文章编号:1001 - 893X(2011)03 - 0089 - 04

基于相对熵理论的网络 DoS 攻击检测算法*

李涵秋[1],马 艳[2],雷 磊[2]

(1.解放军 63778 部队,黑龙江 佳木斯 154002;2. 解放军 63780 部队,海南 三亚 572427)

摘 要:针对日益严重的 DoS(拒绝服务)网络攻击行为,提出了一种 RED(相对熵检测)算法。该算法基于相对熵理论,利用网络流量的自相似特性,通过判断相邻时刻流量之间的相对熵值是否发生突变来进行 DoS 攻击检测。实验结果表明,与传统的信息熵 DoS 攻击检测算法相比,该算法具有较高的检测率。

关键词:网络信息安全;拒绝服务攻击;相对熵;信息熵;检测算法

中图分类号:TN918;TP393.08 **文献标识码**:A doi:10.3969/j.issn.1001 - 893x.2011.03.020

DoS Attack Detection Based on Relative Entropy Theory

LI Han-qiu[1], *MA Yan*[2], *LEI Lei*[2]

(1.Unit 63778 of PLA, Jiamusi 154002, China; 2.Unit 63780 of PLA, Sanya 572427, China)

Abstract: Based on the theory of relative entropy and the self - similarity of network traffic, a Relative Entropy Detection(RED) algorithm is proposed to detect the increasingly serious DoS(Denial of Service) attacks, according to judging the changes of relative entropy values at adjoining times. The experimental results show that the RED algorithm has a higher detection precision compared with information entropy detection algorithm.

Key words: network information security; DoS attack; relative entropy; information entropy; detection algorithm

1 引 言

随着近年来 Internet 的迅速普及,网络已深入到社会生活的方方面面,在经济、文化、教育、军事等各行业和部门中发挥着重大作用。然而,随之而来的网络安全问题也日益受到了人们的关注。在众多的网络攻击手段中,拒绝服务(Denial of Service, DoS)攻击是一种极具破坏力的攻击方式,它消耗可用系统、带宽资源迫使 Web 站点停止服务[1]。在现代信息战中,利用 DoS 攻击手段可以使重要的政府机构甚至军事站点瘫痪。

网络上所承载的信息是人们关心的重点,信息熵作为信息论中所提出的核心概念之一,学术界已经开始关注将其应用于网络攻击检测,并取得了一定的进展。Lakhina 等人[2]在信息熵理论基础上对网络数据包的端口和地址的分布特性进行分析,检测异常的网络行为。Rahmani 等人[3]利用流的数量与流中包含的数据包的数量之间的分布关系来检测 DoS 攻击,但时间间隔的选取对检测效果有较大影响。信息熵作为系统有序化程度的一个度量,能检测出 DoS 攻击对于正常网络流量随机性所产生的影响,然而网络流量的变迁是一个动态连续的过程,只取其中的一点而不去分析其与相邻时间序列之间的关系,对于未能引起流量分布显著变化的 DoS 攻击也很难检测。本文利用了信息论中另一个重要的理论——相对熵理论,通过分析 DoS 攻击发生时所造成的网络流量中 IP、端口等属性分布特性的变化规律及其对相邻时刻网络流量序列之间的相对熵值的影响,提出了一种基于相对熵的网络 DoS 攻击检测(Relative Entropy Detection,RED)算法。

* 收稿日期:2010 - 12 - 14;修回日期:2011 - 02 - 22

2 RED 算法检测原理

2.1 相对熵理论

相对熵是指两个随机序列之间距离的度量,从统计学角度出发它是指两个随机序列之间的相似程度[4]。相对熵可定义为

$$D(P \| Q) = \sum_{i=1}^{n} p_i \mathrm{lb} \frac{p_i}{q_i} \quad (1)$$

式中,$P = \{p_1, p_2, \cdots, p_n\}$,$Q = \{q_1, q_2, \cdots, q_n\}$。

相对熵的重要性质有 $D(P \| Q) = 0$ 当且仅当 $P = Q$,即 P、Q 两个分布序列完全相同时它们之间的相对熵值为 0。$D(P \| Q)$的值越小表示分布序列 P 和 Q 越相似,反之则表示其相差越大。

基于相对熵的 RED 检测算法是建立当前时段 T_{cur}的分布序列作为 P,上一个时段 T_{pre}的分布序列作为 Q,通过计算 T_{cur}和 T_{pre}时段数据之间的相对熵用于 DoS 的攻击检测。

