互联网五种典型应用的平均包长概率分布研究*

杜锡寿[1]，陈庶樵[1]，张建辉[1]，马润华[2]
（1. 国家数字交换系统工程技术研究中心，郑州 450002；2. 海军 92564 部队，广东 汕头 515064）

摘　要：作为流量识别的一个重要手段，深度流检测使用的统计特征中屡屡包含包长信息。从互联网五种典型应用的平均包长入手，利用滑动窗口模型探索五种应用在平均包长概率分布上的差异。对 FTP、Foxmail、WWW、迅雷、Emule 五种应用的实验表明：设置相同的滑动窗口，五种应用平均包长的均值有明显区别；设置不断增大的滑动窗口，五种应用平均包长的均值稳定，标准差逐渐减小。仅用包长信息可识别该五种应用。

关键词：流量识别；平均包长；概率分布；滑动窗口；离散随机过程
中图分类号：TP393.0　　**文献标志码**：A　　**文章编号**：1001-3695(2012)05-1884-03
doi:10.3969/j.issn.1001-3695.2012.05.076

Research on probability distribution of mean packet size for five representative applications in Internet

DU Xi-shou[1], CHEN Shu-qiao[1], ZHANG Jian-hui[1], MA Run-hua[2]
(1. *National Digital Switching System Engineering & Technological Research Center*, *Zhengzhou* 450002, *China*; 2. 92564*th Navy Unit*, *Shantou Guangdong* 515064, *China*)

Abstract: As a important method of traffic identification, DFI(deep flow inspection) employs statistical features which contains packet size information. This paper started with packet size belonging to five representative applications, utilized sliding window model to explore difference on probability distributution of mean packet size of different applications. The experiments on FTP, Foxmail, WWW, Thunder and Emule show: with the same size of sliding window, they behave different expectation of the mean packet size; with being greater size of sliding window, they behave stable expectation and gradually decreasing variance of the mean packet size. It is feasible to identify the five applications only based on packet size information.

Key words: traffic identification; mean packet size; probability distributution; sliding window; discrete stochastic process

0　引言

近年来，随着互联网的高速蓬勃发展，互联网内容得到极大的丰富和发展，网络媒体、互联网信息检索、网络通信、网络娱乐等各种新应用层出不穷；应用模式上从邮件服务、Web 应用等传统应用扩展到流媒体、迅雷等基于 P2P(peer-to-peer)的新应用，但由于 P2P 应用极具吞噬带宽的特性，极易造成网络拥塞，对传统应用的服务质量造成巨大的威胁。因此，对互联网的典型应用，尤其是对 P2P 应用进行识别研究不仅可确保各应用的服务质量，也是网络带宽精细管理的前提条件。当前 P2P 流量识别的主要方法之一是深度流检测(deep flow inspection, DFI)技术，而 DFI 的关键之一在于使用合适的流统计特征作为识别标准。目前，只有少数学者对各应用的平均包长分布进行了初步的研究，但已有众多学者直接将包长的相关统计量(如平均包长、包长度等)用于流量识别中，并达到了满意的识别效果。Este 等人[1]基于信息论中互信息的相关理论对网络数据包分析后认为，包长对互联网应用识别具有重要的意义，特别是 TCP 流的前几个数据包的包长，并且包长在各种环境下含有的信息量相对稳定，适用于对各种应用进行识别，因此建议始终将包长作为流量分类的统计特征之一。Dedinski等人[2]对包到达的时间间隔和包长分布使用小波分析，并分别对比了 eDonkey 和 FTP 数据流包长和控制流包长在频率分布上的区别。Bernaille 等人[3]明确指出不同应用的控制流的包长分布是有区别的。Bernaille 等人[4]另辟思路，提出一种通过分析包长和建立 TCP 链接时的前四个数据包的方向的流量识别方法，该方法对已知应用的识别准确率超过 90%，对未知应用进行识别的准确率为 60%。在对单向应用流的识别中也屡屡出现使用包长的相关统计量的识别方法。Erman 等人[5]在使用聚类的方法对互联网流量进行识别时，采用的统计特征中有平均包长(mean packet size)，并在对流量识别时，对双向应用流和单向应用流都进行了详细的识别分析，且在两种情况下都获得了不错的识别结果。Li 等人[6]仅仅使用短时间窗内的包长分布作为特征对 P2P-TV 应用识别，文中采用基于 SVM 的分类方法，对 PPLive、PPStream、QQLive、SopCast、UUSee 五种基于 P2P 的流媒体应用进行识别，并在 3 s 内获得了超过 99% 的识别准确率。因此，对互联网典型应用的平均包长进行系统的分析不仅可以探索 DFI 采用包长作为流统计特征的内在本质原因，也可为今后的流量识别作出有意义的研究。

