文章编号: 1674-7070(2011) 04-0365-06

多源异构日志综合分析技术研究与实践

刘必雄[1]

摘要

多源异构日志分析技术是目前国内外网络安全领域的研究热点. 首先,提出了一种包括聚焦分析、统计分析和因果关联分析在内的多源异构日志综合分析模型,引入重要度评价方法对日志信息进行聚焦分析,并通过实例加以说明; 然后探讨了多源日志因果关联分析算法; 最后利用网络实例数据,对所提出的综合分析模型和算法进行了验证. 结果表明该模型和算法是可行的和有效的.

关键词

多源异构日志; 重要度评价; 因果关联

中图分类号 TP393. 08
文献标志码 A

0 引言

网络系统在运行过程中会产生大量的系统日志、应用日志、安全日志和网络日志,这些日志信息记录着网络系统发生了各种安全事件. 通过对日志进行分析,不但可以发现网络系统的运行状况,而且还能发现当前系统存在的漏洞以及可能出现的攻击. 传统对各种日志进行单独分析和处理的方法,由于其忽略各种类型日志之间的相关性,使其分析结果无法准确地反应网络系统的安全状况[1]. 因此,多源异构日志分析技术已经成为目前网络安全领域的研究热点.

近年来,国内外研究人员从不同角度对多源异构日志进行研究,并取得一定的成果. Asif-Iqba 等[2] 提出了一种异构日志事件过滤的方法,利用数据挖掘工具 Weka 对 Apache 服务日志、IP Table Syslog、Snort IDS 日志以及 Linux Syslog 进行解析,并利用聚类算法过滤冗余的日志事件,最后对日志事件进行聚合,从而有利于多源日志事件关联分析; Robiah 等[3] 提出了一种基于异构日志的入侵报警关联分析方法; 文献[4] 通过对日志文件的交集进行分析来发现用户的恶意行为,从而提高系统的安全性,但该方法只能对防火墙日志与应用系统日志进行分析; 文献[5] 提出了一种日志关联分析模型,通过对不同来源的日志文件进行采集、过滤、规范化以及关联分析,来重构攻击序列; 韦勇等[6] 利用日志审计技术对各种类型日志进行相关性分析,获取网络安全事件信息,并以此来计算节点的理论安全威胁值,从而获得网络安全态势.

以上方法为多源异构日志分析工作提供了可行的解决思路,为日志分析模型及算法奠定了良好的基础,但也普遍存在分析数据源不够广泛、分析技术单一等问题,为此本文提出了一种包含聚焦分析、统计分析和关联分析在内的多源异构日志综合分析方法,通过对分析模型及分析算法的探讨,进一步推动了多源日志的研究和探索.

1 多源异构日志综合分析模型

网络系统运行过程中产生大量的日志数据,这些日志记录与网络系统安全的相关信息,具有极高的价值. 为了辅助安全分析人员对日志数据进行多层次、多角度的分析,并从中找出重要的、反映攻击方法和技术的有效信息,并以友好的方式展示给安全分析人员,本文提出一种多源异构日志综合分析模型,能够对各种异构日志进行聚

收稿日期 2011-05-01
资助项目 福建省教育厅科技项目(JB09299)
作者简介
刘必雄,男,硕士,讲师,研究方向为网络安全. bxliu@163. com

1 福建农林大学 计算机与信息学院,福州, 350002

焦分析、统计分析和关联分析，从而为安全分析人员提供高层的攻击场景和统计趋势视图，使其全面地掌握整个网络系统的安全状况.

提出的多源异构日志综合分析模型框架如图 1 所示. 其首先通过日志采集 Agent[7] 收集网络系统运行过程中所产生的系统日志、应用日志、安全日志和网络日志，再对各种类型的日志数据经过数据清理、数据归并和数据映射等预处理[8]，转换为统一格式的日志信息; 然后对日志进行聚焦分析，即对日志信息进行规则库匹配得到安全事件，再结合漏洞信息、服务信息和网络拓扑结构对安全事件进行重要度评价，得到高风险的安全事件，并将重要度高的安全事件呈现给安全分析人员，为进一步分析提供切入点; 接着对日志进行统计分析，即根据安全分析人员所关注的某些特征对大量的日志数据进行汇总统计，给出数据分布和发展趋势，同时将统计结果与正常统计模式进行比较，来发现异常行为，并以图形化方式直观地将网络系统的整体状况和趋势展示给安全分析人员，从而使他们对整个网络安全状况有个全面且快速的了解; 最后根据因果关联知识库对多种来源的安全事件进行因果关联分析，发现安全事件所体现的攻击行为之间的关联关系，从而识别攻击者对整个网络系统所实施的攻击过程.

