ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.4, April 2007, pp.967–977
DOI: 10.1360/jos180967

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于 Web 用户浏览行为的统计异常检测*

谢 逸+, 余顺争

(中山大学 电子与通信工程系,广东 广州 510275)

Anomaly Detection Based on Web Users' Browsing Behaviors

XIE Yi+, YU Shun-Zheng

(Department of Electrical and Communication Engineering, Sun Yat-Sen University, Guangzhou 510275, China)

+ Corresponding author: Phn: +86-20-84113303, E-mail: xieyicn@163.com

Xie Y, Yu SZ. Anomaly detection based on Web users' browsing behaviors. *Journal of Software*, 2007, 18(4):967–977. http://www.jos.org.cn/1000-9825/18/967.htm

Abstract: This paper proposes an anomaly detection based on Web user access behavior for the defense of application layer Distributed Denial-of-Service (DDoS) attack. Based on the hyperlink characteristics of Web pages and the HTTP responding effect of different proxies in the Internet, this paper uses hidden semi-Markov model (HsMM) to describe the Web user browsing behavior observed at Web server, and employs likelihood of the observation sequence on user browsing behaviors fitting to the model as a measure of user's normality. A parameterized model and its recursive formulae are derived and an on-line anomaly detection approach is introduced. Some issues involved in practical implementations of the model and the anomaly detection approach are discussed. Finally, an experiment is conducted to validate the model and the algorithm, which is based on a set of data colleted from a heavy-loaded Web server and an emulated DDoS attack that launches HTTP flooding to the Web site. The experimental results show that the model is effective in measuring the user behaviors and in detecting the application layer DDoS attacks.

Key words: hidden semi-Markov model; large-scale Web site; browsing behavior; DdoS (distributed denial-of-service); anomaly detection

摘 要: 提出一种基于 Web 用户访问行为的异常检测方案,用于检测应用层上的分布式拒绝服务攻击,并以具有非稳态流特性的大型活动网站为例,进行应用研究.根据 Web 页面的超文本链接特征和网络中各级 Web 代理对用户请求的响应作用,用隐半马尔可夫模型来描述服务器端观测到的正常 Web 用户的访问行为,并用与大多数正常用户访问行为特征的偏离作为一个流的异常程度的测量.给出了模型的参数化方法,推导了模型参数估计与异常检测算法,讨论了实际网络环境下异常检测系统的实现方法.最后用实际数据验证了模型和检测算法的有效性.仿真结果表明,该模型和检测算法可以很好地描述 Web 用户的正常浏览行为,有效地检测应用层分布式拒绝服务攻击.

* Supported by the National Natural Science Foundation of China under Grant No.90304011 (国家自然科学基金); the Natural Science Foundation of Guangdong Province of China under Grant No.04009747 (广东省自然科学基金); the Research Fund for the Doctoral Program of Higher Education of China under Grant No.20040558043 (高等学校博士学科点专项科研基金)

Received 2005-11-05; Accepted 2006-04-03

关键词: 隐半马尔可夫模型;大型活动网站;浏览行为;分布式拒绝服务;异常检测
中图法分类号: TP393 文献标识码: A

基于洪水式(flooding)的分布式拒绝服务攻击(distributed denial-of-service,简称 DDoS)[1]是一种简单、有效的攻击手段.目前,大部分 DDoS 检测方案[1]主要是针对网络层和传输层的攻击,例如 ICMP flooding 和 SYN flooding 等,很少关注到基于应用层的 DDoS 攻击.应用层 DDoS 攻击是一种与高层服务相结合的攻击方法,有以下两种形式:(1) HTTP Flooding 型.攻击者向目标 Web 服务器发送大量 HTTP 请求,请求内容可以是正常页面、重定向页面(redirected page)、头信息(header information)或某些错误文档(error document).更复杂的可以是对动态内容、数据库查询的请求.攻击者甚至可以模拟搜索引擎,从一个给定的 HTTP 链接开始,以递归的方式顺着指定网站上所有的链接进行访问.2004 年的蠕虫病毒 MyDoom 就是一种基于 HTTP Flooding 的应用层 DDoS 攻击;(2) 主机资源耗尽型.攻击者用少量的 HTTP 请求促使服务器返回大文件(例如图像、视频文件等),或促使服务器运行一些复杂的脚本程序(例如复杂的数据处理、密码计算与验证等).这种方式无须很高的攻击速率就可以迅速耗尽主机的资源,而且更具有隐蔽性.由此可见,与传统的基于低层的 DDoS 攻击相比,应用层 DDoS 具有更加显著的攻击效果,而且更加难以检测.这主要是由于:(1) 从请求内容无法检测出攻击的异常性;(2) 现有的 DDoS 检测方案都是针对工作于 TCP(transmission control protocol)层及其以下层次的攻击方式而设计的,面对应用层的 DDoS 攻击,这些检测系统所看到的都是正常的 IP 分组、正常的 TCP 连接和正常的流特征.因此,攻击请求可以顺利穿越低层的检测与防御体系直接到达 Web 服务器.可以预计,未来基于应用层的 DDoS 攻击可以有更多种不同的形式,所以,研究应用层 DDoS 的检测方法是非常重要的.

大型活动网站是指那些用于报道大型活动的网站,例如将为 2008 年北京奥运会、2010 年上海世博会特设的网站,以及其他一些涉及重大商务/政治活动、大型文艺表演的专题网站.这些非常重要而又特殊的网站具有与一般网站不同的特点:(1) 大部分用户按照大型活动的时间表访问大型活动网站.例如在比赛进行期间,业务量非常巨大且显突发性、峰值流显非稳态变化特性;(2) 大型活动进行期间,由于大部分用户所关注的目标都非常近似(例如比赛结果),因此用户的访问行为相似、访问内容非常集中.这就导致某些被关注的页面被高频率访问,而其他页面则较少被访问.大型活动网站的这些特点使其更容易受到应用层 DDoS 攻击.若使用为一般网站建立的统计异常检测方法[1],则难以有效区分到达大型活动网站的具有突发性强、业务量大等特点的正常流和具有类似特点的 DDoS 攻击流,从而导致高误检率和高漏检率.因此,现有的异常检测方法对于大型活动网站不再适用.