本文首先描述了滑动窗口模型的相关原理，进而给出了平

收稿日期：2011-09-05；**修回日期**：2011-10-09　　**基金项目**：国家“863”计划资助项目(2009AA01A346)

作者简介：杜锡寿(1988-)，男，安徽来安人，硕士，主要研究方向为互联网流量识别(dxshardwork@126.com)；陈庶樵(1973-)，男，黑龙江肇县人，教授，博士，主要研究方向为宽带信息网络；张建辉(1977-)，男，讲师，博士，主要研究方向为网络测量；马润华(1987-)，女，湖北宜昌人，主要研究方向为通信信号分析与处理.

均包长的计算方法,并根据实验情况将不同滑动窗口大小下的平均包长概率分布抽象为离散随机过程,给出了离散随机过程的均值和标准差的计算方法,最后对互联网上典型应用的平均包长的概率分布进行系统的实验研究。本文的贡献在于从统计学的角度总结分析了两种现象:a)五种应用在相同的滑动窗口大小下,其平均包长的概率分布不同,且均值差别较大;b)随着滑动窗口大小不断增大,互联网典型业务平均包长概率分布的均值保持稳定,其标准差逐渐减小。

1　研究基础

1.1　滑动窗口模型

众多学者在发表的论文中并没有对包长进行明确的定义,但是无论是从链路层还是从网络层定义分析的数据包长度,对研究平均包长的概率分布并无影响。本文的包长指的是从链路层算起解析出的数据包长度的简称。

互联网的典型应用在进行数据传输前,多数需要建立 TCP 链接,在数据传输时由于链路抖动、链路拥塞等原因会造成 TCP 链接的消亡,从而又需重新建立链接。因此,在抓取的应用数据中可能夹杂 TCP 控制报文,由于 TCP 控制报文和内容报文在包长大小上存在较大的差异,需要对控制报文的包长进行平滑处理减少计算误差。因此,本文在对包长度处理时使用滑动窗口模型[7]。窗口大小 T_W 可自定义,滑动步幅为 1。使用滑动窗口的优点是能够平滑突发数据包,并统计出更多的平均包长值,更加真实地反映出平均包长的概率分布,因此本文在计算平均包长时是基于滑动窗口的平均包长。

如图 1 所示,在互联网某典型应用接收包的个数为 N 的前提下,滑动窗口大小为 5,记为 T_5,在第一次对平均包长进行计算时,使用序号为 1 ~ 5 的数据包计算。在第二次计算平均包长时,使用序号为 2 ~ 6 的数据包计算。

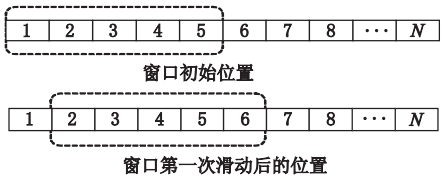

图1　滑动窗口示意图

1.2　基于滑动窗口的平均包长研究

有集合 $w=(w_0,w_1,\cdots,w_m)$,$w_i(0\leqslant i\leqslant m)$ 是互联网中某一典型应用,对该典型应用抓取包的个数为 L,每个包的包长大小为 $P_j(1\leqslant j\leqslant M)$,单位为字节。在滑动窗口为 $T_W=N$ 的条件下,第 k 个平均包长为

$$\overline{P_k}=\frac{\sum_{i=k}^{N+k}P_i}{N}\quad 1\leqslant k\leqslant L-N \tag{1}$$

