基于数据挖掘的入侵检测行为数据辨析

吴 玉[1,2],李 岚[2],朱 明[1]

(1.中国科学技术大学 自动化系,安徽 合肥 230011;
2.安徽交通职业技术学院,安徽 合肥 230051)

摘 要:行为数据辨析的目的是提取大量行为数据以识别趋势及特定活动。行为数据辨析强调入侵检测的判定支持能力的过程。基于数据挖掘关联行为分析方法的入侵检测系统,能够提升安全策略并降低入侵检测系统中的误报率。

关键词:行为数据;入侵检测;数据挖掘

中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2007)07-0139-03

Behavioral Data Forensics in Intrusion Detection Based on Data Mining

WU Yu[1,2], LI Lan[2], ZHU Ming[1]

(1.Department of Automation, University of Science & Technology of China, Hefei 230011, China;
2.Anhui Communications Technology Institute, Hefei 230051, China)

Abstract: The purpose of behavioral data forensics is to draw a large number of behavioral data in order to discern the trend and particular activity. Behavioral data forensics emphasizes the course which supports ability of intrusion detection on the judge. The intrusion detection system related to behavioral analytical method based on data mining can promote the security strategy and reduce the false positive rate in the intrusion detection system.

Key words: behavioral data; intrusion detection; data mining

0 引 言

大多数人考虑计算机辨析时,考虑的是从已经受损的计算机中重新产生数据及配置信息。他们想象着从一个崩溃的驱动器中提取数据,重新产生其内容或恢复被删除的文件。传统的计算机辨析也包括寻找临时文件及其他能拼凑出整个故事的线索。传统的辨析的目的是查明失效的计算机发生了什么。而行为数据辨析的目的是查明由运转着的计算机组成的网络中发生了什么。

行为数据辨析是分析行为的。它查看历史事件而不是静态配置信息,从而检测错误及根除错误。它的两个重要好处是显示趋势及毁坏情况评估。行为数据辨析强调入侵检测的判定支持性能的过程。基于主机的或基于网络的入侵检测系统都能进行行为数据辨析,但用基于主机的系统检测内部人员误用时,行为数据辨析尤为有效[1]。

入侵检测系统把传统的电子数据处理、安全审计、最优模式匹配及统计技术融合在一起,成为计算机和网络安全技术的重要组成部分。入侵检测监测计算机网络和系统以发现违反安全策略事件的过程,入侵检测系统包括:

(1)提供事件记录流的信息源;

(2)发现入侵迹象的分析引擎;

(3)基于分析引擎的结果产生反应的响应部件[2]。

入侵检测系统的核心在于其第二部分,在其数据分析功能中,信息被同步、分类和使用各种类型详细检查去识别安全意义的活动模式。特别是在大规模的网络环境中,由于网络流量和受攻击概率的增加,不但要解决局部环境中传统入侵检测技术的快速检测和分析问题,还要解决大规模主干网络的检测和分析问题,行为数据分析方法的应用,将有助于提升入侵检测的性能。

1 行为数据辨析

行为数据辨析是指在入侵检测系统中分析其数据库时所采用的数据挖掘技术。行为数据分析的目的是提取大量的行为数据以识别趋势及特定活动,行为数

收稿日期:2006-09-29

作者简介:吴 玉(1965-),男,安徽霍丘人,硕士,高级工程师,主要研究方向为计算机网络;朱 明,教授,博士生导师,主要研究方向为网络安全、数据挖掘。

据辨析是强调入侵检测的判定支持能力的过程。

行为数据辨析通常被用于攻击预测、显示趋势及损失评估。但行为数据几乎能表示商业运作的每个方面,所以除了计算机安全之外,其好处还包括重建经营过程、平衡资源负载及传统的审计。

在传统的模式下,人们在考虑计算机数据分析时,是想到如何在受到入侵破坏的计算机中重新恢复数据或系统配置,试图从数据盘中提取数据,重新产生其内容或恢复被删除的文件,以及通过查看系统文件等等各种方式弄清事情的来龙去脉,总而言之,是想了解受损的计算机中发生什么。行为数据分析是分析行为的,它查看历史事件而不是静态地配置信息,从而检测错误及根除错误,其目的是查明由运转着的计算机组成的网络中发生了什么,它的两个重要性能是显示趋势及毁坏情况评估。

行为数据辨析的核心是在一个由不同信息组成的数据库中进行的一系列查询与调查,以查找感兴趣的模式。行为数据辨析工具通常包括一个绑定到图形报表书写程序上的分析工具、一个用于二进制数据源的原始审计观察器、一个 SQL 接口以及一沓纸。