为此,针对大型活动网站峰值期间可能存在的利用 HTTP 请求产生的 DDoS 攻击,本文从应用层的角度,提出了一种基于 Web 浏览行为的 DDoS 攻击检测方法.它通过 Web 用户浏览行为模型来检测用户访问的正常性,然后根据正常性的高低程度来对用户后续的 HTTP 请求进行排队,使正常用户可以优先得到 Web 服务器的响应,同时抑制异常用户对 Web 服务器资源的消耗,从而最大限度地降低大型活动网站被攻击的可能性.采用这种方法的原因在于,每个正常用户的浏览行为和目的性,除了受网络时延和服务器响应速度的影响之外,与访问该网站用户的多少并没有直接的关系.因此,单个用户的访问行为与总业务量相比,具有更好的稳态特性.另外,攻击程序由于受代码长度的限制,一般很难伪装正常用户的、隐含有智能活动的浏览行为来产生攻击流.因此,通过高层的用户行为分析可以有效检测利用 HTTP 请求对大型活动网站发起的 DDoS 攻击.

为了有效描述用户高层访问行为并实现异常检测,本文采用隐半马尔可夫模型(hidden semi-Markov model,简称 HsMM)[2,3]描述 Web 用户浏览行为的随机变化过程.隐马尔可夫模型(hidden Markov model,简称 HMM)已经在语音识别、手写体/文字识别、数字通信编解码、DNA 序列分类等许多重要领域获得了广泛和成功的应用[2].与 HMM 相比,HsMM 更适合于描述状态持续时间为任意分布的隐马尔可夫过程.我们将在前期研究[4]的基础上,进一步考虑用户的浏览时间,并提出一种参数化的 HsMM 来实现用户行为的描述与检测.

1 用户行为与模型

从高层协议看,用户的浏览过程是通过一系列 HTTP 请求/响应构成的.由于一个页面通常包含多个内嵌链接,例如图片、广告条、背景音乐和框架页面等,因此,用户的每一次浏览行为(例如点击页面链接、前进、后退、刷新等)都会触发浏览器发出一系列的 HTTP 请求(如图 1 所示).这些 HTTP 请求到达服务器后,其属性(源地址、请求时间、请求对象等)会被记录在服务器的日志文件中.因此,通过 log 文件可以分析出用户的浏览行为,也就是说, log 文件中的 HTTP 请求记录是反映用户浏览行为的“轨迹”.

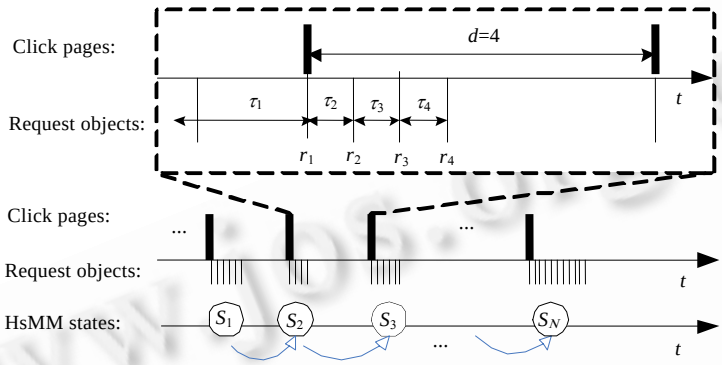

Fig.1 Users’ browsing behaviors and HsMM

图 1 用户浏览行为与 HsMM

但是,在实际网络环境下,由于以下两个原因,常导致 log 文件的记录不能精确反映出用户的浏览行为:首先,log 记录一般不能直接分清收到的 HTTP 请求是直接由用户点击所产生的,还是浏览器自动发出的对内嵌链接对象的请求.因此,从 log 记录无法精确知道用户的点击行为.服务器端精确获取客户点击行为的一种方法是使用脚本程序和 Cookies,但这种方法会给客户端带来一定的风险,不一定得到所有客户的支持.因此,用户的浏览行为是“隐藏”在 log 文件记录的 HTTP 请求序列中的;其次,网络中的各级代理(proxy)、用户浏览器的高速缓存会对用户的 HTTP 请求作出响应,因此,用户浏览行为触发浏览器发出的 HTTP 请求可能不会全部到达 Web 服务器(如图 2 所示),而且 Web 服务器也不知道有多少请求被高速缓存.因此,完全相同的用户浏览(或点击)行为,由于用户浏览器及中间各级代理缓存程度的不同,log 文件记录的用户浏览器发出的 HTTP 请求序列会存在差异.这种差异使服务器更难以了解用户的全部浏览行为.保证用户所有 HTTP 请求都能被 log 记录的一种方法是使服务器的每个响应都包含有禁止高速缓存,或要求已被高速缓存的响应重确认的标头(HTTP header).但是,这种“高速缓存补救”技术将以减少高速缓存的有效性为代价,从而增加了 Web 服务器的流量与负担,因此并不实用.所以,从 Web 服务器的 log 去研究用户的浏览行为,最合适的方法之一就是利用 HsMM 对“隐”状态的描述能力.