平均包长$\overline{P_k}$的编号最多到 $M-N$ 是由于在最后 N 个数据包时,滑动窗口已经移动到末尾。

1.3　离散随机过程

若以 X_n 代表某一应用平均包长在滑动窗口大小 $T_W=n$ 的条件下的概率分布,则 $\{X_n,n=1,2,\cdots\}$ 是一个离散随机过程。设 $f_x(x,n)$ 为滑动窗口大小 $T_W=n$ 的条件下的平均包长概率分布,则随机过程 $\{X_n,n=1,2,\cdots\}$ 的均值为

$$m_{x_n}=E[X_n]=\int_{-\infty}^{+\infty}x f_x(x,n)\,\mathrm{d}x \tag{2}$$

随机过程 $\{X_n,n=1,2,\cdots\}$ 的方差为

$$\sigma_{x_n}^2=E[X_n^2]=\int_{-\infty}^{+\infty}x^2 f_x(x,n)\,\mathrm{d}x-m_{x_n}^2 \tag{3}$$

其中:σ_{x_n}称为标准差。

2　实验仿真及结果分析

2.1　数据来源

Moore 在文献[8]中将网络流量分为 Bulk、Database、Interactive、Mail、Services、WWW、P2P、Attack、Games、Multimedia 九类,但是从用户使用的角度出发,Database、Interactive、Attack 应用互联网终端用户较少,Games 应用多种多样,这里不予采集分析。因此本文从剩余的五类中选择了五种目前互联网使用较为广泛的典型应用,数据采集点位于校园网出口处,分为四个时间点基于 Wireshark 进行采集,数据量大小如表 1 所示。

表 1　五种典型应用的采集时间点和数据大小　　/GB

典型应用	0 点	6 点	12 点	18 点	total
FTP	0.81	0.36	0.96	1.05	3.18
Foxmail	0.42	0.30	0.52	0.63	1.87
WWW	0.5	0.43	0.77	2.32	4.02
迅雷	6.21	1.96	5.02	2.18	15.37
Emule	5.37	2.13	2.27	4.89	14.66

2.2　实验平台

由于处理的数据量较大,本文在实验中采用的机器配置为 4.12 GHz 的 Pentium4 CPU,3.87 GB 的内存,500 GB 硬盘。数据处理软件使用的是 Eclipse 3.5 和 MATLAB 7.9。

2.3　实验内容与结果分析

1)不同应用的平均包长概率分布

实验中,将各应用数据分别用 Eclipse 和 MATLAB 处理后,设置的滑动窗口大小 window size = 30 Byte,得到图 2(a) ~ (e),图中横坐标单位为 Byte。

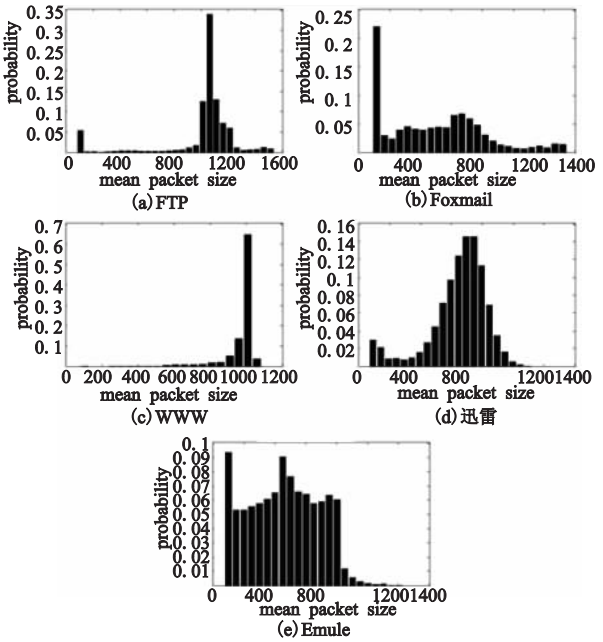

图2　五种典型应用的平均包长在T_w=30时的概率分布

从图 2 中可以看出,(a)FTP 平均包长的峰值出现在 1050 左右,但在 100 时有明显概率分布,这是由于无论是 FTP 服务器还是客户端,它们的进程都可分为控制进程和数据传送进程,控制进程产生的数据包长集中分布在 100 左右,而数据传

送进程产生的包长集中分布在 1050 左右;(b) Foxmail 平均包长的峰值出现在 100 左右,且平均包长落入区间[200,1000]的概率分布基本一致,这是由于 Foxmail 是在邮件服务器和用户代理之间基于 TCP 的可靠传输,因而 TCP 控制报文较多;(c) WWW 的平均包长的峰值出现在 1000 的位置,其他点的概率分布基本为 0,这是由于 WWW 应用的请求报文较短,而响应报文较长,分布集中在 1000 左右,且响应报文的数量要远远多于请求报文;(d)迅雷平均包长的概率分布在[500,900],较集中且概率值较大,而其他值概率较小;(e) Emule 平均包长的概率分布具有双峰值的特性,在区间[100,800]的概率分布较为平稳,且概率分布最大值为 0.09。