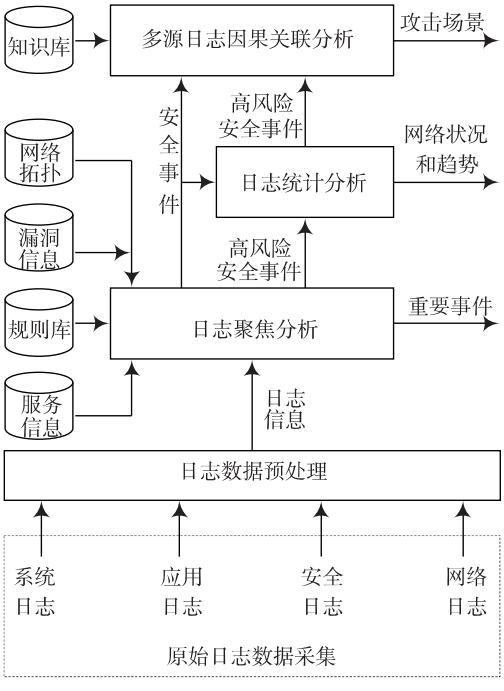

图 1　多源异构日志综合分析模型

Fig. 1　Multi-source heterogeneous log analysis model

2　基于重要度评价的日志聚焦分析

本文在文献[9]给出的观察数据重要度评价方法基础上，对各种类型的日志信息进行评价，并将重要度高的日志事件呈现出来，为安全分析人员进一步分析提供切入点.

2.1　重要度评价方法

文献[9]在风险评估公式的指导下，给出了观察事件的重要度评价方法，如图 2 所示. 该方法在获得目标资产重要度、攻击行为严重性以及观察事件相关度的相关评价指标的基础上，通过加权计算观察事件重要度. 图 2 中椭圆表示观察事件的各种指标，以 x_i 来表示第 i 个评价指标，第 i 个指标在重要度评价中的关键度用 w_i 表示，然后由公式(1)计算观察事件的重要度得分:

$$p = \sum_{i=1}^{n} x_i \times w_i. \tag{1}$$

通过计算观察事件重要度得分后，根据设定好的重要度阈值将当前观察事件的重要度等级为 H、M 和 L 3 种类型.

2.2　日志聚焦分析实例

本文以 IDS 日志数据为例来讨论基于重要度评价的日志聚焦分析方法的实现. 根据观察事件的重要度评价方法，在对 IDS 日志数据的属性进行分析的基础上，从目标资产重要度、攻击行为严重性以及观察事件相关度 3 个方面设置 IDS 日志事件重要度指标，如表 1 所示. 结合 IDS 日志数据分析方面的实践经验，并设置 IDS 日志事件重要度得分大于或等于 15 为“H”，得分在 4 和 15 之间的为“M”，得分在 4 分以下的为“L”.

表 1　安全事件重要度评价指标

Table 1　Importance evaluation index of security event

重要度评价项	评价因素	取值	权值
目标资产重要度	目标设备重要度	$\{h,m,l\}$	1
攻击行为严重度	日志事件安全等级	$\{h,m,l\}$	2
	数据包数量	$\{h,m,l\}$	1
	总字节数	$\{h,m,l\}$	1
观察事件相关度	目标主机是否活跃	$\{y,n\}$	2
	目标端口是否开放	$\{y,n\}$	2
	目标服务是否相关	$\{y,n\}$	2
	是否存在相关漏洞	$\{y,n\}$	2

注: 取值中 h,m,l 分别对应整数 5,2,0; y,n 分别对应整数 1,0.