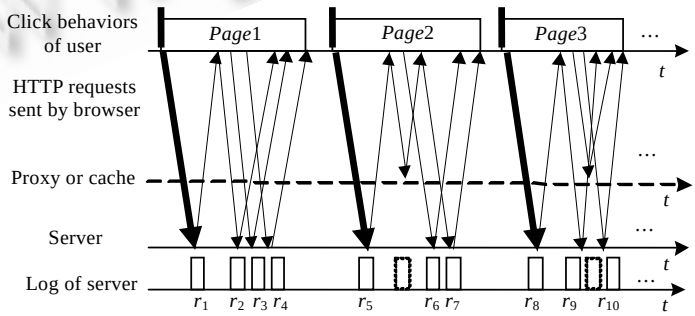

Fig.2 Responses of HTTP requests and HsMM

图 2 HTTP 请求的响应与 HsMM

由于大部分用户浏览行为的统计特征(点击速度、浏览的内容或请求对象、浏览时间、浏览过程等)具有一定的相似性,因此,我们可以把这种统计特征看作是用户正常的、合法的行为.相对于正常的普通用户来说,攻击者发出的攻击流通常具有以下特点:(1) 攻击者为达到攻击的目的,必须向目标服务器发送大量 HTTP 请求.因此,从服务器端来看,攻击者浏览行为的切换频率远大于普通正常用户,这是时间上的差异;(2) 攻击者可以模仿浏览器发送请求,但无法模拟正常用户的浏览行为.因此,攻击者通常只能随机生成或重复一些简单的 HTTP 请求(例如对主页的请求、对高频访问对象的请求)来形成攻击流,这就导致攻击流的请求序列与正常用户的请求序列不同,其中包含了内容和次序上的差异.利用这些统计特征上的差异,通过分析用户行为与代表正常访问行为的 HsMM 的偏离程度,可以从用户浏览行为的角度区分出攻击流.

基于上述考虑,本文选取以下两个观测量来描述用户的浏览行为:(1) 用户向服务器请求的对象序列;(2) 到达服务器的相邻 HTTP 请求的时间间隔.由于正常用户的浏览行为通常是从一个页面到另一个页面,因此,假定用户的浏览行为符合马尔可夫链的特性,并可以用一个浏览状态链来描述.一个状态代表用户的一次(或多次)点击行为.从浏览器端看,就是浏览器发出一组给定的 HTTP 请求.不同状态代表用户对不同链接内容的点击行为,或者浏览器发出的不同集合的 HTTP 请求.所有状态的集合表示为 $S=\{1,\ldots,M\}$.对于一种给定的状态,由于受各级 cache 的影响,到达服务器端的 HTTP 请求及其个数是不同的.这些 HTTP 请求可以看作是在给定状态下的观测值,它以一定的概率出现,所有 HTTP 请求的集合表示为 $V=\{1,\ldots,K\}$.来自于该用户的相邻 HTTP 请求之间的时间间隔是给定状态下的另一个独立随机变量,它可以看作是在给定状态下的另一个观测值.为了分析方便,并考虑到实际的 Web 服务器都是以秒为单位记录 HTTP 请求到达的时间,我们将时间间隔离散为整数秒,并将其集合表示为 $I=\{1,2,\ldots\}$;对于用户的某一类典型浏览行为,服务器能够收到的 HTTP 请求的个数是另外一个随机变量,它可以认为是在给定状态下输出的观测值的个数,其集合表示为$\{1,\ldots,D\}$.把用户发出的 HTTP 请求序列表示为 $O=\{(r_1,\tau_1),\ldots,(r_T,\tau_T)\}$,其中:$r_t\in V$ 表示用户向 Web 服务器请求的对象;$\tau_t\in I$ 表示服务器收到 HTTP 请求 r_t 与 r_{t-1} 之间的时间间隔,O 是模型的二维观测值序列.用 $\boldsymbol{B}=\{b_m(v,q)\}$表示模型的输出概率矩阵,$b_m(v,q)$表示对于给定状态 $m\in S$,Web 服务器收到请求 $r_t=v\in V$ 且与前一个到达请求之间的时间间隔为 $\tau_t=q\in I$ 的概率,且满足$\Sigma_{v,q}b_m(v,q)=1$.用 $\boldsymbol{P}=\{p_m(d)\}$表示在给定状态 m 下输出观测值个数为 $d\in\{1,\ldots,D\}$的概率,且满足 $\Sigma_d p_m(d)=1$,即 $\boldsymbol{P}$ 是 HsMM 模型中的状态停留时间概率矩阵.用 $\boldsymbol{\Pi}=\{\pi_m\}$表示初始状态的概率向量,$\pi_m$ 表示初始状态为 $m\in S$ 的概率.用 $\boldsymbol{A}=\{a_{mn}\}$表示状态转移概率矩阵,$a_{mn}$ 表示从状态 $m\in S$ 转移到 $n\in S$ 的概率.

系统的实现方法如图 3 所示.首先采集用户 HTTP 请求数据作为系统的观测序列,经过预处理后,形成训练序列对模型进行训练.在模型参数确定之后,该模型即可用于入侵检测.实测数据通过预处理后即成为所需的观测值,通过调用 HsMM 算法模块,计算得到平均对数或然概率.然后,在正常度判决模块中得到该用户行为的正常度值.如果该用户的正常度处于正常范围,则用户数据将被加入到训练数据集中用于在后台更新 HsMM 模型参数,并进入服务队列;否则,该用户将被认为是异常,并交给其他模块(异常处理模块)进行处理.

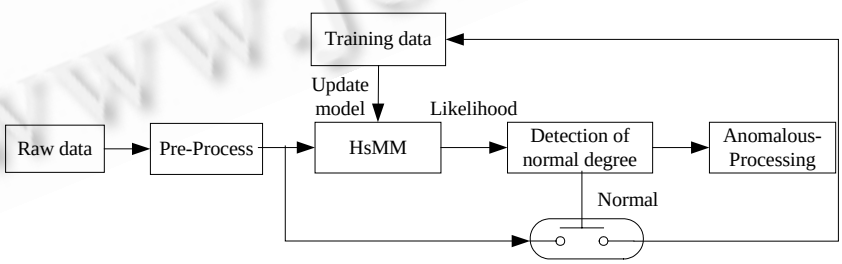

Fig.3 The framework of anomaly detection

图 3 异常检测系统框图

考虑到在二维观测值序列中 r_t 和 τ_t 相互独立,即

$$b_m(v,q)\equiv\Pr[r_t=v,\tau_t=q|s_t=m,\Omega]=b_m(v)b_m(q) \quad (1)$$

其中:$\Omega=\{\boldsymbol{\Pi},\boldsymbol{A},\boldsymbol{B},\boldsymbol{P}\}$为 HsMM 模型参数集;$v\in V,q\in I,m\in S$,且满足$\Sigma_v b_m(v)=1$ 和$\Sigma_q b_m(q)=1$.

