典型业务的包长分布规律

李为民，刘晓楠，缪 晨，陈陆颖，雷振明

(北京邮电大学信息与通信工程学院 北京 海淀区 100876)

【摘要】网络流量的识别和分类是网络管理、流量工程等应用的重要前提。该文针对占据主要互联网流量的bitTorrent、HTTP、PPStream、QQ和迅雷5种典型业务进行研究。对不同时间段采集数据的分析结果表明：各典型业务的包长分布均有独特的分布规律，且随时间不同仅存在微小变化。对数据集按不同比例进行采样分析结果表明：随样本总数的递减，各典型业务包长分布形态没有显著改变，且随着样本总数的递减，包长分布曲线虽有所变化，但整体形态和趋势并没有显著改变。该文对研究互联网各业务流量特点，分析和掌握网络流量规律等方面具有重要价值。

关 键 词 计算机网络; HTTP; 包长分布; 采样; 流量
中图分类号 TP393 文献标志码 A doi:10.3969/j.issn.1001-0548.2014.02.017

Packet Size Distribution of Typical Internet Applications

LI Wei-min, LIU Xiao-nan, MIAO Chen, CHEN Lu-ying, and LEI Zhen-ming

(School of Information and Communication Engineering, Beijing University of Posts and Telecommunications Haidian Beijing 100876)

Abstract Identification and classification of network traffic are an essential prerequisite of network traffic, network management, traffic engineering, and other applications. This paper mainly studies five typical Internet applications, including bitTorrent, HTTP, PPStream, QQ, and thunder, which occupy a majority of the Internet traffic. Analysis of data collected during different time periods shows that each application has a unique typical packet size distribution, and there are only minimal changes in packet size distribution among different durations. Moreover, the analysis on data of discrete sample proportions shows that with decreasing the number of samples, the packet size distribution patterns do not change significantly. With the decreasing of the total number of samples, the packet size distributions curve varies, but the overall shape and the trends have not changed substantially.

Key words computer network; HTTP; packet size distribution; sampling; traffic

网络流量的识别和分类是网络流量工程、区分业务服务、异常流量检测等的重要前提。分析和掌握网络流量特点，对于了解网络运行状态，区分网络应用，进行各种网络工程等具有重要意义。文献[1-3]根据网络中数据的流统计特性，对业务识别进行研究和尝试，并取得了较高的业务识别率，而UCR与Intel[4]则研究了综合运用流统计特性以及报文载荷关键字的流量识别。文献[5]证实了在网络业务识别方面，朴素贝叶斯、C4.5决策树、贝叶斯网络与朴素贝叶斯树等算法在准确率方面并无较大差异，但从算法复杂度分析，有明显差别。文献[6]对仅使用TCP连接建立最初5个报文的业务识别进行了实验。文献[7]对比了DBSCAN、K-Means与AutoClass在流量聚类方面的性能。文献[8]则从区分服务，保证服务质量(QoS)的角度对流统计特性作出了分析。文献[9]将流量管理机制应用于多播蜂窝网络获得了更高的性能。文献[10]通过智能网络减少流量进而减少网络建设成本，提高了运营商的市场份额。

通过对互联网现网采集的流记录数据进行分析，本文统计和总结了各类主流业务的包长分布规律。文献[11-12]也做了类似的研究，但都未在业务层面进行分析。本文在统计各业务包长分布规律后，通过对不同时间段采集数据的分析结果进行对比，说明了业务的包长分布随时间不同仅存在微小变化。并且，对数据集按不同采样比进行采样、分析和统计的结果表明，随着样本总数的递减，包长分布曲线有所变化，但整体形态和趋势并没有显著改变。本文的研究对于分析和掌握网络流量规律具有重要价值。

1 数据采集

本文针对各业务的流记录信息进行分析，统计

收稿日期：2011－07－07；修回日期：2013－10－02
基金项目：国家自然科学基金(61072061)
作者简介：李为民(1981－)，男，博士，主要从事网络流量监控、宽带管理、P2P网络技术等方面的研究.

每条流记录中的平均出入向包长度，即流记录中相应方向报文数与字节数的比值，总结典型业务的包长分布规律。流记录数据易于在互联网现网中获得，各通信设备商的网络设备，经过简单配置，都能向指定服务器发送流记录。本文的数据采集于自行研发的在线流量监测设备。该设备事先部署在国内某运营商现网中的8条10G链路上，每天能监控约400 TB的互联网双向流量。该设备除了具有流记录统计和上报能力外，还有基于报文载荷关键字与简单流量特征进行业务识别的功能，并在流记录中标记流量被识别为何种业务。从数据源看，本文具备进行互联网业务包长分布规律研究的条件。

研究数据采集于2010年09月19日09 : 04 : 34至2010年09月19日10 : 01 : 58 CST，持续3 444 s，共采集到流记录225 469 440条。定义上行数据为客户端向服务器发送的数据，下行数据为服务器向客户端发送的数据。其中，上行流量2 350 816 469 802字节，下行流量1 983 423 421 458字节。由于互联网上的业务有上千种之多，为了有效地缩小研究范围，使结果更有针对性，选取BitTorrent、HTTP、PPStream、QQ和迅雷这5类典型应用进行研究。经统计，这些业务的流记录数占总体的70.77%。