行为辨析包括图形解释、原始审计分析及数据查询。

2 数据挖掘

数据挖掘本身是一项通用的知识发现技术,将它应用于入侵检测的目的是从大量数据中提取出有用的数据信息,发现未知攻击。应用到入侵检测系统中的数据挖掘算法,目前主要集中在关联、序列和分类这 3 种类型上[3]。

(1)关联分析算法。关联分析算法由 R. Agrawal 等人提出,是数据挖掘的一个重要课题。其目的是挖掘事务集中满足给定支持度的项集,然后产生关联规则。比较主流的算法有 Apriori 算法和 AprioriTid 算法[4]。

(2)序列分析算法。关联分析用于挖掘数据记录中不同数据项之间的关联性,而序列分析则是发现不同数据记录之间的相关性。序列分析的目标是在事务中挖掘出序列模式,即满足用户指定的最小支持度要求的频繁序列,并且该序列模式不被任何其它序列所包含。代表算法是 AprioriAll, AprioriSome, PSP, GSP 等。

(3)分类算法。数据分类的目的是提取数据库中数据项的特征属性,生成分类模型,该模型可以把数据库中的数据映射到给定类别中的一个。常用的分类算法有:RIPPER, ID3, C4. 5, NaiveBayes, 神经网络等。RIPPER 是由 W. Cohen 提出来的,是一种通用的分类规则生成算法。它在对包含大量噪声数据的数据集进行处理时得到很好的性能,而且 RIPPER 算法中的规则优化模块可以循环调用,从而进一步提高了分类的准确性。

数据挖掘被定义为从数据中提取隐含的、以前不知道的、有潜在作用的信息。它利用统计与可视化技术以易于理解的形式发现并表现信息。

用于数据挖掘的技术包括与上下文相关的解释、数据表达精练、深入分析、合并不同的数据源以及使用带外资源。数据挖掘结果有助于实现损失评估、攻击预测、监视及警告。采用的挖掘技术、使用的工具将直接与入侵检测需求联系在一起。

在入侵检测中,数据挖掘被定义为处理大量在中央位置收集到的数据,从而查看感兴趣的模式。笔者特别提到感兴趣的模式而不是误用模式,这是因为行为数据辨析并不限于计算机安全。感兴趣的模式包括表示个人、团体及经营过程的行为数据。处理这些感兴趣的数据可以提供大量信息(包括重建经营过程、平衡资源负载及传统的审计)。

2.1 形式与格式

数据将以许多不同的形式与格式存在,所以能够处理这些不同的数据类型是重要的。下面是可能会遇到的数据类型列表:

* 捕获网络事件期间收集到的原始 TCP/IP 数据。

* 原始二进制操作系统数据。

* ASCII 应用程序数据(如 Syslog)。

* 存储在关系数据库中的被检测到的标志。

* 存储在关系数据库中的行为统计。

2.2 数据量

由于数据的多样性及数量,挖掘行为数据可能是个挑战。如果从目标服务器中收集原始数据,可以指望每天收集 5~10MB 数据。有些目标机产生的数据量还要更多。如果只有 100 台服务器,那每天产生的数据就超过 1GB。

收集了数据之后,就必须考虑数据的价值与收集的花费及越来越大的存储量。过多的数据可能是坏事。数据量可能太大以至于实际上不可能使用或处理所有的数据。

2.3 用户中心监控和目标中心监控

行为数据辨析及数据挖掘很大程度上要依赖于目标数据库及数据源。用户中心监控是指数据库被优化,从而提供用户数据,而目标中心监控则是指数据库被优化,从而提供目标数据。尽管大多数数据是基于

每个目标机收集、组织的,但大多数分析员监控的是用户而不是目标机。传统上数据是以这种方式组织的,因为基于主机的传感器通常驻留在目标机上。入侵检测系统必须采用一组额外的步骤以沿着用户线路重新组织数据。用户中心监控对有效的入侵检测系统来说是极其重要的。

阐明这点的最好的方式是比较搜索大的报表所用的时间。在传统的入侵检测数据库结构中,标志数据是存储在表示目标机的表中的。搜索归结于单个目标机上所有用户的标志将非常有效,这可从单个表得到。但搜索归结于所有目标机上单个用户的标志将需要搜索每个目标机表中的单个用户——效率非常低。

当一个用户已被认定为误用者,现在安全人员必须确定损失的程度,这时需要考虑一个标准的损失评估活动。安全人员必须确定是否有任何其他的系统受损。如果采用用户中心监控,有可能产生单个报表并显示非法用户在短时期内访问过的所有服务器。如果采用目标中心监控,安全人员必须进行非常低效的搜索,可能要花费数小时,这依赖于企业的规模。

