基于数据挖掘的用户安全行为分析*

缪红保, 李　卫

(西安交通大学 计算机系 网络化系统与信息安全研究中心, 陕西 西安 710049)

摘　要: 通过对用户网络流量进行协议投影,获得其行为的具体信息,然后采用关联规则等方法,将上面得到的信息进行统计学习,从而得到该用户所特有的行为模式。利用这种模式,可以进行网络用户的身份识别。实验结果表明,该方法为进行用户网络行为特征提取和身份识别提供了一种新思路,另外也有助于发现蠕虫或其他大规模入侵行为。

关键词: 行为分析; 身份识别; 数据挖掘; 关联规则; 网络安全

中图法分类号: TP309.2　　**文献标识码**: A　　**文章编号**: 1001-3695(2005)02-0105-03

Analysis of Network User Behaviors Based on Data Mining Theory

MIAO Hong-bao, LI Wei

(Center for Networking Systems & Information Security, Dept. of Computer Science & Technology, Xi an Jiaotong University, Xi an Shanxi 710049, China)

Abstract: By using protocol decoding technology, the system can get detailed information about monitored user. Then, the method based on data-mining theory is taken to extract association rules from the information. These rules combined with other statistical data will form the pattern of the monitored user. Empirical results show that we can provided with the capability of extracting user patterns and identifying them in the network. On the other hand, this method also can detect large scale intrusion such as worms.

Key words: Behavior Analysis; User Identification; Data Mining; Association Rules; Network Security

随着计算机网络的深入普及,网络技术开始成为国民经济快速发展的一个重要的助推器。与此同时,网络信息安全也逐渐成为全社会共同瞩目的焦点。然而,实践表明,仅仅依靠现有的网络安全技术,诸如防火墙、入侵检测、身份认证等,并不能阻止网络中的所有攻击,也无法确保系统的安全。其原因在于现有的安全技术大多将主要精力集中在"防外",对外来攻击进行响应,而对于内部网络受外界未知攻击后,产生的后果以及内网用户可能进行的破坏性行为考虑得较少。这方面比较典型的例子是垃圾邮件、有害信息泛滥以及今年出现的 Slammer, Mblast 蠕虫等。

那么,如何从"管内"的角度来提高系统的安全性呢? 在允许用户正常使用网络的同时,尽量限制其有害行为(被攻击产生的后果或攻击他人)是我们的基本出发点。因此,发现并定位用户的网络有害行为,然后进行适当的控制就是我们要完成的主要任务。本文将主要讨论如何发现、识别不同的用户网络行为。

我们知道网络用户都具有各自的兴趣爱好,不同的人在上网时,关注的内容也不完全相同,那么相应的用户网络行为也不可避免地带有各自的特征。近年来,已经有不少研究人员开始对这种特征进行了有益的探索。但是,从已有的文献来看,他们中的大多数都是通过对服务器端的用户访问日志进行分析来实现的。比如在电子商务网站中的推荐系统,就是根据日志中记录的用户历次登录所访问的页面、点击的项目,来归纳出该用户的购物倾向。

对于大规模园区网络的管理者,上述方法无法发挥作用。受控子网用户可能访问大量外部站点,而这些被访服务器的日志多数是管理者无法得到的。因此我们考虑通过流量分析的技术,尝试将这种属于用户自身所特有的特征从网络流量中提取出来,并存入特征库中。此后,当用户再次开始网络访问,我们都可以通过对其行为进行检测、分析和匹配,以评估安全状况,识别用户身份,实现网络的安全管理。此外,在分析行为的同时,我们还能够得到一系列"副产品",文中将会介绍用户安全行为分析,也为蠕虫防治提出了一种新思路。

1　行为分析模型的设计

1.1　分析对象的选择

用户行为可以通过这样一个五元组{ *T, S, O, C, Q*} 来描述,即用户使用网络的时间(*T*)、访问的站点(*S*)、访问操作的分类(*O*)、访问的内容类型(*C*)和该访问操作的数据流量(*Q*)。说明如下:

(1) 用户使用网络的时间(*T*)。用于标志行为发生的时间,可按一定标准进行分类,诸如早、中、晚等。

(2) 访问的站点(*S*)。可划分为门户类、新闻类、教育类、娱乐类、安全类等。

(3) 访问操作的分类(*O*)。WWW——普通浏览,搜索,下载,发表; 邮件——发送,接收; Telnet——浏览,发表; 黑客行为——扫描等。

收稿日期: 2003-12-08; 修返日期: 2004-04-22

基金项目: 国家自然科学基金资助项目(59937150); 国家"863"计划资助项目(2001AA413910)

(4) 访问的内容类型(*C*) 。文本、图片、动画、音/视频、普通字符流等。

(5) 访问操作造成的数据流量(*Q*) 。上传、下载的流量等。

1.2 模型结构设计

考虑将该系统设计为一个三层结构的模型,如图 1 所示。

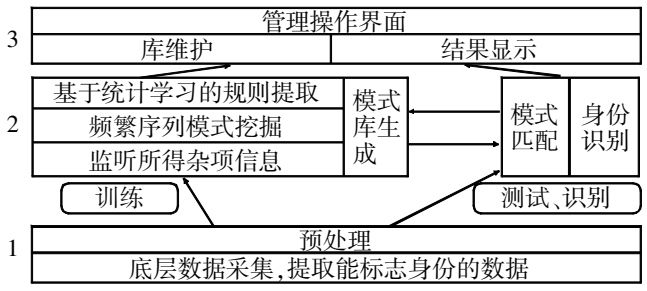

图 1 系统结构模型

2 数据获取及预处理

获取原始数据信息是整个系统的基础。对于园区网络的管理者而言,受控子网的流量是比较容易得到的。例如,我们可以通过主交换机上的 Port Mirror 或者分光器等方法,获得所需要的用户网络流量。

对数据进行采集,可以采用 PCAP 软件包。如果在 Windows 平台下,则可采用由芬兰 Politecnico di Torino 开发的 WinPcap 类库,它基于 Bpf(Berkeley 分帧过滤器) 内核,妥善封装之后对外提供调用接口,可以用来进行数据包捕获与分析。

从用户网络流量中得到源地址、源端口、目的地址、目的端口、包长、时间戳、传输层标志位等基本信息,然后通过对应用层协议进行投影,得到诸如 HTTP 的 URL, SMTP 中的 Sender, Rcpt, Subject 等扩展信息。前面介绍的各个分析对象都可以从这些通过协议解码得到的信息中获得。

部分统计学习的工作可以在预处理的同时进行。对某一指定的被监控主机,统计向外连接的目的地址、对应端口,以及对各地址、各端口发送的包数和包长。一方面,这些统计信息能够反映用户行为的倾向;另一方面,它们对于发现网内正在扩散的蠕虫或部分黑客行为也有一定的帮助。

蠕虫的发作总有一定的规律可循。蠕虫通常都是利用系统的某种漏洞作为切入点,以扩散为其特征之一。当蠕虫将源机攻克后,势必会在网内进行扫描,以寻找新宿主。这种扫描包一般都有固定的目标端口和包长。因此当我们统计时,一旦发现多台主机均有上述现象出现,与已知行为模式差别极大,即可大致判定网内有蠕虫正在发作。

3 网络行为模式挖掘

源数据采集并预处理之后,按一定比例被划分为两大部分:训练集与测试集。注意,在训练时我们必须确定每个数据包对应的发出者,这样才能建立最初的规则和模式。而在利用测试集测试时,则将结果与实际属主进行比较,对所得模式的有效性进行评估,从而发现挖掘阶段存在的问题,利于后期的改进。

3.1 基于统计学习理论的规则提取

这部分主要用于获得用户网络行为的概貌。统计的内容主要包括以下几个方面:

(1) 某个时间段内的连接统计(*BS*) 。不同 Dp(目的地址 + 目的端口) 的个数;各 Dp 上对应的出入流量;

(2) 某个时间段内协议流量的统计(*FS*) 。各协议层对应的流量。

(3) 各个时间段的流量与行为统计(*TS*) 。目标是找到在对应时间段,用户都倾向于从事哪些类型的活动。具体区段包括: am(上午) 、noon(午间) 、pm(下午) 、nt(晚上) 。

构建的规则形式为: R = address + behavior + time, 计算其支持度 N_R/N, 置信度 N_R/N_A。以上统计完成之后,结合训练前提,我们可以获得如下结论:

If $BS_i \in [\ V_{jb}\ ]$ and $FS_i \in [\ V_{jf}\ ]$ and $TS_i \in [\ V_{jt}\ ]$ Then Operator = i(注: 这里的 $X_i \in [\ V_j\ ]$, 意思是被监控对象 i 的统计项 X 满足阈值 V_j 的要求。)

3.2 频繁向量序列模式挖掘

3.2.1 挖掘方法概述

在进行频繁序列模式挖掘时,可以参考主成分分析[1] 的技术,将向量中对模式生成贡献较大的若干分量(如地址、操作) 抽取出来,单独进行挖掘,最后再将所得结果汇总进行综合分析。

3.2.2 挖掘前的准备工作

首先是构造行为向量。行为向量记作 (*T*, *A*, *B*, *F*) , 包括时间 *T*: 行为发生时间; 地点 *A*: 源 *IP* + 源端口 + 目标 *IP* + 目标端口; 行为 *B*: 协议 + 操作,例如 Http + Get; 流量 *F*: 行为过程中的流量。

然后按时序构造事务集。事务集是由诸多行为向量组成。在这里,我们把被监控主机在 10min 内产生的行为向量组合成一个事务。在后面的分析中,主要是对行为向量的分量组成的子事务进行挖掘。

3.2.3 基于关联规则技术的频繁向量模式挖掘

关联规则[2] 用于寻找给定数据集中项与项之间的联系。经典的关联规则挖掘算法是 Apriori 算法[3] 。目前,诸多研究成果都是在该算法的基础之上进行的。Apriori 算法存在一些固有的缺点: ①算法在运行过程中会产生大量的候选项集; ②算法需要重复扫描数据库,通过模式匹配的方法对候选项集进行筛选。当分析对象规模庞大的时候,上述两条带来的运算开销将很大。为弥补该算法的不足, Han J. 等人提出了一种基于 FP-TREE 挖掘的 FP-grow 算法[4] , 将发现长频繁模式的问题转换成递归发现一些短模式,然后进行连接,从而提高了运算效率。

我们希望,当记录集发生更新需要重新挖掘的时候,可以充分利用上一次挖掘所得到的成果,减少运算开销。因此决定采用 FP-grow 算法的改进算法 $FIUA_2$ 算法[5] , 并修改了其中的错误。假定 *D* 为某用户的已有事务数据库, *d* 为新监听获得的事务数据库,操作如下:

扫描 d 求 D 中的强频繁项集 $L_{D'}$, 以及 d 的 1-强频繁项集 $L_{d1'}$, L_{n1}, 并利用 FP-grow 得到 L_{dk}; 调用 AdjustFP-tree(FP-tree, L_{n1}) 建立模式树 P-tree;

C_{d2} = Apriori_gen($L_{d1'}$) ;

for (k = 2; $C_{dk} \neq \phi$; k + +) do {

扫描 d 求 C_{dk} 中的各项集在 d 中的支持数,删除 d 中的非频繁项集得到 C_{dk};

$C_{nk} = C_{dk} \cap L_{dk} - L_k$;

if ($C_{nk} \neq \phi$) then {

for each $\rho = \{\ \alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_t\} \in C_{nk}$ 调用 computercount (P-tree, ρ) ; }

