基于频繁模式挖掘的报警关联与分析算法

董晓梅[1], 于 戈[1], 孙晶茹[1], 王丽娜[1,2]
(1. 东北大学信息科学与工程学院, 辽宁沈阳 110004; 2. 武汉大学计算机学院, 湖北武汉 430072)

摘 要: 提出了一个入侵检测与响应协作模型, 结合入侵容忍的思想扩展了入侵检测消息交换格式 IDMEF, 增加了怀疑度属性. 除了发现的入侵事件外, 一些可疑的事件也会报告给协作部件. 提出了一个基于修改的 CLOSET 频繁闭模式挖掘算法的报警关联与分析算法, 在分布式入侵检测与响应协作系统中, 帮助协作部件对收到的 IDMEF 格式的报警消息进行关联和分析, 以便做出合适的响应. 为此, 修改了 CLOSET 算法来按照最小支持度和最小怀疑度来得到频繁闭模式. 实验结果表明, 应用该算法可以很好地缩减报警数量, 同时对于所有可疑的和入侵事件, 都可以做出适宜的响应.

关键词: 入侵检测; 协作; 报警; 入侵容忍; 频繁模式
中图分类号: TP393.08 文献标识码: A 文章编号: 0372-2112 (2005) 08-1356-04

An Alert Correlation and Analysis Algorithm Based on Frequent Pattern Mining

DONG Xiao-mei[1], YU Ge[1], SUN Jing-ru[1], WANG Li-na[1,2]
(1. *School of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110004, China;*
2. *School of Computer Science and Technology, Wuhan University, Wuhan, Hubei 430072, China*)

Abstract: An intrusion detection and response cooperation model was proposed. Incorporating the intrusion tolerance idea, the Intrusion Detection Message Exchange Format (IDMEF) was extended and a suspicious degree attribute was added. So suspicious events as well as intrusions can be reported to the cooperation components. An alert correlation and analysis algorithm was proposed, which was based on the modified CLOSET frequent close pattern mining algorithm. The algorithm can help the cooperation components in a distributed intrusion detection and response cooperation system to correlate and analyze the alerts received to make appropriate responses. To meet this purpose, the CLOSET algorithm was modified to obtain frequent close patterns according to a minimum support and a minimum suspicion degree. Experimental results show that when applying the algorithm, the amount of alerts can be effectively decreased. And appropriate responses can be made according to all the suspicious and intrusion events.

Key words: intrusion detection; cooperation; alert; intrusion tolerance; frequent pattern

1 引言

在分布式入侵检测系统中, 一般由多个检测部件分别监控不同的主机和网络资源, 各部件之间有时需要互相协作来完成较复杂的检测任务[1]. 然而, 为解决部件间进行协作的复杂性, 需要一种通用且高效的入侵检测协作机制. 通用入侵检测框架(Common Intrusion Detection Framework, CIDF)[2]是在美国国防部高级研究项目署(DARPA)领导下开发的一套通用的语言、协议和 API(Application Programming Interface), 允许异构入侵检测部件互操作和共享信息. IETF(Internet Engineering Task Force)的入侵检测工作组 IDWG 在 CIDF 的基础上, 提出了一系列入侵检测信息交换标准草案. 其中, 在入侵检测消息交换格式 IDMEF(Intrusion Detection Message Exchange Format)[3]中, 以 UML(Unified Modeling Language)描述了信息交换中的数据模型, 以 XML(eXtensible Markup Language)来定义入侵检测系统中要交换的消息(如告警).

同时, 仅有入侵检测技术是不够的, 一个系统能够容忍一般的攻击也很重要[4~7]. 传统的入侵检测主要集中在底层的攻击的检测上, 但是往往各个报警信息不是孤立的, 一次攻击通常需要几个步骤来完成. 每个检测部件每天将产生大量的报警消息, 因此需要对这些消息进行关联和分析, 一方面可以缩减报警的数量, 缩短响应时间. 另一方面, 也可以从中得到一些有价值的隐含信息, 从而发现一些复杂的攻击.

本文提出的入侵检测与响应协作模型中, 修改了 IDMEF 消息交换格式, 提出了怀疑度的概念, 将发现的所有可疑的入侵行为都报告给协作部件, 以便及时进行隔离和监控. 为了便于协作代理分析收到的报警消息, 本文提出了一个基于频繁模式挖掘的报警关联与分析算法, 按照一个最小怀疑度和最

收稿日期: 2003-09-23; 修回日期: 2005-05-01
基金项目: 国家 863 计划 CIMS 主题(No. 2003AA414210); 国家自然科学基金(No. 60473073); 教育部高等学校博士学科点专项科研基金(No. 20030145029); 教育部优秀青年教师科研教学奖励计划资助项目

小支持度对报警消息进行预处理,然后使用修改的频繁闭模式挖掘算法,对报警信息进行挖掘,得到频繁闭模式序列.

