第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

一种基于属性哈希的告警日志去重方法

胡 倩 罗军勇 尹美娟 曲小美
(信息工程大学网络安全空间学院 郑州 450002)

摘 要 网络安全防护设备产生的告警日志中存在大量重复告警,影响实时的网络威胁态势分析。为解决告警日志的实时准确去重问题,提出了一种基于属性哈希的告警日志去重方法。该方法采用属性哈希实现重复告警的快速检测,并采用哈希表同时解决了大量非重复告警日志的存储问题。在基于Darpa数据集构建的告警日志上进行了实验,结果表明该方法在保证较低时间复杂度的同时,去重准确率可以达到95%以上。

关键词 告警日志,重复告警,属性哈希
中图法分类号 TP393 文献标识码 A

Method of Duplicate Removal on Alert Logs Based on Attributes Hashing

HU Qian LUO Jun-yong YIN Mei-juan QU Xiao-mei
(Network Security Space Academy, Information Engineering University, Zhengzhou 450002, China)

Abstract Alarm logs generated by network security equipment have a large number of repeated alarms, which impact real-time network situational threat analysis. In order to solve real-time accurate de-duplication problem of alarm logs, we proposed a method of duplicate removal on alert logs based on attributes hash. The method uses attribute hash for duplicate alarms quick detection and uses the hash table to solve the storage problem of a large number of non-repeating alarm logs at the same time. Conducted experiments results in the alarm log based on Darpa data set show that the method ensures lower time complexity, while deduplication accuracy rate can reach 95%.

Keywords Alert log, Repeat alert, Property hash

1 引言

网络安全设备产生的告警日志是网络安全态势感知系统的一个重要的要素来源。安全设备的告警日志中存在大量的重复告警[1],影响了基于告警日志的安全事件分析的效果和效率,阻碍了对网络威胁的及时响应。如何有效去除安全设备告警日志中的重复信息,是基于告警日志的网络安全态势感知系统在要素获取阶段要解决的首要问题。

现有的基于告警日志属性的去重方法可大致分为两类:基于属性匹配的去重方法和基于属性相似性的去重方法。刘夏龙等人[2]指出了关键属性值相同,并且两条告警之间的时间间隔在阈值范围内,则为重复告警;此类方法有很高的准确度,但是属性匹配的高时间复杂度是此类方法的不足。SRI的Andersson等[3]和Valdes&Skinne[4-6]在EMERALD项目中提出了用手工定义的入侵事件间的概率相似度来进行安全事件的关联分析,从而融合相似告警;该类方法的优点是具有较好的实时性,但对于复杂的告警信息,仅减少了告警数量,并未真正消除去重现象。

而在去重方法中基于哈希值去重的方法比较常见。王源[7]提出基于特征词获取哈希值的Simhash算法,来判断相似网页之间的差异程度;张曼等人[8]提出一种基于SHA-1的邮件去重算法,其将邮件按大小分开处理,根据哈希值快速去除正文相同或相似的重复邮件。该类方法对重复信息进行检测,能够实现快速去重。

本文提出一种基于全部属性哈希的告警日志去重方法,该方法通过属性哈希实现重复告警的快速检测,并利用哈希表同时解决了非重复告警日志的存储问题。与现有的去重方法相比,所提方法在保证算法准确性的同时,有效解决了网络安全态势感知系统中的告警日志去重问题。

2 重复告警

一个网络行为由网络事件时间序列构成,它反映网络行为从发生到逐步形成所经历的过程。通过观测网络获得网络事件,并按一定规则对其进行描述,综合分析网络事件时间序列形成网络行为。一个网络事件通常由事件的名称、发生时间、参与者、出现场景、类型、性质、检测依据等要素组成,通常用日志来记录各种网络事件,日志中的每一条记录描述一个网络事件。