由前向-后向(forward-backward)算法[3]及 World Cup 1998 的访问数据[5]的分析可以知道:$b_m(q)$在 log-log 图上近似一条直线,接近 Pareto 分布,因此,进一步假定 $b_m(q)$服从 Pareto 分布:

$$\sum_{q'=1}^{q-1} b_m(q') = P(X < q \mid \text{state } m) = 1 - q^{-\lambda_m},\ q=1,2,\ldots,\infty,\lambda_m>0,m\in S \tag{2}$$

或

$$b_m(q) = q^{-\lambda_m} - (q+1)^{-\lambda_m},\ q=1,2,\ldots,\infty,\lambda_m>0,m\in S \tag{3}$$

由此得到一个参数化的 HsMM,参数集记为$\Omega=\{\boldsymbol{\Pi},\boldsymbol{A},\boldsymbol{B},\boldsymbol{\lambda},\boldsymbol{P}\}$,其中:$\boldsymbol{\Pi}=\{\pi_i\};\boldsymbol{A}=\{a_{mn}\};\boldsymbol{B}=\{b_m(v)\};\boldsymbol{\lambda}=[\lambda_1,\ldots,\lambda_M];\boldsymbol{P}=\{p_m(d)\}$.令 o_t 代表第 t 个观测向量,它包括第 t 个请求对象 r_t 和 r_t 与 r_{t-1} 之间的时间间隔 τ_t,即 $o_t=(r_t,\tau_t)$, o_a^b 代表从第 a 个到第 b 个观测向量序列, o_1^T 则代表整个观测向量序列,其长度为 T,s_t 代表在 t 时刻所处的状态;ε_t 代表当前状态还将输出观测值的个数,$1\le t\le T$.于是,HsMM 的部分参数($\pi_i;a_{mn};b_m(v);p_m(d)$)、用户观测序列的或然概率($\Pr[o_1^{T_l}\mid\Omega]$)及状态序列($\hat{s}_t^{(l)}$)可由前向-后向算法[3]及多观测序列 HsMM 模型参数估计方法[4]计算得到.

与已有的模型不同,本文二维观测向量中的另一个观测量的输出概率密度函数 $b_m(q)$是一个参数化的函数,因此,其参数($\boldsymbol{\lambda}=[\lambda_1,\ldots,\lambda_M]$)估计方法与其他模型参数不同.具体估计方法如下:

定义多观测序列下的联合概率密度函数为

$$\gamma_t^{(l)}(m) \equiv \Pr[o_1^{T_l}, s_t = m] \tag{4}$$

参数化后的 $b_m(q)$可由下式计算:

$$\hat{b}_m(q) = \sum_{l=1}^{L}\sum_{t:\tau_t=q}^{T_l}\gamma_t^{(l)}(m) \Big/ \sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^{(l)}(m),\ v\in V,q\in I,m\in S \tag{5}$$

于是,参数 λ_m 的最大或然估计为

$$\begin{aligned}\hat{\lambda}_m &= \arg\max_{\lambda_m}\sum_{q\ge1}\hat{b}_m(q)\ln b_m(q)\\ &= \arg\max_{\lambda_m}\sum_{q\ge1}\hat{b}_m(q)\ln(q^{-\lambda_m}-(q+1)^{-\lambda_m})\end{aligned} \tag{6}$$

令 $z_q=\ln(q+1/q)$,则 λ_m 可以由下式求得:

$$\frac{\partial}{\partial\lambda_m}\left[\sum_{q\ge1}\hat{b}_m(q)\ln(q^{-\lambda_m}-(q+1)^{-\lambda_m})\right] = \sum_{q\ge1}\hat{b}_m(q)\left(-\ln q + \frac{1}{\lambda_m}\frac{\lambda_m z_q}{\mathrm{e}^{\lambda_m z_q}-1}\right) = 0 \tag{7}$$

当 λ_m 不大时(通常小于等于 2),可以利用麦克劳林(Maclaurin)公式对式(7)中的 $\lambda_m z_q \big/ (\mathrm{e}^{\lambda_m z_q}-1)$ 进行展开,从而得到 λ_m 的估计值:

$$\begin{aligned}\hat{\lambda}_m &\approx \sum_{q\ge1}\hat{b}_m(q) \Big/ \sum_{q\ge1}\hat{b}_m(q)(\ln q + z_q/2)\\ &= 2\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^{(l)}(m) \Big/ \sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^{(l)}(m)(\ln\tau_t^{(l)}(\tau_t^{(l)}+1))\end{aligned} \tag{8}$$

2 HsMM 模型的应用

(1) 训练与实测

模型的应用分训练和检测两步进行.在训练阶段,首先把观测到的数据预处理为二维观测序列 $O^{(l)}=\{(r_1^{(l)},\tau_1^{(l)}),\ldots,(r_{T_l}^{(l)},\tau_{T_l}^{(l)})\},l=1,\ldots,L$,然后采用上述训练算法迭代计算模型参数,直到模型的或然概率的积(或对数和 $\sum_{l=1}^{L}\ln(\Pr[o_1^{T_l}\mid\Omega])$)收敛到一定的区间内.模型参数确定后,即可应用于对用户的 HTTP 请求序列的在线统计异常检测,即实时计算各个请求序列对于给定模型的或然概率.本文定义平均对数或然概率(即熵) $\log\Pr[o_1^{T_l}\mid\Omega]/T_l$ 作为第 l 个观测序列与模型符合程度的度量.具体做法:当收到一个用户的第 t 个 Web 请求时,记录下请求对象和到达时间,并计算出它与前一请求的时间差;最后计算该用户在[1,t]区间内对应于该模型的平均对数或然概率.

(2) 正常度判断

本文以训练数据集内所有序列的平均对数或然概率的均值 *lkh* 作为正常行为的参考点.可以预先定义一个

观测序列长度门限值 T_0,当 l 用户的 HTTP 请求序列的长度达到 T_0 时,可计算该用户的平均对数或然概率 $lkh^{(l)}$.通过比较 $lkh^{(l)}$和 lkh,可以得到该用户相对于模型的偏离度.差值的绝对值越小,偏离越低,正常度越高,优先权也越高.依据正常度,可以将该用户的后续 Web 请求送入相应的队列进行排队服务.最低优先权的 Web 请求分组,在网络资源不够时,将被过滤或拒绝.由此达到保护正常 Web 请求、化解和抑制攻击流占用服务器资源的目的.