2.2 检测原理

近年来,对网络业务流量的实际测量和统计分析结果表明,互联网业务流量具有突发性(Bursty)、长相关、自相似性等主要特性。流量在宏观上的流量变化较大,但在微观层面,单位流量具有较为稳定的分布结构[5]。DoS 攻击往往破坏网络流量稳定的分布结构,而采用相对熵就能很好地检测出这种变化。

从 DoS 的攻击手段方面进行分析:正常网络流量的目的 IP 的分布较均匀,当发生攻击时,被攻击 IP 的流量会显著增加,导致这个 IP 在统计中出现较大比例。同样,当 DoS 对某一特定的服务漏洞进行攻击时,也会引起目的端口属性、协议属性和包大小属性的分布发生较大变化。

从 DoS 的攻击效果方面进行分析:某一服务器向外提供服务所发送的数据包的源地址都为服务器地址,当 DoS 攻击造成服务器停止服务的后果时,会对网络流量中的源地址属性和源端口属性的分布造成较大影响。DoS 攻击所瘫痪的服务主机越重要,这种分布变化就会越明显。

无论是从 DoS 的攻击手段还是从攻击效果上来分析,这些都可以通过相对熵值的跳变来检测攻击。

3 RED 算法描述

3.1 RED 算法步骤

基于相对熵理论的 RED 算法是一个以时间为循环变量的步进计算过程,每次循环收集 T_{cur}时段的分布序列,并需要上一次循环中所收集的分布序列作为 T_{pre}序列,计算 T_{pre}和 T_{cur}两个序列之间的相对熵值作为 T_{cur}时段的处理结果,最后将 T_{cur}时段的分布序列作为下一次循环的 T_{pre}序列进入下一次循环。以计算源 IP 地址属性的相对熵值为例,算法步骤描述如下:

步骤 1:参数初始化。定义 *step* 为时段区间的长度,T_{cur}时段出现的 IP 存入 ipSourceArrayCur 结构中,计算这些 IP 出现的概率结果存放在 ipSourceProbaArrayCur 结构中。

步骤 2:取一条数据记录,若所有数据记录已取完则跳转到步骤 10。判断该条数据记录是否在当前 T_{cur}时段内,如在则跳转到步骤 3,如不在则跳转到步骤 4。

步骤 3:从该条数据记录中提取出 IP 源数据,存入 ipSourceArrayCur 结构中并更新该 IP 源地址对应的计数变量。跳转到步骤 2。

步骤 4:表示 T_{cur}时段内的所有数据已经统计完毕。通过 ipSourceArrayCur 结构分别计算其中每个源 IP 地址出现的概率并将计算结果存入 ipSourceProbaArrayCur 结构中。

步骤 5:取 T_{pre}时段即上一次循环中所产生的 ipSourceArrayPre 结构和 ipSourceProbaArrayPre 结构。

步骤 6:将 T_{pre}和 T_{cur}时段是数据合并。由 ipSourceArrayPre 和 ipSourceArrayCur 构成 mergeSourceArray 结构,由 ipSourceProbaArrayPre 和 ipSourceProbaArrayCur 构成 mergeSourceProbaArray 结构。

步骤 7:通过 mergeSourceProbaArray 结构计算 T_{cur}时段的相对熵值,并将结果输出。

步骤 8:将变量 T_{cur}、ipSourceArrayCur 结构、ipSourceProbaArrayCur 结构分别赋给变量 T_{pre}、ipSourceArrayPre 结构和 ipSourceProbaArrayPre 结构。

步骤 9:将 T_{cur}以 step 步进,ipSourceArrayCur 和 ipSourceProbaArrayCur 结构清零。跳转到步骤 2。

步骤 10:算法结束。

该 RED 算法虽然计算相对熵时需要 T_{pre}和 T_{cur}两个时段的数据,但并不需要每次计算过程都分别运算 T_{pre}和 T_{cur}的相关数据,每次循环只需要填充 T_{cur}的数据即可。当前循环的 T_{cur}数据作为下一循环的 T_{pre}数据使用,很有效地避免了数据的重复计算。