$L_{dk'} = \{ c \in C_{dk} \mid c.sup\ D_d >= s \}$;

$C_{d(k+1)}$ = Apriori_gen($L_{dk'}$) ;

$L_{nk} = \{ c \in C_{nk} \mid c.sup\ D_d >= s \}$; }

$L_{Dd} = L_{nk} \cup L_{D'}$

由此,即使在数据库发生更新的时候,我们都能迅速从中

提取出频繁项集。

4　模式比较与用户识别

模式比较从本质上讲就是计算序列相似度的过程。已有文献中介绍的序列相似度比较算法包括 SAM 算法[6]、基于欧式距离算法等。传统算法对序列间整体相似程度比较关注,但却经常忽略子序列之间的相似关系。比如对序列 S_1(1, 2, 3, 4, 5, 6) 和 S_2(1, 6, 2, 3, 9, 4, 5, 6),常规用最大公共子序列来判别相似度的算法通常都只能发现(4, 5, 6) 这个子序列,但事实上子序列(2, 3),(4, 5, 6) 在两个序列中出现的顺序是一致的,这说明其实际相似程度应该大于常规算法计算所得结果。

为描述这种子序列之间的顺序关系,更精确地计算序列之间的相似程度,设计算法 Deep-Simi 如下:

对模式序列 S_p 和测试序列 S_t,

①求出最大公共子序列 S_c,求得其长度为 $L(S_c)$;

②从原序列中分别删除最大子序列,从而得到新的 S_p, S_t;

③判断其 2 阶或 2 阶以上最长公共子序列是否为空,以及 S_p, S_t 本身是否为空,空则结束循环转入④,否则重复步骤①~③,并将得到的最大子序列加入到子序列集中;

④当循环结束后,对最终生成的子序列集,设集合中元素个数为 n,判断其元素在两个序列中是否顺序一致,每两两一致,则计数器 Count 加 1,然后将最终得到的 Count 值对 $n(n-1)/2$ 作归一化,结果为 ρ

⑤由此可得两序列 S_p, S_t 的相似度为

$$\text{Deep-Simi}(S_p, S_t) = L(S_c) / \max(L(S_p), L(S_t)) + \rho$$

当然,整个用户识别的过程并不仅仅包括模式比较,此外还应该考虑统计学习过程中获得的规则和监听过程中其他一些能描述用户个人身份的资料,如邮件系统的用户名、口令等。对统计学习获得的规则,判别方式比较简单,在满足阈值要求的前提下,与规则内容完全一致,则认为身份相符。

5　结论

本文介绍了基于数据挖掘理论的用户安全行为分析模型。它通过搜集用户在上网过程中生成的大量数据信息,采用统计学习和关联规则挖掘技术,充分发挥数据挖掘理论能从超大规模数据集中发现知识的优势,对搜集得到的行为数据进行分析,提炼用户行为模式,并进一步将这些模式应用于对网络的管理和监控工作之中。

目前系统还只是初步实现,并对数台机器组成的局域网流量进行了简单测试。其中,针对用户 A 挖掘所得部分行为模式如表 1 所示。

表 1　针对用户 A 的行为数据统计学习所得规则

编号	规则内容	规则说明	支持度
1	AM 202.117.7.218:* ~202.112.11.200: 110 POP3	早晨,该用户去 202.112.11.200 收取邮件	31%
2	NOON 202.117.7.218: * ~202.117.1.8: 23 Telnet	中午,该用户去 202.117.1.8 访问 BBS	22%

实验结果表明,该系统能有效地发现用户的行为模式,能够识别出网内突然发作的蠕虫。在未来的工作中,为了使系统能运行于大规模园区网络上,还需要对数据预处理和挖掘算法作进一步的修改,提高系统的处理能力。源数据库的维护和模式的更新机制也有待深入考虑。此外,作为一个网络安全相关的项目,系统自身也需要加强安全防范,提高其健壮性。以上这些都需要在今后的工作中进一步得到改善。

参考文献:

[1] and D, Mannila H, Smyth P. 数据挖掘原理[M]. 张银奎, 廖丽, 宋俊, 等. 北京: 机械工业出版社, 2003. 48-54.

