

基于相异度分析的网络通信异常行为识别方法

荆 涛^{1,2}, 李 俊¹

(1 中国科学院 计算机网络信息中心, 北京 100190; 2 中国科学院大学, 北京 100049)

摘 要: 提出了一种基于相异度分析的网络通信异常行为识别方法,以 Netflow 网管数据为基础,设计了具体实现步骤,实现了对网络中异常网络攻击、突发流量异常以及网络通信行为规律显著变化等表征现象实时监测,取得良好效果,大幅提升了网络运营单位的网络安全支撑服务能力。

关键词: 相异度;异常行为;网络安全

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2014)06-0012-04

An Identification Method Based on Analysis of Dissimilarity Degree for Network Communication Behavior

JING Tao^{1,2}, LI Jun¹

(1 Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China;

2 University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: This paper presents an analysis meathod based on dissimilarity abnormal behavior for network communication identification . With Netflow network management data, we designed concrete steps for implementation, then we can achieve significantly changes on real-time monitoring for realization of abnormal network attacks, emergency traffic anomaly, and phenomenon of network communication behavior rules. This meathod show significantly implement results,and increase network security support service capabilities of ISP.

Key words: dissimilarity degree; abnormal behavior; network security

1 引言

目前,网络安全技术已经得到了极大的发展,产生了诸如防火墙、入侵检测系统、入侵防御系统等安全措施,但是网络攻击与防范是共同向前发展的.如何在网络安全事件发生之前及时、快速地发现网络异常行为,成为一个很重要的研究课题.

同时,对于高速互连网络中,通信对象的行为分析已经成为当前网络行为分析的热点,而如何针对网络通信行为,对通信主体自身进行推测分析,具有重要意义.随着网络规模的不断扩大,网络数据已经越来越多,网络协议和服务日趋复杂,网络通信连接

往往不再是端到端的简单通信,随着网站服务和 P2P 等业务的广泛开展,网络中多连接网络通信越来越多,对于数据的处理逐渐有超过人工处理能力的趋势^[1].如何处理高速互连网络环境下的海量数据,并合理推测分析通信对象的行为,正是本文探讨的重点.

2 算法分析

2.1 基于地址的网络通信行为的时空分析

2.1.1 地址相异度分析方法

NetFlow 本身主要是一套网络流量统计协议,路由器通过源 IP 地址、源端口号、目的 IP 地址、目的端口号、协议类型、服务种类以及路由器输入界面

收稿日期: 2013-08-18; 修回日期: 2013-10-09

基金项目: 中国科学院“十二五”信息化专项——中国科学院网络安全保障与服务工程

这些控制信息来区分 NetFlow 流. 任何时间当路由器接受到新的数据包时, 路由器将检查数据流中的这些控制信息来判断这个数据包是否属于任何已经记录的 NetFlow 流, 如果有的话则将新收集的数据包相关信息整合到相应的 NetFlow 流记录中^[2-3], 如果没有则产生一条新的记录.

我们对网络中通信的两个地址进行地址相异度分析, 可以按照 IPv4 地址格式, 将其进行划分, 并写成四元组的形式, 按照各个属性值的差异进行比对两个地址间的实际差异, 由于 IP 地址是点分十进制格式, 而每个隔点的差异, 代表着地址类别的不同. 而高位数值的差异, 代表着网络范围的显著差异, 因此, 可以把地址相异度进行抽象描述, 并进行量化分析^[4-5].

在我们的研究中, 地址相异度的定义是关键问题, 而一个 IP 地址在规定的单位时间内, 与相异度较大的地址通信连接的数量和度量数值, 一定程度上, 可以表征该地址的通信活动特征, 由此, 在分析通信地址对的相异度的同时, 也可以考量地址在空间维度上的离散程度.

可以做如下假设: (1) 给定的 IP 地址与其通信的目的地址对数量不为小数; (2) 选定的单位时间片可适合反映给定地址的相异度特性, 但时间片长度不为大时间长度.

根据以上假设, 可以得到适合条件下的网络通信特征数据.

地址相异度分析: IPv4 地址的构成可以记为: a, b, c, d , 可将地址写作四元组 $\{a, b, c, d\}$, 对于每一个 IP 地址 A 来说, 都有 4 个数值属性 v_a, v_b, v_c, v_d , 那么, 一个地址 A 的属性集合可记为 $m_i \in \{v_a, v_b, v_c, v_d\}, 1 \leq i \leq 4$.

