

# 基于 CURE 算法的网络用户行为分析

孙燕花, 李 杰, 李 建

(中南大学 信息科学与工程学院, 湖南 长沙 410075)

**摘 要:**从安全的角度分析网络用户行为, 建立了一个基于 Netflow 统计的用户行为向量数据模型, 提出了一个网络用户行为的分析框架, 建立了一个分析流程。针对存储网络用户行为的大型数据库选用了合适的聚类算法即 CURE 算法, 并对 CURE 算法进行了基于实际应用的改进。实验结果表明, 改进后的 CURE 算法不仅能很好地聚类, 而且能区分出正常行为和异常行为, 通过危害行为评价体系分析, 聚类得到的异常行为是危害行为的检测率非常高。对于实时网络上的增量数据, 文中也给出了增量挖掘的算法, 符合网络实时分析的需要。

**关键词:**网络安全; 数据挖掘; CURE 算法; 异常行为; 增量挖掘

**中图分类号:** TP31

**文献标识码:** A

**文章编号:** 1673-629X(2011)09-0035-04

## Network Users Behavior Analysis Based on CURE Algorithm

SUN Yan-hua, LI Jie, LI Jian

(School of Information Science and Engineering, Central South University, Changsha 410075, China)

**Abstract:** For analysing network user behavior based on network security, a network user behavior data model based on Netflow statistics is established. A framework of analysis is put forward. An analysis process is established. According to the consumer behavior of large storage network database, an appropriate clustering algorithm, called CURE algorithm, is chosen, which is improved based on actual application. Experiment results show that the improved algorithm is not only able to cluster, but also can distinguish the normal and abnormal behaviors. Analysed by harm behavior evaluating system, most of the abnormal behaviors belong to harm behaviors. For increment data on real net, it also gives the method of increment mining, which accords with the need of real time network analysing.

**Key words:** network security; data mining; CURE algorithm; abnormal behavior; increment mining

## 0 引 言

网络用户行为是指行为主体为实现某种特定目标, 采用基于计算机系统的电子网络作为手段和方法而进行的有意识的活动<sup>[1]</sup>。网络用户行为从对网络影响来说分为正常行为和异常行为。异常的网络用户行为可能对网络产生以下三类影响<sup>[2]</sup>: 影响网络的安全, 如信息窃密、冒充等信息安全方面的攻击; 影响网络的性能, 如 DoS、ARP 泛洪等耗尽网络资源方面的攻击; 既影响网络安全又影响网络的性能, 如 ARP 欺骗攻击既窃取用户信息又消耗网关的资源。现有的安全技术有以下两方面具有脆弱性:

①外来未知攻击对内部网络产生的可能影响的控制;

②内网用户的可能威胁的感知和控制。

因此通过网络用户行为分析来监控网络安全成了一个重要的研究课题。

## 1 网络用户行为分析方法

数据挖掘<sup>[3]</sup>是一种特定应用的数据分析过程, 可以从包含大量冗余信息的数据中提取出尽可能多的隐藏知识, 从而为做出正确的判断提供基础。将数据挖掘技术应用到网络用户行为分析中, 可以对海量的流量数据进行智能化和快速处理。

稳定网络内的网络用户行为基于各种应用呈现多类别。聚类算法可以按照相似性将网络用户行为聚类。

对正常行为进行分析, 可以提取其行为模式<sup>[4]</sup>, 用于预测网络行为。对异常行为进行分析, 可以准确定位到网络中出现问题的主机, 方便网络管理者采取一定的控制策略来控制后续行为, 确保网络安全、畅通<sup>[5]</sup>。