[2] Han J, Kamber M. 数据挖掘: 概念与技术[M]. 范明, 孟小峰, 等. 北京: 机械工业出版社, 2001. 150-161.

[3] Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Items in Large Databases[C]. Proceedings of ACM SIGMOD International Conference on Management of Data, Washington DC, 1993. 207-216.

[4] Han J, Pei J, Yin Y. Mining Frequent Patterns without Candidate Generation[C]. Proceedings of ACM SIGMOD International Conference on Management of Data, Dallas, TX, 2000. 1-12.

[5] 朱玉全, 孙志挥, 季小俊. 基于频繁模式树的关联规则增量式更新算法[J]. 计算机学报, 2003, 26(1).

[6] Hay B, *et al*. Clustering Navigation Patterns on a Website Using a Sequence Alignment Method[C]. Proc. of In telligent Techniques for Web Personalization: IJCAI 2001 17th Inter national Joint Conference on Artificial Intelligence, Seattle, Washington DC, 2001. 1-6.

作者简介:

缪红保(1981-),男,硕士研究生,研究方向为网络安全;李卫(1967-)男,主任,副教授,博士,研究方向为网络安全与管理。

(上接第 112 页)海量影像文件,只能显示文件前面 2GB 部分的数据,在实时缩放和显示的速度方面相差不大。图 4 是 8GB 的影像文件在自主开发的中国海岸带及近海卫星遥感综合应用系统(Max_Deskpro) 和 ESRI 的 ARCGIS 8.2 上的显示结果。

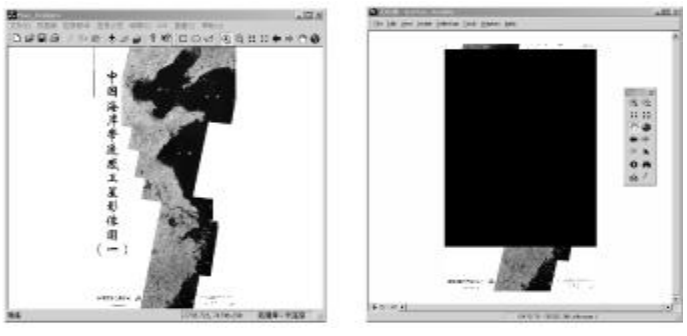

图 4　在 Max_Deskpro 和 ARCGIS 8.2 上的显示效果

5　结束语

实践证明,采用内存映射文件技术读取海量图形文件的速度比采用 Win32 API 和 MFC 提供的文件操作函数和类读取文件的速度要快得多,因为经它映射后,文件中的数据直接就存放在缓存中了,省去对文件执行 I/O 操作的时间,此外它还可以正确读取大于 2GB 的图像,有效地解决了传统的文件指针只能读取 2GB 数据的限制;其次由于本程序使用了缓存技术,即每次在内存中的数据,只是影像文件中的一块,所以占用内存很小。总之,采用本方法进行海量遥感图像的读取具有一定的优势。

参考文献:

[1] 希望图书创作室. Visual C++ 6.0 技术内幕[M]. 北京: 北京希望电子出版社, 1999. 215-218.

[2] 孙家炳, 舒宁, 关泽群. 遥感原理、方法和应用[M]. 北京: 测绘出版社, 1997. 191-194.

[3] 吕京国, 黄国满. 用 Visual C++实现大数据量的快速存取[J]. 测绘科学, 2002, (9): 29-31.

作者简介:

胡伟忠(1979-),男,浙江湖州人,硕士研究生,主要从事海洋 GIS 和图形图像处理方面的研究;刘南(1944-),男,教授,博士生导师,主要从事 GIS 理论研究;刘仁义(1960-),男,教授,主要从事时态 GIS 方面研究。