3 模型的验证

3.1 实验数据

本文使用 WorldCup1998[5]的实际流来验证模型的有效性.首先从数据集中随机、不重叠地选取两组用户,每组约 30 000 个用户序列,分别记为 *DS*1 和 *DS*2,其中:*DS*1 用于构造 HsMM 模型;*DS*2 用于模型测试.由于该数据集中不包含 DDoS 攻击数据,为了验证模型对 DDoS 攻击流的识别能力,我们根据 Internet 上提供的 Mydoom 分析报告及已有的模拟 DDoS 攻击流的方法[6],模拟了基于 HTTP 请求的 DDoS 攻击,并采用其数据验证 HsMM 的异常检测能力.模拟攻击的具体方法如下:(1) 假设 200 个攻击节点通过向目标 Web 服务器发送 GET 请求来实现 DDoS 攻击;(2) 攻击者发送 GET 请求的时间间隔设为均值 20ms 的随机数(HTTP 请求的发送速率约 50 个/s),攻击时间长度为 30 分钟;(3) GET 请求的内容使用两种方法生成,并由此生成两种不同的攻击流作为比较:第 1 种,截取一段正常用户高频发出的 HTTP 请求序列片段构成;第 2 种,随机生成每个 GET 请求的内容.

通过对 WorldCup1998 数据集的分析,可以得到正常用户发送 HTTP 请求的速率分布,如图 4 所示,其中约 70%每秒仅发出一个 HTTP 请求,而据 Internet 上的分析报告可知,Mydoom 在实施攻击时的请求包发送速率约为 64 个/s,因此,上述的实验可以产生具有一定效果的 DDoS 攻击流.

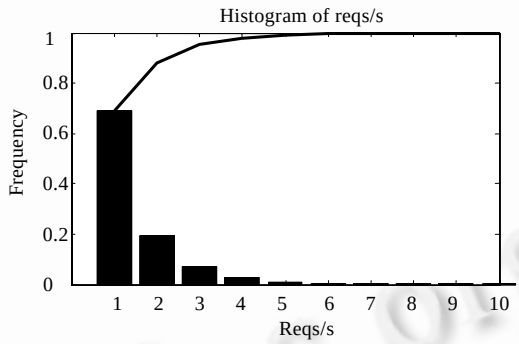

Fig.4 Request rate of normal users

图 4 正常用户请求发送速率分布

3.2 正常性测量

使用 *DS*1 训练 HsMM,然后使用该模型分别计算 *DS*1 和 *DS*2 的平均对数或然概率,并得到它们的直方图分布(如图 5 所示中的 *a* 曲线和 *b* 曲线).可见,它们的分布基本一样.而且由于个体用户访问的多样性,这两条曲线的变化比较平缓,这说明个体用户的“正常性”变化范围比较大.为验证该模型检测异常流的能力,我们使用该模型计算 DDoS 仿真攻击流的平均对数或然概率值的直方图分布(如图 5 所示中的 *c* 曲线和 *d* 曲线).其中:*c* 曲线代表的 DDoS 攻击流采用上述第 1 种方法生成 GET 请求的内容.由于这些攻击者发送 HTTP 请求的间隔都很短,而且 GET 请求的内容仅局限于少量正常用户高频请求的对象(例如主页或 logo 图片),从用户行为模型的角度来看就是高频率地重复相同的操作(这是与正常用户不同的浏览特征),因此,它们的平均对数或然概率值的分布相当集中,而且由于生成的 GET 请求内容是正常用户高频访问的片段,因此,平均对数或然概率偏大,即偏移到正常用户分布的右边;而曲线 *d* 代表的是攻击者采用上述第 2 种方法生成 GET 请求的内容,它们发送请求的间隔同样非常短,所不同的是,GET 的内容是随机选取.从用户行为模型角度看,就是这一类用户的访问不具有正常用户的目的性,由于 GET 请求内容具有随机性,因此,相当一些请求序列的内容和次序与有目的性进行访

问的正常用户不同,因此,这一部分请求序列的平均对数或然概率偏低,最终导致攻击者的平均对数或然概率值的分布偏移到正常用户分布的左边.另外,由于 GET 请求的内容是随机选取的,*d* 曲线的动态变化范围大,平均对数或然概率值的分布范围也较大,因此,比 *c* 曲线平缓,不会出现像 *c* 曲线那样的"尖峰"状.由此可见:无论是哪一种类型的攻击流,模型对正常与异常的识别能力都是非常好的.

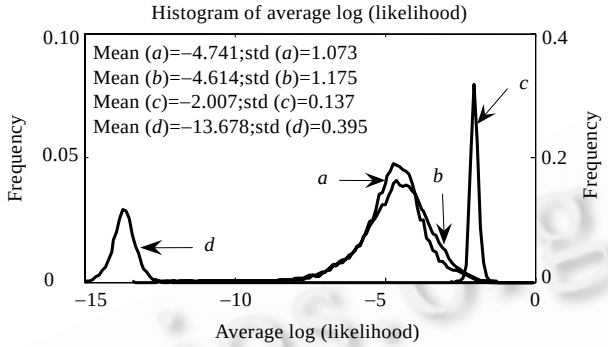

Fig.5 Distribution of average log (likelihood)

图 5 平均对数或然概率分布

图 6 是分别对 4 个数据集中用户的"正常"程度的度量(即 $\log(\Pr[o_1^{T_l} \mid \Omega])/T_l$).可以看到:*DS*1~*DS*2 中的序列随着序列长度的增长,其"正常性"收敛于–5 附近的范围内,其中的短序列比较分散;而攻击流的变化范围非常小,收敛很快,分别在–2 和–13 附近.由此,在实测过程中,可以通过前向算法逐步计算并检测用户的正常度,一旦发现其变化趋势逐渐偏离正常范围,则可以及时进行处理.