在计算机网络中,通过安全设备发现一系列网络事件并以告警日志的形式存储。告警日志记录可定义为:$L=(a_t, a_1, a_2, \cdots, a_i, \cdots, a_k)$,其中,$a_t$ 代表该告警日志记录的产生时间;a_i 代表告警日志记录的属性,包括攻击事件类型、告警规则、协议类型、描述信息等。

假设有两个告警日志记录 x 和 y,则表示为:

胡 倩(1988—),女,硕士生,主要研究方向为网络信息安全,E-mail:huqian-07@qq.com;罗军勇(1964—),男,教授,主要研究方向为信息安全、数据挖掘;尹美娟(1977—),女,博士,讲师,主要研究方向为数据挖掘、社会网络分析;曲小美(1988—),女,硕士生,主要研究方向为数据挖掘。

$$L_x=(a_t^{(x)},a_1^{(x)},a_2^{(x)},\cdots,a_i^{(x)},\cdots,a_k^{(x)})$$

$$L_y=(a_t^{(y)},a_1^{(y)},a_2^{(y)},\cdots,a_i^{(y)},\cdots,a_k^{(y)})$$

若除去时间属性，可表示为：

$$\tilde{L}_x=(a_1^{(x)},a_2^{(x)},\cdots,a_i^{(x)},\cdots,a_k^{(x)})$$

$$\tilde{L}_y=(a_1^{(y)},a_2^{(y)},\cdots,a_i^{(y)},\cdots,a_k^{(y)})$$

由于安全设备采用的检测技术与策略因素，导致同一网络事件在告警日志中被重复记录；或者由于网络行为自身特点，使相同网络事件反复出现而导致告警日志中重复记录[9]。一个网络行为具有一定的持续时间，若告警日志中出现了重复记录且其出现的时间间隔在某个特定范围内，则可判定为同属一个网络行为，即可去重；否则从属于不同网络行为，即保留。

定义 1　对于构造的告警日志去重函数 f，如果 L_x 与 L_y 满足 $|a_t^{(x)}-a_t^{(y)}|<Timestamp$（Tiemstamp 为时窗阈值），且使得 $f(\tilde{L}_x)=f(\tilde{L}_y)$，则表示告警日志记录 L_x 与 L_y 为重复告警。

定义 2　假设 L_x 与 L_y 是一个网络行为开始与结束时产生的两条告警日志记录，$TimeWin=a_t^{(y)}-a_t^{(x)}$ 表示网络行为的时窗大小，则时窗阈值 $Tiemstamp$ 满足：$0<Timestamp\leqslant TimeWin$。

本文中的时间窗是指安全设备检测到一个网络行为从开始到结束时的持续时间。不同网络行为的时间窗大小不同，时窗阈值范围也就不一样。

3　告警日志去重方法描述

有了判定时窗阈值内两条告警日志记录是否重复的条件，就可以建立判断函数。然而，告警日志中不同的重复记录可能会交错出现，就需要快速的随机存储结构来存储和定位非重复的告警日志记录。因此，基于属性哈希的告警日志去重需要解决两个关键问题：重复告警判定和哈希表构造。

3.1　基于属性哈希的重复告警判定

根据对重复告警的定义，需要建立告警日志去重函数 f 并确立时窗阈值。本文采用哈希函数对告警日志属性进行散列[8]，时间阈值采用统计经验值。

(1) 构建哈希函数

哈希函数构造需要遵循的原则[10]是：

a) 函数本身便于计算；

b) 计算出来的地址分布均匀，即对任一关键字 k，$f(k)$ 对应不同哈希值的概率相等。

哈希函数值的构造可采用平方取中法，即取关键字平方的中间若干位作为哈希值。通过平方取中法得到的哈希值与关键字的每一位都有关，使得哈希值具有较好均匀性。构造一个哈希函数，输入为一个 m 位的关键字，哈希值也为 m 位，哈希函数 f 的计算过程如下：

Step 1　输入 m 位的 Key；

Step 2　计算 Key^2；

Step 3　若 Key^2 不足 $2m$ 个位，在前补 0，取这个数中间 m 个位的数，即 $10^{[m/2]}$ 至 $10^{[m/2]+m}$ 的数，将结果作为哈希值。