整个网络用户行为分析流程图如图 1 所示。

## 2 网络用户行为分析的数据模型

### 2.1 数据采集

数据采集是网络用户行为分析的基础, 只有选择

收稿日期: 2011-01-14; 修回日期: 2011-04-21

作者简介: 孙燕花(1985-), 女, 湖南娄底人, 硕士研究生, 研究方向为网络管理; 李 杰, 教授, 研究方向为网络管理。

合适的采集方式,才能兼顾到效率和质量。

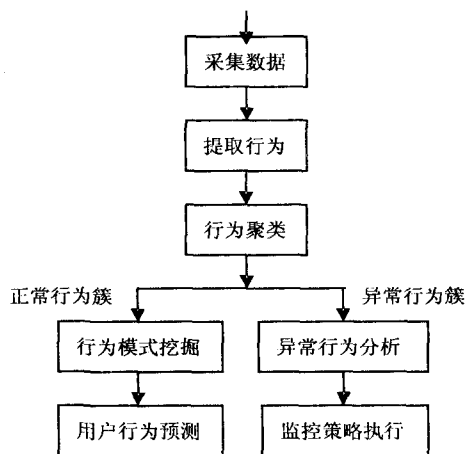


图 1 网络用户行为分析流程图

数据采集技术从粒度大小来分有两种。一种是数据包捕获,将采集主机的网卡设置为混杂模式,那么该采集主机所属网段内的所有数据包都能被采集到;另一种是流量数据获取,将经过网络设备的所有数据包顺序组成流量数据进行采集,如 Cisco 的 Netflow<sup>[6]</sup> 技术可通过 Cisco 交换机的所有数据包按照四属性字段相同组成流数据进行汇集。

从分析的角度看,分析网络上的单个数据包难度非常大,耗时多,且分析出来的结果零散没有意义;而分析流量<sup>[7]</sup>能够得到大量关于网络运行状况的有用信息,且难度不大,针对性非常强,更有效。现有的网络监控体系一般都是基于流量分析。

## 2.2 网络用户行为表示

网络用户行为可以用某些特征量<sup>[5]</sup>的统计特征或特征量的关联关系定量或定性地表示。网络用户行为一般用四元组(源 IP、目的 IP、统计参数、统计参数值)<sup>[9,10]</sup>来表示。其中统计参数的选取可以根据研究的目的而定。不同类型的变量需要进行相异度计算。

文中用到的数据源为 Netflow 流数据,因此行为向量可以根据 Netflow 流数据的格式进行设计。

Netflow 协议目前有 1、5、6、7、8、9 六个版本,而版本 5 是当前主要的实际应用版本。表 1 是版本 5 的流记录格式。

根据表 1 中 Netflow 流的记录格式,可以设计基于 IP 地址的行为向量属性。设计的方法是统计采集周期内各 IP 发生的流的主要信息。表 2 是经过采集、统计一定时间粒度下的 Netflow 流数据之后得到的每个 IP 的行为向量属性。

## 3 CURE 算法

### 3.1 算法描述

绝大多数聚类算法(如 K-means, K-medoid,

表 1 Netflow 版本 5 的流记录格式

字节	内容	描述
0 ~ 3	srcaddr	源 IP 地址
4 ~ 7	dstaddr	目的 IP 地址
8 ~ 11	nextthop	下一跳的路由器 IP 地址
12 ~ 13	input	输入接口的 SNMP 索引
14 ~ 15	output	输出接口的 SNMP 索引
16 ~ 19	dPkts	流中的报文
20 ~ 23	dOctets	在流的报文中的第 3 层字节总数
24 ~ 27	First	流开始处的 SysUptime
28 ~ 31	Last	流中最后一个报文被接收时的 SysUptime
32 ~ 33	sreport	TCP/UDP 源端口号或等价值
34 ~ 35	dstport	TCP/UDP 目的端口号或等价值
36	pad1	未使用的字节(0)
37	tcp ~ flags	TCP 标记的累积 OR
38	prot	IP 协议(如 6=TCP, 17=UDP)
39	tos	服务的 IP 类型
40 ~ 41	src ~ as	源的 AS, 原来的或对应的
42 ~ 43	dst ~ as	目的 AS, 原来的或对应的
44	src ~ mask	源地址前缀的掩码位
45	dst ~ mask	目的地址前缀的掩码位
46 ~ 47	pad2	未使用的字节(0)