在 $T_W=30$ 的条件下,各应用平均包长分布的均值和标准差如表 2 所示。

表 2 五种应用在 $T_W=30$ 下的均值和标准差

参数	FTP	Foxmail	WWW	迅雷	Emule
均值	1 000.4	498.7	928.7	619.4	432.5
标准差	305.8	347.2	145.6	189.4	226.3

2)窗口大小 T_W 变化下的包长概率分布

由于滑动窗口的尺寸越大,在计算一个平均包长值时使用的数据包会越多,平滑作用更加明显,分布越能刻画典型应用之间的本质区别。因此,实验 2 针对当前互联网上较为流行的迅雷应用,在不同的 T_W 下观测其平均包长的概率分布情况,同时计算在每个 T_W 下平均包长分布的均值和标准差,结果如图 3、4 所示。

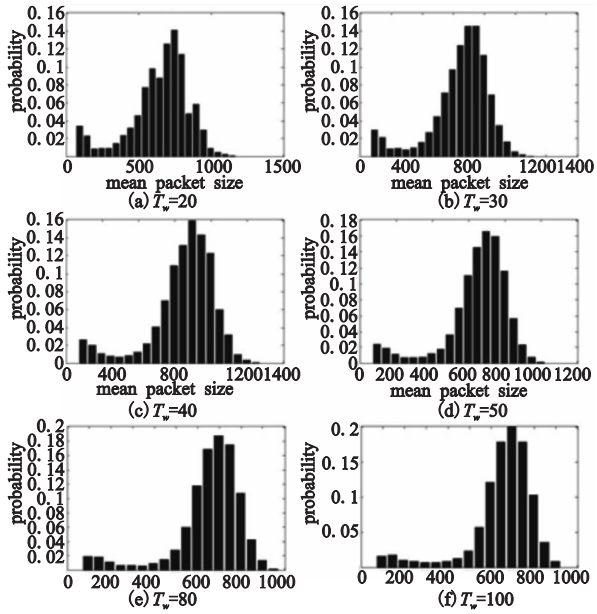

图3 相同T_w下的平均包长分布

从图 3 和 4 中可以看出,随着滑动窗口尺寸增大,概率分布趋于平滑,分布中心值两端的概率逐渐减小,总体分布趋于稳定,因此,标准差随着 T_W 的增加而逐渐减小。但在图 3 中,在不同的窗口大小下,均值的分布一直保持稳定,仅有细微差别,可依赖这一特性对互联网的应用进行流量识别。下面对均值保持稳定的原因加以数学分析。

在 $T_W=N$、抓取的包个数为 L 的条件下,由式(2)知,平均包长的均值为

$$E[X_N]=\frac{\frac{\sum_{i=1}^{N}p_i}{N}+\frac{\sum_{i=2}^{N+1}p_i}{N}+\cdots+\frac{\sum_{i=L-N+1}^{L}p_i}{N}}{L-N+1}=$$

$$\frac{\frac{1}{N}\left(\sum_{i=1}^{N}p_i+\sum_{i=2}^{N+1}p_i+\cdots+\sum_{i=L-N+1}^{L}p_i\right)}{L-N+1}=$$

$$\frac{\frac{1}{N}(P_1+2P_2+\cdots+NP_N+NP_{N+1}+\cdots+NP_{L-N+1}+(N-1)P_{L-N+2}+P_L)}{L-N+1}=$$

$$\frac{\frac{1}{N}(P_1+2P_2+\cdots+(N-1)P_{N-1}+(N-1)P_{L-N+2}+\cdots+P_L)}{L-N+1}+$$