提取的告警日志属性均等重要，散列均匀分布。首先利用一个输出为 m 位的随机函数对告警日志记录中除时间之外的全部属性分别进行哈希计算，然后将各个属性的基础哈希值进行相加作为 m 位的 Key，输入构造的哈希函数 f 获得最终的告警日志记录的哈希值。实例计算如表 1 所列。

表 1　平方取中算法的结果

Key	Key^2	补足 2m 位后的值	H(k) 哈希值
526987	277715298169	277715298169	715298

(2) 时窗阈值的选择

不同网络行为的时窗大小不同，时窗阈值范围也就不一样。本文根据对已知攻击行为的时窗统计来计算时窗阈值，时窗阈值计算方法定义为：

$$\bar{T}=\frac{\sum_{i=0}^{n}T_i}{n}$$

式中，T_i 表示第 i 种网络攻击行为的持续时间；$\bar{T}$ 是平均网络攻击行为时间，即时窗阈值。

基于属性哈希的告警日志去重方法的基本思路是采用构造的哈希函数计算告警日志记录的哈希值，通过比较哈希值来判断出重复告警，具体步骤如下：

(a) 判断两告警日志记录出现的时间间隔是否在时窗阈值内；

(b) 若不在时窗阈值内，则判为非重复告警日志记录；

(c) 若在时窗阈值内，则计算两告警日志记录的哈希值并对其进行判断，若哈希值相等则为重复告警日志记录，否则为非重复告警日志记录。

3.2　哈希表的构建

哈希表是指在一块连续的存储空间中，数据的存储位置可按数据的哈希值进行定位的存储结构。在哈希表的设计中，主要考虑两点：1) 确立哈希表的大小；2) 处理冲突。

哈希表可以实现数据的快速、随机的访问，在哈希表不产生冲突的前提下，查找长度是常量，当哈希表产生冲突时，查找长度就与哈希表长度以及哈希表装入的告警记录数有关。若哈希表的取值越大，产生冲突的机会就越小。如果哈希表取值过大，存储空间的利用率很低，则会造成较大的空间浪费。

因此，哈希表的大小是哈希表设计的关键。而本文中哈希表规模与时窗阈值的取值有关，可根据不同网络行为的时窗大小来确定哈希表的最大长度，也可依据平均时窗阈值大小 $\bar{T}$ 来决定哈希表的固定长度。

用哈希表来存储非重复的告警日志记录，告警日志记录在哈希表中的存储位置即哈希表索引，其计算方法为：告警日志记录的哈希值关于哈希表大小求模取余。为了观测时间窗口中的记录，设计两个指针：指针 F 指向当前时窗中最早出现在哈希表中的记录位置，且此位置的告警记录时间属性用 t_0 表示；指针 R 指向当前时窗中最晚出现在哈希表中的记录位置，此位置的告警记录时间属性用 t_r 表示，且满足 $t_r-t_0\leqslant\bar{T}$。生成的哈希表如图 1 所示。

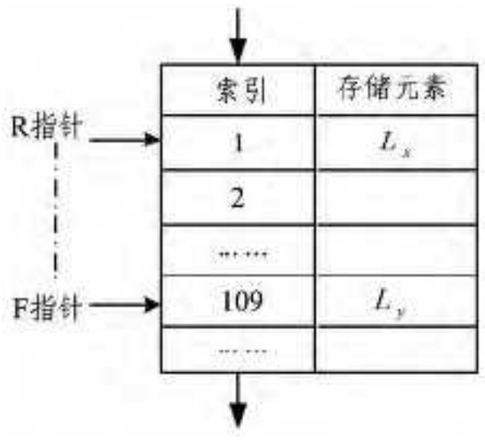

图 1　哈希表

3.3 算法描述

基于属性哈希的告警日志去重方法的基本步骤如下。

(1)从原始告警日志中读取一条告警日志记录,将其作为写入哈希表的初始告警日志记录,并将指针 F 和 R 均指向该告警日志的存储位置。其中,用 t_0 表示指针 F 所指向的告警日志记录的时间属性,用 t_r 来表示指针 R 所指向的告警日志记录的时间属性。