表 2 网络用户行为向量特征内容

IP 统计属性	说明
$X_{1S}$	作为源数据包字节数
$X_{1D}$	作为目的数据包字节数
$X_{2S}$	作为源数据包数量
$X_{2D}$	作为目的数据包数量
$X_{3S}$	作为源数据流目的子网个数
$X_{3D}$	作为目的数据流目的子网个数
$X_{4S}$	作为源源端口个数
$X_{4D}$	作为目的源端口个数
$X_{5S}$	作为源目的端口个数
$X_{5D}$	作为目的目的端口个数
$X_{6S}$	作为源目的 IP 个数
$X_{6D}$	作为目的源 IP 个数
$\{X_{7S}, X_{7S+1}, \dots, X_{7S+N}\}$	作为源前 N 个协议所占总流量比例
$\{X_{7D}, X_{7D+1}, \dots, X_{7D+N}\}$	作为目的前 N 个协议所占总流量比例
$\{X_{8S}, X_{8S+1}, \dots, X_{8S+N}\}$	作为源前 N 个源端口所占总流量比例
$\{X_{8D}, X_{8D+1}, \dots, X_{8D+N}\}$	作为目的前 N 个源端口所占总流量比例
$\{X_{9S}, X_{9S+1}, \dots, X_{9S+N}\}$	作为源前 N 个目的端口所占总流量比例
$\{X_{9D}, X_{9D+1}, \dots, X_{9D+N}\}$	作为目的前 N 个目的端口所占总流量比例

CLARANS, BIRCH)擅长处理球形和相似大小的聚类,而 CURE 是一种针对大型数据库的高效的聚类算法,通过采用固定数目的代表对象来表示每个类,使得 CURE 能够识别非球形和大小变化较大的类,并且对孤立点不敏感。为了处理大数据集, CURE 算法通过随机抽样的方法从大数据集中抽取一个随机样本。

一定规模的网络数据量大而繁杂,呈现多维特性,

事先无法确定各类数据的形状,因此 CURE 算法适用于网络用户行为数据的挖掘。

但是,CURE 算法只是针对聚类,其对异常点进行了一次过滤、两次剔除,而在网络用户行为中异常数据往往代表异常行为,对异常行为的分析有助于检测网络内发生的异常。因此必须将异常点收集起来,建立异常行为库。

### 3.2 CURE 算法的改进

加入“收集异常点”后的 CURE 算法步骤如下:

Step1. 对源数据库进行抽样,得到一个随机样本  $S$ ;

Step2. 将样本  $S$  分割为一组划分;

Step3. 对每个划分局部地聚类;

Step4. 第一步收集异常点:给定增长阈值  $\varepsilon$ ,如果某个簇的增长速度始终小于  $\varepsilon$ ,则将该簇汇入异常簇,并从样本中去掉;

Step5. 对局部的簇再进行聚类,整个样本数据聚类完毕;

Step6. 第二步收集异常点:对所有类进行孤立点集探测,将孤立点集合收集起来,汇入异常簇。

### 3.3 孤立点集探测

经过 Step6,聚类之后需要对簇进行孤立点集探测<sup>[11]</sup>。

几个相关的概念定义如下:

定义1(中位数):对于一维空间中的  $n$  个数据来说,排序后的数据计为:  $num_1, num_2, \dots, num_n$ 。mid 为这排序后的  $n$  个数据中位于最中间位置的数,如果  $n$  为偶数,  $mid = num_{n/2}$ , 如果  $n$  为奇数,则  $mid = num_{\lceil n/2 \rceil}$ , 其中  $\lceil \cdot \rceil$  为取整,例如  $\lceil 3.5 \rceil = 4$ 。

定义2(异常临界点):  $|C_1| \geq |C_2| \geq \dots \geq |C_k|$  分别为聚类后得到的类  $C_1, C_2, \dots, C_k$  中数据的数量,给定一个数值参数  $\sigma$ , 如果其满足下面两个条件:

$$|C_{mid}| / |C_b| \geq \sigma \quad (1)$$

$$|C_{mid}| / |C_{b-1}| \leq \sigma \quad (2)$$

(其中  $b$  为整数,且  $mid < b \leq k$ )

则称  $b$  为异常临界点,簇  $C_b$  后的均可视为异常簇。

将聚类后的簇由大到小进行排序,标类为  $C_1, C_2, \dots, C_k$ , 则可以根据给定合适的参数  $\sigma$  得到异常临界点,进而找出异常簇。

### 3.4 增量数据挖掘

CURE 算法本身只能处理静态样本,而网络数据不断更新,如果在新增数据和原有数据组成的大数据库中再应用 CURE 算法,则会耗费时间。可以对新增数据采用增量式挖掘<sup>[12]</sup>的办法。

定义3(簇宽度):设簇  $C_k$  的代表点是  $d_1, d_2, \dots,$

$d_c, c$  为簇  $C_k$  的代表点个数,则簇  $C_k$  的宽度等于代表点之间距离的最大值,即  $wid(C_k) = \max[\text{dist}(d_i, d_j)]$  ( $1 \leq i \leq c, 1 \leq j \leq c, \text{dist}$  为欧几里得距离)。