鉴于 IP 地址的特殊性, 每一个相邻分类属性间的区别, 将决定地址类别差别的划分, 对于给定的 IP 地址 x 和 y , 在属性 m_i 上的差异程度(或距离)目前有几种类似的计算方法, 如: 欧几里得距离、曼哈顿距离和闵可夫斯基距离, 由于闵可夫斯基距离是欧几里得距离和曼哈顿距离计算模型的推广, 针对 IP 地址的数据格式情况, 我们考虑采用闵可夫斯基距离来进行数学描述. 即对于两个地址对象 x 和 y , 其地址间的距离(差异)可表示为

$$\text{dif}(x, y) = (\left| x_1 - y_1 \right|^p + \left| x_2 - y_2 \right|^p + \left| x_3 - y_3 \right|^p + \left| x_4 - y_4 \right|^p)^{\frac{1}{p}},$$

式中, p 为闵可夫斯基距离的范数, 为正整数. 当 $p = 1$ 时, 表示为曼哈顿距离, 当 $p = 2$ 时, 则表示为欧几里得距离.

但显而易见, 地址的数值属性 m_i 的不同, 对真实地址间的差异程度显然是不同的, m_1 的差异显然要比 m_2, m_3, m_4 的差异影响要大, 依次类推, 对应每一个 m_i 的影响, 都应该赋予一个权重, 因此我们提出一个改进的闵可夫斯基距离的数学描述, 即地址距离可以描述为

$$\text{dif}(x, y) = (\omega_1 \left| x_1 - y_1 \right|^p + \omega_2 \left| x_2 - y_2 \right|^p + \omega_3 \left| x_3 - y_3 \right|^p + \omega_4 \left| x_4 - y_4 \right|^p)^{\frac{1}{p}},$$

对每一个变量根据其重要性赋予一个权重, 由于 IP 地址的格式结构和地址分配定义, 我们可以将权重 ω_i 表示为 $(2^8)^{4-i}$, 其中 $i \in N, 1 \leq i \leq 4$ 鉴于 ω_i 数值可能很大, 我们引进一个参数 k , 来控制结果的数值规模, k 为正整数, 那么上述公式我们可以写为 $\text{dif}(x, y) = ((2^8)^3 \left| x_1 - y_1 \right|^p + (2^8)^2 \left| x_2 - y_2 \right|^p + (2^8)^1 \left| x_3 - y_3 \right|^p + (2^8)^0 \left| x_4 - y_4 \right|^p)^{\frac{1}{p \cdot k}}$, 即:

$$\text{dif}(x, y) = \left(\sum_{i=1}^4 (2^8)^{4-i} \left| x_i - y_i \right|^p \right)^{\frac{1}{p \cdot k}} \quad (1)$$

对于现有的 Netflow 数据集所包含的信息而言, 能够比较充分地涵盖当前网络节点及节点关系间的关键属性数据, 通过构造的网络异常扫描行为分析模型, 来发现新的网络行为和异常, 也都是可行的.

2.1.2 基于地址的平均相异度分析

地址相异度的提出, 为进行下一步质心分析提供了理论手段.

假设一个网络节点 a , 在某一时间片 Δt 内, 与 m 个目的地址有网络通信行为, 目的地址的集合可以记做 B , 通过计算式(1), 我们可以得到节点 a 与集合 B 的 m 个元素间的地址相异度 $\text{dif}(a, b_i), i \in N, i \leq m$, 那么, 我们可以把 $\text{dif}(a, b_i)$ 重新定义为一个新的集合 C , 即 $C = \{\text{dif}(a, b_1), \text{dif}(a, b_2), \text{dif}(a, b_3), \dots, \text{dif}(a, b_k)\}$, 那么, 我们可以计算节点 a 到所有 m 个目的地址的平均相异度, 记做

$$\text{dif}(a, C)_{\text{avg}} =$$

$$\frac{\text{dif}(a, b_1) + \text{dif}(a, b_2) + \text{dif}(a, b_3) + \cdots + \text{dif}(a, b_k)}{m}$$

$$= \frac{\sum_{i=1}^k \text{dif}(a, b_i)}{m} \quad (2)$$

如果我们以节点 a 作为原点, 那么它到其所有的目的地址的平均相异度为 $\text{dif}(a, C)_{\text{avg}}$.