2　入侵检测与响应协作模型

对应于大型分布式异构网络环境的要求,一个分布式入侵检测与响应协作系统由若干个易于管理的功能部件组成,分布于组成网络的各个域中.域是相对独立的网络部分,各个域之间通过TCP/IP网络进行通信.系统的结构可以描述为一个5元组:

〈Distributed Cooperative IDS〉= (〈Data_collector〉,〈Detection_agent〉,〈Monitor〉,〈Cooperation_agent〉,〈Intrusion_event_database〉)

其中,〈Data_collector〉:数据采集器,负责采集检测数据并发送给相应的检测代理,可以有多个,分布于网络各处;〈Detection_agent〉:入侵检测代理,可以有多个,使用各种检测算法,分布于网络各处;〈Monitor〉:监控部件,负责监控各代理的工作、控制系统的响应及向管理员报告系统的状态,每个域中唯一;〈Intrusion_event_database〉:入侵事件数据库,每个域中唯一;〈Cooperation_agent〉:协作代理,每个域中唯一,负责对收到的报警进行分析和汇总,如果发现入侵,则将分析结果存入入侵事件数据库,并自动生成XML文档传送给监控部件进行显示,同时通过网络发送给其他主机上的协作代理.

分布式入侵检测与响应协作系统总体结构用UML类图描述,如图1所示.

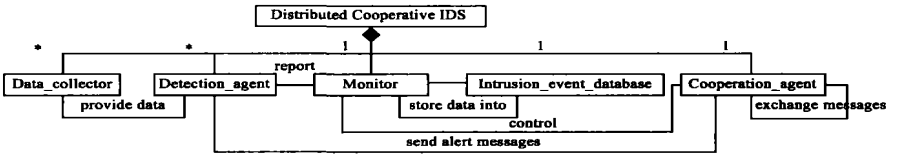

图1　分布式入侵检测与响应协作系统总体结构

本文中扩展了IDMEF格式,在alert类中加入了一个"suspect"(怀疑度)属性,表示报警的可疑程度.怀疑度的取值在0到1之间,用百分比来表示.怀疑度值越大,则行为的可疑程度越高.

3　基于频繁模式挖掘的报警关联与分析算法

协作代理需要合并报警消息中的重复信息,提取怀疑度较高或者频繁发生的事件信息.数据挖掘中的频繁模式挖掘技术可以发现数据中频繁发生的事件,很适合用来分析报警信息.其中,CLOSET[8]算法是一个高效的频繁模式挖掘算法,使用一种称为频繁模式树[9]的数据结构来快速产生频繁模式.但是,该算法只能发现频繁出现的模式,并不能考虑到报警信息中安全事件的可疑程度.例如,当某条报警消息具有很高的怀疑度,但其中包含的信息出现次数很少时,直接应用CLOSET算法进行挖掘,就会忽略这些不频繁出现的信息.因此,我们修改了CLOSET算法.

定义1　对于频繁项集X,如果不存在项y使得每个包含X的事务也包含y,则称X是频繁闭模式/项集[8].

定义2　一颗FP-Tree(Frequent Pattern Tree,频繁模式树)是如下定义的一棵树:

(1)它包含一个标志为空的根节点,一系列的前缀子树作为根节点的孩子,并生成一个常序列头表;(2)每一个节点在项前缀树包括三个部分:item name,count,和node-link,其中item name是代表这个节点代表哪个项,count代表在这个分支上该节点出现的次数,node-link节点连接到下一个FP-Tree节点,如果没有,那么为空;(3)在频繁项集头表包括两项:项名和头连接点,它指向FP-Tree的下一个节点.

3.1　报警关联算法

由于在我们的入侵检测中引入了怀疑度的概念,所以在我们对数据进行预处理时,除了要提取支持度满足最小支持度min_suport要求的报警数据外,还要提取出怀疑度符合我们要求的报警数据.为此,我们要事先设置3个参数值:最小支持度min_support、最小怀疑度min_suspect和最小路径深度min_depth.

数据读出后,对数据中的各项的数目进行统计,排除值小于min_suport,且怀疑度小于min_suspect的项,并对其进行降序排列,生成frequentItemList.根据frequenItemList中的order,重新在事务中对各项进行排序,用于记录所选择事务的项信息,又根据事务中项的个数将事务按降序进行排序,这样我们就获得了建立FP-Tree的数据.读取频繁项数据从root依次存入到树中,生成FP-Tree.然后,再使用CLOSET算法,查找出频繁闭项集,生成频繁闭模式序列.如果某一条报警消息的suspicious_degree=100%,即在检测代理处已经认为该行为是入侵,则向监控部件monitor发出警告信息.

