分类号： TP393.08 单位代码： 10346

密级： 学号

杭州師範大學
HANGZHOU NORMAL UNIVERSITY

硕士学位论文

（学术学位）

基于攻击图的多源安全事件关联分析研究

Research on multi-source security event correlation analysis on the bases of attack graph

申请人姓名：

指导教师：

合作导师：

专业名称： 计算机应用技术

研究方向： 信息安全

所在学院： 杭州国际服务工程学院

论文提交日期 2016 年 4 月

–

基于攻击图的多源安全事件关联分析研究

论文作者签名:

指导教师签名:

论文评阅人 1：_____

评阅人 2：_____

评阅人 3：

评阅人 4：

评阅人 5：

答辩委员会主席：_____

委员 1：

委员 2：

委员 3：

委员 4：

委员 5：

答辩日期：

杭州师范大学研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得**杭州师范大学**或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解**杭州师范大学**有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权**杭州师范大学**可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

致谢

摘要

随着互联网技术的发展和计算机网络的普及，信息系统的安全漏洞和网络恶意行为不断涌现，信息安全事件层出不穷，网络安全形势日益严峻。为保护网络系统的安全，防火墙、入侵检测系统等传统的安全设备被部署到网络环境中，在从多方位保护网络系统的同时，也带来了新的问题：这些安全设备独立运行、自成体系，相互之间缺少协同合作、数据交互，信息资源无法集中、不能共享；安全管理员难以统一管理、配置众多异构的安全设备，无法对业务和系统进行集中管理，资源无法形成合力最优，缺乏从总体上把握网络安全态势的有效手段。此外，安全设备每天产生海量且夹杂了大量不可靠信息的安全报警，使得管理人员被洪流一样的数据所淹没，很难提取出有意义的事件，更无法从中得到真正对系统造成威胁的事件，进而无法评估当前系统的整体安全态势，大大降低了系统的安全性。

针对上述问题，本文拟采用多源安全事件关联分析方法，基于攻击图技术构建实时、准确的攻击场景，利用图上深度挖掘技术对下一步攻击行为进行预警，协助管理员进行网络安全防护。该方法能基于粗糙集理论，挖掘报警属性之间的关系，实现属性权重计算方法，改进报警数据聚类粒度；设计并实现了基于多特征融合分析的报警分析方法研究，该方法从不同维度预处理原始报警，可有效去除原始报警数据集中冗余、错误的报警，形成标准化的安全事件数据集；设计了面向场景的通用树状攻击模型，实现基于攻击图的多源事件分析算法，利用上述层次化关联规则，关联多步骤报警，构建攻击场景，并识别攻击意图，预测下一步攻击行为。依托上述研究内容，我们进行了相关系统的研发，并利用大数据分布式分析技术，为分析海量报警数据提供技术保证。

本文采用 DARPR 数据集和真实网络报警作为实验数据，实验结果显示基于属性权重的聚类算法平均效率达到 84.6%，报警特征分析方法效率达到 86.3%，实验的场景匹配效率验证了多源安全事件关联分析方法有效性。

关键字：特征分析，粗糙集，攻击图，关联分析，攻击场景

ABSTRACT

With the development of Internet technology and the popularity of computer networks, the security vulnerabilities and network malicious behaviors have sprung up, and security events emerge one after another, the situation of network security is becoming more and more serious. To protect network security, traditional security devices, such as firewall, intrusion detection system are deployed into the network environment, which could all-roundly protect network system, but also bring some problems to administrators: the security devices run independently, lack of cooperation and data interaction between each other. It is hard for administrator to manage number of heterogeneous security devices. Increasing devices in the network could lead to the amount of the data growing in geometry level, the overwhelming worthless data may cover the meaningful network information. In addition, if the alerts of different devices are operated dividedly, administrators could miss the key information because of the information island phenomenon. Consequently, it is a hot issue to integrate and manage the network devices.

This paper proposes the multi-source security event correlation analysis method based on attack graph, and the method is consisted of the alert fusion, alert validation, alert aggregation and alert correlation analysis. Based on rough set theory, we first min the relationship between the attributes of alert and put forward weights calculation method, then we design and implement the multi-level characteristics analysis method to reduce the redundancy and false alerts. After that, we design the tree-like attack model, which is extensible, general and easy-understanding rules. Based on the hierarchical rule, we could associate multi-step alerts and construct attack scenario, identify attack intention and forecast the next step of attack.

Finally, we deploy the system's platform, and adopt DARPR data sets and real network alert data set to conduct an experiment, the result verify the effectiveness of the proposed analysis method.

Keyword: Characteristic analysis, Rough set theory, Attack graph, Correlation, Attack scenario

目录

1 绪论

1.1 背景及意义

随着网络技术的发展和计算机网络的普及，信息化、网络化的触角已延伸到了传统行业的各个领域，各类信息系统渗透到了社会生活的方方面面。信息系统日益发展极大推动了社会进步和发展，但同时也伴随着产生了诸多安全威胁。由于系统在社会中的重要性，其往往成为网络不法分子的攻击目标。近些年，各类安全漏洞和网络恶意行为不断涌现，信息安全事件层出不穷，网络安全形势日益严峻。

2007 年 4 月，爱沙尼亚遭受的历时三周的分布式拒绝服务攻击导致国家公共生活全面瘫痪，20lO 年伊朗布什尔核电站里正在工作的 8000 台离心机受“震网”病毒攻击，突然出现大面积故障，数据大量丢失，上千台离心机被物理性损毁。2013 年，我国破获网络犯罪案件 17 万起，造成经济损失约 2300 亿元，受害者接近 3 亿人；2014 年 46.3%的网民遭遇过网络安全问题，在安全事件中，电脑或手机中病毒或木马、账号或密码被盗情况最为严重，分别达到 26.7%和 25.9%[1]。随着人类迈入信息网络时代，恐怖主义活动也由物理空间延伸到了信息网络空间，2015 年 4 月 8 日自称属于“伊斯兰国”的网络黑客攻陷了法国国际电视五台的 Facebook 和 Twitter 账户以及该台的网站，贴出了宣扬“圣战”内容的信息，随后还入侵电视台的电脑系统，导致该频道出现黑屏，整个过程持续数小时。网络安全问题不仅对人民生活、国家经济、甚至是国家安全都有重大影响。

据洛克希德·马丁公司（Lockheed Martin）的调查显示，大多数公司领导者认为，由于缺少足够可依赖的威胁情报，造成无法成功预测网络安全漏洞，并有效对付威胁。网络威胁情报[2]是近两年国外热点，被认为是解决安全防护和 APT 攻击的重要手段之一。在安全厂商、政府机构等组织，我们往往可以看到诸如漏洞信息、网络威胁通告、预警安全信息等，这些信息都属于威胁情报。报告还发现大部分机构靠直觉来评估其安全级别，仅有 32% 的研究对象依靠数据情报证实自己的想法；部分企业领导者认为缺少专业的情报和人才来应对网络威胁。由此可见，根据网络系统信息，如何提供有效的、可信赖的、具有针对性的网络威胁情报，展现网络安全状况是一个迫切需要解决的问题。

现有网络环境中部署的安全产品，如防火墙、入侵检测系统等就是提供威胁情报的数据生成源。这些安全产品能够实现检测网络中的恶意行为，识别网络攻击，检测系统漏洞等功能，从不同角度帮助管理员管理系统网络、理解网络运行状态。但当这些产品被部署到网络中后，产生了一系列内部问题，例如：多源安全产品日志语义和表达方式各异；大量安全报警使管理者无法分辨出攻击等等。从单个安全产品考虑，由于单个报警日志包含的信息量有限，单独报警能否完成有效的入侵检测[3]也受到了质疑；此外，从大量异构的日志数据中，利用简单的统计分析方法很难挖掘网络中直观有效的威胁情报。

SEM（Security Event Management）技术的出现，给网络安全体系带来了新的思考角度。综合各类网络安全设备优势，集中设备信息，统一实时检测入侵行为、关联分析安全事件、安全应急处理指示等安全需求，是安全体系发展的导向。因此，根据安全需求，统一管理各类安全产品及其日志数据，有效利用产品功能、挖掘产品联系，从而保证网络安全性，是信息安全界的热点问题。

本文所依托的“安全事件分析系统”项目是上述问题的一个解决途径。该系统基于分布式计算框架，集成了多种安全产品，利用层次化、模块化功能模块，进行报警数据特征分析，实时关联分析安全事件，发现潜在的攻击行为；从大量异构报警中准确挖掘出攻击行为，实现攻击源定位、事后取证技术，实时、准确地提供多角度威胁情报和响应建议，以提高网络整体安全性。此外，项目基于大数据分析技术，为处理大规模网络中产生的海量报警提供技术保证。

本文将网络监测设备生成的原始日志定义为报警，预处理后的报警定义为安全事件。针对大量报警数据，如何实现融合多源安全信息，形成准确的安全分析；突破安全事件关联分析技术，识别多步骤攻击、构建攻击场景，重点研究安全事件诊断技术，实现快速、有效的攻击预警，是系统实现的难点所在。本文研究题目是基于攻击图的多源安全事件关联分析技术，包括以下两个部分研究内容：

（1）研究多级特征分析方法，集中处理各类安全设备的日志；

（2）研究网络攻击步骤，深入学习层级化攻击建模方法；基于攻击图方法，重点突破安全事件关联分析，实时构建攻击场景。

1.2 国内外研究现状

1.2.1 报警数据预处理研究现状

报警数据预处理是多源报警关联分析的前提，因此，面对网络中产生的多源异构报警数据，如何去除冗余、错误的报警，保证关联分析功能模块的效率和准确性是安全管理技术需要解决的问题。近些年，国内外的学者在报警数据预处理分析方面做了不少研究，主要集中以下几点。

张[4]提出了基于融合规则的报警预处理方法，该方法将符合一定规则的特征属性相似的报警合并，来融合重复的报警，比如多个安全产品对一个攻击产生的报警；方法只完成简单的去冗余，处理速度快，对特定数据集有较好的融合效果。Lee[5] 等人研究并设计了相似度的计算方式聚合报警数据。针对 DDoS 攻击触发的报警，深入分析攻击生成原理，并结合欧几里得距离计算方式，判断报警是否需要被聚合。Xiao[6] 等人设计了多层级报警预处理方法，通过流程化功能模块，逐步达到数据清洗的目的。该多级预处理方案中也详细介绍了相似度值得计算方法，但并未挖掘报警属性之间的关系。

这些文章中[7,8,9,10,11]都提到了关于报警的验证研究。根据报警属性特征，Sadighian[7] 等人和 Xiao[8] 的团队利用自定义的规则过滤错误的报警，但如何有效定义规则是一大难题。Cheng-Yuan[9]等人提出了一种错误报警 P/N 评估机制，通过收集实际相关网络报警，综合统计分析报警实例的真实性。Richhariya[10]等人提出了特定离散化算法减少误报率，该方法使用贝叶斯理论和 K-means 聚合算法，有较好的实验效果，但并没有详细描述模型构造方法。

随着网络规模不断扩大，需要部署的安全产品数量不断增加，使得网络中生成的安全运行数据呈几何级增长。面对海量的日志数据，云计算、分布式等新兴概念，开始结合到网络安全领域中。虽然，Roschke[12]等人设计了一种基于高级存储队列的方法，利用数据库系统，结合存储支持算法，实现 IDS 报警的聚合。此外，Yang 等人[13]利用 Hadoop 分布式计算框架，结合 MapReduce 计算原理处理网络日志，取得了较好的实验成果。在未来，借助大数据、分布式技术完成数据预处理可能会成为一种常态。

1.2.2 攻击图生成研究现状

攻击图生成的研究工作主要集中在两个方面，分别是状态攻击图和属性攻击图。早期的研究主要集中在状态攻击图[14,15]的生成，由于其生成受主机规模等因素影响，无法应用于大规模网络;在系统网络信息已知的基础上，研究人员转换

思路对网络和攻击能力进行建模，生成属性攻击图[16,17]。

表 1-1 攻击图生成方式比较

类型	节点	有向边	优/劣势
状态攻击图	攻击者和目标网络的状态	单一攻击引起的主机状态转换	攻击路径数受目标主机规模和系统漏洞数量限制，无法应用于大规模网络
属性攻击图	系统属性和原子攻击	节点间的因果关系	可应用于大规模网络，但需要预先了解网络基本配置和攻击能力