(2)将原始告警日志记录依次逐条读出。

(3)若新来告警日志记录 L_i 的时间属性 t_i 满足:$t_i - t_0 \leqslant \bar{T}$,则根据哈希函数求得该告警日志哈希值,并查看哈希表中此位置是否有记录存在。

(4)若此位置无记录存在,则将该新来告警日志记录 L_i 写入哈希表,并更改指针 R 的指向,将其指向该新来告警日志的存储位置,同步将 t_i 的时间值赋予变量 t_r;否则,将该报警舍去,认为该告警日志记录为时窗内的重复告警。

(5)若新来告警日志记录 L_i 的时间属性 t_i 满足 $t_i - t_0 > \bar{T}$,则根据对存储在哈希表中非告警日志记录的时间属性 t 进行判断,将满足 $t_0 \leqslant t < t_i$ 的告警日志记录从哈希表中移除,并将其写入数据库存储;之后将新来告警日志记录 L_i 写人哈希表中,并更改指针 F、指针 R 的指向,将其均指向该新来告警日志记录 L_i 的存储位置,同步将 t_i 的时间值赋予变量 t_0。

具体实现代码如下:

```
1. for all alert Li from text
2. create HashTable, database;
3. F as 当前时窗中最早出现在哈希表中的记录位置, R as 当前时窗
   中最晚出现在哈希表中的记录位置, T̄为时窗阈值;
4. t0 as 指针 F 所指向的告警日志记录的时间属性, tr as 指针 R 所指
   向的告警日志记录的时间属性;
5. if(ti − t0 ≤ T̄){
6. h = H(La)
7. if( h){
8. Hashtable.add(Li);
9. F、R 指向 Li 位置;
10. t0 = ti;}
11. else{
12. if(t0 ≤ t < ti){
13.     Remove(t);
14. Database.add(t);}
15.     Hashtable.add(Li);
16.     F、R 指向 Li 位置;}
17. t0 = ti;}
```

4 实验

4.1 实验数据

为了验证所提方法的有效性,需要选择合适的数据集。DARPA 2000[11]数据集模拟了美国 Eyrie 空军基地的网络系统遭受拒绝服务攻击的情形,包括两个攻击场景:LLDoS 1.0 和 LLDoS2.0.2,以及其他的网络边界和网络内部的数据信息、主机日志等。同时采用 DARPA 1999 数据集[12]给出了 5 周的模拟数据,其中前 2 周是提供给参与评测者的训练数据:第 1、3 周为不包含任何攻击的正常数据;第 2 周中插入了属于 18 种类型的 43 次攻击实例;第 4、5 周中包含了属于 58 种类型的 201 次攻击实例,其中 40 种攻击类型并没有在前 2 周的训练数据中出现,属于新的攻击类型。

实验环境中部署的网络安全防护设备是网络入侵检测系统 Snort(最新发行版 2.9.6 和相应的规则库)。为了模拟真实的网络环境,实验过程如下:在攻击测试机上利用 Netpoke 重放工具来重放 DARPA 2000 的 LLDOS 1.0 数据集和 DARPA1999 数据集,详细描述见表 2,agent 代理服务器收集安全设备(Snort2.0 入侵检测系统)所产生的告警数据信息,然后对数据信息用本课题所提出的方法进行实验分析。

表 2 实验数据集

数据集	包含攻击类型	告警日志条数
DARPA2000 (LLDOS 1.0)	DDOS 攻击	63980
DARPA1999(第 1、3 周)	正常数据	12429
DARPA1999(第 4、5 周)	包含多种攻击类型	15640