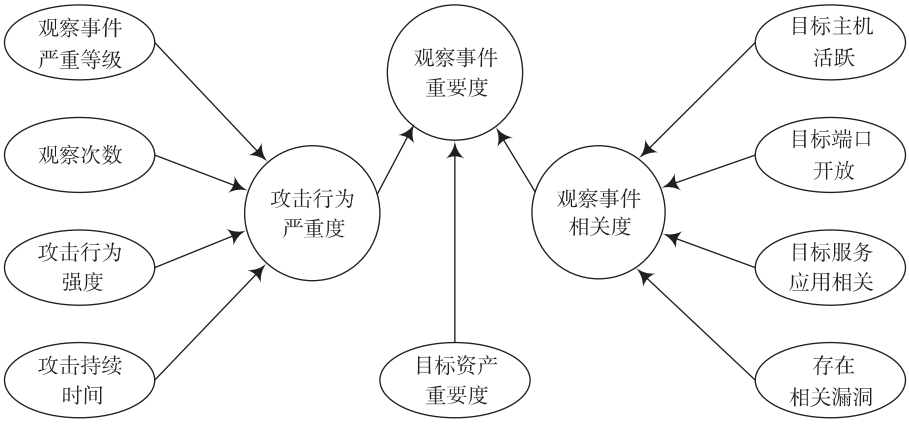

图2 观察事件重要度评价方法[9]

Fig. 2 Importance evaluation method of event

日志聚焦分析模块能够自动对接收到的日志数据进行重要度评价,并根据设置的等级,将日志事件按照重要度等级不同呈现给安全分析人员,这样分析员就可以将大量精力放在那些重要度高的日志事件上,对那些事件进一步进行深入分析. 例如,本文对部署在某单位一个 IDS 在 1 h 内产生的日志数据进行重要度评价,按表 1 的评价指标及式(1)计算日志事件的重要度得分,并按重要度评价阈值[4,15]来进行评价,从评价结果的分布图可以得出重要度评价为"H"的事件仅占 10.35%,而重要度为"L"的事件占 68.8%,这样安全分析人员重点查看的日志事件数量就大为减少,从而提高了日志事件的分析效率.

3 基于因果关系的多源日志关联分析

通过日志聚焦分析与统计分析,可以从大量日志信息中提取出所有描述同一攻击的安全事件,但这些安全事件往往是相互孤立的,不利于安全分析人员了解整个攻击场景以及网络系统的整体安全状况. 因此,通过对不同来源、不同时间、不同层次的安全事件从全局进行关联分析,将原先孤立的安全事件根据其内在的联系整合起来,从而发现各事件之间的因果关系,重构整个攻击过程,这样安全分析人员就能了解整体网络系统的安全状况.

安全事件的关联分析技术是目前网络安全领域研究的热点,研究人员提出了多种关联分析方法,主要可以分为 3 类.

1) 基于属性相似度的关联方法[5]. 即根据事件属性之间的相似度进行融合. 该方法原理简单,实现方便,但无法发现事件之间的因果关系.

2) 基于已知攻击场景的关联分析[10]. 利用已知的攻击场景进行关联分析,分析的准确率较高,但存在攻击场景不易描述、不易发现新的攻击行为等缺点.

3) 基于因果关系的关联分析[11]. 根据安全事件之间的固有联系,即通过对较早发生安全事件的行为的结果和较晚发生安全事件的行为的前提条件进行比较. 该方法无需依赖预定义攻击场景来发现相关攻击序列,能够识别攻击意图.

本文采用因果关联分析方法对多源日志进行关联分析,形成安全事件关联图,从全局角度给出攻击场景,有利于安全分析员更好地理解攻击者的攻击意图.

3.1 因果关联分析技术

因果关联是通过建立一种高阶安全事件的形式,来发现安全事件之间内在因果关系,从而生成多个安全事件之间的攻击序列的一种安全事件关联技术[11]. 由于各攻击步骤所产生的安全事件之间存在的因果关系,因此本文先定义每个单独攻击的前提和后果,然后将先发生的攻击步骤的后果和后发生的攻击步骤的前提进行匹配,从而实现安全事件的关联分析. 本文参考文献[11]中所提的因果关联方法,使用谓词来描述各种类型攻击的前提和结果. 例如,用 UDPVulnerableToBOF(VictimIP,VictivmPort)表示通过扫描攻击发现 UDP 服务漏洞以确定是否进行了缓冲区溢出攻击. 为了便于描述因果关联分析算法,先介绍几个相关的定义.