MulVAL[17]是由堪萨斯州立大学的 Xinming Ou 等人研发的网络安全分析工具，其使用一种可扩展的建模方法来产生属性攻击图。工具采用 Datalog 的建模语言进行元素分析(bug 规范、配置描述、推理规则、操作系统权限和特权模式等等)，该语言能轻松转化系统的漏洞信息为 Datalog 语言描述文本，进而输送到分析工具中。通过在大规模网络中的分析实验，证明了该方法有效。

文章[18, 19]研究了利用攻击图聚合 IDS 报警的方法。该方法构建一系列的攻击图模型，该图对应 IDS 所监测的网络；报警是攻击行为的不确定反映信息，通过报警和攻击图模型的匹配，可以判定攻击的真实性；攻击往往是多步骤的，攻击图模型的层级化构建方式可以轻松确定攻击已经完成的程度。

1.2.3 关联分析方法研究现状

关联分析的目的是研究 IDS 报警等实时安全设备日志，审计安全事件，进而判断网络系统中的攻击行为。日常网络中，我们往往依赖各类入侵检测系统来监测网络威胁，但其生成的大量低级的、无效的报警，使得管理人员无法准确把握网络的整体安全状态。因此，如何处理大量的原始系统日志，提取出攻击相关的安全事件，识别网络中真正的攻击，是安全领域研究的热点。

很多研究者通过构建攻击模型的方式进行报警关联分析[20,21,22,23]。其中，SnIPS[23]是一个利用报警关联分析方法实现入侵分析的开源系统。基于可扩展的 DS 理论，计算关联分析中命题的可信值；SnIPS 实现的优先级排序模型，将根据可信值对分析结果信息排序，把高准确性的结果推送给管理员。Richard[24]等人认为攻击是造成系统状态变化的重要因素，并建立一种基于状态转移分析的攻击描述语言。该语言能适用于各类产品，可用于实现多种分析方法。

1.2.4 存在问题

完成报警数据预处理、攻击图生成研究、关联分析方法研究的调研后，总结当前的安全事件管理分析系统仍存在下述问题：

首先，报警预处理粒度不一致。对于各类安全设备生成的大量异构报警，其数据特征是不一样的。例如 IDS，其日志特点是量大，有冗余，存在误报。因此在预处理过程中要根据产品类别，设置符合网络特征的、特定粒度的处理方法。在报警聚类方法上，这些研究中没有深入挖掘报警属性之间的关系，虽然不影响去除聚类的结果（指效率上），但结果是否合理、有效还需要考证。

另外，关联规则的研究是安全分析领域的关键点。现有的分析技术主要基于攻击模型进行，其模型是否满足可扩展、可通用、易理解的特点还有待检验。在关联分析方法上，基于攻击图理论是一种比较常见的方案，但结合该理论进行的后续衍生工作，如攻击预测、路径展示等，大都处于研究理论阶段，没有将这些工作转化成实际可用的系统模块。

1.3 研究工作及目标

（1）多特征融合分析方法研究

当各类安全产品被部署在网络环境中，其生成的日志数据结构各异，还掺杂着一些错误和冗余的报警。通过本研究内容实现的多特征融合分析模块，可进行多源报警数据的融合、验证和聚类操作三个层面的报警处理。本文重点研究了属性权重计算方法，利用信息理论知识离散化报警属性，并基于粗糙集理论，将报警集合与决策树概念结合，确定条件属性和决策属性，进而挖掘数据中元素间的隐含关系。多特征融合分析能从原始的报警中筛选出必要、正确的安全事件，保证关联分析的准确性提升分析效率。

（2）层次化规则建模方法研究

面对防火墙，入侵检测系统、扫描工具等安全产品，如何从其生成的多源异构日志数据找出关联规则,进行攻击知识建模,是本文依托项目需要解决的问题。本文提出了基于场景的攻击建模方式，设计了树状关联规则，其具有可扩展、可通用、易理解的特点，可消除同类型设备间界限，便于系统的关联分析。

（3）基于攻击场景的多源事件关联分析方法研究

基于场景的关联方法提出了统一的完整攻击描述方式。报警是攻击步骤的最直观的数据反映，但步骤往往是不连续的、不完整的攻击体现；不同产品生成的报警，其表现和表述方式也有较大差异。基于攻击场景关联分析方法将事件关联分析和攻击描述结合在一起，挖掘安全事件之间关系，构建攻击场景进而预测攻

击行为。

1.4 论文章节安排

第一章：描述了论文的背景和意义，分析了国内外学者在报警预处理、攻击图生成、关联分析方法方面的研究现状，总结其成果和存在问题，并给出研究内容和目标。

第二章：描述在大数据时代人们如何理解安全，并研究大数据应用安全技术。

第三章：主要介绍多源报警日志特征分析的研究了基本理论信息、主要问题、实现目标，提出了多级特征分析方法，深入分析各层级的设计思想，并结合实际规则进行说明。

第四章：介绍了网络攻击建模方法，阐述了基本概念、建模原因和建模目标，提出层次化关联规则生成方法，并描述了一个典型的攻击模型；设计并实现了基于攻击场景的多源事件关联分析方法。

第五章：介绍了实验环境和实验数据，根据本文研究内容，提出了具体的实验方案，并以图例描述和分析实验结果。

2 大数据安全技术应用

网络技术、数字化和移动化的发展，使生活更加便利，但信息安全、网络安全的问题也渐渐显现，个人信息很容易被黑客获取，其违法行为也不易被发现。于此同时，信息技术的发展，使得各企业内的数据信息量呈现爆炸式增长，因此在数据保护、安全检测等方面上，企业对安全性和隐私保护上的要求不断增加，传统的安全方案已经无法满足大数据安全的需求。如何利用大数据技术，保护网络安全，减少网络威胁，保护企业、组织和个人的安全利益都迫切需要考虑并解决的问题。

2.1 理解大数据安全

大数据是信息化时代的石油，在这个时代，计算机不再一味追求计算速度，转而追求大量数据的处理能力，这改变了计算机软件的发展方向，从单纯实现业务功能向数据、信息为核心的应用处理。大数据不再仅仅局限在科研机构的内部探索，随着对大数据关注的增加，有关大数据安全的行动也已经展开，诸如科学研究机构、应用企业、安全厂商等组织，都在极力推动大数据安全产业发展，积极参与相关安全标准的设定，为大数据安全技术的广泛应用提供全面、坚实的基础[25]。

2.1.1 大数据安全行业研究

（1）互联网行业

在云计算和物联网等概念的刺激下，“对大数据进行集中处理、分析和挖掘能够为企业创造更多价值”，这一理念已得到互联网行业的普遍认可，众多互联网企业相继开始研究大数据相关的应用。

企业在应用大数据时，常常会涉及诸多安全问题，例如数据安全和个人信息保护等，由于其涉及到企业核心技术和国家法律法规，因此很难有专家可以明确界定出因安全问题给企业带来的损失。例如 2015 年 5 月 28 日发生的携程官网瘫痪问题，数据系统和对应的灾备系统全部失效，而详细的安全事故过程由于涉及公司机密，并未对外告知真实的原因。此外，几乎所有数据安全都是由网络运营服务商负责，但安全责任却并未受到法律的严格保护。

在互联网行业，大数据安全的需求是遵循企业数据安全存储守则，严格执行大数据安全监管和审批管理，保证用户隐私和企业安全利益，在此基础上利用大

数据挖掘分析技术，处理海量数据，发现和发掘合理的企业需求。

（2）电信行业

电信行业的特点给运营商带来了巨大的数据“福利”，包括用户基本信息、消费水平、行走轨迹、日常偏好等内容[26]。随着互联网技术的发展，运营商面对着众多互联网公司的挑战，例如微信、imessage 等通信软件，对传统的通信服务造成巨大的冲击。对于运营商来说，要应对这些冲击，必须改变原有的服务模式，避免沦为数据通信管道。相比互联网企业，现阶段运营商在用户量、用户信息上仍有较深厚的基础，但在技术研发能力上仍存在差距。利用这些数据更加精确地洞察客户需求，提供符合用户需求的新业务及服务，是行业的发展趋势。

目前，在利用大数据的过程中，运营商需要应对以下两方面的问题：首先，电信企业的数据来源分散而庞杂，运营商需要在众多系统中进行数据采集、处理和分析，并保障数据的机密性，完整性和有效性；其次，运营商在使用利用系统工具分析内部业务数据时，必然需要向外部开放一部分的内容，这个过程需要面临一系列的安全问题，如数据保密、用户隐私等。因此，在电信行业，其安全发展需求是保证行业核心数据安全，保护用户隐私，利用大数据技术，充分挖掘数据价值。

（3） 金融行业

随着互联网行业进驻金融行业，金融业务的载体与电子商务、互联网概念的融合越来越紧密，互联网金融服务在金融行业发挥越来越大的作用[27]。面对行业中的大量业务数据，常用的结构化数据分析方法已经不能满足企业发展的需求。利用大数据、云计算等技术，整合所有可利用的结构化和非结构化数据，打破数据壁垒，为企业翻开全面的、多角度数字化服务新篇章。与此同时，大数据正在改变着银行的运作方式，形成了一些较为典型的业务类型，对理解和洞察市场和客户方面产生着深远影响。

金融信息行业系统服务对象广、牵涉面大，因此对网络有较高的稳定性和安全性的要求，为保证能在任何情况正常运行，系统需要提供数据备份和容错功能，并保证数据高速稳定处理的能力；同时，面对实际中多样的应用场景，该行业往往希望系统能具有较强的灵活性和良好的管理能力。在大数据、云计算模式的普及下，金融企业信息系统必然面对更加复杂的业务内容，系统功能和对外服务变

得更加复杂，这会增加金融业的安全风险。

在金融行业，安全需求是利用大数据安全技术，满足数据管理、访问控制等安全需求，提高机构内部管理门槛，改进行业服务能力，防范和化解机构的安全风险。

（4） 医疗行业

医疗行业开始引入了大数据技术实现行业信息数字化，该技术可应用于医疗诊断、远程监察、新药物研发、分析由生活方式和行为引发的疾病等。据麦肯锡研究结果显示，在美国通过对医疗大数据的分析，产生的价值将近 3000 亿美元，有效减少居民医疗支出。医疗离不开数据，数据用于医疗，大数据的基础为医疗服务行业提出的“生态概念”的实现提供了有力的保障。

随着医疗数据存储压力的变大，医疗信息中心的关注点由传统的计算领域转移到存储领域。数据安全存储是保证医院业务的基础，作为社会重要的基础机构，医疗系统出现故障后，如果数据不能迅速恢复，不仅会影响医院正常的业务，减少患者对医院满意度，还会给社会稳定带来危害。此外医疗数据属于个人隐私数据，多数人并不希望自己的诊疗数据作为研究资料被使用，而真实的数据资源有限，因此如何保护个人隐私，又不造成数据资源的浪费，是医疗行业面对的问题。

在医疗行业，大数据安全的需求是安全可靠地存储和管理数据，在保证用户隐私的前提下，丰富医疗行业治疗数据，以提升医疗行业在诊断疾病，管理决策等方面的水平。

2.1.2 大数据安全产业动态

大数据是一把双刃剑，一方面能为社会形成巨大发展机会和便利，另一方面也带来了前所未有的挑战。在各类型企业中，由于发展特点和行业性质，互联网企业往往是大数据技术应用的先驱，这也意味着需要应对更多的挑战。本文调研了几家国内领先的互联网企业在大数据安全形势下的发展动态。

（1）腾讯大数据安全应用动态

腾讯是一家提供大型互联网服务的公司，其在全世界拥有大量的用户，通过汇聚所有用户的信息，能够分析出用户有价值的信息。腾讯在保障数据安全的前提下，提供自动化服务平台，隐藏技术开发中的技术细节，不断优化用户体验，降低数据分析和挖掘门槛，从人肉服务模式转向平台自动化服务方式，帮助数据

分析人员通过自主服务的方式，降低人工成本，满足业务量快速增长的需求。

腾讯数据挖掘体系分为 5 层，分别是：数据层、分析层、算法层、输出层和投放层。在算法层具有定向规则过滤，输出层严格控制细分人群的精准投放概率，投放层控制投放频次进行算法配置。同时，腾讯还通过安全监控对数据挖掘应用的效果进行评估，从效果分析层、基础数据层、算法层、系统层进行安全保障。

在大数据安全应用方面，腾讯也在进行积极探索，并且在 2013 年推出了全国首个“全景网络安全防御系统”，该系统依托腾讯电脑管家安全云库的海量数据，通过腾讯即时通讯软件 QQ、搜索引擎 SOSO 等上网入口，可以迅速捕捉到钓鱼网站的踪迹。