3.2 RED 算法说明

RED 算法中 T_{cur}时段的数据相当于 P 分布序

列,T_{pre}时段的数据相当于 Q 分布序列。将 T_{cur}时段和 T_{pre}时段的数据进行合并时就会出现两种情况:对于在 T_{cur}时段出现而 T_{pre}时段没有出现的元素其对应的 $pi \neq 0, q_i = 0$,此时定义 $p_i \ln \frac{p_i}{q_i} = p_i$,对于随着时间的推移而消失的 IP 等元素,通过下降的幅度来影响相对熵值;对于在 T_{pre}时段出现而 T_{cur}时段没有出现的元素定义其对应的 $p_j = 0$,$q_i \neq 0$,此时定义 $p_i \ln \frac{p_i}{q_i} = q_i e$,对于本时段突然出现的 IP 等元素赋予较高的权值,通过出现的频率与权值的乘积来影响相对熵值。

4　实验与讨论

实验使用了 MIT 实验室在美国国防先进技术研究计划署(DARPA)和美国空军研究所共同资助下所公布的 DARPA99 数据集,该数据集已经作为研究领域所认可并广泛使用的评测标准。

为了分析 RED 算法与信息熵检测算法相对比所具有的优势,选取了含有较多种 DoS 攻击的 Apr-5-outside.dump 数据包计算 IP 地址属性结果,如图 1 所示。

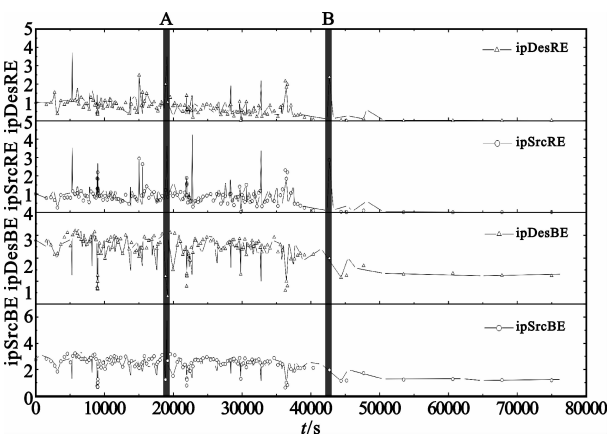

图 1　IP 属性的信息熵值与相对熵值对比图

Fig.1　Comparison between relative entropy and information entropy value of IP property

图 1 中分别使用了 RED 算法和信息熵算法对目的 IP 地址和源 IP 地址属性进行了计算。ipSrcBE 和 ipDesBE 分别表示源 IP 地址和目的 IP 地址的信息熵值,ipSrcRE 和 ipDesRE 分别表示源 IP 地址和目的 IP 地址的相对熵值;标记 A 表示在19 092 s爆发,持续了120 s的 Smurf 攻击;标记 B 表示在43 227 s爆发,持续了900 s的 Udpstorm 攻击。通过这两个攻击实例具体分析 RED 算法与信息熵检测算法的共性与区别。

信息熵算法和 RED 算法都可以检测 Smurf 攻击。Smurf 攻击爆发时大量主机同时向受害主机发送 ICMP 报文。信息熵算法中 ipSrcBE 值明显增大是因为该时段内的源地址数目急剧增多分布发散使系统处于无序状态。ipDesBE 值明显降低是因为该时段目的 IP 地址数目相对集中。RED 算法中 ipSrcCE 值明显增大是因为与前一时段相比出现了大量新的源 IP 地址且各地址的概率分布都发生了较大变化。ipDesCE 值也明显增大是因为与前一时段相比,原有的大量目的 IP 地址消失,同时受害主机的 IP 地址出现概率明显增加,从而也使网络流量的分布结构发生较大变化。

信息熵算法无法检测 UdpStorm 攻击而 RED 算法可以。UdpStorm 攻击是通过伪造的 UDP 数据包使多个受害主机之间不停的相互发送数据从而造成网络的拥挤消耗有效带宽。该时段内各主机之间在进行通信,源 IP 地址和目的 IP 地址都未出现明显的集中或发散的现象,信息熵值不会产生较大的变化,从图 1 中标记 B 处的 ipSrcBE 和 ipDesBE 值未发生变化也验证了这一点。RED 算法可以检测 UdpStorm 攻击是因为虽然 IP 地址并未发生较多变化而每个 IP 地址的出现概率都发生较大的变化使相对熵值明显增大。选取攻击前一时刻与攻击发生时刻源 IP 地址的相对熵值进行对比,结果如表 1 所示。