假设样本数据中聚类到的正常簇是:  $C_1, C_2, \dots, C_m$  ( $m$  为正常簇个数)。新增数据归类步骤如下:

Step1. 计算行为  $d'$  与各簇代表点的平均距离;

Step2. 如果平均  $\text{dist}(d', d_i) \leq wid(C_k)$  ( $1 \leq i \leq c, 1 \leq k \leq m$ ), 则点  $d'$  属于  $C_k$ , 否则转到 Step3;

Step3. 进行异常点分析。

## 4 实验结果分析

实验采用中南大学铁道校区网一个工作日早上 9:00 到晚上 9:00 从 Cisco 路由器得到的 Netflow 流数据,并将其按照表 1 中总结的格式进行行为建模,统计时间粒度为 0.5min,得到 5816726 条行为信息。随机抽取 50000 条数据作为样本数据。由于 CURE 算法本身的聚类结果受用户输入的聚类数目  $K$ 、收缩因子  $a$  以及代表点个数  $c$  的影响,因此在试验中表现出了不同的结果。通过 5 组实验,分别对 5 个输入参数影响进行观察,如表 3 所示。表中  $a$  表示 CURE 收缩因子,  $K$  为聚类个数,  $c$  为代表点个数,  $\sigma$  为孤立点集探测的临界因子,  $\varepsilon$  为簇增长阈值,  $N1, N2$  分别为改进 CURE 算法中 Step4、Step6 得到的异常簇个数。

表 3 CURE 算法用于网络用户行为聚类的实验结果

序号	$a$	$K$	$c$	$\varepsilon$	$\sigma$	$N1$	$N2$	危害行为	检测率
1	0.3	30	15	20	100	27	42	57	82.6%
2	0.3	30	15	5	100	21	38	54	91.5%
3	0.2	30	15	15	10	25	46	62	87.3%
4	0.3	30	20	5	10	30	55	76	89.4%
5	0.2	25	15	5	10	38	52	82	90.1%

收集到所有异常簇后,进行分析评价,参照文献[2]中提出的评价体系,存在危害级别的行为检测率结果如下。

从表 3 第 1 组和第 2 组实验数据可以看出,  $\varepsilon$  越大,被判断为异常行为总量越多,检测率低。第 2 组和第 4 组表明,  $\sigma$  越大,被判断为异常行为簇总量少,检测率高。第 4 和第 5 组中  $a, K, c$  三个参数在异常检测中影响不明显,这说明 CURE 算法对孤立点敏感度低,有利于收集异常点。

## 5 结束语

文中建立了一个网络用户行为的数据分析模型,设计了一个网络用户行为分析的方法流程。将 CURE 算法中对异常点进行删除的操作改成收集,使之适应网络用户行为分析的研究。对于如何识别异常点,设

计了一个算法,实验证明该算法具有较好的健壮性。

CURE 算法涉及的实际参数如样本大小、正常簇的聚类数目、收缩因子  $\alpha$ 、簇增长阈值  $\varepsilon$  以及判别临界异常簇的  $\sigma$  都需要技术人员手工输入,不同的输入对聚类结果和异常点收集产生不同影响。

CURE 聚类得到的用户行为为正常行为,对正常用户行为后续的工作是通过关联规则和序列模式分析等手段挖掘其行为模式及预测用户的行为。预测用户行为能够提前感知网络的异常行为<sup>[13]</sup>,及时阻断危害性网络连接<sup>[14]</sup>。