上述系统目前包括“十大恶意网址类型”、“当日已拦截恶意访问”等几大纬度指标，力求能够通过“视觉化”和“数据化”的模型来呈现全球互联网的风险情况。

（2）阿里巴巴数据防护开展情况

阿里云梯集群采用了 HDFS 和 MapReduce 技术，承载着淘宝、天猫、一淘等应用服务。为了实现原始表、中间表、元素数据的共享，避免数据操作，还在阿里云梯集群上构建了阿里巴巴数据交换中心。

阿里巴巴数据交换平台的功能是打通、整合集团数据，为用户提供个性化服务，构建统一的大数据平台。交换平台的安全防护策略主要有：数据作为资产进行管理和开放，在数据管理层，设置了安全预警、质量监控、元数据、逻辑及生命周期管理等策略；在数据开放层，设置了审计、计量、监控等策略。其关键技术包括：离线数据分析服务及源头数据质量监控和元数据管理。

根据上面描述，可以看到阿里巴巴在阿里云梯集群中对数据安全防护集中在数据流向管控上，通过严格的数据流向控制，外加客户端与服务端的访问控制来保证数据应用安全，但是仅对数据流向进行处理还会存在诸多安全隐患。因此阿里巴巴将在未来克服数据安全、数据质量和数据化运营的问题。

（3）百度基础大数据平台的安全防护进展

百度是全球最大的中文搜索平台，在数据处理、分析、挖掘领域进行了积极地探索。自 2011 年开始，百度的发展思想逐渐从以计算为中心转变为以数据为中心。2013 年，百度重点在大规模数据存储、数据分析以及数据索引等方面做

了研究与应用。

百度认为基础架构能力的强弱是决定分析的重要因素。目前支撑着百度快速迭代的正式其强大的基础架构能力。现在，百度数据仓库的基本原理是：数据要求准确、全面、一致，并且易于分析理解；数据写入有 ETL 效率决定。在数据仓库中，平台对非结构化高价值内容进行治理，并进行层次化存储。同时，百度还是用数据集市，增加特定领域数据和逻辑，与数据仓库搭配成为数据管理基本结构。

百度建立敏感数据防泄漏系统进行数据保护。该系统以数据安全管理平台为基础，结合数据加解密、身份认证、权限控制等技术，防范了非法泄露百度内部重要数据，同时确保应用系统安全加固，管控范围达到了各类业务的全面覆盖。

2.2 大数据安全应用技术

面对海量的数据信息，企业需要通过对信息的整合与分析，才能更深入理解自身的业务，进而挖掘出新的需求。对于企业，信息安全是为信息化服务的，而信息化服务于业务增长，因此利用大数据提升企业信息安全防护水平，能够间接为企业带来效益。

2.2.1 安全大数据挖掘技术

安全大数据是指与网络安全、系统安全等有关的各类数据，例如脆弱性数据、报警数据、资产数据等。针对安全数据的挖掘与分析，可以从大量的、不完全的、随机的真实数据中，分析出系统、网络的整体运行状况，及时发现安全威胁，进而反映出整体的安全运行趋势。

安全检测与大数据的融合能够及时发现潜在的威胁，提供安全分析与趋势预测，加强应对威胁的能力。首先需要对数据进行分类、过滤与筛选，其次采用信息安全监测技术，对系统环境和数据环境进行检测，之后通过关联分析和数据挖掘构建安全威胁模型，并经过数据分析预测安全趋势[28]。此模块主要根据以下四个方面进行描述。

（1） 数据提炼与处理

系统处理平台在收集到原始数据后，往往会执行分类、过滤、筛选等操作。当根据数据的敏感程度、应用场景，对数据进行分级分类的预处理。然后过滤敏感数据、保密数据、非法数据，得到有价值的可使用的数据。基于大数据技术的

智能过滤和内容审计，能够快速准确地匹配大量自定义的关键字，过滤违反相关法律法规的内容，确保信息内容安全；同时还可以按照企业的安全需求，进行安全数据关键信息的提炼和处理，为报警关联分析提供技术保障。

（2） 信息安全检测

根据实现原理与应用机制的差异，目前常用的安全检测方法包括漏洞检测、入侵检测、审计追踪等。入侵检测可以实现实时监控目标系统的安全状态，任何对系统有企图的行为都会触发生成报警，报警能够体现部分攻击的步骤、企图和结果等信息，向网络管理者提供详细安全状态检测结果。漏洞扫描工具可以利用特定手段扫描目标网络主机，获取其脆弱性数据。安全审计追踪可以对系统安全和数据安全进行实时检测，实时展现整个环境的运行状态，反映外部入侵意图和非法行为。

（3） 威胁模型构建

大数据环境下，需要从广泛的数据来源获取、度量、处理、分析大量数据。大数据的威胁模型能够明确指出哪些地方可能会被攻击或者利用。一个典型的大数据安全模型由数据流程图、入口点和退出点、潜在威胁列举组成，其核心的元素是应用程序入口点的描述。良好的威胁模型可以捕获网络可访问性和接口的身份验证/授权要求，包括基于 IP 地址的网络可访问性、身份验证和授权等内容。

通过数据挖掘操作，可以从低层次、不完整、不统一的数据中，挖掘出隐含的关系以辅助完成决策支持。图 2 展示了几类典型的数据挖掘技术。

表 2-1 典型数据挖掘方法列举

方法名称	挖掘目标	描述
关联分析	隐藏在数据间的关联或相互关系	描述数据项之间所存在关系的规则，分析数据之间的关系
决策树	决策支持，预测	决策树是一个树结构，通过分析各种情况发生的概率，评估风险结果
粗糙集	数据分类	体现不确定、不完整的建模理论，能够完成分析和推理工作，挖掘数据中元素的隐含关系。
聚类分析	数据分类	按照相似性或差异性，将原始数据分为几个类别
回归分析	数据关系、变化趋势	利用数据统计法，建立多元变量间的关系
特征分析	数据特征	挖掘数据集的特征关系，总结数据集的特征
神经网络	数据分类	建立在自学习的基础上，分析复杂的数据，可完成极为复杂的模式提取
遗传算法	数据评估、搜索	基于特定的候选度值，筛选数据集中的候选解，并更新当前集合中的候选值，提高各个数据适应能力

目前，众多研究者进行了数据挖掘技术相关的研究，产出很多研究成果[29,30,31,32]，也存在一些像挖掘算法设计、模式识别和解释的问题。对于基于大数据技术的业务系统，数据挖掘技术需要结合业务数据的特点处理数据，并能够从时序、逻辑等多角度将知识表达出来。

安全大数据挖掘可以被理解成表现网络安全现状、预测变化趋势及行为，根据网络知识形成决策的过程，其目标是深入挖掘目标数据集，获取有效的、隐含的信息。其中，数据之间的关联性是数据挖掘需要寻找和利用的重要知识。通对数据的关联分析，能找出海量数据中的存在的隐含关系，并总结出可利用、易理解的结果。

2.2.3 基于大数据的网络态势感知

网络态势感知源于 ATC 态势感知这一项目[33]。1999 年，Tim Bass[34]首次提出这一概念，并将其与 ATC 的概念进行了类比，旨在把 ATC 态势感知的成熟理论和技术借鉴到网络态势感知中，他认为基于融合的网络态势感知必将成为网络管理的发展方向。网络态势表示整个网络环境当前的运行状态及变化发展方向的综合指标，该指标包括各种网络安全产品如入侵检测系统，漏洞监测系统，防火墙等反馈的攻击行为、运行状况因素。

网络态势感知目前还没有全面、标准的定义[35,36,37]。但大多数学者[38]认为网络态势感感知是指在大规模网络规模中，对能够引起网络态势发生变化的安全要素进行获取、理解、显示以及预测未来的发展趋势。网络态势感知研究包括网络数据采集，态势数据分析处理、态势发展预测这三个阶段。通过特定的网络安全态势评估体系，集中各类产品生成的安全报警，进行分析处理，并将分析结果以文本、web 等形式呈现。

本节还研究了席荣荣等[35]提出的网络态势感知概念框架，该框架包括多源异构数据采集、数据预处理、关联与目标识别、态势评估、威胁评估、响应与预警、态势可视化显示以及过程优化控制与管理。孤立网络安全事件的无法体现出网络态势。只有综合多方的报警与流量数据，设立特定的分析流程或构件标准的处理体系，才能达到准确、有效评估网络安全态势的目标。

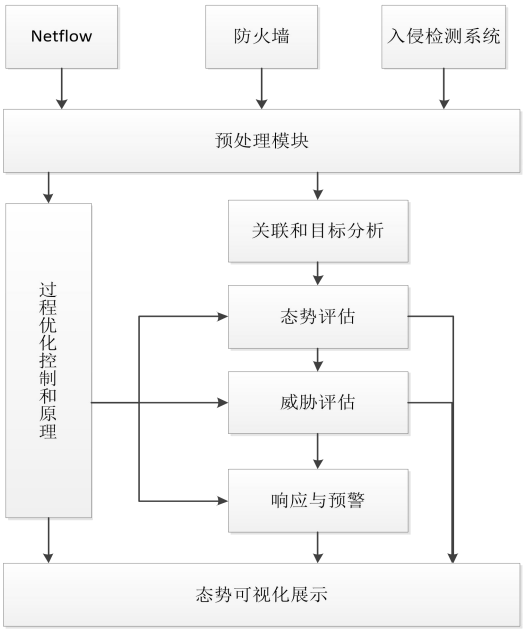

图 2-1 态势感知感念框架

随着网络技术的迅速发展，网络互联速度变快，这给 IDS 的工作带来困难，其检测可信性和有效性无法得到保障。而且，由于行为识别规则不统一、分析技术不完整，形成了大量虚假报警。因此，IDS 目前大多部署于中小规模的分支网络中。目前，监控带宽主干网一般采用网络流量分析技术，以检测流量的变化趋势和突变，有助于快速定位异常、采取后续响应措施。主干网络反映的大规模网络状态和趋势也需要从流量中分析，因而这是网络态势感知的重要组成部分。

目前网络流量研究基本都是基于流量采样的分析，主流采样方式是 NetFlow，该技术是由思科公司在 1996 年开发完成，已成为业界主流的流量计费方法。在工作时，NetFlow 会采集流经交换设备的所有流式数据，然后按预设的格式生成流量报告，并发送给目标服务器。流缓存技术相比传统的流量采集模式有分组丢失率低的特点，保证了能够提供比传统 SNMP 更加丰富的流量信息，可以回答更精细的问题。获得网络数据之后，由于其体量巨大、内容复杂的问题，网络管理人员单单观察原始数据，很难得到有用的信息。这些网络流量数据必须经过分析形成简明的、易于理解的网络状态，即通过网络流量判断状态正常与否，异常情况在什么时间和位置发生。

针对上述几种实际中异常情况，可选用的方法有：分类筛选、统计分析、TOP 排序、模式匹配等方法。由于网络流量本身具有突发性和快速变化的特点，因此，在实际使用时需要结合网络拓扑、流量特点、采集协议等情况，是当选择相应方

法。

2.3 本章小结

现阶段，大数据安全在信息安全领域内是一个复杂的命题。本章首先介绍了大数据安全的概念，并调查研究了国内外大数据安全相关的行业需求和发展动态；然后，研究了安全检测技术和大数据处理技术的融合方法；阐述了面向安全大数据的挖掘技术，对发掘方法和目标进行研究；分析了大数据场景下态势感知技术的应用，重点研究了基于流量交换的态势评估方法。

3 基于多特征融合分析的报警分析方法

随着网络威胁多元化和规模复杂化，各类安全产品被部署在网络环境中，对于同一个网络攻击，可能会有多个产品同时触发报警，而且这些报警往往存在一定的相似性，甚至出现完全重复的情况。此外，这些报警中还可能存在一些错误的报警。因此，在关联分析之前，需要对这些原始的报警中进行特征性分析，筛选出必要、正确的安全事件，保证关联分析的效率。

3.1 研究内容概述

当各类安全产品被部署到网络中，网络管理员可以通过这些产品的特点，多角度检测网络攻击，发现网络受到的安全威胁。其中入侵检测系统是一款常用的安全产品，该产品可以帮助管理员监控网络环境中的运行状态，分析异常行为，但其存在一个问题：入侵检测系统会产生大量的实时报警。据统计，在中型网络中，一个入侵检测系统每天至少能产生 200M 的最简化的日志；同时，这些报警日志中，可能还存在大量重复或者错误的报警。

在大规模网络环境下，往往会部署多个入侵检测系统，这会进一步加剧报警数据量和冗余、误报的问题。如果不能去除数据集中无效的报警，不仅无法保证关联分析的有效性，还会增加分析计算的压力；此外，面对海量的报警数据，传统的并发处理技术能提供的计算能力有限，一旦超出运算负荷，可能会引起预处理功能失效甚至系统瘫痪。因此，如何高效、稳定地去除原始数据中冗余的、错误的报警就成为报警分析的首要问题。