表 1　相对熵值与信息熵值对比表

Table 1 Comparison between relative entropy and information entropy

时间/s	T_{pre}时段源 IP 数量	T_{cur}时段源 IP 数量	共同 IP 数量	信息熵	相对熵
43 200	25	27	22	2.119 06	0.095 75
43 230	27	28	23	1.960 95	2.857 04

从表 1 中可以看出,在攻击发生时刻与攻击还未发生的前一时刻相比,IP 数量没有明显变化,并且也没有发生涌现出较多新 IP 地址的情况,所以信息熵值比较稳定,无法检测出 UdpStorm 攻击。UdpStorm 攻击造成网络的拥挤和减慢使各主机之间的通信无法正常进行,导致各地址的概率分布均发生了较大变化,从而导致了相对熵值的明显增大。

分析 RED 算法与信息熵算法的本质区别在于信息熵只关心宏观上 IP 地址的分布情况而相对熵不仅考虑宏观的分布变化并且还考量其中各个元素

的变化情况。

为了进一步验证 RED 算法的有效性，将 DARPA99 数据集中所列出的 17 种 DoS 攻击组合为 170 次攻击数据，并随机插入到第一周和第三周的正常的网络流量数据中。RED 算法和信息熵算法的检测结果对比如表 2 所示。

表 2 RED 算法和信息熵算法检测结果对比

Table 2 Comparison between the results of RED and the information entropy detection

算法	攻击总数量	检测率/%	漏报率/%	误报率/%
RED 算法	170	92.9	7.1	4.7
信息熵算法	170	70.6	29.4	4.2

由表 2 可知，基于相对熵的 RED 算法相比信息熵检测算法对于 DoS 攻击具有更高的检测率，但 RED 算法将一些正常网络状态的转变误判为攻击导致了一定的误报率。

5 结 论

本文针对 DoS 网络攻击行为提出了一种基于相对熵理论的 RED 检测算法。相比传统信息熵检测法，RED 算法不仅考虑宏观的分布变化而且从微观角度考量其中各个元素的变化情况，实验结果表明攻击检测率有了较大提升。未来研究的方向是如何在 RED 算法基础上更好地区分正常突发性流量和攻击流量，从而有效降低误报率。

参考文献：

[1] Loukas G, Oke G. Protection Against Denial of Service Attacks: A Survey[J]. Computer Journal, 2010, 53(7): 1020 – 1037.

[2] Lakhina A, Crovella M, Diot C. Mining anomalies using traffic feature distributions[J]. Computer Communication Review, 2005, 35(4): 217 – 228.

[3] Rahmani H, Sahli N, Kammoun F. Joint entropy analysis model for DDoS attack detection[C]//Proceedings of the Fifth International Conference on Information Assurance and Security. Xi'an: IEEE, 2009: 267 – 271.

[4] Thomas M, Joy A. Elements of Information Theory[M]. New York: John Wiley & Sons Inc., 2006: 19 – 22.

[5] 石江涛，王永纲，戴雪龙，等. 自相似网络业务流量的研究与实现[J]. 通信学报，2005，26(6)：112 – 120.
SHI Jiang-tao, WANG Yong-gang, DAI Xue-long, et al. On study and implementation of self-similar network traffic[J]. Journal on Communications, 2005, 26(6): 112 – 120. (in Chinese)

作者简介：

李涵秋(1985 –)，女，河南信阳人，2007 年获学士学位，现为助理工程师，主要研究方向为通信系统、信息安全；

LI Han – qiu was born in Xinyang, Henan Province, in 1985. She received the B.S. degree in 2007. She is now an assistant engineer. Her research interests include communication system and information security.

Emai: hanhanxjtu@163.com

马 艳(1983 –)，女，陕西渭南人，2006 年获学士学位，现为助理工程师，主要研究方向为通信系统、信息安全；

MA Yan was born in Weinan, Shaanxi Province, in 1983. She received the B.S. degree in 2006. She is now an assistant engineer. Her research interests include communication system and information security.

雷 磊(1984 –)，男，陕西合阳人，2006 年获学士学位，现为助理工程师，主要研究方向为通信系统、信息安全。

LEI Lei was born in Heyang, Shaanxi Province, in 1984. He received the B.S. degree in 2006. He is now an assistant engineer. His research interests include communication system and information security.