对于异常行为,根据不同危害级别,可以采取相对应的管理方法,如对于严重级别的,可以报警通知网络管理员强制查封用户的 MAC 和 IP。

#### 参考文献:

- [1] 吴勇. 网络环境下用户行为研究与实现[D]. 南京:南京理工大学,2007.
- [2] 梅震琨. 基于用户行为的园区网网络管理模型[D]. 长沙:中南大学,2009.
- [3] Han Jiawei, Kambe M. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2001.
- [4] 江伟,陈龙,王国胤. 用户行为异常检测在安全审计系统中的应用[J]. 计算机应用,2006,26(7):1637-1639.
- [5] 张锋军,牟其林,江泓. 用行为分析技术来增强网络管理的能力[J]. 信息安全与通信保密,2009(6):72-77.
- [6] 刘璇,张凤荔,叶李. 基于 Netflow 的用户行为挖掘算法设计[J]. 计算机应用研究,2009,26(2):713-715.
- [7] 李崇东,肖晓强,李达,等. 基于参数测量的园区网可靠性分析系统的实现[J]. 计算机技术与发展,2010,20(10):21-25.
- [8] Tao Qin, Xiao Hongguan, Yi Long. Users' Behavior Character Analysis and Classification Approaches in Enterprise Networks [C]//2009 Eighth IEEE/ACIS International Conference on Computer and Information Science. [s.l.]: [s.n.], 2009: 323-328.
- [9] 董富强. 网络用户行为分析及其应用[D]. 西安:西安电子科技大学,2005.
- [10] 杨铮. 基于流量识别的网络用户行为分析[D]. 重庆:重庆大学,2009.
- [11] 曹文平. 基于聚类的孤立点集探测算法[J]. 现代计算机,2008,297:35-37.
- [12] 冯兴杰,黄亚楼. 增量式 CURE 聚类算法研究[J]. 小型微型计算机系统,2004,25(10):1847-1849.
- [13] Li Wen, Ping Lingdi, Lu Kuijun, et al. Trust Model of Users' Behavior in Trustworthy Internet [C]//2009 WASE International Conference on Information Engineering. [s.l.]: [s.n.], 2009.
- [14] Wu Hanching, Huang S S. User Behavior Analysis in Masquerade Detection Using Principal Component Analysis [C]//8th International Conference on Intelligent Systems Design and Applications. [s.l.]: [s.n.], 2009.

(上接第 34 页)

其中  $i = 1, 2, \dots, m$

从数学意义上说,该优化算法是一个等式约束最优化算法的问题<sup>[12]</sup>,通过该优化算法可以大大提高仿真系统的效率。大量静态与动态物体发生碰撞,碰撞的效率是必须考虑的问题。同时也可以利用已知物体运动参数,如速度和加速度,来预知物体未来碰撞的位置,提高仿真系统的实时性。

## 5 结束语

文中在 OSG (OpenSceneGraph) 场景管理软件中,对碰撞检测进行了初步的研究,通过 VS2008 编程,实现了多个静态物体(树)与动态物体(坦克)的碰撞检测,提出了大量静态物体与动态物体碰撞优化算法的数学模型,满足了实时性要求,取得了不错的效果。

#### 参考文献:

- [1] 石教英. 虚拟现实基础以实用算法[M]. 北京:科学出版社,2002.
- [2] OpenSceneGraph [EB/OL]. 2008. <http://www.openscenegraph.org>.
- [3] Martz P. OpenSceneGraph [EB/OL]. 2007. [http://www.osg-books.com/books/osg\\_qs.html](http://www.osg-books.com/books/osg_qs.html).
- [4] The MultiGen Creator Desktop Tutor [M]. [s.l.]: MultiGen-Paradigm, Inc, 2003.
- [5] 王志强. 碰撞检测问题研究综述[J]. 软件学报,1999,10(5):545-551.
- [6] 高军峰,徐凯声,崔劲,等. 一个基于包围盒技术提高光线与物体求交效率的算法[J]. 交通与计算机,2004,22(6):65-68.
- [7] 喻家龙,姜太平,汪光阳. 在 GPU 上基于物体空间的碰撞检测[J]. 计算机技术与发展,2009,19(9):83-86.
- [8] 周之平,吴介一,白伟冬,等. 基于矩形包围盒的多边形碰撞检测算法[J]. 中国图象图形学报,2004,9(11):1294-1303.
- [9] 朱元峰,孟军,谢光华,等. 基于复合层次包围盒的实时碰撞检测研究[J]. 系统仿真学报,2008,20(2):372-377.
- [10] Hubbard P M. Approximating Polyhedra with Spheres for time-critical collision detection[J]. ACM Transactions on Graphics, 1996,15(3):179-210.
- [11] 马登武,叶文,李瑛,等. 基于包围盒的碰撞检测算法综述[J]. 系统仿真学报,2006,18(4):1059-1064.
- [12] 唐焕文,秦学志. 实用最优化方法[M]. 大连:大连理工大学出版社,2004.