在我们的实验中,根据需要对CLOSET算法做了适当的修改:首先查找出所有支持度值大于最小支持度,或者怀疑度值大于最小怀疑度的频繁闭项集,生成频繁闭模式序列.最后,在频繁闭模式序列中排除长度小于min_depth的序列.具体步骤如下:

报警关联分析算法:

Step1　将报警消息存入事务数据库TDB;
　　计算各项目数量:Count_items();
　　for each alert in TDB {
　　　if (alert.suspect==100%) 向监控部件发送报警;
　　　for each item in alert
　　　　if (item.count >= min_support)||(alert.suspect >= min_suspect)
　　　　将item加入频繁项列表FrequenItemList;}

Step2　将频繁项列表FrequentItemList倒排:
　　Sort_descend(FrequentItemList);

Step3　根据frequentItemList,获得事务中的频繁项列表,并排序:
　　sort_transaction_item();

Step4　根据事务长度对事务进行索引,并返回事务列

表,生成构建FP-Tree 的数据:

frequent_data_list= sort_transaction();

Step5 初始化频繁闭项集 *FCI*:

FCI= Φ;//

Step6 调用 CLOSET 子程序,获得频繁闭项集 *FCI*:

CLOSET(Φ,TDB, FrequentItemList, *FCI*);

Step7 for each list in *FCI*

if(list.length〈min_depth) delete list;

子程序 CLOSET(*X*, DB, FrequentItemList, *FCI*)

(1) if(∃ *Y*∈FreqentItemList && *Y*.count == DB.count) && ({*X*∪*Y*}⊄ *FCI*)

FCI= *FCI*∪{*X*∪*Y*};

(2) 从数据库 DB 构建 FP-Tree:

Fptree= build_FP-Tree(DB);

(3) 从 FP-Tree 直接抽取频繁闭项集:

single_seg_Itemset= extract_fptree(Fptree);

FCI= *FCI*∪{single_seg_Itemset};

(4) for each i in frequentItemList && (i∉ *FCI*){

为每个频繁项 i 建立条件数据库 DB_i: create_cond_database(DB_i);

建立局部频繁项列表 $fList_i$: create_flist($fList_i$);

if({iX}⊄ *FCI*) 递归调用子程序 CLOSET(iX, DB_i, $fList_i$, *FCI*);}

通过挖掘频繁闭模式,对报警信息进行了合并. 之后,可以根据获得的频繁闭模式序列构造汇总后的报警消息,我们称之为汇总报警. 这样,监控部件只需处理这些汇总报警,而忽略其他怀疑度较低且出现次数较少的报警. 每个汇总报警可以由一个频繁闭模式序列表示:

(attribute$_1$: value$_1$, attribute$_2$: value$_2$, …, attribute$_n$: value$_n$, suspect: maxSuspect, support: support)

频繁闭模式序列的怀疑度 maxSuspect 按如下方法计算,其中 n 为列表中事务个数.

$$maxSuspect = max(suspect_i) + 0.01 * (n - 1)$$

3.2 数据预处理

我们从报警消息中选择了 24 个属性进行分析:"impact","alertname","targetident","targetdecoy","targetnodename","targetnodelocation","targetnodecategory","targetaddress","targetaddrcategory","targetUsercat","targetuseridtype","targetuserid","targetprocessname","sourceident","sourcespoofed","sourcenodename","sourcenodelocation","sourcenodecategory","sourceaddress","sourceaddrcategory","sourceusercate","sourceuseridtype","sourceuserid" 和"sourceprocessname".

为了避免报警的细节信息过于分散,我们先将目标节点的 IP 地址进行了处理,只保留网络地址,不分析主机地址. 然后,从报警数据库中选出上述 24 个属性,并按照目标 IP 地址进行分组,分别进行挖掘. 例如,如表 1 所示为部分对应于目标地址"192.168.1.*"的经过预处理的数据.

