TB
网络环境	吉比特网卡
OS	RedHat Linux 6.3
Java 虚拟机	jdk-6u24-linux-i586
Hadoop	Hadoop-1.1.2

在测试环境中,利用 MapReduce 的并行处理机制对其进行事件模糊等量约束因果关联分析,并对分布式序列图模式挖掘规则间关联的算法进行处理,在测试环境中能初步实现对大数据环境下网络安全海量规则的加速并行处理,对于基本的安全事件行为进行分析和判别,在最小支持度和最小置信度分别设置为 60% 和 50% 时,大数据安全监控平台关联分析的误报率和漏报率分别为 11% 和 8%。同时,为了对异构网络安全监控平台的功能、性能等进行测试和全面验证,模拟搭建了一个大数据安全监控综合测试环境,如图 2 所示。在测试环境中能初步实现对大数据环境下异常行为的分析,对于基本的安全事件行为进行分析和判别,实验结果证明了此理论框架的有效性,但系统需进一步修改和完善。

6 结束语

大数据时代,国家的核心竞争力在很大程度上依赖于

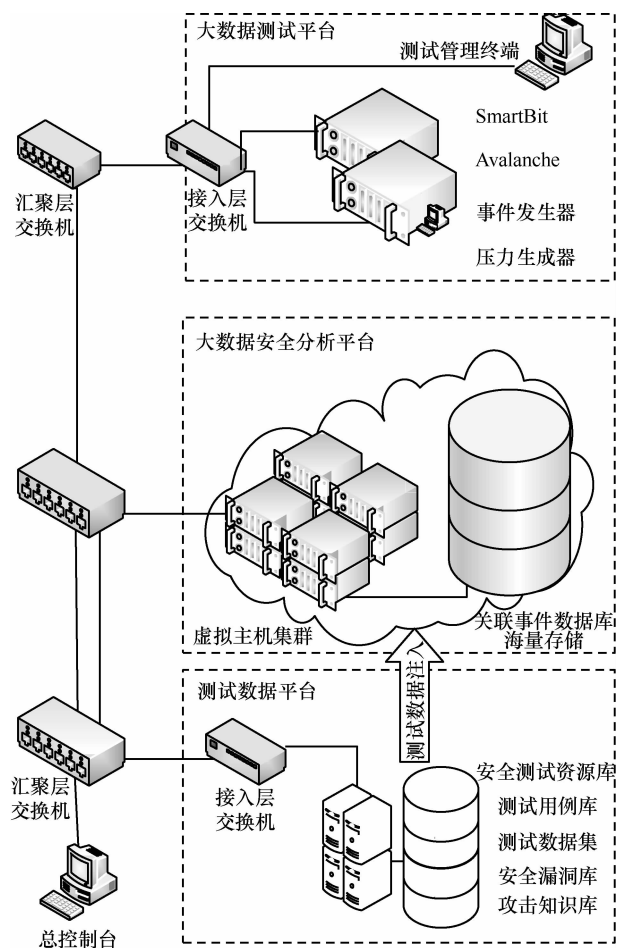


图2 大数据异构网络安全监控综合测试平台

把海量数据转化成信息和知识的能力与速度,本文提出了面向大数据环境的异构网络安全监控系统的相关技术问题及一些解决思路,异构网络包含不同类型的互相关联的大量安全行为事件,从整体的体系模型和处理过程讨论了大数据环境下异构网络安全监控的行为机制和安全策略,对系统的核心关联算法和知识表示做了较详细的分析。实现了原型系统的搭建,并对相关算法进行了模拟测试和分析,原型系统能对一些典型安全事件做出反应,其算法和知识表示需进一步改进分析。

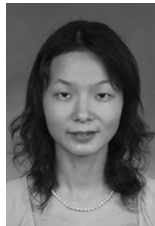
参考文献

- 1 Big_data. http://en.wikipedia.org/wiki/Big_data, 2012
- 2 王珊,王会举,覃雄派等. 架构大数据:挑战、现状与展望. 计算机学报, 2011(10): 1741~1752
- 3 王元卓,靳小龙,程学旗. 网络大数据:现状与展望. 计算机学

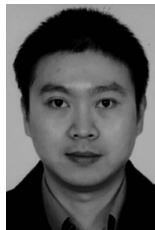
报, 2013, 36(6): 1125~1138

- 4 Boschetti A, Salgarelli L, Muelder C. Tvi: a visual querying system for network monitoring and anomaly detection. Proceedings of the 8th International Symposium on Visualization for Cyber, Security, Pittsburgh, PA, USA, 2011
- 5 Fischer F, Mansmann F, Keim D A, *et al.* Large-scale network monitoring for visual analysis of attacks. Proceedings of the 5th International Workshop on Visualization for Computer, Security, Cambridge, MA, USA, 2008
- 6 Bahadori M T, Liu Y, Zhang D. Learning with minimum supervision: a general framework for transductive transfer learning. Proceedings of IEEE International Conference on Data Mining (ICDM), Vancouver, Canada, 2011
- 7 刘兰,李之棠,李家春等. 基于异构系统的统一网络安全监控体系模型. 小型微型计算机系统, 2006(9): 1682~1686
- 8 邹柏贤. 网络异常的检测与诊断方法. 小型微型计算机系统, 2004, 25(4): 506~510
- 9 Liu L, Xu X P. A kind of fuzzy constraint correlation algorithm based on prerequisites and consequences of security event. Proceedings of IEEE International Computer and Information Technology, 2009(10):225~229

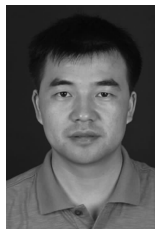
[作者简介]



刘兰,女,博士,广东技术师范学院副教授,主要研究方向为网络安全、电子通信。



林军,男,工业和信息化部电子第五研究所高级工程师,主要研究方向为软件测试与评价、信息安全。



蔡君,男,博士,广东技术师范学院讲师,主要研究方向为网络安全、未来网络。

(收稿日期:2014-03-25)