2.2 基于地址的网络通信行为的时间分析

在某一时间片 Δt 内的基于地址的相异度已经进行了分析, 但当考虑连续时间情况下, 我们需要对上述模型进行时间维度的扩展:

对于一个连续的时间序列 $T = \{t_1, t_2, t_3, \cdots, t_i, \cdots, t_n\}$, 每一个时刻 t_i 与前一个时刻 t_{i-1} 之间的时间间隔 Δt_i 内, 节点 a 与 n_i 个目的地址有通信行为, 那么根据(1)和(2)可以计算出节点 a 在此时间间隔内的目的地址平均相异度 $\text{dif}(a, C_i)_{\text{avg}}$. 在连续时间序列 T 中的每个时间间隔, 都可得出平均相异度 $\text{dif}(a, C_i)_{\text{avg}}$. 那么整个时间序列 T 下, 可通过计算每个时间间隔的平均相异度计算整个 T 的平均相异度, 那么对相异度进行加权平均有

$$\text{dif}(a, C)_{T\text{-avg}} = \frac{\text{dif}(a, C_1)_{\text{avg}} + \text{dif}(a, C_2)_{\text{avg}} + \text{dif}(a, C_3)_{\text{avg}} + \cdots + \text{dif}(a, C_n)_{\text{avg}}}{n}$$

$$= \frac{\sum_{i=1}^k \text{dif}(a, C_i)_{\text{avg}}}{n} \quad (3)$$

为了计算相异度之间差异的表征, 我们引入偏移方差 σ 的概念. 表示在某一时间间隔 Δt_i 内, 节点 a 与该间隔内所有目的地址的平均相异度 $\text{dif}(a, C_i)_{\text{avg}}$ 与整个时间序列 T 下的平均相异度 $\text{dif}(a, C)_{T\text{-avg}}$ 间的差距, 即偏移方差 σ_i . 按照欧几里得距离并加权平均, 其数学表达写作

$$\sigma = \sqrt{\frac{\sum_{i=1}^n [\text{dif}(a, C_i)_{\text{avg}} - \text{dif}(a, C)_{T\text{-avg}}]^2}{n}}$$

$$(4)$$

引入偏移方差 σ 的意义在于, 我们可以约定在地址差异度区间 $[\text{dif}(a, C)_{T\text{-avg}} - \sigma, \text{dif}(a, C)_{T\text{-avg}} + \sigma]$ 内, 当节点 a 在某一规定时间内的平均地址相异度 $\text{dif}(a, C)_{\text{avg}}$ 的值落在该区间中时, 可以认为 a 节点当前的地址通信行为在正常范围, 当平均地址相异度 $\text{dif}(a, C)_{\text{avg}}$ 落在该区间以外时, 可以认为 a 节

点当前的地址通信行为不在正常范围, 网络通信存在异常^[6].

3 框架设计

我们可以采用 Netflow 数据作为训练集的数据来源, 构建合理的训练集, 可以采用协议生成法来构造. 协议生成法: 利用程序模拟各种协议的通讯模式, 通过流量记录的方式, 将制定行为的网络流量特征保存下来, 这样就会构造已知的不同类别网络行为的训练集合. 这种训练集可以一定程度上降低误报率. 用协议生成法模拟网络通讯后, 将路由器生成的有关 netflow 数据收集起来, 就完成了训练集的构造, 具体步骤如下:

(1) 提取训练集数据中的地址信息和时间信息, 进行预处理.

(2) 将有关数据信息导入数据库, 并根据有关算法计算相关导出属性.

(3) 选取一个较为合理的时间长度, 由于 netflow 通常为每 5 分钟一次采样, 因此时间上考虑以 5 分钟为间隔长度, 计算各个源 IP 地址对应的各类数学参数的值, 带入分析模型得出正常相异度区间.

(4) 计算全部时间长度的数学参数, 计算偏移方差的值, 考虑时间长度不少于 12 小时.

(5) 建立各个源 IP 地址的基本对照集, 并供验证使用.

(6) 选取某一时间间隔的网络数据, 分别代入模型, 验证结果.

4 实验分析

我们选取某时间区间一台网站服务器的 Netflow 数据做为实验数据, 收集该时间区间内的主机 Netflow 数据在 1 小时内的访问信息可了解, 在此期间的主机连接次数和 IP 地址通信对的 IP 地址构成信息, 以各个地址通信期间的时间信息. 可以采用本文中的基于地址和时间相异度分析方法进行该网站访问情况的比较分析. 对比每次通信连接的 IP 地址对的对应关系进行地址度. 为了控制计算数据规模, 我们设定 $p=2, k=6$. 而同时, 可以根据每个时间节点近 10 天内的地址差异度的平均对照值, 如表 1 所示.