表 1 部分经过预处理的数据

AlertId	suspect	Impact	…	Targetnode category	Target address	…
33	80%	bad_unknown	…	dns	192.168.1.*	…
34	60%	bad_unknown	…	dns	192.168.1.*	…
35	60%	attempted_admin	…	dns	192.168.1.*	…
36	60%	attempted_admin	…	dns	192.168.1.*	…

3.3 挖掘频繁闭模式

对于每个不同的目标 IP 地址的数据组,应用上述算法来挖掘频繁闭模式. 例如,对于表 1 的数据,当设置最小支持度为 60%,最小怀疑度为 70% 时,进行关联分析得到 1 个频繁闭模式:(Targetdecoy: Yes, targetnodecategory: dns, targetaddrcategory: unknown, targetuseridtype: uer_privs, sourcenodecategory: dns, sourceusercategory: unknown, sourceuseridtype: current_user, targetnodename: northeast, targetnodelocation: local, targetusercategory: os_evice, sourcespoofed: no, sourceusercategory: application, impact: bad_unknown, alertname: Vendor_specific, targetuserid: se93, sourcenodelocation: console, sourcenodeaddress: 192.168.11.112, sourceuserid: 2, sourceprocessname: devent, targetprocessname: revent, suspect: 80%, support: 25%)

4 实验结果

(1) 算法的运行时间 在我们的实验中,分别模拟了 100、120、140、160、180 和 200 条报警消息,对算法的运行时间进行了测试. 测试环境为: CPU: Pentium III, 1G Hz; RAM: 256MB; 硬盘: 40GB.

如图 2 所示为数据量不同时算法的执行时间. 从中可以看出,随着数据量的增大,算法的执行时间会增加,这一点很好理解.

(2) 参数设置对结果的影响 图 3 所示为处理 150 条报警消息时,最小怀疑度和最小支持度设置不同的值,对得到的结果数量的影响.

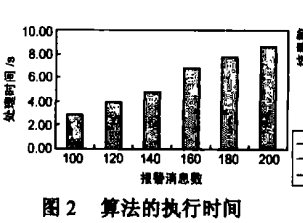

图 2 算法的执行时间

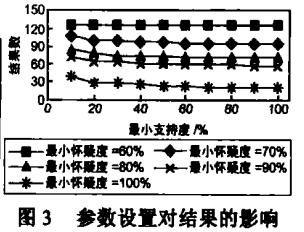

图 3 参数设置对结果的影响

从结果可以看出,报警数量大大减少. 随着最小怀疑度阈值的增大,得到的频繁闭模式数量会显著减少. 同时,随着最小支持度阈值的增大,得到的频繁闭模式数量也会逐渐减少,但是递减到一定程度后就基本保持不变. 这就保证了除了频繁出现的报警信息外,怀疑度高的报警即使不频繁出现也会得到及时响应.

(3) 算法的实际应用 为测试算法实际应用的可行性,

我们设计了如下实验:使用 ddosping 和 udpflood 两种工具在网络中进行模拟攻击,用 2 台 PC 机分别运行入侵检测程序,检测结果发送到第 3 台PC 机.在第 3 台 PC 上运行报警关联分析算法,得到的测试结果如表 2 所示.

表 2　测试结果

项目	数值
测试时间(秒)	180
报警总数	16086
预处理后报警数	96
关联分析时间(秒)	2.8
得到频繁模式数	52

5　相关工作

国外一些学者提出了一些报警关联算法. Ning P 等[10~12]提出了一种方法,用谓词作为攻击的先决条件及后续的结构表示法,定义了 hyper alert 消息类型,一个 hyper alert 可以相对于几个关联警告,当几个入侵检测器检测到从一个攻击者处发起的一系列的相同目标的攻击时,就可以将这些报警对应到一个 hyper alert 上.这种方法提供了一种高层关联报警信息的表示法,可以潜在地压缩错误报警的影响. Cuppens F 等[13~15]也提出了类似的方法,对于 IDMEF 格式的消息进行关联分析.以上这些方法需要事先定义每种攻击类型的先决条件和后续结果.因此只能对已知攻击方法的报警信息进行关联.这些算法一般都是进行离线的分析,速度一般都相对较慢,因此关联分析的处理效率不高,难以对攻击行为采取及时的响应.此外,这些研究中都没有考虑入侵容忍技术.

本文提出的基于频繁模式挖掘的报警关联与分析算法,是一种实时的处理算法,对收到的报警信息进行挖掘,得到频繁闭模式序列.本文的算法不仅速度快,而且结合了入侵容忍的思想,更有利于对系统的保护.

6　结论

由于入侵行为向分布式协作化入侵方向发展,所以对分布式入侵检测与响应机制的研究在今天具有重要意义.本文提出的基于频繁模式挖掘的报警关联与分析算法,使用数据挖掘技术,将来自各检测代理和其他域的报警消息进行关联分析,可以将大量的报警消息相互关联起来,并可以有效地缩减报警数量,使系统能够及时处理重要的事件,忽略一些无关紧要的事件.这不仅可以提高入侵检测与响应系统的效率,而且通过各入侵检测部件和入侵检测系统的互相协作,可以充分发挥各检测部件的优势,检测较复杂的协同攻击,增强对系统的保护.本文的算法不需要事先对各种攻击行为的先决条件和后果进行定义,可以自动挖掘出隐含在大量报警信息中的规律和特征.实验表明,我们的算法执行速度快,实时性好.