在异常检测中,检测率(detection ratio,简称 DR)与误报率(false positive ratio,简称 FPR)是两个重要的性能衡量指标.在我们的模型中,DR,FPR 与模型的判断阈值有关,需要进一步分析三者间的关系以获得一个较好的或然概率判决门限.图 7 是正常访问行为或然概率的左、右门限值与 DR,FPR 的关系.由图 5 可知,两种不同方式的攻击流的或然概率分布分别处于正常行为分布的左、右两侧.因此,从或然概率的角度去识别异常攻击,应该有左、右两个门限值.由图 7 可见,当右门限取–2.2 时,对第 1 种攻击的 DR 为 96%,FPR 为 0.85%;而当左门限取为–12.3 时,对第 2 种攻击的 DR 为 99%,FPR 为 0.003%.因此,对于该模型,如果把或然概率落在(–12.3,–2.2)的用户行为视为正常,则系统的综合检测率约为 97.5%,误检率约为 1%.

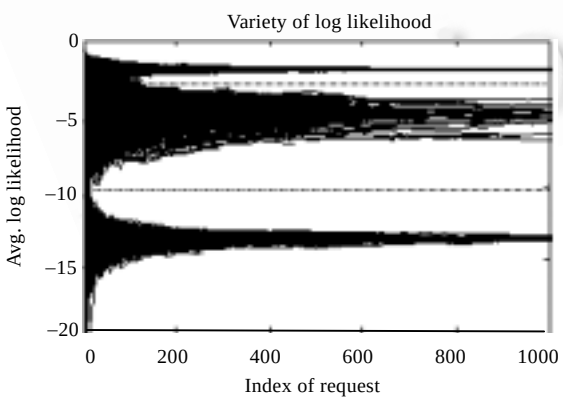

Fig.6 Average log (likelihood) vs. index of request

图 6 平均对数或然概率随序列长度变化趋势

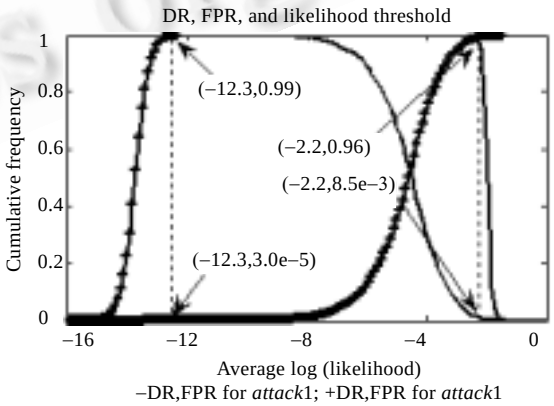

Fig.7 DR, FPR vs. likelihood

图 7 检测率、误检率与或然概率

由上述结果可见,HsMM 模型可以有效地描述正常用户访问的行为特征,并区分出异常用户.由于实际的 DDoS 攻击是通过大量的傀儡机向攻击目标发送海量的请求以促使目标服务器瘫痪,而且发送的请求内容通常

都非常单一(例如网站的主页),因此,本文的 HsMM 模型更容易识别这些攻击.

3.3 状态序列统计特征

由前文的定义可知,HsMM 的一个状态表示用户的一次或连续多次点击所导致的一组请求.为此,我们通过分析各个数据集中各个状态的分布情况,可以了解用户的浏览行为特点.图 8 对 6 个频繁出现的状态进行统计,可见训练组用户和检验组用户的各个状态出现的概率分布基本相同;而对于攻击流,状态出现的概率分布与正常用户的分布有明显的差异.其中:第 1 种攻击的差异主要是由于 GET 请求的内容非常单一,因此只有一个状态是高频率出现的;而第 2 种攻击的差异是由于 GET 请求内容是随机生成的,变化范围大,因此,高概率出现的状态数比前者要多;而且攻击序列中高概率出现的状态数和状态索引号也与正常访问用户序列不同.这种差异恰恰反映出攻击者的访问行为与正常用户的不同.由此可见:通过对状态出现概率的统计,同样可以分辨出不同用户的浏览行为特点.由于模型状态数通常远小于访问对象个数,因此结果更加简单、明了.

状态的具体停留时间长度可以近似用户某一类浏览行为的持续时间.图 9 显示了不同数据集中状态持续时间长度(秒)的 log-log 分布.它们总体上服从 Pareto 分布,而且在 100 秒附近都出现一个拐点.从 *a* 曲线、*b* 曲线可见:对于正常用户序列,状态持续的时间在(1,100)秒区间内的概率分布比较均匀,即拐点以左的曲线比较平缓;而对于攻击流,从 *c* 曲线、*d* 曲线可见:状态持续的时间在(1,100)秒区间内的概率分布呈明显的下降趋势,即 *c* 曲线和 *d* 曲线在拐点以左的曲线下降速度比 *a* 曲线和 *b* 曲线要快,这是由于攻击流的发送速率比正常用户高,从而导致其典型浏览行为的切换频率也比正常用户高.这说明浏览行为的持续时间分布也可以为异常检测提供依据.