针对上述问题，本章提出多特征融合分析方法，方法包括了多个级别的融合分析步骤。在安全事件分析系统中，报警数据的生成、采集、分析等功能模块，是以流水线方式加工数据。系统采集的原始数据报警集，首先要进入融合分析模块，去除冗余的、错误的报警，并将生成的安全事件集合（处理后）送入关联分析模块。多级特征分析方法能从不同层面清洗报警，根据 3.2.2 中描述的问题，该分析方法能做到：

- 报警融合。基于特定融合规则，去除原始数据集中重复报警；
- 报警评估。基于网络系统基本资源配置信息库和预先定义的判定因素、报警范围，报警评估算法可实现报警过滤功能，去除错误报警。
- 基于属性权重的报警聚类。基于历史报警数据集合，挖掘报警属性间的

联系，计算属性权重；结合相似度模糊理论方法，深度分析属性特征间的关系，最终形成可供关联分析的安全事件。

此外，针对网络中产生的海量报警数据，研究者开始利用大数据技术进行分析处理。本章也将描述特征分析方法与大数据技术的联合，并设计基于大数据技术框架的特征分析实现方案。

3.2 粗糙集理论简介

1982 年，Zdzislaw Pawlak[39]提出了粗糙集理论，该理论是一体现不确定性和不完整性的建模理论。从非一致、不精确、不完整的数据中，它能够协助完成决策支持和理论推导工作，挖掘数据中元素间隐藏的关系。该理论现已被应用于决策支持、工程、环境、银行、医药等领域。

粗糙集理论中有以下重要概念：

信息系统(Information systems)[39]:进行数据表述方式的基本概念，其被定义为一个二元组 S = (U,A)，其中 U 和 A 都是有限非空的集合，名称分别是论域(universe)和属性(attribute)集。

决策表(Decision tables) [39]： 决策表是一种特殊的信息系统，被定义为。其中 C 条件（condition）属性，D 是决策（decision）属性，在决策表中，A 被划分为 C 和 D 两个不相交集合。

对于安全事件分析系统来说，该理论和系统有两个契合点。首先，从大量不一致、不精确、不完备的原始报警数据中，分析攻击行为，推测攻击意图，其本身就是一个利用决策理论的分析过程；其次，系统所处理的报警数据是由一些属性组成，其可以对应信息系统中的属性集，也可以根据研究内容的需要对应决策表中的特定集合。

3.3 多级融合分析方法设计

安全事件分析系统通过关联分析模块，可以从大量、多源的报警数据中，挖掘出网络的攻击行为，进而识别攻击意图。但安全产品产生的原始数据中，往往包含大量冗余、低级、错误的报警信息，这会给关联分析模块带来效率和准确性的问题。为此，本节提出了多级特征分析方法，从不同角度对报警进行分析。模块化的设计便于流程和需要的重组，可以满足安全事件分析系统的分析需求。

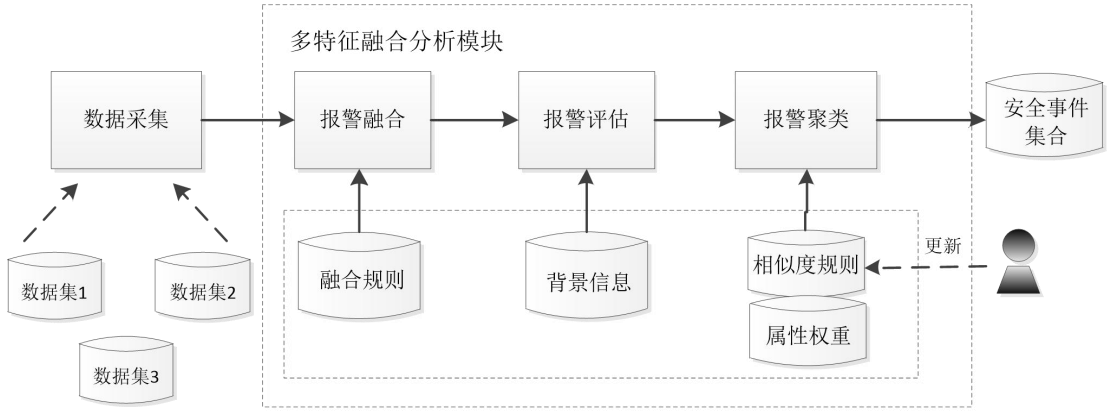

图 3-1 特征分析示意图

3.3.1 报警融合

该方法基于特定规则实现分析目标，该规则实现对众多本质一样或相近的报警进行规则判断，方法目标是过滤数据集中的重复报警,形成无重复的报警数据集，减少后续模块的计算压力。

3.3.1.1 融合规则设计

现假设一个报警队列 Queue，新的报警 a_{new} 按生成的时间顺序进入 Queue，若 $\{a_1, a_2, \cdots, a_n\}$ 是时间跨度为†的一个报警子队列 Q，方法按照以下四个规则进行融合比较：

Rule 1，假设†=10s，且 a_{new} 和 $\{a_1, a_2, \cdots, a_n\}$ 都在该时间跨度内，当检测到 a_{new} 和 $a_i, (\mathrm{i}=1,2,\cdots,\mathrm{n})$ 的重要属性相同时，判定 a_{new} 和 a_i 重复，公式如 3-1 所示：

$$\ell = \begin{cases} 0, & \longrightarrow Q = \{a_1, ..., a_n, a_{new}\} \\ 1, & \longrightarrow Q = \{h, ..., a_n\} \end{cases}, \qquad (3\text{-}1)$$

$$starttime_h = starttime_{a_1},\quad endtime_h = endtime_{a_{new}}$$

其中，ℓ 为 a_{new} 和 a_i 的判断结果，0 表示不需要融合，1 则相反。

Rule 2：假设†=10s，且 a_{new} 和 $\{a_1, a_2, \cdots, a_n\}$ 都在该时间跨度内，当检测到 a_{new} 和 $a_i, (\mathrm{i}=1,2,\cdots,\mathrm{n})$ 的重要属性相同，但 PluginID 和 sid 不全一样那么进行如下操作：

$$\ell = \begin{cases} 0, & \to Q = \{a_1, ..., a_n, a_{new}\} \\ \mathrm{pluginID_{new}} = \mathrm{pluginID_i} \,\|\, \mathrm{sid_{new}} = sid_i, & \to Q = \{h, ..., a_n\} \end{cases}, \qquad (3\text{-}2)$$

$starttime_h = starttime_{a_1}$, $endtime_h = endtime_{a_{new}}$

Rule 3：假设Ŧ=100s，且a_{new}和$\{a_1, a_2, \cdots, a_n\}$都在该时间跨度内，若检测到$a_{new}$和$a_i, (i = 1,2,\cdots,n)$的其它重要属性都相同，只有目的 ip 不相同，那么进行如下操作：

$$\ell = \begin{cases} 0, & \rightarrow Q = \{a_1, ..., a_n, a_{new}\} \\ \text{to}_{\text{new}} = \text{to}_i & \rightarrow Q = \{h, ..., a_n\} \end{cases}, \quad (3\text{-}3)$$

$starttime_h = starttime_{a_1}$, $endtime_h = endtime_{a_{new}}$

Rule 4：假设Ŧ=100s，且a_{new}和$\{a_1, a_2, \cdots, a_n\}$都在该时间跨度内，若检测到$a_{new}$和$a_i, (i = 1,2,\cdots,n)$的其它重要属性都相同，只有原始 ip 不一样，则进行如下融合操作：

$$\ell = \begin{cases} 0, & \rightarrow Q = \{a_1, ..., a_n, a_{new}\} \\ \text{from}_{\text{new}} = \text{from}_i & \rightarrow Q = \{h, ..., a_n\} \end{cases}, \quad (3\text{-}4)$$

$starttime_h = starttime_{a_1}$, $endtime_h = endtime_{a_{new}}$

3.3.1.2 融合过程设计说明

结合上述规则，该算法实现过程如图 3-2 所示。

（1）当特征分析模块接受到一条报警a_{new}时，先判断报警队列是否为空。如果队列为空，将a_{new}插入 Queue，进入步骤（2）；若不为空，则直接进入下一步；

（2）按顺序遍历队列中的所有报警；

（3）若队列不为空，从 Queue 中取出报警a_i，进入下一步；若为空队列，则把a_{new}插入 Queue 中，报警融合结束。

（4）根据上一个模块定义的规则。若需要融合，则将融合后的报警插入队列，并则跳回步骤（3）；若不需要，直接跳回步骤（3）。

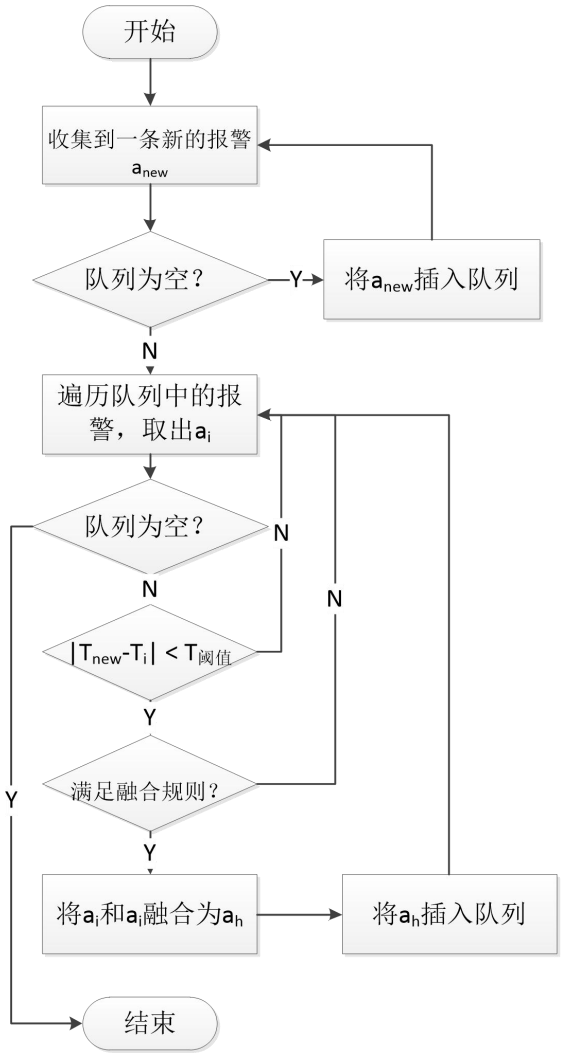

图 3-2 报警融合过程

本小节设计了报警融合的规则，并描述了融合实现方法，该方法可以去除原始报警集中重复的报警。本节详细介绍了融合方法中运用的四种特征分析规则，定义了每个规则的去工作的过程，最后，详细介绍了报警融合算法的实现过程。

3.3.2 报警评估

基于网络系统基本资源配置信息库和预先定义的判定因素、报警范围，报警评估算法可实现报警过滤功能，去除错误报警。该算法将从以下两个方面判断有效性：

（1）报警能否成功研判攻击。若该攻击行为不存在，则判断为误报；

（2）攻击目标是否有效。通过报警评估模块，多级特征分析方法能够去除错误的、无效的报警，将能够反映攻击行为的安全事件提交给关联分析模块。

报警评估算法通过匹配属性类型的方式，计算可信度值，根据该结果评估报

警的可信度。报警将进行五种形式的评估，分别是操作系统、端口、协议、服务名字、服务版本号。

- 操作系统评估：这里将检查目的主机的 OS 类型，并且尝试判断攻击报警是否是针对那个 OS 的。报警本身并没有 OS 信息，所以检查在数据库中是否有针对这个报警和操作系统特定的关联信息。为了简化匹配，目前只匹配以下的操作系统类型。如 Windows、Linux、Cisco、Solaris 等。
- 端口评估：进入交叉关联时，首先进行端口和协议匹配。显然，没有完成三次握手之前不能产生一个 TCP 报警，也就是说没有完成连接之前，无法进行对应用程序的攻击。一旦完成了 3 次握手，目的主机的端口必定是开启的。如果报警中包含的端口和数据库中保存的开启端口列表中有匹配，那么这个报警的可信度将保存不变。
- 协议评估：除了端口关联，还有一类是 UDP 流量。如果报警来源是 UDP 流量，不需要进行握手，攻击只需要一个数据包。所以如果目的 IP 只有一个 TCP 端口打开，攻击类型是 UDP 的，那么该报警的可信度将被降为 0。因为这个报警对目的主机基本没有任何影响。如果目的 IP 只有 UDP 端口打开，那么尝试 TCP 连接的攻击也不会有任何结果。
- 服务评估：为了进行服务关联，需要完成主机端口和对应协议对照。为了查询报警和服务之间的关联，需要查询相应数据库。
- 版本评估：这是交叉关联中最为精确的一个关联类型。版本指的是目的主机上安装的应用程序版本，版本关联之前首先需要进行服务关联。因为服务对不上，匹配版本号就没有意义了。如果版本匹配上了，就确认这是一个刻意的针对目的主机的攻击，可将这个报警的可信度设置为 9。