定义1 复合安全事件类型(Compound Security Event Type,CET):在因果关联知识库中用来表示每

一种安全事件类型的前提和后果. 复合安全事件类型用一个三元组 CET = (Fact, Prerequisite, Consequence) 来表示. 其中: Fact 是一系列安全事件的属性名,如 VictimIP、VictivmPort 等; Prerequisite 和 Consequence 表示该类安全事件的前提集合和后果集合,是由谓词和逻辑运算符组成的表达式. 对于每一类复合安全事件 CET,用 P(CET) 代表 CET 的前提集合 Prerequisite,用 C(CET) 代表 CET 的后果集合 Consequence.

例如,缓冲区溢出攻击可以用 SadmindBufferOverflow = ({ VictimIP, VictivmPort} , ExistHost (VictimIP) ∧ VulnerableSadmind (VictimIP) , { GainRootAccess(VictimIP) }) 来描述,这样就有 P(SadmindBufferOverflow) = { ExistHost(VictimIP) , VulnerableSadmind (VictimIP) } 和 C(SadmindBufferOverflow) = { GainRootAccess(VictimIP) } .

定义 2 复合安全事件(Compound Security Event, CE) : 根据安全事件中包含的信息来给 Fact 的各个属性赋值,即对一个复合安全事件类型实例化. 例如,某一低阶的缓冲区溢出安全事件,其 VictimIP = 152. 1. 19. 5, VictivmPort = 1235,则实例化后复合安全事件可以表示为 CE = { (152. 1. 19. 5, 1235) , ExistHost(152. 1. 19. 5) ∧ VulnerableSadmind (152. 1. 19. 5) , { GainRootAccess(152. 1. 19. 5) } } .

定义 3 事件类型关联(Event Type Correlation) : 设 CET_A 和 CET_B 表示关联事件库中 2 个不同的事件类型,当 $C(CET_A) \cap P(CET_B) = I$ 且 $I \neq \Phi$, 则称事件类型 CET_A 和 CET_B 是可关联的,记为 Event_Type_Correlation(CET_A, CET_B) .

定义 4 事件关联(Event Correlation) : 设 ET_A 和 ET_B 表示 2 个复合安全事件,其在关联事件库中分别对应于事件类型 CET_A 和 CET_B 且有 Event_Type_Correlation(CET_A, CET_B) ,如果存在某个谓词 $p \in C(ET_A)$ 且 $p \in P(ET_B)$, p 满足最一般合一置换 θ,则称事件 ET_A 和 ET_B 是可关联的,记为 Event_Correlation(ET_A, ET_B) .

3.2 多源日志事件因果关联分析

多源日志事件因果关联分析的过程如图 3 所示. 首先,对因果关联知识库进行解析,为生成复合安全事件做准备; 然后将经过聚焦分析和统计分析之后得到的每一条高风险安全事件和一般安全事件跟知识库中的复合安全事件类型名字进行匹配,如果匹配,就将安全事件中真实的属性值,填写到复合安全事件的前提集合和结果集合的谓词中,从而完成复合安全事件的生成; 接着根据复合安全事件之间的因果关系,识别出安全事件之间的联系,形成多步攻击场景图,使安全分析人员能够清晰了解攻击的整个过程和详细步骤. 下面介绍日志事件因果关联分析算法.

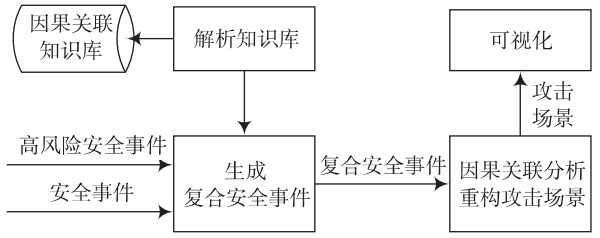

图 3 因果关联分析过程

Fig. 3 The process of causality correlation analysis

输入: 安全事件库 SE,高风险安全事件库 HE,事件可关联间隔时间 interval. 输出: 攻击场景 G.

```
Begin
  CSE = Generate( SE) ; //生成复合安全事件
  CHE = Generate( HE) ; //生成复合高风险安全事件
  For each h_i( i←1 to n) in CHE{
    t = h_i. time; //第 i 条高风险安全事件发生的时间
    Select E from CSE where time < t
    //找出事件发生在时间 t 之前的所有安全事件,并按时间顺序存入 E 中
    e = h_i; e′= e_m;
    For each e_j( j←m - 1 to 1) in E and ( e. time - e′. time < interval) {
      If Event_Correlation( e′,e) then
        Add e′→e to G;
        e = e;
      End If
      e′= e_j;
    }
    Select E from CSE where time > t
    e = h_i; e′= e_1;
    For each e_j( j←2 to m) in E and ( e′. time - e. time < interval) {
      If Event_Correlation( e,e) then
        Add e→e′to G;
        e = e;
      End If
      e′= e_j;
    }
  }
  Return G; //返回攻击场景
END
```