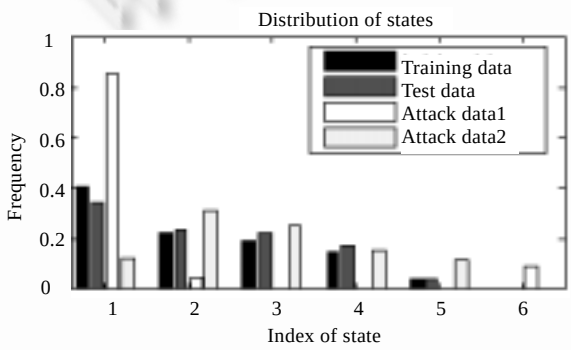

Fig.8 Distribution of states

图 8 状态出现频率分布

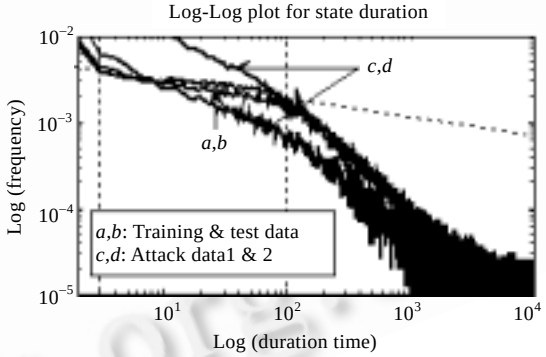

Fig.9 Log-Log plots of states' duration

图 9 状态停留时间的 log-log 图

4 讨 论

4.1 相关研究

从收集到的文献来看,目前专门研究大型活动网站的文章并不多.Arlitt 等人[7]对 FIFA'98 世界杯网站的 HTTP 流量进行分析,并从流量特性上比较了该网站与普通网站的区别,指出 Cache 虽然减少了服务器响应突发流请求的数据量,但并没有降低服务器需要处理的请求数.Liu 等人[8]从请求层、页面层和会话层 3 方面对 2000 年悉尼奥运会、2000~2002 年澳大利亚网球公开赛的 Web 服务器流进行分析,并指出请求层和页面层的流在 30 秒附近具有很强的周期性,而且请求层的到达序列在非峰值区和峰值区分别具有短距离相关和有长距离相关的特点.Iyengar 等人[9]分析了 1998 年长野冬季奥运会期间不同地理位置的 Web 服务器的访问日志,使用 ARIMA(autoregressive integrated moving average)建立 Web 用户请求模型,并用于分析和特征化高流量大型 Web 服务器的访问模式,比较了 1996 年亚特兰大奥运会和 1998 年长野冬季奥运会的流量特征差异.可见:目前对大型活动网站的研究主要在流量特征方面,目的是提高 Cache 和服务器性能,还未涉及到安全性问题.

DDoS 攻击检测与防御一直是网络安全领域的研究热点.现有的检测方法主要有以下几种:(1) 通过 IP 包或 TCP 段的头信息来实现异常检测.这包括:IP 地址与 TTL 值[10]、TCP SYN/FIN 分组检测[11];(2) DDoS 回溯.Park 等人[12]通过分组随机标记来识别攻击分组的路由,然后对攻击源实现速率限制;(3) 基于“puzzle”的方案.Kandula 等人[13]利用攻击程序不具有人的智能性的特点,使服务器在受到攻击威胁时生成一些简单的问题要求用户回答,如果返回的结果正确说明是正常用户;否则就是攻击源.但是,这些检测大部分局限于 IP 层和 TCP 层,这适合于传统的 DDoS 攻击,对于应用层 DDoS 攻击则显得不足.而且现有的统计异常检测方法都隐含了一个前提条件:攻击分组与正常分组的流特征存在统计上的差异.但是这种假设对应用层 DDoS 攻击并不成立,因为这一类攻击在低层不存在特别的异常特征,因此,攻击可以穿越检测系统.另外,应用层 DDoS 的攻击节点只需要按正常速率(甚至低于正常速率)发送具有特殊功能的 HTTP 请求(例如数据库查询或下载大文件)就可以达到攻击的目的,因此,发送速率也不足以检测这一类的攻击.而基于“puzzle”的方法同样具有一些不足之处:(1) 需要得到客户端的支持;(2) 会干扰 Web 用户的正常浏览;(3) 没有办法解决攻击者与正常用户同处于一个终端的情况;(4) 这种方法有可能导致 Internet 中的搜索引擎和缓存/代理等无法正常工作.

用户行为分析在异常检测的应用主要有 UNIX 系统程序调用[14,15]和键盘输入的异常检测[16],但它在 Web 服务器异常检测中的应用并不多.基于用户行为的系统调用、键盘输入异常检测方法与 Web 访问行为异常检测存在明显差异:前者可以采集到完整的观测序列(例如命令序列、输入字符序列);而后者的 HTTP 请求序列有可能被中间代理所响应,因此,其观测序列是不完整的.所以,基于 Web 用户行为的异常检测更加复杂.

4.2 模型性能

HsMM 的状态数与最大状态停留次数共同决定模型结构,也影响着模型的精度和计算复杂度.从分类的角度看:状态数越大,则模型的分类应该越精确,性能应该越好.但实验表明:状态数和最大状态停留次数无须取太大,主要是基于以下几点原因:(1) 由实验可见(如图 10 所示),当状态数增加到一定程度以后,状态数的增加并没有明显改善模型的性能(或然概率),模型平均熵的增量逐渐趋于 0.这说明,当状态数达到某一个值后,模型趋向收敛,因此,没必要继续加大状态数;(2) 大量的研究表明,大部分用户的访问总是集中在少量的页面上.所以,用于描述用户行为的状态数并不需要太多,只要足够描述用户的典型浏览行为即可.

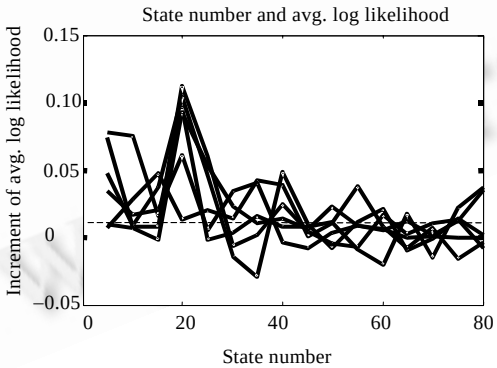

Fig.10 Increment of log (likelihood) vs. state number

图 10 状态数与或然概率增量

用户相对于模型的或然概率可以直接反映用户浏览行为的正常程度,而模型的每个状态代表用户的一种典型浏览行为(包括若干个请求对象和请求的互到达时间),状态的排列次序表明了用户浏览各种典型页面的次序,状态的持续时间代表了用户在浏览典型页面时发出请求的个数.由实验可见,或然概率、状态分布、次序及持续时间等都可以识别出异常访问行为.这进一步表明,基于 Web 用户访问行为的模型可以有效检测和过滤应用层 DDoS 攻击.考虑到不同的时间段其用户行为往往差别很大,随着时间的推移,用户的行为也有可能逐渐偏离已有模型,这种偏离的正常用户行为会导致系统出现高误报率.为此,本文在系统中增加了参数实时更新模

块,如图 3 所示,其中所有被认为正常的实测数据都被加入到训练数据集中,后台程序因此可以周期性地更新模型参数,使得模型参数既可以动态更新又避免遭受攻击者的训练.