3　行为数据辨析在现实世界中的实例

在入侵检测系统中使用数据挖掘技术,通过分析历史数据可以提取出用户的行为特征、总结入侵行为的规律,从而建立起比较完备的规则库来进行入侵检测[4]。文中构建了一个基于数据挖掘关联分析方法的入侵检测系统,该系统主要用于异常检测。

该系统的数据来源是基于网络的,通过在网络中安放嗅探器来获取用户的数据包,然后采用协议分析的方法,丢弃有效负荷,仅保留包头部分,按特定的方法预处理后得到的数据包含 7 个字段:时间、源 IP、源端口、目的 IP、目的端口、连接的 ID、连接状态。

由于 TCP 的连接建立包含 3 次握手过程,所以在所有收集的训练数据中会包括一些未能成功建立的连接,它们将对后面的数据挖掘过程产生负面影响,故应当去掉,仅保留那些反映网络正常情况的数据。对于 UDP 则不存在此问题,只需将每个 UDP 包都视为一次连接即可。采用 APRIORI 算法[5]对数据进行挖掘。

APRIORI 算法常用在购物篮分析中,它用于发现“90% 的客户在购买商品 A 时也会购买商品 B”之类的规则。它通常的输入分为两列:

规则输出的形式为 I1&12aI5(support = 2%, confidence = 60%)。其中 support 是支持度,confidence 是可信度。

将前面收集到的网络流量数据格式化成为 APRIORI 算法的输入形式,用连接 ID 代替客户 ID,其他属性替代购买的商品。在给定了支持度和可信度之后,可以得到一组规则,形式为 192.168.0.50&202.117.80.8a80(support = 6%, confidence = 95%)。

规则的含义为源 IP 为 192.168.0.50 且目的 IP 为 202.117.80.8 则目的端口是 80,该规则的支持度为 6%,可信度为 95%。

一段时间的采样不能够完全代表用户的行为,因此有必要多次采样,并重复上述过程,然后用归并的方法将多次得到的规则集合并起来,直至不再产生新的规则为止。笔者采用此方法从大量的网络流量数据(28.8MB)中可以提取出 100 多条规则(支持度 2%,可信度 85%),发现其中有很多是明显无意义的,这就需要管理员通过个人经验加以精简,最终得到可以用于检验的规则集。至此,产生的规则集已经可以比较完整地描述用户的行为特征了。将得出的规则集用于入侵检测。例如,规则库中的一条规则为 192.168.0.50&202.117.80.8a80 (support = 6%, confidence = 95%)。

而在检测的过程中发现网络数据中的一个连接源 IP 地址是 192.168.0.50 且目的 IP 地址为 202.117.80.8,访问的端口为 1000,则说明违反规则的小概率事件发生,该连接的可疑度随之增加。在实际过程中,来自同一 IP 地址的异常的连接可能会违反多条规则,当多个可疑度之和超过一个阈值时系统就产生报警。

采用了两组数据对此系统进行了实验。一组是已知不含任何攻击的正常数据(约 30MB,包含 35 万余条记录),该数据用于训练系统,采用以上介绍的方法,在设定支持度为 1%,可信度为 85% 情况下,得到了 17 条检验规则。然后将得到的规则用于检测另一组已知包含攻击的数据(约 54MB,包含 63 万条记录),实验结果证明此方法可以有效地发现 PROBING 攻击。

4　小　结

在入侵检测系统中运用数据挖掘技术,可以有效地从各种数据中提取出有用的信息,降低 IDS 中的误报率。通过上面完整的实例,可以看到行为数据辨析在入侵检测系统中的应用,它在改善网络性能、加强系统安全性、减少工作负载、提升安全策略等方面将发挥重大的作用。

参考文献:

[1] Proctor P E. 入侵检测实用手册[M]. 邓琦皓等译. 北京:中国电力出版社,2002:104 - 118.

(下转第 144 页)

得到认证。如果等式(5) 不成立,则委托过程失败。

Step3:代理签名的产生过程。对任何消息 $m(o < m < n)$,A 的代理签名人B 可以按照下面的方法产生关于消息的代理签名。B 首先选取随机数K,$o < K < n$,然后计算 K_P,记 $K_P = (x,y)$,其中 $x,y \in F$,接着 B 计算:①$r \equiv x \bmod N$;②$S \equiv k^{-1}(m + rQ_0)\ m_0 \bmod N$,则$(m,r,s,Q_0)$ 一起构成了代理签名人及消息的代理签名。

Step4:验证过程。任何一个验证人 C 收到代理签名(m,r,s,Q_0) 后,利用原始签名人的公钥 P_A,进行下列计算:

①$c \equiv s^{-1} \bmod N$;

②$u_1 \equiv mc \bmod N$;

③$u_2 \equiv rc \bmod N$;

④ 计算 $u_1 p + u_2(PA + r_0 Q_0)$;设 $u_1 p + u_2(PA + r_0 Q_0) = (x,y)$;

⑤ 计算 $x \bmod N$,如果 $r = x \bmod N$,则代理签名得到认证。

在整个代理签名方案中,A,B 的密钥管理是前提,要充分考虑时效性。同时在委托过程中 A,B 双方要认证对方身份是否合法。验证过程严格要求计算步骤执行。文中验证过程的安全性证明可参考文献[4]。

3 代理签名性能分析

3.1 基本性能

(1) 基本不可伪造性。由 shamon 信息理论知[5],在未知关于 $h(m_w)a_1$ 的信息情况下,从等式 $s_1 = ux_A + h(m_w)a_1$ 中不能得到任何关于 X_A 的信息,但若求出 a_1,则可解出 X_A。但从中很难解出 a_1,从而不能伪造 A 的普通签名。

(2) 代理签名的不可伪造性。只有 B 知道 K 的随机独自秘密数,故只有 B 能生成代理签名。

(3) 不可抵赖性。由于任何人都不能伪造 A 的签名,所以 A 不能否认一次有效的签名,因授权给 B,因此只有 B 才能代理 A 签名。

(4) 密钥依赖性。因 A,B 的私钥都具有时效性即 $SK_i = f(SK_{i-1})$,所以密钥之间存在依赖性,但生成后自动消失。

(5) 可注销性。A 的授权消息包括授权时间及代理的有效性。对一个诚实的代理来说,授权消息保证方案的可注销性。目前大多数具有证书的签名方案其可注销性都基于此,但代理人若不诚实,可通过自动更换私钥解决。

3.2 代理签名方案特有的优点

能防止代理签名人滥用自己的签名权。

①签名方案附有授权消息 m_w,ID_A,ID_B,代理签名消息的许可范围,且 m_w 受安全Hash函数保护。这样 B 既不能将自己的代理权转移给他人,又不能对任何消息(如有损 A 利益的消息) 进行签名。

② 证书上的 N 为A 规定B 在T(T 为授权时间及代理的有效期限) 内签名的最大次数。若 B 违反了规定,在进行 $N+1$ 或 $N+1$ 次以上的签名后,由代数知识知,方程组 $n_i u = m_i h(nt_i)X_B + s_i k + h(m_{i1}) + h(m_{i2}) + \cdots + h(m_{iN-2})(i = 1,2,3,\cdots,N)$ 共有 $N+1$ 个未知量,任何人都可以收集 $N+1$ 或 $N+1$ 个以上有效签名,带到方程组里使可解出 X_B(但此对原始签名人毫无影响),故 B 不会冒危险而滥用其代理权。这样,B 在有效期内至多签名 N 次后其代理签名权会终止。

4 结束语

文中提出了一种前向安全性的可证实代理数字签名方案,讨论了该方案的正确性和安全性,分析结果表明该方法有很强的理论和实用价值。

参考文献:

[1] 白国强,黄 谆. 基于椭圆曲线的代理数字签名[J]. 电子学报,2003,31(11):1659-1663.

[2] 李方伟,王 建,陈广辉. 前向安全的基于椭圆曲线密码体制的签密方案[J]. 北京邮电大学学报,2006,29(1):22-25.

[3] 眭新光,罗 慧. 基于S盒的数字图像置乱技术[J]. 中国图像图形学报,2004,9(10):1223-1226.

[4] 秦 波,王尚平,王晓峰,等. 一种新的前向安全可证实数字签名方案[J]. 计算机研究与发展,2003,40(7):1016-1020.

[5] 李淑静,赵远东. 基于椭圆曲线的 ELGamal 加密体制的组合公钥分析及应用[J]. 微计算机信息,2006(12):69-71.

(上接第 141 页)

[2] Denninc D. An Intrusion-detection Model[C]//In IEEE Symposium on Security and Privacy. Oakland, USA:[s.n.], 1986.

[3] 李守国,李 俊. 基于数据挖掘的入侵检测系统设计[J]. 计算机技术与发展,2006,16(4):212-214.

[4] Lee W, Stolfo S J. Data Mining Approaches for Intrusion Detection[C]//In: Proceedings of the 7th USENIX Security Symposium. San Antonio:[s.n.],1998:6-9.

[5] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范 明,孟小峰等译. 北京:机械工业出版社,2001:147-158.