4 实验与分析

为了验证综合分析模型和算法的适用性,本文使用2000年DARPA资助MIT林肯实验室构造的攻击场景关联评测数据集作为实验数据,该数据集被普遍用于关联算法的有效性验证. DARPA 2000除了提供LLDOS 1.0和LLDOS 2.0.2两个攻击场景实例之外,还提供网络边界和网络内部的数据信息以及主机日志,然而该数据集中并没有提供网络拓扑结构以及主机的服务信息和漏洞信息等,给本文日志综合分析带来一定的困难. 因此,本文在文献[12]的基础上,根据报警信息获取关键主机的网络拓扑结构,如图4所示.

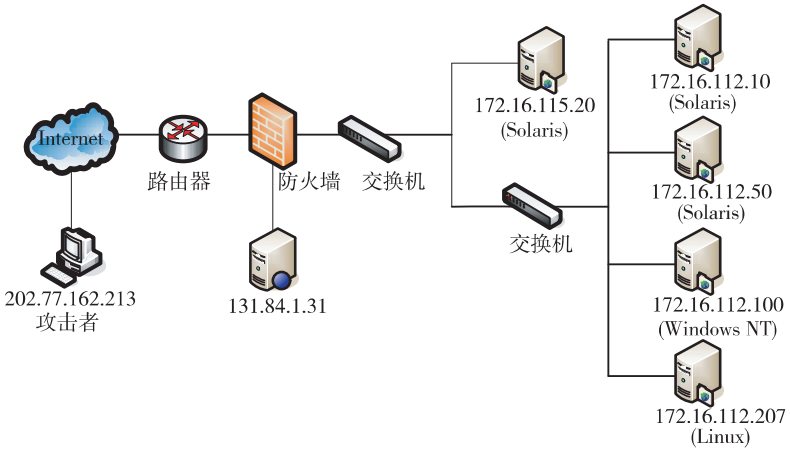

图4 网络实例拓扑

Fig. 4 The topology of network sample

本文在LLDOS 2.0.2数据集上进行了实验,用RealSecure检测数据集得到的报警信息作为原始安全事件,并结合图4的网络拓扑结构,从中选择相关的安全事件SE作为实验数据,然后,结合文献[12]提供的主机的漏洞信息和服务信息,对日志事件通过聚焦分析和统计分析,得到高风险安全事件HE. 使用上述算法进行因果关联分析,得到如图5所示的攻击场景. 图5中节点与安全事件的对应关系如表2所示.

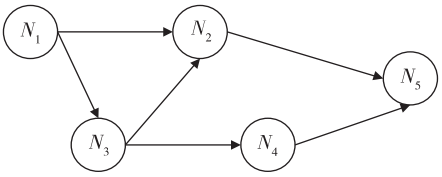

图5 攻击场景

Fig. 5 The graph of attack scenario

表2 节点与安全事件的对应关系

Table 2 Correspondence of the node and the meta alert

节点	安全事件
N_1	Sadmind_Amslverify_Overflow
N_2	Telnet
N_3	FTPUpload
N_4	Mstream_Zombie
N_5	Stream_DoS

事实上,LLDOS 2.0.2包含了一个复合攻击的过程,其完整的攻击序列分为5个攻击阶段:1)通过DNS服务器进行HInfo查询获取主机相关信息,从而找到可能存在Sadmind漏洞的Solaris主机;2)利用主机漏洞进行系统入侵,获得172.16.115.20的root控制权限;3)通过FTP下载安装Mstream handler和agent程序;4)以172.16.115.20为跳板,入侵另外一台主机172.16.115.50,并安装Mstream agent程序;5)攻击者Tenet到172.16.115.20,发出DDoS攻击指令,使172.16.115.20和172.16.112.50一起向131.84.1.31发动DDoS攻击. 图5重构了该攻击过程,但也存在不吻合的地方,即没有描述攻击者DNS_HInfo_Query的攻击行为,其原因在于Realsecure没有检测到DNS_HInfo_Query的攻击行为.