4.2 实验结果与分析

该实验的目的在于在相同数据集上实现多种算法和不同数据集上实现本文算法,来验证本文方法在去除重复告警日志上的高效性。

4.2.1 方法对比实验

为了验证去重效果,以告警日志属性比对算法为参照,以 DARPA1999(第 4、5 周)数据集中的 50s 数据记录作为测试数据,分别在 SNM 算法和本文提出的算法上进行比较实验。实验参数设定如下:时窗阈值为 50s,而根据时窗阈值大小设置表长为 6719。去重率和准确率[13]为常用的去重效果的验证指标,如式(1)、式(2)所示。

$$\text{去重率}:R=\frac{A+B}{A+B+C+D} \tag{1}$$

$$\text{准确率}:P=\frac{A}{A+B} \tag{2}$$

其中,A,B,C,D 由二值列联表确定,如表 3 所列。去重率用来反映告警去重算法除去重复告警的效率,去重率越接近人工判断的去重率,就表示该算法效果越好。准确率反映了去重质量,准确率越高,去重性能越好。实验结果见表 4。

表 3 准确率和去重率二值列联表

系统判断	人工判断	
	属于重复	不属于重复
属于重复	A	B
不属于重复	C	D

表 4 3 种算法的结果比较

方法类型	原始告警条数	检测重复条数	重复率(%)	去重率(R)(%)	准确率(P)(%)	时间(ms)
属性比对算法	1582	1388	87.7	87.7	100	128476
SNM 算法	1582	1402	87.7	88.6	76.8	18329
本文算法	1582	1379	87.7	87.1	99.35	6172

算法均使用同样的测试样本。通过表 4 可以看出,属性比对算法的准确率最高,然而最耗时;SNM 算法在时间效率上有了一定的提高,但是准确率却大大下降。从本文算法与

(下转第 360 页)

Key Cryptography(PKC 2011). 2011:53-70

[4] 马丹丹,陈勤,党正芹,等. 基于多属性机构的密文策略和加密机制[J]. 计算机工程,2012,38(10):114-116

[5] Lai Lun-zuo, Deng R H, Guan Chao-wen, et al. Attribute-Based Encryption With Verifiable Outsourced Decryption[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(8):1343-1354

[6] Green M, Hohenberger S, Waters B. Outsourcing the Decryption of ABE Ciphertexts[C]// Proceedings of the 20th USENIX Conference on Security. 2011

[7] Gamal T E. A public key encryptosystem and a signature scheme-based on discrete logarithms[M]// Advances in Cryptology: Proceedings of CRYPTO 84. 1985:10-18

[8] Goyal V, Pandey O, Sahai A, et al. Attribute-based encryption for fine-grained access control of encrypted data[C]// ACM Conference on Computer and Communications Security. 2006: 89-98

[9] 石岳蓉,郭俊,赵晶晶,等. 高效数据外包的基于多授权中心的ABE方案[J]. 信息技术,2015(2):97-100

[10] Beimel A. Secure Schemes for Secret Sharing and Key Distribution [D]. Israel Institute of Technology, Technion, Haifa, Israel, 1996

(上接第 334 页)

两种传统算法的对比结果可以看出,本文算法在去重率接近传统算法的基础上,具有较高的准确率(达 0.9 以上),同时显著提高了去重时间效率,具有很好的去重效果。

4.2.2 基于不同数据集的方法进行去重效果实验

在不同的情况下 Snort 产生重复告警的比例并不太一样,通常在有攻击行为发生时重复告警比例很高,而正常网络中的重复告警数量较少。为了验证在重复告警比例不同的情况下本文算法的有效性,采用 Dapar1999 第 1、3 周不包含任何攻击的正常纯净数据集、第 2 周中插入了包含 18 种类型的 43 次攻击实例的数据集和 DARPA2000 的 LLDOS 1.0 数据集中截取纯攻击数据集进行测试,实验结果如图 2 所示。