算法名称 报警评估算法
程序名称 FlowMatch(Set_{back}, Alert)
输入:
Set_{back} 报警相关主机的背景信息集合，包括操作系统、端口号等，用于验证报警；
Alert 报警数据，包括 pluginID，sid 等属性。
输出:
V_r 报警评估算法计算结果，至于为[0,10]

```
1: n ← Set_back 包含的属性个数
2: 初始化数组 re[1...n] ← 0, index ← 0
3: table 为报警对应的背景知识对应表，rule 为信息匹配表
4: for all g = (table(sid,(flow,os,version,port)),id_table) ∈ Set_back do
5:     if id_table∈ Set_back then
6:         re[index] ← re[index] + rule(sid,id_table), index ← index+1
7:     end if
8: end for
9: r ← 0
10: for i ← 1 to n do
11:     if re[i] ≥ r then
12:         r ← re[i]
13:   end if
14: end for
15: return r
```

报警评估算法具体步骤以下：

（1） 扫描事件中目标主机，获取主机的漏洞信息和主机信息，如操作系统类型、端口号、应用等；

（2） 找出报警对应的攻击类型，然后将步骤 1 中得到的信息和该报警所属的依赖信息进行匹配，匹配类型如表 3-1 所示；

（3）利用数据库中的报警的背景知识对应表，分别进行匹配，并设置可信度。

表 3-1 信息匹配表

类型	匹配	不匹配	没有足够信息	例子
OS	0	+1	不变	“OpenBSD”
Port	不变	0	不变	“80”
Protocol	不变	0	不变	“TCP”
Service	+2	不变	不变	“Apache”
Version	9	不变	不变	“1.3.33”

可信度值可以帮助网络管理员从大量报警中评估报警，提取可信的安全事件信息，并将安全事件送入关联分析模块,为关联分析带来便利。

3.3.3 报警聚类

上一章中，报警融合方法设计了特定融合规则，能够去除原始报警数据中大量的重复报警，但没有考虑报警类别属性。报警信息往往能体现攻击的一些特征，如攻击类别，因此对于同一攻击，其触发的报警极有可能是相似的。针对上述情况，本章提出了基于属性权重的报警聚类方法，该方法根据报警类别，深入分析报警属性之间的关系，可以减轻去冗余计算压力，并达到更好的预处理效果。

现有许多基于属性权重的聚类操作，其权重值很大程度依赖于经验。而本节提到的聚合算法，是通过挖掘历史报警数据集，结合决策理论获得各属性的权重，为报警相似度计算和报警聚合方法提供客观支持。下面将详细描述分析方法的步骤：

3.3.2.1 基于粗糙集理论的属性权重计算方法

属性权重计算方法如图 3-3 所示，主要包括报警属性离散化和属性权重计算两块内容。

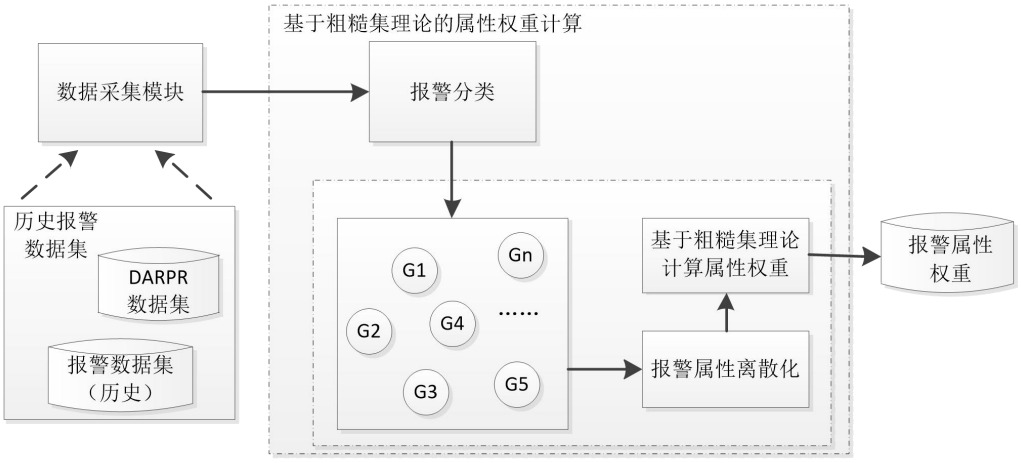

图 3-3 报警属性权重计算流程示意图

首先将对一些重要的属性，包括目的 ip、源 ip、目的端口、源端口、sid 和报警类型进行属性权重计算方法设计。该方法包括报警属性离散化和属性权重计算两个步骤.

（1）离散化报警日志属性

（一）属性定义：根据 3.1 中设定的概念，定义 C 为条件属性，其包括 from，to，sourcePort，destPort，sid，分别记为条件属性值 $C_i(i=1,2,3,4,5)$；定义 D 为决策属性，即报警类别，包括 Scan，Dos，U2R，R2L，Misc，UE 六种类别；

（二）样本空间分类：将样本空间按照 Scan，Dos，U2R，R2L，Misc，UE 分为六类，分别记为决策属性类 $D_j(i=1,2,3,4,5,6)$；

（三）区间构建：条件属性值 $C_i(i=1,2,3,4,5)$ 分别对每类 D_j 进行由大到小进行排序，假设样本集中所有数据的 C 值都不相同，可构造 C 的 n 个不同的数据区间，分别记为 $I_1, I_2, ..., I_n$，其中，n 为样本个数；

（四）当合并任意区间时，计算相应损耗值 Infor_Loss：信息损耗值 Infor_Loss 的计算方法如下：

$$\text{Infor\_Loss} = \text{E(I)} - \text{E}(\text{I}_{\text{p+1}}, \text{I}_{\text{p+2}}, \ldots, \text{I}_{\text{p+m}}) \tag{3-5}$$

其中，E(I) 表示合并其中 m 个区间后的信息熵，$\text{E}(\text{I}_{\text{p+1}}, \text{I}_{\text{p+2}}, \ldots, \text{I}_{\text{p+m}})$ 表示合并其中 m 个区间的熵值；

（五）确定区间样本，实现条件属性离散化：合并区间的规则如下：设定一阈值，当信息损耗值 Infor_Loss 小于该阈值时，则接受该区间的合并；若信息损耗值 Infor_Loss 大于该阈值时，则放弃此次区间合并；这样可以得到一个合理的区间样本，实现条件属性的离散化。

（2）动态分配报警事件属性权值

（一）计算 C_i 对 D 的重要程度 $\sigma_D(C_{\text{i}})$：属性重要程度计算公式如下：

$$\sigma_D(C_{\text{i}}) = 1 - \frac{\gamma_{C_{\text{i}}}(\text{D})}{\gamma_{\text{C}}(\text{D})} \tag{3-6}$$

其中，$\gamma_D(C)$ 表示 D 对 C 的依赖程度，$\gamma_{C_{\text{i}}}(C)$ 表示 D 对 C_i 的依赖程度；

（二）属性权值化处理的处理公式如下：

$$\eta(C_{\text{i}}) = \frac{\sigma_{\text{D}}(\text{C}_{\text{i}})}{\sum \sigma_{\text{D}}(\text{C}_{\text{i}})} \tag{3-7}$$

其中，对于每一个 C_i，都可以得到 i 的权值 $\eta(C_{\text{i}})$，i 表示报警属性。

3.3.2.2 相似度计算定义

结合报警属性权重计算结果，本章定义相似度计算方法。其中 S 代表两个属性的相似度。

（1）属性 IP 比较

$$S(ip_i, ip_{new}) = \begin{cases} 1, & ip_i = ip_{new} \\ 1 - L(ip_i, ip_{new}) / H, & ip_i \neq ip_{new} \end{cases}, \quad (3\text{-}8)$$

其中 L 表示两台主机到达共同父节点的最长路径，H 表示层次数。

（2）端口属性比较

$$S(port_i, port_{\text{new}}) = \begin{cases} 1, & port_i = port_{\text{new}} \\ 1 - L(port_i, port_{\text{new}}) / H, & port_i \neq port_{\text{new}} \end{cases} \quad (3\text{-}9)$$

（3）sid 属性比较

$$S(sid_i, sid_{\text{new}}) = \begin{cases} 1, & sid_i = sid_{new} \\ 0 & sid_i \neq sid_{new} \end{cases} \quad (3\text{-}10)$$

由上述定义可得，对于两条 Scan 类型的报警 a_{new} 和 a_{i}，其相似度的值分别为 $\delta(x_{i,1}, x_{\text{new},1}) = S(from_i, from_{new}) \times \eta(C_{from})$，$\delta(x_{i,2}, x_{new,2}) = S(\text{to}_i, to_{new}) \times \eta(C_{to})$，

$\delta(x_{i,3}, x_{new,3}) = S(srcPort_i, srcPort_{\text{new}}) \times \eta(C_{srcPort})$，

$\delta(x_{i,4}, x_{\text{new},4}) = S(destPort_i, destPort_{new}) \times \eta(C_{destPort})$，

$$\delta(x_{i,5}, x_{\text{new},5}) = S(sid_i, sid_{new}) \times \eta(C_{sid})，且 \sum_{\text{i=1}}^{\text{n}} \eta(C_{\text{i}}) = 1 \quad (3\text{-}11)$$

其中 x 表属性集合 $\{from, to, srcPort, destPort, sid\}$。故总的相似度计算和为:

$$sim(a_i, a_{new}) = -\sum_{k=1}^{n} \delta(x_{i,k}, x_{new,k}) \quad (3\text{-}12)$$

根据上述公式的计算结果，若报警安全事件相似度 $sim(a_i, a_{new})$ 小于阈值 λ，则判断为不需要聚合；否则，则判断为需要；阈值 λ 的计算方法如下：多个样本区间中每一个条件属性值 C_i 和决策属性类 D_j 的报警安全事件相似度 $sim(a_i, a_{new})$ 组成一个收敛数列 $\{sim(a_i, a_{new})\}$，该数列收敛于 λ，即为阈值。

3.3.2.3 基于属性权重分析的报警聚类实现流程

基于 3.3.2.1 中设计的属性权重分配方法和相似度计算方法，首先将历史报警数据集作为实验集合中，用于计算报警属性权重；然后，将经过报警融合模块

和评估模块的结果送入聚类处理模块；最后，基于属性权重，根据报警相似度算法处理进行聚类操作。处理流程如图 3-4 所示。

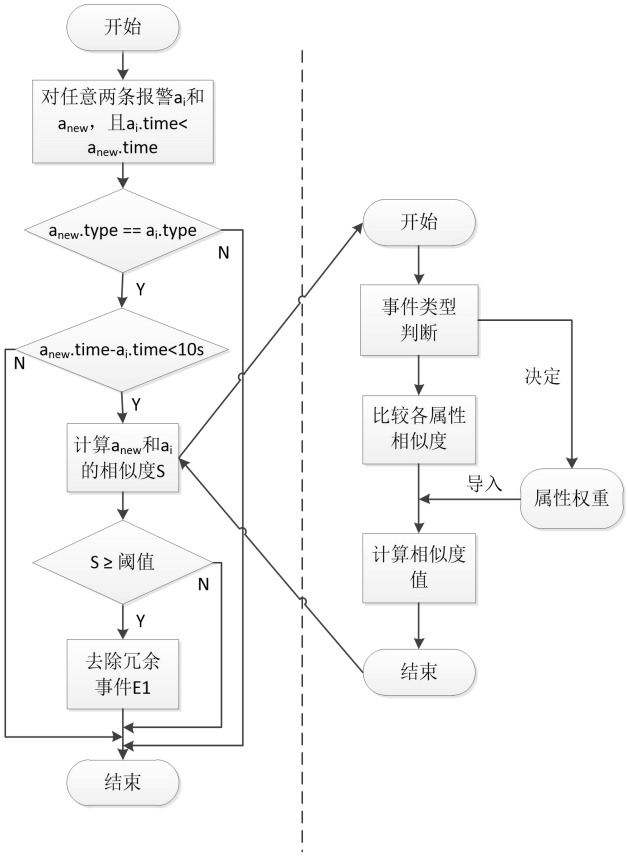

图 3-4 基于属性权重分析的报警聚类流程（单流程）

（1）任意两条报警 a_{new} 和 a_i，假设 $Time_i<Time_{new}$；

（2）若报警的类型不相同，则结束聚类操作；若相同，判断两条报警的时间间隔，若超过 10s，则结束，若不超过，则进入下一步；

（3）根据报警类型，获取对应的属性权重数据，按照公式计算相似度值 S；

（4）若相似度值大于或等于阈值，则进行报警聚类操作；否则不进行聚类；最后结束该流程。

3.4 基于 Spark 分布式技术实现研究

3.3 节中描述了特征分析方法的实现方法和具体流程，从传统角度来看，该设计方案流程合理，可以有效的完成预处理工作。但随着网络技术快速发展，大数据概念的兴起，该方案会出现一些处理效率上的问题。下面进行具体分析。