5 结束语

本文在已有工作[1,7-8]的基础上,针对多源异构日志的相关性和复杂性,提出了一种日志综合分析模型,该模型能够对各种类型的日志进行聚焦分析、统计分析和因果关联分析,适合多源异构日志多层次、综合性的需求. 本文详细描述了日志综合分析模型的结构,并给出日志聚焦分析的实例和多源日志

因果关联分析算法的详细描述,最后结合网络实例的分析进一步验证了本文所提模型、重要度评价方法和因果关联分析算法的适用性和有效性.

参考文献

References

[1] 刘必雄,杨泽明,吴焕,等. 基于集群的多源日志综合审计系统[J]. 计算机应用,2008,28(2):541-544
LIU Bixiong, YANG Zeming, WU Huan, et al. Multi-source log audit system based on cluster [J]. Computer Applications,2008,28(2):541-544

[2] Asif-Iqbal H, Udzir N I, Mahmod R, et al. Filtering events using clustering in heterogeneous security logs [J]. Information Technology Journal,2011,10(4):798-806

[3] Yusof R, Selamat S R, Sahib S. Intrusion Alert correlation technique analysis for heterogeneous log [J]. International Journal of Computer Science and Network Security, 2008,8(9):132-138

[4] Herrerias J, Gomez R. A log correlation model to support the evidence search process in a forensic investigation [C] //Proceedings of the 2nd International Workshop on Systematic Approaches to Digital Forensic Engineeriog, 2007:31-42

[5] Myers J, Grimaila M R, Mills R F. Log-based distributed security event detection using simple event correlator [C] //Proceedings of the 44th Hawaii International Conference on System Sciences,2011:1-7

[6] 韦勇,连一峰. 基于日志审计与性能修正算法的网络安全态势评估模型[J]. 计算机学报,2009,32(4):763-772
WEI Yong, LIAN Yifeng. A network securitysituational awareness model based on log audit and performance correction [J]. Chinese Journal of Computers,2009,32(4):763-772

[7] 刘必雄,魏连,许榕生. 基于 Agent 技术的多源日志采集系统的设计与实现[J]. 计算机系统应用,2008,17(2):71-74
LIU Bixiong, WEI Lian, XU Rongsheng. Design and implementation of multi-source log collect system based on agent technology [J]. Computer Systems & Applications, 2008,17(2):71-74

[8] 刘必雄,许榕生. 基于 XML 的综合日志预处理模型设计[J]. 莆田学院学报,2007,14(5):65-68
LIU Bixiong, XU Rongsheng. The design of the integrative log pre-processing model based on XML [J]. Journal of Putian University,2007,14(5):65-68

[9] 诸葛建伟. 网络入侵检测与行为关联分析技术研究[D]. 北京: 北京大学计算机科学技术研究所,2006
ZHUGE Jianwei. Research on technologies for network intrusion detection and behavior correlation analysis [D]. Bejing: Institute of Computer Science & Technology of Peking University,2006

[10] Herrerias J, Gomez R. Log analysis towards an automated forensic diagnosis system [C] //Proceedings of the 5th International Conference on Availability, Reliability and Security,2010:659-664

[11] 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,2002:245-254

[12] 韦勇,连一峰,冯登国. 基于信息融合的网络安全态势评估模型[J]. 计算机研究与发展,2009,46(3):353-362
WEI Yong, LIAN Yifeng, FENG Dengguo. A network security situational awareness model based on information fusion [J]. Journal of Computer Research and Development,2009,46(3):353-362

Research and practice on comprehensive analysis technology for multi-source heterogeneous log

LIU Bixiong

1 College of Computer & Information Science, Fujian Agriculture and Forestry University, Fuzhou 350002

Abstract The multi-source heterogeneous log analysis technology is one of the hottest topics in the area of network security in recent years, which attracts the interest of more and more domestic and abroad researchers. According to the characteristics of multi-source log in network system, a multi-source heterogeneous log analysis model which composed of focused analysis, statistical analysis and causality correlation analysis is proposed in this paper. Importance Evaluation method is introduced to the focused analysis for log information and an example is given to illustrate it, then causality correlation algorithm for multi-source log is discussed. Finally an example of actual network data is given to validate the comprehensive analysis model and algorithm. The results show that this model and algorithm is feasible and effective.

Key words multi-source heterogeneous log; importance evaluation; causality correlation