$$\frac{P_N+P_{N+1}+\cdots+P_{L-N+1}}{L-N+1} \tag{4}$$

由于第 i 个包长与第 j 个包长相差在有限的范围内,即 $0<|P_i-P_j|<1500$,因此可将式(4)写为

$$E[X_N]=\frac{2(P_1+P_2+\cdots+P_{\frac{N}{2}})+\delta}{L-N+1}+$$

$$\frac{P_N+P_{N+1}+\cdots+P_{L-N+1}}{L-N+1} \tag{5}$$

其中:δ 是将$\frac{1}{N}(P_1+2P_2+\cdots+(N-1)P_{N-1}+(N-1)P_{L-N+2}+\cdots+P_L)$合并为 $2(P_1+P_2+\cdots+P_{\frac{N}{2}})$时的有限的误差值。由于在抓取数据包时数据量巨大的情况下(大于 1 GB),有 $L\to\infty$,可不考虑 δ 的影响。因此当 N 值在较小范围内变化时(如[0,100]),有

$$\lim_{L\to\infty}\frac{2(P_1+P_2+\cdots+P_{\frac{N}{2}})+\delta}{L-N+1}=0 \tag{6}$$

此时式(5)可改写为

$$E[X_N]\approx\frac{P_N+P_{N+1}+\cdots+P_{L-N+1}}{L} \tag{7}$$

由于窗口大小 N 值不同造成在取不同的 N 值时,$P_N+P_{N+1}+\cdots+P_{L-N+1}$有微小差别,因此 $E[X_N]$有微小差别,但由于数据量巨大,平均包长的均值分布总体保持稳定,所以出现图 4 所示的曲线。

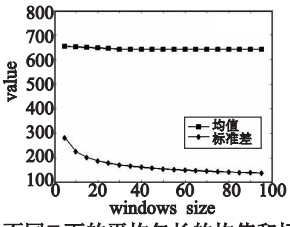

图4 不同T_w下的平均包长的均值和标准差

从式(4)~(7)的数学分析中可以看出,在分析过程中没有指定互联网应用类型,因此,对于任何一种互联网典型应用来说,只要获取的数据量足够大,在不同的滑动窗口大小下,其平均包长的均值是稳定的。

3 结束语

本文从实际网络数据出发,通过分析对比 FTP、Foxmail、WWW、迅雷、Emule 这五种典型应用的平均包长,为在流量识别中将平均包长作为流统计特征提供了实验依据。研究发现,在相同的滑动窗口下,不同应用具有不同的均值和概率分布;在不同的滑动窗口,对于迅雷应用,随着窗口大小的增加,其均值大小稳定,标准差逐渐减小。但是,由于本文采集的采集点单一,数据量还不够,导致在平均包长的分布上可能会出现细微偏差。下一步将通过在不同的数据集中加大数据采集量以进一步验证分析平均包长的概率分布情况,同时研究快速有效的基于平均包长分布的互联网流量识别算法。

(下转第 1900 页)

(上接第 1886 页)

参考文献:

[1] ESTE A, GRINGOLI F, SALGARELLI L. On the stability of the information carried by traffic flow features at the packet level[J]. ACM SIGCOMM Computer Communication Review, 2009, 39(3): 13-18.

[2] DEDINSKI I, De MEER H, HAN L, *et al*. Cross-layer peer-to-peer traffic identification and optimization based on active networking [C]//Proc of International Working Conference on Active and Programmable Networks. [S. l.]: Springer, 2005: 13-27.

[3] BERNAILLE L, TEIXEIRA R, AKODKENOU I, *et al*. Traffic classification on the fly[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(2): 23-26.

[4] BERNAILLE L, TEIXEIRA R, SALAMATIAN K. Early application identification [C]//Proc of the 2nd ADETTI/ISCTE CoNEXT Conference. New York: ACM Press, 2006: 1-12.

[5] ERMAN J, ARLITT M, ANIRBAN M. Traffic classification using clustering algorithms[C]//Proc of SIGCOMM Workshop on Mining Network Data. New York: ACM Press, 2006: 281-286.

[6] LI Jin, ZHANG Xin, ZUO Xiao-liang, *et al*. Using packet size distribution to identify P2P-TV traffic[C]//Proc of International Conference on Cyber-enabled Distributed Computing and Knowledge Discovery. Washington DC: IEEE Computer Society, 2010: 150-155.

[7] LI Hua-fu, LEE S Y. Mining frequent item sets over data streams using efficient window sliding techniques[J]. Expert Systems with Applications, 2009, 36(2): 1466-1477.

[8] MOORE A, ZUEV D. Internet traffic classification using Bayesian analysis techniques [C]//Proc of ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. New York: ACM Press, 2005: 50-60.