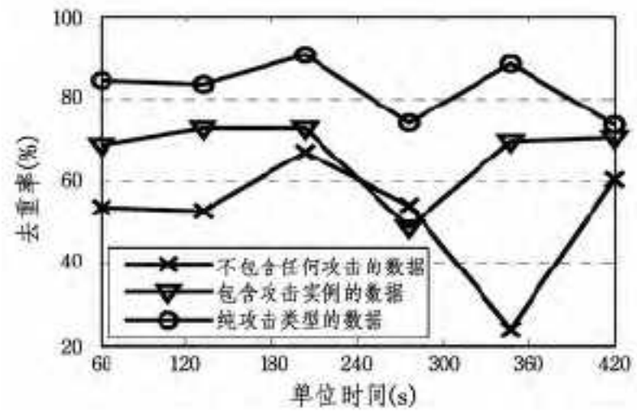

图 2 不同类型数据集在单位时间内的去重率

本文算法采用不同的测试样本,以去重率为指标对效果进行比较分析。对不包含任何攻击的正常纯净数据集的 Dapar1999 第 1、3 周数据来说,其去重率比较低;从 Snort 检测分析捕获的数据包匹配入侵行为的特征原理来说,网络中正常流量包的告警次数以及重复告警频率都远没有 DARPA2000 的 LLDOS 1.0 数据集中截取纯攻击数据集的高。通过分析 Snort 系统告警日志,本文方法去重率与测试样本中的重复日志比例基本一致。

结束语 本文针对大规模网络的安全态势感知系统中的告警日志去重问题,提出了采用基于属性哈希的日志去重方法,其充分考虑了告警日志去重操作的准确性、高效性和大量数据对存储内存容量的需求。实验结果表明,与传统属性比对算法相比,本文算法基本保持了原有算法的准确率、去重率等性能,并显著缩短了运行时间,有效降低了时间复杂度;同时,可处理大容量数据的特性使本文算法适用于大型网络的告警日志去重操作。

参 考 文 献

[1] 郭帆,叶继华,余敏. 一种分步式 IDS 告警聚合模型的设计和实现[J]. 计算机应用研究,2009,26(1):325-330

[2] 刘夏龙. 入侵检测告警数据的过滤与聚合技术研究[D]. 北京:中国科学院研究生院,2012

[3] Andersson D, Fong M, Valdes A. Heterogeneous Sensor Correlation: A Case Study of Live Traffic Analysis[C]// Proceeding of Third Ann. IEEE Information Assurance Workshop: IEEE Computer Society. StuartFeldman, MikeUretsky, New York, USA, June 2002:198-207

[4] Valdes A, Skinner K. Adaptive, Model-Based Monitoring for Cyber Attack Detection[C]// Proceeding of RAID2000 Conf: RAID 2000. 2000:204-217

[5] Valdes A, Skinner K. An Approach to Sensor Correlation[C]// Proceeding of Int'l Symp: Recent Advances in Intrusion Detection: IEEE Computer Society. 2000:197-201

[6] Valdes A, Skinner K. Probabilistic alert correlation[C]// Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection (RAID 2001). Davis, CA, USA, 2001, London, UK: Springer, 2001:54-68

[7] 王源. 一种基于 Simhash 的文本快速去重算法[D]. 吉林:吉林大学,2014

[8] 张曼等. 基于 SHA-1 的邮件去重算法[J]. 计算机工程,2008,34(11):270-272

[9] 黄思斯. 基于多 IDS 系统的攻击场景重建方法的研究[D]. 武汉:华中科技大学,2007

[10] 黄汉永,肖杰,张驹. 一种基于 Hash 函数抽样的数据集流聚类算法[J]. 计算机系统应用,2009,18(3):73-75

[11] Mit L L. DARPA 2000 intrusion detection evaluation datasets [OL]. (2000). http://ideval.ll.mit.edu/2000 index.html

[12] Mit L L. DARPA1999 intrusion detection evaluation datasets [OL]. (1999). http://www.ll.mit.edu/2st/ideval/data/1999/1999-data-index.html

[13] 尹美娟. 基于邮件正文的邮箱用户别名抽取[J]. 计算机科学,2011,38(12):200-202