对于消息队列，在特定时间段内，任意两条报警都要进行特征分析操作。在不考虑分析分析方法复杂度为前提，假设队列中存在 N 条报警，遍历所有报警

最坏情况下的时间复杂度为 $O(N^2)$，在小规模网络环境中，还能通过改进程序实现或提升分析机器性能的手段保证分析的效率，但当大数据量的报警涌入系统，成千上万条数据需要在秒级的时间段内处理，如何实时接收数据、有效完成分析处理是系统实现的技术难点。

本文借助 Spark Streaming 分布式技术框架来应对这个大数据安全问题。Spark Streaming 是核心 Spark API 的扩展，可实现可伸缩、高吞吐量、可容错的实时数据的流处理。该框架提供 map, reduce, join 或 window 等高级算法，通过任意的组合，可以帮助开发人员实现目标处理流程。在处理过程中，我们将系统特定时间段内接收到的数据组成一个集合，结合强大的分布式处理能力，能够高效完成分析处理这些时序相关的数据集。

图 3-5 描述了具体的处理流程。JavaStreamingContext 调用 socketTextStream(ip, port) 方法，接收获取经 Json 序列化的报警信息 Line DStream。Spark 提供 JavaReceiverInputDStream 类作为流式数据写入存储对象 JavaDStream，该类是 DStream 的子类。一个 DStream 对象包括一系列连续的 RDD，每个 RDD 存储时间间隔内接收到的数据，该时间间隔在 JavaStreamingContext 对象初始化时设置。

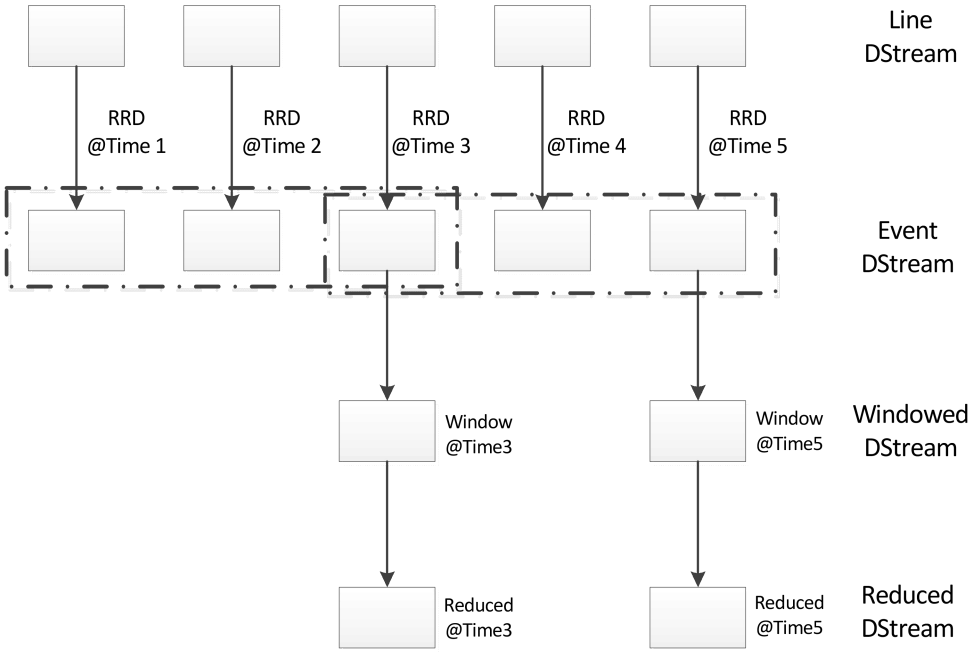

图 3-5 基于 Spark Streaming 的特征分析流程

任何应用在 DStream 的操作都会转化为对 RDD 的操作。在报警预处理过程中，对 Line DStream 的处理步骤如下：

- 调用 flatMap 操作，将流式化数据转化为为 Event 对象，生成 Event

DStream。

- 调用 window 方法，选取特定时间段的 RDD；由于相似度算法对时间属性的特殊要求（两条报警时间间隔大于一定值，判断为非冗余报警），故设置时间段为 5s；
- 按照步骤 b，针对特定 RDD，调用 mapPartitions 方法，获取 RDD 中 Event，与结果集中的 Event 进行相似度计算，完成报警去冗余操作。

3.5 本章小结

安全事件分析系统采集的原始数据集中，包含大量重复、错误的报警，这无疑会给系统在关联分析上带来不便。针对这个问题，本章提出的了多级特征分析方法。该方法根据报警融合规则，能过滤掉大量的冗余报警；其次，通过报警评估算法，能有效去除融合数据操作后的错误报警；经过前两个模块后，报警集合还需要通过基于属性权重分析的聚类模块，去除数据集中的冗余报警，向系统的关联分析模块提供有效的安全事件数据集合。

4 基于攻击图的多源事件分析方法

本章提出了分析方法，包括攻击建模方法设计和面向攻击场景的多源事件关联方法研究。该方法基于研究者对网络攻击的理解，将攻击分装为格式化、步骤化的攻击场景，并设计了层次化的关联分析规则；基于特征分析方法研究结果，该方法能关联多源安全事件，有效构建多步骤网络攻击，研判网络攻击行为，进而分析攻击意图，并进行攻击预警。

4.1 攻击建模方法

进行网络攻击是一种系统性的行为，沿着一定的途径或步骤，最终达到攻击者的目的。一个完整的攻击，不仅仅只是一个报警事件所能描述体现的，换言之，攻击触发的报警事件不是孤立存在的，而是有一定时间上、逻辑上关联的。之前的一系列报警可能预示着后面更深入的攻击步骤。针对不同攻击类型的特点，结合通用的攻击描述方式，描述网络攻击过程中体现的逻辑性，为基于攻击场景的关联分析研究提供理论支持，提高关联分析能力。

4.1.1 基本概念

定义 1：事件。将经过报警聚合处理的数据定义为事件（Event），即 Event(pluginID, sid, from, to, sourcePort, destPort, measure)。

其中 pluginID 和 sid 表示事件的类型。pluginID 是插件编号，告知我们事件的来源；sid 是该插件对应的事件编号，代表着具体的事件信息及所属的类别。表 4-1 中列举了一些典型的插件编号信息。

表 4-1 插件编号映射表

插件名	编号
Snort	1001
Ntop	1501
Nmap	1502

为便于攻击构建研究及攻击图的表达，事件的 from, to, sourcePort, destPort 字段有以下两种描述方式：

① 一般形式：from, to 可取值特定的 IP；sourcePort, destPort 取值可为端口号。

② 特殊形式：from, to 可取值为 n:SOURCE_IP 或者 n:DEST_IP，sourcePort, destPort 取值可为 n:SOURCE_PORT 或者 n:DEST_PORT，其中 n 表示攻击场景

的层数，例如 3：SOURCE_IP，2：DEST_PORT，分别表示第 3 层攻击的源主机 IP 和第 2 层攻击的目的主机端口。当然也可以用 ANY 关键字表示任何的 IP 或端口。

定义 2：背景事件。是一种特殊的事件，除上述事件定义的元素以外，可用（condition，timeout，value，interval）表示，这类信息由监测器提供，通过扫描、嗅探等安全服务，可以获取目标主机的操作系统信息，主机状态等信息。

表 4-2 背景事件属性描述信息

属性名称	描述
condition	需要符合的条件
timeout	规则匹配的等待时间,如果超出，匹配将失败
value	特定条件的数值
interval	表示监测的时间间隔

定义 3：攻击场景。表示具有相同攻击行为的事件集合，可表示为（id, name, rule, category）。Id 是场景的编号，具有唯一性；name 为场景的名称；rule 标识组成场景的规则，每个场景包含的 rule 数量大于 1；category 是场景所属的攻击类别，其包括表 4-3 中的内容：

表 4-3 攻击场景归类情况

名称	描述
Scan	扫描类
Buffer Overflow	缓冲区溢出攻击
Denial of Service	拒绝式服务攻击
Web Attack	Web 网站攻击
Malware	木马蠕虫威胁
Miscellaneous	混合攻击

4.1.2 问题描述

在大规模网络环境中，网络安全产品种类各异，虽然特征分析模块清洗原始报警，并生成安全事件集合，但事件的表现形式略有差别。因此，如何有效地分析这些安全事件是安全分析系统面临的挑战。

一次典型的攻击过程通常经历五个步骤：信息收集、漏洞探测、权限提升、实施破坏和攻击痕迹清除[40]。以信息收集为例，攻击者会通过扫描工具，全面了解收集目标主机的信息，对于基本的可访问查询，服务可用查询等网络行为，不同的安全产品触发报警的阈值不同。攻击往往是由多个步骤组成，安全产品根据

通用或者自定义规则，触发的报警事件，其表述的攻击粒度是特定的；当各类安全产品部署在网络，面对不同类别、不同粒度的报警，管理人员无法理解发现攻击行为，无法重构出攻击步骤，网络防御、保证网络安全便无从谈起，因此必须解决统一攻击描述和粒度的问题。

4.1.3 建模目标

研究者网络攻击的目的是减少网络系统受到攻击，因此需要研究者对攻击有具体、实际的认识。黑客获取并利用网络系统内的漏洞信息，逐步完成攻击行为，并达到攻击目的。如果不知道如何攻击，不知道攻击的目的，网络安全就无法保证。从攻击所触发一系列的多源安全事件中，挖掘出有意义的信息，进而识别攻击者的意图，甚至提出应对措施，这个过程能让我们理解网络安全中“攻击”、“防御”之间密切的关系。因此，结合网络攻击知识，高效构建网络攻击模型，实现攻防行为具体化是我们的研究目标。

攻击建模有以下特征：

- 可扩展：能进行原有攻击模型的延伸，也可增加不同的攻击建模结果；
- 可通用：支持来自不同产品安全事件的接入；
- 易理解：保证能准确、易于理解地描述攻击行为。

在安全事件分析系统中，关联分析模块利用攻击规则和相应算法进行攻击建模，实现生成层次化关联规则的目标，进而关联安全事件，研判网络攻击行为，识别攻击意图。

4.2 攻击模型设计

通过攻击建模方法的研究，将网络攻击转化为具体的攻击场景，把不同时间、不同层次的安全事件联系起来，是实现基于攻击场景的多源关联分析的前提。按照网络安全攻击表现方式，本小节描述了攻击场景具体结构。攻击场景以 XML 格式存储，作为关联规则安全事件分析系统的关联规则，研判网络攻击行为，识别攻击意图。

如图 4-1 中的规则所示，该攻击场景名为“微软的 Remote Desk 服务发现的网络扫描”，编号为 10，具体的攻击描述是在网络上检测到一个 RDP 扫描，表明可能有攻击者正在寻找网络中的系统漏洞；该攻击能够影响的系统为 Windows。