模型计算复杂度是影响实际应用的一大因素.本文采用的 HsMM 分训练和实测两步执行.在模型训练阶段,需要估计的模型参数是 $M^2+M(V+D+2)$,计算复杂度受状态数、被请求对象个数、状态最大输出值个数的影响.根据本文实验所采用的数据,在配置为 P4 2.8GHz,1GB 内存的微机上,几分钟之内就可以完成模型的训练.在实测阶段,由于或然概率的计算只需要用到模型的前向算法,因此,一个普通的在线设备一秒钟之内处理几千个用户的记录是没有问题的.主要的时间开销是在内存搜索用户前一次记录的前向变量.而搜索时间可以通过其他一些方法来减少,例如建立搜索树、分流处理等.实际上,我们并不需要对每个用户的访问流都进行检测,可以按照一定的随机抽取策略,或者按照流量大小,或者根据新老用户,选择一部分用户进行检测,从而提高在线处理的效率.所以,用本文提出的统计异常检测算法完全可以实时在线应用.

5 总 结

本文应用 HsMM 来描述 Web 用户的浏览行为,给出了相应的模型参数估计算法和统计异常检测算法,并采用 WorldCup98 的实际数据验证了该模型.模拟分析结果显示,HsMM 可以用于刻画具有非稳态流特性的大型活动网站的正常用户访问行为特征,其中使用了平均对数或然概率来描述用户访问行为的正常程度.基于本文的 HsMM 模型,正常访问序列的平均对数或然概率在均值、方差、分布方面都与异常序列存在明显的差异.这种差异也可以通过状态的概率分布和状态的持续时间长度分布观察到.从平均对数或然概率的收敛上看,正常用户的正常度随序列长度的增长收敛于一个固定的区间内;而异常序列的平均对数或然概率则落在该区间以外.这一点使得我们在实际应用中使用 HsMM 的前向算法就可以实时跟踪用户的正常度变化,并对出现异常或逐渐偏离正常访问行为的用户及时作出处理.本文提出的模型和算法既可以应用于对具有非稳态流特性的大型活动网站的用户浏览行为的异常检测,也可以推广到对普通网站用户浏览行为的异常检测.

References:

[1] Douligeris C, Mitrokotsa A. DDoS attacks and defense mechanisms: Classification and state-of-the-art. Computer Network, 2004, (44):643−666.

[2] Rabiner LR. A tutorial on hidden markov models and selected applications in speech recognition. Proc. of the IEEE, 1989,77(2): 257−286.

[3] Yu SZ, Kobayashi H. An efficient forward-backward algorithm for an explicit duration hidden Markov model. IEEE Signal Processing Letters, 2003,10(1):11−14.

[4] Xie Y, Yu SZ. A detection approach of user behaviors based on HsMM. In: Liang XJ, Xing ZH, Iversen VB, Kuo GS, eds. Proc. of the 19th Int'l Teletraffic Congress (ITC19). Beijing: Beijing University of Posts and Telecommunications Press, 2005. 451−460.

[5] http://ita.ee.lbl.gov/html/contrib/WorldCup.html

[6] Agarwal S, Agarwal S, Gloden B. DDoS attack simulation, monitoring, and analysis. Technical Report, CS 590D, West Lafayette: Purdue University, 2004. 1−27.

[7] Arlitt M, Jin T. A workload characterization study of the 1998 world cup Web site. IEEE Network, 2000,14(3):30−37.

[8] Liu Z, Squillante MS, Xia C, Yu SZ, Zhang L, Malouch N. Analysis of measurement data from sporting event Web sites. In: Proc. of the IEEE Global Telecommunications Conf. New York: IEEE Press, 2002. 2543−2547.

[9] Iyengar AK, Squillante MS, Zhang L. Analysis and characterization of large-scale Web server access patterns and performance. World Wide Web, 1999,2(1-2):85−100.

[10] Jin C, Wang HN, Shin KG. Hop-Count filtering: An effective defense against spoofed DDoS traffic. In: Proc. of the ACM Conf. on Computer and Communications Security. New York: ACM Press, 2003. 30−41.

[11] Wang HN, Zhang DL, Shin KG. Detecting SYN flooding attacks. In: Proc. of the INFOCOM 2002, the 21st Annual Joint Conf. of the IEEE Computer and Communications Societies. New York: IEEE Press, 2002. 1530−1539.

[12] Park K, Lee H. On the effectiveness of route-based packet filtering for distributed DoS attack prevention in power-law internets. In: Proc. of the ACM SIGCOMM. New York: ACM Press, 2001. 295−306.

[13] Kandula S, Katabi D, Jacob M, Berger AW. Botz-4-sale: Surviving organized DDoS attacks that mimic flash crowds. Technical Report, TR-969, MIT, 2004. http://www.usenix.org/events/nsdi05/tech/kandula/kandula.pdf

[14] Hoang XD, Hu JK, Bertok P. A multi-layer model for anomaly intrusion detection using program sequences of system calls. In: Proc. of the 11th IEEE Int'l Conf. on Networks. New York: IEEE Press, 2003. 531−536.

[15] Lian YF, Dai YX, Wang H. Anomaly detection of user behaviors based on profile mining. Chinese Journal of Computers, 2002, 25(3):325−330 (in Chinese with English abstract).

[16] Song DXD, Wagner D, Tian XQ. Timing analysis of keystrokes and timing attacks on SSH. In: Proc. of the 10th USENIX Security Symp. 2001. http://citeseer.ist.psu.edu/song01timing.html

附中文参考文献:

[15] 连一峰,戴英侠,王航.基于模式挖掘的用户行为异常检测.计算机学报,2002,25(3):325−330.

谢逸(1973－),男,博士生,主要研究领域为网络安全,入侵检测,用户行为.

余顺争(1958－),男,教授,博士生导师,主要研究领域为 Internet 流量测量、分析、建模,统计异常检测,隐半马尔可夫模型算法,无线网络.