参考文献:

[1] Spafford E H, Zamboni D. Intrusion detection using autonomous agents [J]. Computer Networks, 2000, 34(4): 547- 570.

[2] Staniford-Chen S, Tung B, Schnackenberg D. The Common Intrusion Detection Framework (CIDF) [EB/OL]. http://www.isi.edu/gost/cidf/papers/cidf isw.txt, 1999- 09- 10.

[3] IETF. Intrusion Detection Message Exchange Format Data Model and Extensible Markup Language (XML) Document Type Definition[EB/OL]. http://www.ietf.org/internet-drafts/draft-ietf-idwg-idmef-xml-10.txt, 2003- 01- 10.

[4] Pal P, Webber F, Schantz R E, et al. Survival by defense-enabling [A]. Proceedings of the New Security Paradigms Workshop (NSPW 2001) [C]. New York: ACM Press, 2001. 71- 78.

[5] Wu T, Malkin M, Boneh D. Building intrusion tolerant applications [A]. Proceedings of the 8th USENIX Security Symposium[C]. Washington, D. C: USENIX Association. 1999. 79- 91.

[6] Gong F, Goseva-Popstojanova K, Wang F, et al. Characterizing Intrusion Tolerant Systems Using A State Transition Model [EB/OL]. http://www.anr.mcnc.org/projects/SITAR/papers/darpa00.pdf. 2000.

[7] 彭文灵,王丽娜,张焕国,等.基于角色访问控制的入侵容忍机制研究[J].电子学报,2005,33(1):91- 95.
PENG Wen-ling, WANG Li-na, ZHANG Huan-guo, et al. Research on intrusion tolerant architecture based on role- based access control [J]. ACTA ELECTRONICA SINICA, 2005, 33(1): 91- 95.

[8] Pei J, Han J, Mao R. CLOSET: An efficient algorithm for mining frequent closed itemsets[A]. Proc 2000 ACM-SIGMOD Int. Workshop on Data Mining and Knowledge Discovery (DMKD' 00) [C]. New York: ACM Press, 2000. 21- 30.

[9] Han J, Pei J, Yin W. Mining frequent patterns without candidate generation[J]. ACM SIGMOD Record, 2000, 29(2): 1- 12.

[10] Ning P, Cui Y, Reeves D S. Analyzing intensive intrusion alerts via correlation[A]. Proceedings of the 5th International Symposium on Recent Advances in Intrusion Detection (RAID 2002) [EB/OL]. http://infosec.csc.ncsu.edu/pubs/raid-02.pdf, 2002.

[11] Ning P, Cui Y, Reeves D S. Constructing attack scenarios through correlation of intrusion alerts[A]. Proceedings of the 9th ACM Conference on Computer and Communications Security[EB/OL]. http://infosec.csc.ncsu.edu/pubs/ccs02.pdf, 2002.

[12] Ning P, Xu D. Adapting query optimization techniques for efficient intrusion alert correlation[A]. Proceedings of the 17th IFIP WG 11.3 Working Conference on Data and Application Security [EB/OL]. http://discovery.csc.ncsu.edu/~pning/pubs/FastCorrelation.pdf, 2003.

[13] Cuppens F, Miege A. Alert correlation in a cooperative intrusion detection framework[A]. PROC IEEE COMPUT SOC SYMP RES SECUR PRIVACY[C]. Washington, DC: IEEE Computer Society Press, 2002. 202- 215.

[14] Cuppens F, Autrel F, Miege A, et al. Correlation in an intrusion detection process [EB/OL]. http://www.lsv.ens-cachan.fr/~goubault/SECI-02/Final/actes-seci02/pdf/014-cuppens.pdf, 2002.

[15] Cuppens F. Managing Alerts in a Multi-Intrusion Detection Environment[EB/OL]. http://www.acsac.org/2001/papers/70.pdf, 2001- 10- 10.

作者简介:

董晓梅　女,1970 年生于辽宁省沈阳市,博士,现为东北大学副教授,研究方向为计算机网络及信息系统安全. E-mail: xmdong@mail.neu.edu.cn.

于　戈　男,1962 年生于辽宁省大连市,博士,现为东北大学教授,博士生导师,研究方向为数据库、信息安全、嵌入式与实时系统等.