表 1 近 10 天某主机 9 点至 10 点间的地址差异距离对照值

时间	9:05	9:10	9:15	9:20	9:25	9:30
距离值	7.437	7.441	7.439	7.461	7.458	7.533
时间	9:35	9:40	9:45	9:50	9:55	10:00
距离值	7.512	7.435	7.356	7.315	7.432	7.398

对于某日上午 9 点至 10 点间某网络服务器 A, 在每 5 分钟的时间间隔内, 主机 A 的地址相异距离平均值 $\text{dif}(a, C)_{\text{avg}}$ 如表 2 所示。

表 2 某主机 9 点至 10 点间的地址差异距离

时间	9:05	9:10	9:15	9:20	9:25	9:30
距离值	7.438	7.552	7.467	7.455	7.434	8.557
时间	9:35	9:40	9:45	9:50	9:55	10:00
距离值	9.331	10.714	9.858	7.423	7.516	7.476

我们选取前 5 个时间节点作为实验样例, 经过计算可知, $\text{dif}(a, C)_{T\text{-avg}} = 7.469$, 而由此得出的 $\sigma = 0.019$, 因此得到的该时间间隔内的地址差异度区间为 $[7.45, 7.488]$ 之间。

根据以上数据, 可将观测数据与近期的对照数据在各时间点间的数值变化以图 1 形式展现, 我们可以明显看到时间在 9:25 至 9:45 间有关距离数值明显超过差异度可信区间, 并且显著区别于对照曲线, 由此我们可怀疑该时间点内的主机流量存在异常。后经初步分析判断, 在此期间该主机遭受分布式拒绝服务攻击, 由此可判断该主机在该时间段内存在网络通信行为异常。

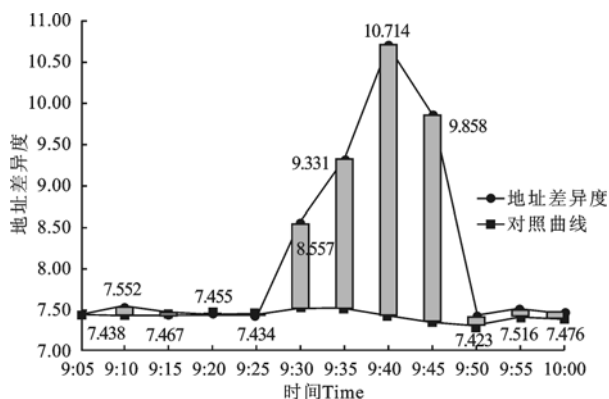


图 1 某主机在某时刻内地址差异分析图

我们利用该分析方法, 共分析测试的 30 个主机

样本中显示流量异常的有 21 个, 经过后期实际分析验证, 有 19 个有通信行为异常, 另外两个误报是由于新增其他网络应用服务后, 造成的网络通信访问行为的差异明显异常, 但观测 10 天后, 数据趋于平稳。因此该算法在样本测试中的准确率在 90.5%, 也验证说明了该算法的有效性总体良好。

5 结束语

本文将基于相异度分析方法应用到网络通信异常行为的识别中, 提出了一种针对海量网管数据对异常网络行为的筛查和甄别的有效方法, 提升了互联网运营单位的服务质量和支撑保障能力, 但该方法目前仍存在计算复杂度相对较高, 且对有关控制参数的考量、引入访问端口差异因素及模型优化方面还需进行深入的研究, 仍需在未来研究工作中进行完善和充实。

参考文献:

- [1] 蒋卓明, 周旭, 许榕生. P2P 网络行为安全监控研究综述[J]. 计算机科学, 2009, 36(6): 1-4.
- [2] 牛国林, 管晓宏, 龙毅, 等. 多源流量特征分析方法及其在异常检测中的应用[J]. 解放军理工大学学报: 自然科学版, 2009, 10(4): 350-355.
- [3] 邓冠男. 聚类分析中的相似度研究[J]. 东北电力大学学报, 2013, 33(1/2): 156-161.
- [4] 肖宇, 许炜, 张晨, 等. 社交网络中用户区域影响力评估算法研究[J]. 微电子学与计算机, 2012, 7(7): 58-63.
- [5] 郭小芳, 李锋, 刘庆华. 一种有效的多元时间序列相似性度量算法分析[J]. 江西师范大学学报: 自然科学版, 2013, 37(1): 56-59.
- [6] 谢信喜, 王士同. 适用于区间数据的基于相互距离的相似性传播聚类[J]. 计算机应用, 2008, 28(6): 1441-1443.

作者简介:

荆涛 男, (1979-), 博士研究生, 工程师. 研究方向为网络安全、入侵检测。

李俊 男, (1968-), 博士生导师, 研究员. 研究方向为网络安全、下一代互联网。