针对这个攻击，提出以下建议：

（1） 安装应用必要的补丁，以消除这种威胁；

（2） 禁用不常用的服务，创建访问列表，阻止未知的计算机访问该服务；限制远程访问。

```
<scenario id="10" name="AV-FREE-FEED Network scan, Microsoft Remote Desktop service discovery from SRC_IP">
    <step type="detector" name="MRDS0" occurrence="1"
     from="ANY" to="ANY" sourcePort ="ANY" destPort="3389" pluginID="1001"
     sid ="2001972" protocol="TCP" >
          <steps>
               <step type="detector" name=" MRDS1" reliability="4" occurrence="2"
               from="1:SRC_IP" to="ANY" time_out="600" sourcePort ="ANY"
               destPort ="3389" pluginID ="1001" sid ="2001972" protocol="TCP"
               measure=”Apply necessary patches to mitigate this threat”>
                    <steps>
                         <step type="detector" name="MRDS2" occurrence="20"
                          from="1:SRC_IP" to="ANY" time_out="1800" sourcePort ="ANY"
                          destPort="3389" pluginID ="1001" sid ="2001972" protocol="TCP"
                          measure=”disable unusued services”>
                           <steps>
                              <step type="detector" name=" MRDS3" occurrence="200"
                               from="1:SRC_IP" to="ANY" time_out="7200" sourcePort="ANY" destPort
                               ="3389" pluginID ="1001" sid ="2001972" protocol="TCP"
                               measure=”disable unusued services” >
                                 <steps>
                                    <step type="detector" name="MRDS4" occurrence="20000"
                                     from="1:SRC_IP" to="ANY" time_out="21600" sourcePort ="ANY"
                                     destPort="3389" pluginID ="1001" sid ="2001972" protocol="TCP"
                                     measure=”create an access list to prevent unknown computers
                                     accessing this service and restrict remote access” />
                                 </steps>
                              </step>
                           </steps>
                         </step>
                    </steps>
               </step>
          </steps>
     </step>
</scenario>
```

（3） 访问软件供应商网站的建议解决方案或计划更新。

图 4-1 规则实例

图 4-1 所示场景中包括 5 个攻击步骤，详细信息如下：

（1） 主机 H 通过 TCP 协议对任意 IP 的 3389 端口的进行访问，该访问将触发测器插件 Snort 检测到 1 次 sid 为 2001972 的报警；

（2） 在十分钟内，若 Snort 检测到 H 主机的该访问行为，且次数超过 2 次；

（3） 三十分钟内，若 Snort 检测到 H 主机的该访问行为，且次数超过 20 次；

（4） 两个小时内，若 Snort 检测到 H 主机的该访问行为，且次数超过 200 次；

（5） 六个小时内，若检测到 H 主机的该访问行为，且次数超过 2000 次；

4.3 面向攻击场景的多源事件关联分析

一个攻击往往是由多个步骤组成的，因此其触发的事件在时间上或逻辑上是有一定相关性的，这些事件构成了一个攻击场景。本节提出了基于面向场景的关联分析方法，该方法对事件聚合后的事件进行关联分析操作，能深入挖掘安全事件之间的关系，构建出已匹配的攻击，帮助管理者识别攻击意图，并实现攻击预警功能。

4.3.1 基于攻击图的关联分析算法

一个完整的攻击，一般由多个相互联系的步骤组成。后一个步骤的触发条件，往往是前一步骤满足数量、时间、类型等要求。在安全事件分析系统中，报警就是攻击步骤的最初始、直观的数据反映，这些报警格式各异，表达方式、粒度也不一样，通过特征分析模块将所有报警转化为统一的安全事件数据集，再由关联分析算法构成攻击场景。

如 4.1 小结所述，关联规则由不同步骤的安全事件组成，具有可扩展、可通用、易理解的特点。面向层次化规则的关联分析方法能够关联安全事件，构建匹配的攻击场景，挖掘出层次化、多步骤的攻击行为；此外，相比于安全事件，面向场景的攻击描述能更生动形象地描述网络中发生的入侵行为。该关联流程如图 4-2 所示。

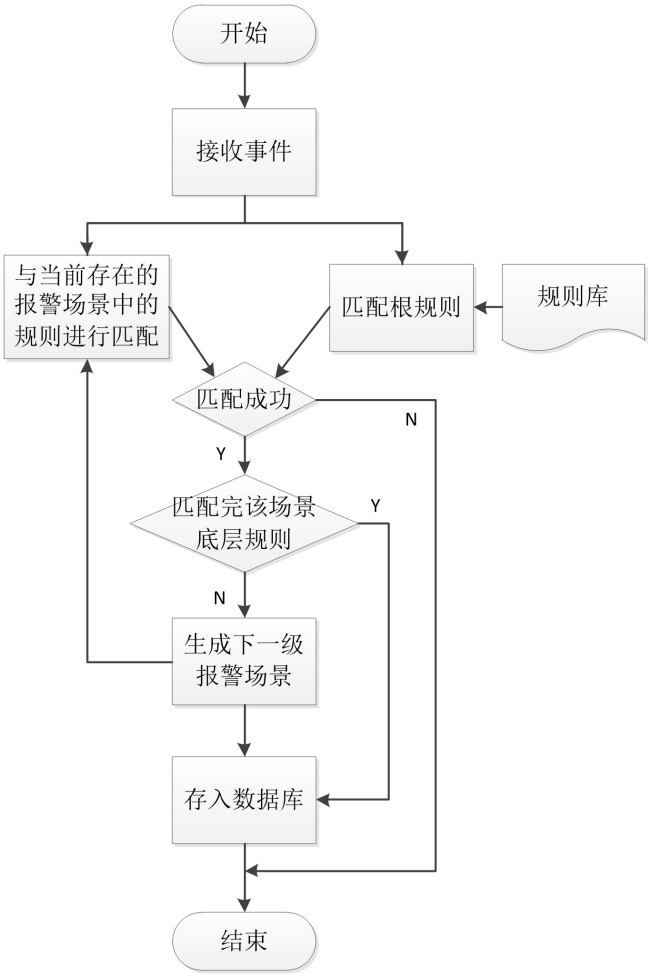

图 4-2 关联流程示意图

（1）进入分析模块的事件首先将和 backlog 集合完成匹配检查。backlog 就是已经和事件进行匹配，但未完成的规则。每一个到达关联分析模块的事件都会和 backlog 集合进行匹配；

（2）对于存储在 backlog 列表的事件会被存储在数据库中的 backlog 表中供以后参考查询使用；一个关联规则事件产生后，这个事件将会被匹配为一个具体的攻击；关联规则事件创建后将会被插入到 backlog 和 backlog_event 表中；

（4）关联规则事件重新进入到事件队列，和其他事件一样，可能会进行优先级划分，储存，转发。至此，事件到达了关联规则的第一层，当其它事件到达服务器端时，backlog 队列将不为空，新的事件将会按照另外一个流程进行匹配检查。过程如下：

（5）首先匹配检查该事件是否和 backlog 中的储存的关联规则的子节点之间是否有匹配。如果是，事件数据和事件将会被储存在关联规则列表中，此过程

和第一条事件的处理过程一样。

（6）如果上一步中事件有匹配，一个新的关联规则事件被匹配出来并且被插入到 backlog 和 event 表中以作为后续的引用。

（7）规则的等级被审查。如果匹配的是最后一条规则，那么储存在 backlog 中的关联规则保持不变。如果接收到了匹配一个关联规则事件，并且它们进入了关联规则的内部层次，那么该关联规则将会被插入到一个队列，在稍后将被删除。

4.3.2 面向攻击场景的攻击预警

基于层次化规则的攻击场景能够挖掘多源安全事件的关系，其目的是描述具体的、完整的攻击路径。然而，攻击行为包括多个步骤，在分析过程中，场景往往不是完全匹配的，部分场景只匹配了一两步，或者某个攻击在一个时间段内完整地匹配了很多次。本节将基于 4.3.1 的研究结果，进一步丰富攻击场景的用途，通过预测和回溯攻击步骤，识别攻击意图进行攻击预警。

如图 4-3 所示，当管理员需要查看某一个检测到的攻击场景时，分析模块会显示层次化攻击，利用攻击图技术的表述方式，生动展示攻击已触发的步骤，并提出相对应的预警信息和预防措施。

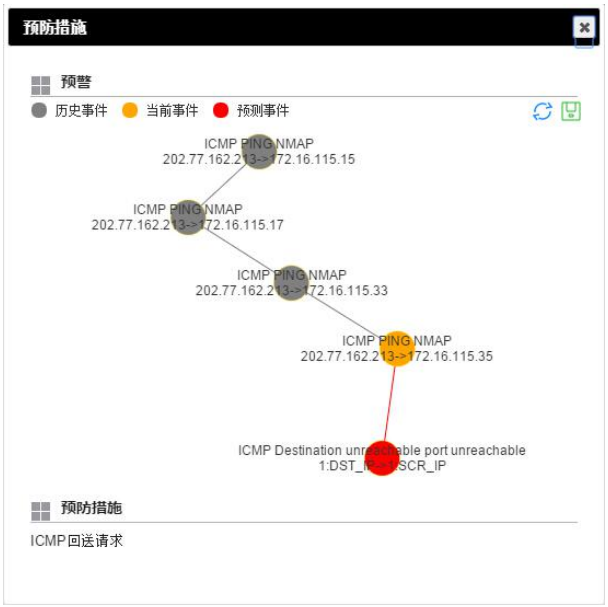

图 4-3 面向攻击场景的攻击预警效果图

4.4 本章小结

攻击往往是由多个步骤组成的，面对不同类别、不同粒度的单一报警，很难从中发现网络中真正攻击行为，也无法重构出攻击步骤。针对上述问题，本文提

出了基于攻击图的多源事件分析方法。通过构建攻击模型，封装网络攻击知识，并以树状文本形式存储；基于层次化关联规则，关联特征分析后的安全事件数据集，利用攻击图方法构建完整的、多步骤的攻击路径，并利用攻击图技术实现攻击预警。

5 实验与测试

第三、四章节中研究成果都已应用在此安全事件分析系统中，该系统可对网络安全设备生成的数据进行集中处理和分析。基于数据采集框架和分析引擎，本章将搭建安全事件分析系统，并以此作为实验环境对本文的研究内容进行研究和验证。

5.1 实验数据与环境

实验中使用的数据源由以下两个个部分组成：

（1） 真实的网络数据集。

（2） DARPR 2000 数据集。

其中第一部分的数据是利用 Snort 安全插件采集的真实数据集，该插件部署于某高校的一个中型网络和外网之间。该数据集积累了超过 17 天的数据，数据集已通过映射的方式将真实 ip 匿名化。第三部分仿真数据是通过人工方式生成，用于模拟攻击，触发攻击场景。DARPR[41]是目前最为全面的攻击测试数据集，也是作为研究领域共同认可及广泛使用的基准评测数据集。本文选用 2000 年的 DARPR 数据作为实验数据。

本实验方案中使用了部分上述的三类数据集进行实验分析。这部分数据集由 5 万条报警事件组成，去除没有类别的报警，剩余可用 40435 条。

5.2 实验结果分析

5.2.1 属性权重计算实验结果

首先，实验按照第三章所述展示属性权重分配结果。不同的报警类型，对应不同的权重分配结果。根据报警类型，现将其分为：Probe and Scan, Denial of service(Dos), User privilege gain(R2L), Super-user privilege gain(U2R), Misc 和 Unknown Execution。

如表 5-1 所示，分配结果基本符合网络安全常识。例如，Dos 攻击一般由大量的肉机发动攻击，因此相比源信息，关联分析需要相对关注攻击的目标信息。表中 Dos 类的分配结果中，原始 ip 和端口的权重都比较低，接近于零。

表 5-1 报警属性权重分配表

	Src_ip	Dst_ip	Src_port	Dst_port	Sid
Scan	0.3	0.15	0.4	0.15	0
Dos	0	0.4	0.05	0.25	0.3
U2R	0.5	0.2	0	0.1	0.2
R2L	0.2	0.2	0.1	0.5	0
Misc	0.25	0.35	0.3	0.1	0
UE	0.05	0.3	0.4	0.2	0.05

5.3.2 多级特征分析实验结果

根据第三章描述的聚合过程，实验将采用表 5-1 的权重数据来计算任意一对报警之间的相似度值。报警去冗余的结果图表 5-2 所示。

表 5-2 报警去冗余结果

Threshold value	Total alerts	Left alerts	Reduction efficiency
0.6	40435	3522	91.3%
0.7	40435	4647	88.5%
0.8	40435	6546	83.9%
0.9	40435	10140	74.9%

实验分别选用四种阈值对相同数据集进行去冗余操作。如表 5-3 所示，选用的阈值越小，去冗余率越高。当选取 0.6 为阈值时，去冗余率达到了 91；实验结果表明平均去冗余率为 84.6%。

表 5-3 报警去冗余结果（续）

	Scan	Dos	U2R	R2L	Misc	UE
Total	16070	30	1556	301	9761	12717
Left(threshold=0.6)	1575	10	533	60	644	700
Left(threshold=0.7)	1875	10	658	69	991	1044
Left(threshold=0.8)	2456	12	702	69	1072	2235
Left(threshold=0.9)	3022	12	801	114	3291	3544

根据上述的实验结果，在表 5-4 中，本小节详细分析了方案 1、方案 2 和本方案。三个方案都将属性数量作为去冗余操作的指标，并根据方案特点排除无关属性；本方案和方案 1 将类别属性作为前置条件，即相同类别的报警需要进行相似度计算，不同类别则不用，方案 2 没有考虑类别属性；至于去冗余效率，三个方案的结果相近，本方案的效率稍高于另外两个方案。

表 5-4 现有去冗余方案比较

	选用的属性个数	基于攻击分类和权重	效率
本方案	5	是	约为 84.6%
方案 1[42]	6	是	约为 80%
方案 2[6]	7	否	约为 80%

5.3.3 关联分析实验结果

本小节完成基于攻击图的多源事件分析实验，并将这三个小节的实验结果进行统计，形成一个完整流程的分析结果，实验结果如表 5-5 所示。

表 5-5 整体实验结果

报警数量	特征分析/阈值	场景匹配		攻击匹配
40435 条	5516 条/0.8	57 条	8 个/30 次	12 个

从实验结果看，报警经过多级特征分析后，排除率为 86.3%，主要是端口扫描等初步的探测行为；剩下的 5516 条安全事件被送入场景关联分析模块，匹配了 8 个攻击场景，成功关联分析了 57 条安全事件，其中部分场景匹配了多次，如 DDOS attack；构建了 12 个状态攻击图，并实现了 18 个攻击的预警。表中展示的结果和实验数据介绍的资料基本一致。

6 结论与展望

6.1 本文总结

本文首先介绍了大数据安全的科学内涵，描述了大数据技术的安全需求，并强调了大数据技术在保障系统安全和应用安全上的重要性；当前网络安全设备管理面临众多挑战，例如原始数据存在大量、重复、误报的低级报警，报警语义不统一等。本文提出基于攻击图的多源安全事件关联分析方法，该方法包括基于多特征融合分析报警分析方法和基于攻击图的多源事件分析方法。下面将详细描述本文的研究内容：

（1）从报警数据预处理、攻击图生成方法和关联分析方法等角度进行研究现状调研，并提出了存在的问题和本文研究目标；

（2）针对原始报警大量重复、误报的问题，设计并基于多特征融合分析的报警分析方法，该方法包括报警分析的一系列操作，包括融合、验证和聚类。报警融合基于特定规则，删选去除报警集合中的冗余报警；结合真实网络环境中的配置、漏洞信息，验证报警是否可信；设计并实现基于模糊理论的聚类算法，有助于聚合相似报警。

（3）面对海量报警数据，传统报警融合方法无法保证处理能力和效率，本文设计了基于大数据技术的特征分析方法。将安全产品生成的报警转化为流式数据，利用 Spark Streaming 框架的分布式能力，为处理大数据安全数据提供技术保障。

（4）为达到更好的报警聚类效果，设计并实现了基于粗糙集理论的属性权重计算方法，在相似度计算的基础上给属性赋予权重。权重计算需要利用信息理论知识离散化报警属性，并基于粗糙集理论，将报警集合与决策树概念结合，确定条件属性和决策属性，进而挖掘数据中元素间的隐含关系。

（5）描述攻击建模的目的和方法，并基于报警与攻击步骤的关联关系，设计层次化关联规则，该规则具有可扩展、可通用、易理解的特点；结合层次化的关联规则，提出了基于攻击场景的多源事件关联分析方法，匹配对应攻击场景，分析攻击者的潜在目的，还可以利用攻击图实现攻击预警。

（6）搭建实验环境，根据研究内容设计实验方案，采用网络真实报警数据集和 DARPR 数据集作为实验分析样例，分析结果显示该方案能去除原始数据中

的重复、错误报警，构建完整攻击，识别攻击者真正的意图。

6.2 展望

针对本文提出了研究课题，今后可在以下几个方面继续开展：

（1）深入多级特征融合分析方法。报警预处理分析是网络安全分析领域一个关键研究点。本文提出的多特征融合报警分析方法从多冗余性和正确性上进行分析研究，是提高分析准确性和效率的一个有益的尝试。这三个模块化的分析方法是处理流程更灵活，但也使得模块联系不紧密。此外，报警聚类方法的阈值关系到分析效果，因此阈值的设置一直都是一个棘手的问题。

（2）研究关联分析规则更新问题。安全事件关联分析模块是安全事件分析系统的核心内容，该模块需要依据关联规则库的关联规则，形成多步骤攻击场景。其中关联规则库由项目组成员总结已有攻击场景形成，如何使得安全事件分析系统能够根据未知攻击场景进行预设或者新出现的攻击场景能够智能学习添加是网络安全事件分析的难题。

（3）研究如何实时关联分析海量安全事件。由于入侵行为越来越广泛、复杂，海量安全事件关联分析给服务器带来巨大计算压力。面对海量安全事件的各类关联分析的计算操作和攻击场景构建的需求，传统的数据处理方式无法适应多数据源，高并发的分析场景。如何利用分布式技术或云计算的优越性处理海量的安全事件将是项目研究的将是一个挑战。

（4）本文所依托的安全事件分析系统诣在展现网络的安全状况，如何有效地将系统分析结果传达给管理员，也是需要考虑和优化的问题。

参考文献

[1]360 安全白皮.2015. http://zt.360.cn/report/

[2]王培. 中国信息安全市场现状与未来展望[J]. 中国信息安全, 2015(3):79-81.

[3]Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In Proceedings of the 6th ACM

[4] 张玉刚. 基于模糊聚类和因果关联的攻击场景构造方法的研究与实现[硕士学位论文]. 华中师范大学, 2009.

[5] Keunsoo Lee, Juhyun Kim, Ki Hoon Kwon, Younggoo Han, and Sehun Kim. Ddos attack detection method using cluster analysis. Expert Systems with Applications [J], 2008, 34(3):1659–1665,.

[6] Shisong Xiao, Yugang Zhang, Xuejiao Liu, and Jingju Gao. Alert Fusion Based on Cluster and Correlation Analysis. International Conference on Convergence and Hybrid Information Technology[C], 2008, pages 163 - 168.

[7] Alireza Sadighian, José M Fernandez, Antoine Lemay, and Saman T Zargar. Ontids: A highly flexible context-aware and ontology-based alert correlation framework. In Foundations and Practice of Security[C], 2014,pages 161–177.

[8] Fu Xiao, Shi Jin, and Xie Li. A novel data mining-based method for alert reduction and analysis. Journal of Networks [J],2010, 5(1):88–97.

[9] Ho C Y, Lai Y C, Chen I W, et al. Statistical analysis of false positives and false negatives from real traffic with intrusion detection/prevention systems[J]. Communications Magazine IEEE, 2012, 80(1):18-23.

[10] Richhariya V, Sharma N. Optimized Intrusion Detection by CACC Discretization Via native Bayes and K-Means Clustering[J]. International Journal of Computer Science & Network Security, 2014.

[11] Pietraszek T. Using Adaptive Alert Classification to Reduce False Positives in Intrusion Detection[C]. Recent Advances in Intrusion Detection: 7th International Symposium, RAID 2004:102-124.

[12] Roschke S, Cheng F, Meinel C. A Flexible and Efficient Alert Correlation Platform for Distributed IDS[C]// Proceedings of the 2010 Fourth International

Conference on Network and System Security. IEEE Computer Society, 2010:24-31.

[13] Shun-Fa, Yang, Wei-Yu, et al. ICAS: An inter-VM IDS Log Cloud Analysis System[C]. Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on. IEEE, 2011:285-289.

[14] Swiler L. P, Phillips C, Ellis D, et al. Computer-Attack graph generation tool[C]. Proc. Of the 2nd DARPA Information Survivability Conf. & Exposition. Los Alamitos, USA, IEEE Computer Society Press, 2001:307-321.

[15] Lippmann R P, Ingols K W. An Annotated Review of Past Papers on Attack Graphs[J]. An Annotated Review of Past Papers on Attack Graphs, 2005.

[16] Jajodia S, Noel S, O’ Beny B. Topological analysis of network attack vulnerability[C].Proc. of the Managing Cyber Threats: Issues, Approaches and Challenges. Boston, USA,Kluwer Academic Publisher, 2003:3-4.

[17] Ou Xinming, Boyer W F, McQueen M A. A Scalable Approach to Attack Graph Generation[C]. Proc of the 13th ACM Conf on Computer and Communications Security,New York, USA, ACM Press, 2006:336-345.

[18] Cuppens F, Miege A. Alert correlation in a cooperative intrusion detection framework[C]. IEEE Symposium on Security & Privacy IEEE Computer Society. IEEE, 2010:202-215.

[19] P. Ning and D. Xu, Learning attack strategies from intrusion alerts, Proceedings of the 10th ACM Conference on Computer and Communications Security, New York: ACM Press, 2003, 200–209.

[20] Ning P, Cui Y, Reeves D S, et al. Techniques and tools for analyzing intrusion alerts.[J]. Acm Transactions on Information & System Security, 2012, 7(2):274-318.

[21] P. Ning, Y. Cui, and D. S. Reeves. Constructing attack scenarios through correlation of intrusion alerts. In Proceedings of the 9th ACM Conference on Computer & Communications Security (CCS), 2002.

[22] Noel S, Robertson E, Jajodia S. Correlating Intrusion Events and Building Attack Scenarios Through Attack Graph Distances[C]. Computer Security Applications Conference. 2004:350--359.

[23] Ou X, Rajagopalan S R, Sakthivelmurugan S. An Empirical Approach to

Modeling Uncertainty in Intrusion Analysis[C]. Computer Security Applications Conference, 2009. ACSAC '09. Annual. IEEE, 2009:494-503.

[24] Li W, Vaughn R B. Cluster Security Research Involving the Modeling of Network Exploitations Using Exploitation Graphs[C]. IEEE International Symposium on CLUSTER Computing & the Grid. 2006:26-26.

[25] 徐赐发.大数据时代金融业面临的挑战[J].金融科技时代, 2012(10):54-54.

[26] 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1):246-258.

[27] 王志军, 黄文良. 大数据在移动用户上网记录查询中的应用研究[J]. 信息通信技术, 2013(6):29-34.

[28] 耿冬旭. “大数据”时代背景下计算机信息处理技术分析[J]. 网络安全技术与应用, 2014(1):19-19.

[29] 基于粗糙集的入侵检测方法研究.史志才,夏永祥.计算机工程与科学 Vol.34,No.2,2012

[30]辛义忠. 基于数据挖掘的网络安全审计技术的研究与实现[D]. 沈阳工业大学, 2004.

[31] 姚钦锋. 基于数据挖掘的网络安全态势分析[D]. 上海交通大学, 2012.

[32] 高彩容. 基于数据挖掘的网络安全审计技术研究[D]. 西安电子科技大学, 2008.

[33] Endsley M R. Design and Evaluation for Situation Awareness Enhancement[C] Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, 1988:97-101.

[34] Bass t, arbor a. Multisensor data fusion for next generation distributed intrusion detection systems. Proceeding of iris national symposium on sensor and data fusion. laurel, 1999: 24-27.

[35]席荣荣, 云晓春, 金舒原,等. 网络安全态势感知研究综述[J]. 计算机应用, 2012, 32(1):1-4.

[36] 苏忠, 林繁, 陈厚金,等. 网络安全态势感知系统的构建与应用[J]. 信息网络安全, 2014(5):73-77.

[37] 谢丽霞, 王亚超. 网络安全态势感知新方法[J]. 北京邮电大学学报, 2014(5):31-35.

[38] 金爽. 基于 NetFlow 的实时安全事件检测技术研究[D]. 哈尔滨工程大学, 2006.

[39] Pawlakb Z. Rough Set Approach To Knowledge-Based Decision Support[J]. European Journal of Operational Research, 1997, 99(1):48－57.

[40] 刘雪飞, 马恒太, 张秉权,等. NIDS 报警信息关联分析进展研究[J]. 计算机科学, 2004, 31(12):61-64.

[41] MIT Linconln Lab. DARPA Intrusion Detection Evaluation Data Sets. 2000. http://www.ll.mit.edu/IST/ideval/data/2000/2000_data_index.html

[42] Li Hongcheng, Fu Yu, Ye Qing, Wu Xiaoping. Network Security Situation Element Extraction Method Based on Rough Set[J]. Computer&Digital Engineering, 2014, pages 436-439.

作者简介

攻读硕士学位期间，发表两篇论文：

1 “Rough Set Theory based Multi-source Alarm Data Analysis for Security Incidents Discovery”. CCNIS 2015, 第一作者, 已收录

2 “Efficient authentication and access control of message dissemination over vehicular ad hoc network”. Neurocomputing，2015，第一作者，已收录