2 数据分析

2.1 不同时间段包长分布规律

为了分析各业务包长分布规律随时间的变化，将数据按照采集时间分割为3个数据集。这3个数据集(Trace1, Trace2, Trace3)的详细信息如表1所示。

表1 不同采集时间数据详细信息

数据	流记录数/个	采集时间	各业务流记录数分布	
			业务	流记录数/个
Trace 1	75 156 608	09:04:34～09:24:12	BitTorrent	3 753 434
			HTTP	34 731 305
			PPStream	4 196 071
			QQ	1 944 329
			Thunder	1 698 367
Trace 2	75 156 272	09:24:12～09:43:16	BitTorrent	3 667 266
			HTTP	34 339 796
			PPStream	4 594 980
			QQ	2 137 431
			Thunder	1 714 105
Trace 3	75 156 560	09:43:16～10:01:58	BitTorrent	3 558 432
			HTTP	33 835 828
			PPStream	5 017 607
			QQ	2 294 826
			Thunder	1 747 587

将完整数据集记为Trace 4，作为基准分别对Trace1～Trace3进行分析，并对结果进行对比。统计这4个数据集的流记录中典型业务上下行包长分布的百分比，分布情况如图1～图10所示。

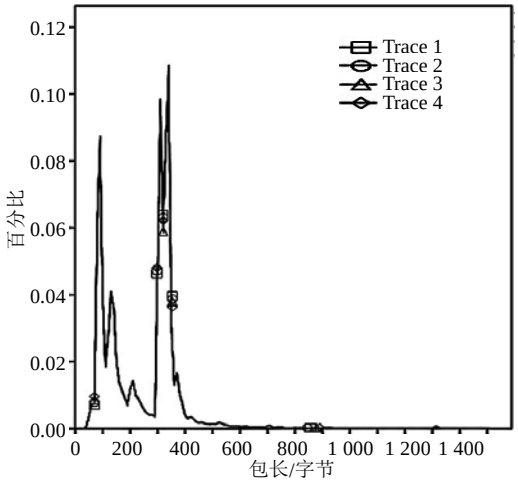

图1 BitTorrent下行包长分布

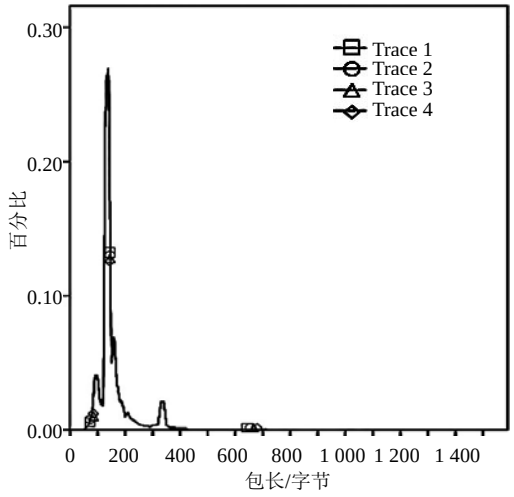

图2 BitTorrent上行包长分布

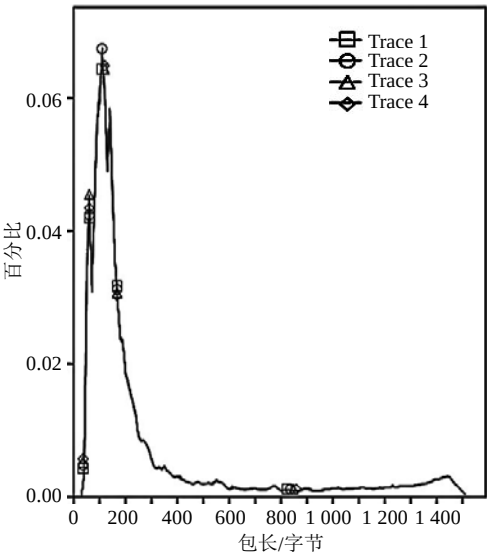

图3 HTTP下行包长分布

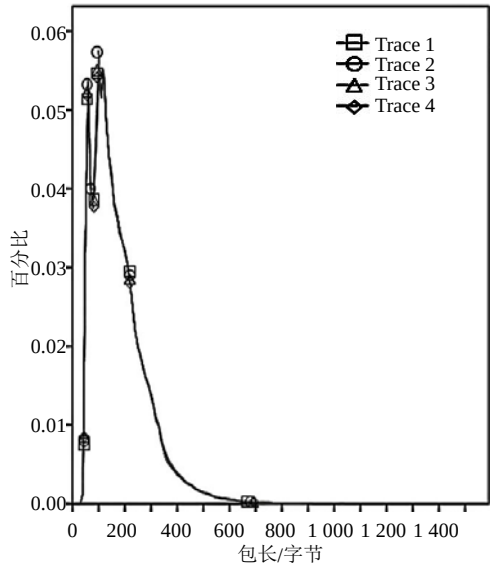

图4 HTTP上行包长分布

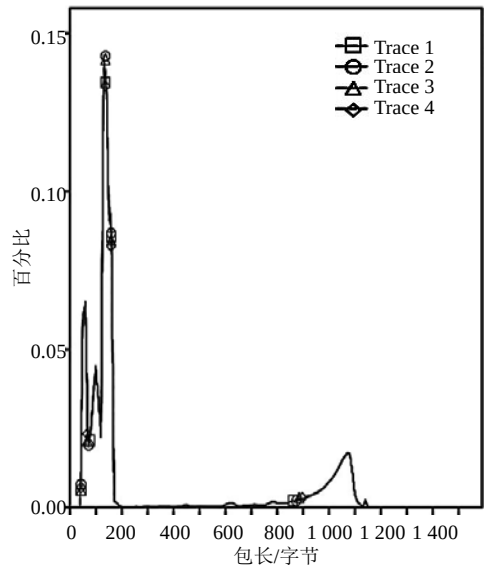

图5 PPStream下行包长分布

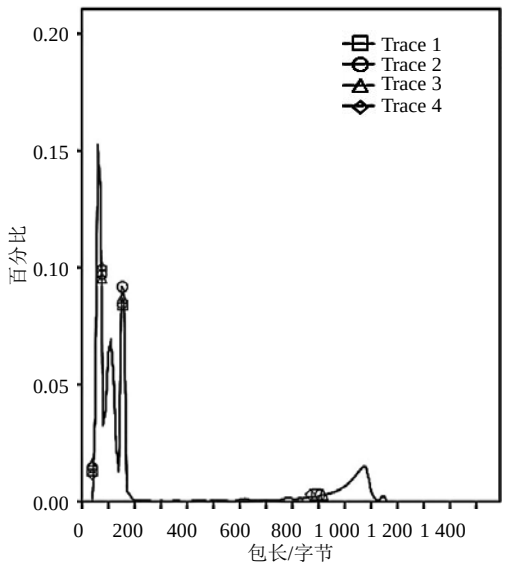

图6 PPStream上行包长分布

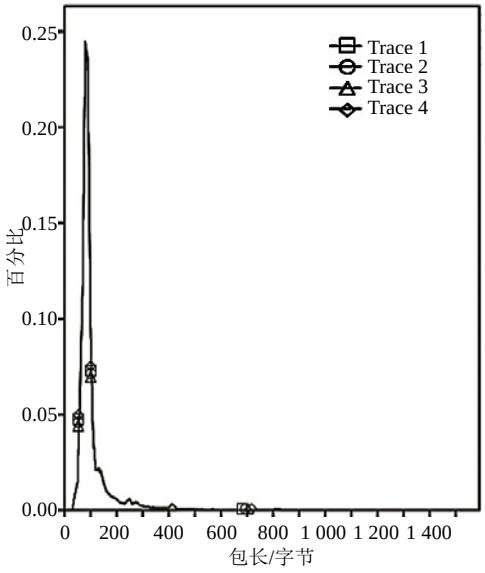

图7 QQ下行包长分布

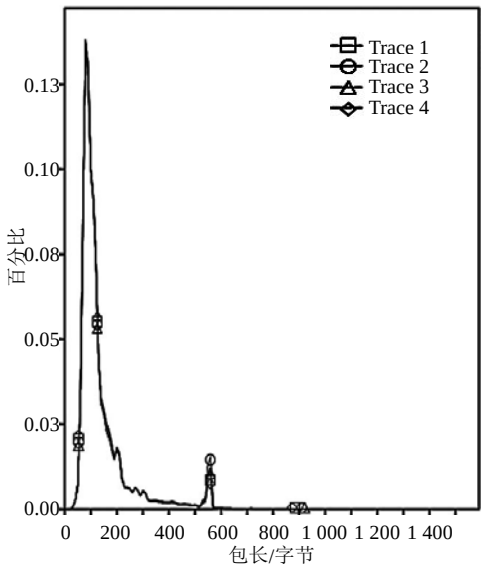

图8 QQ上行包长分布

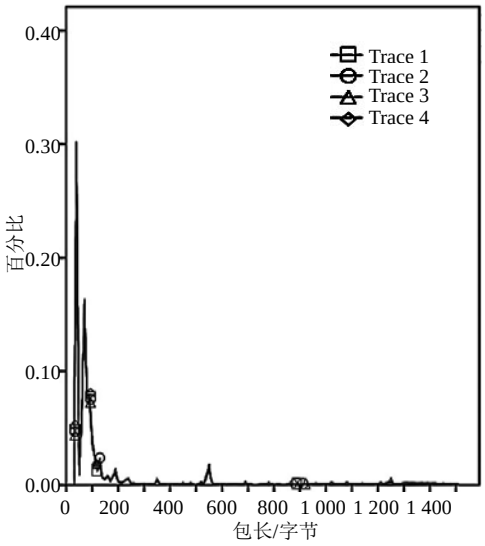

图9 迅雷下行包长分布

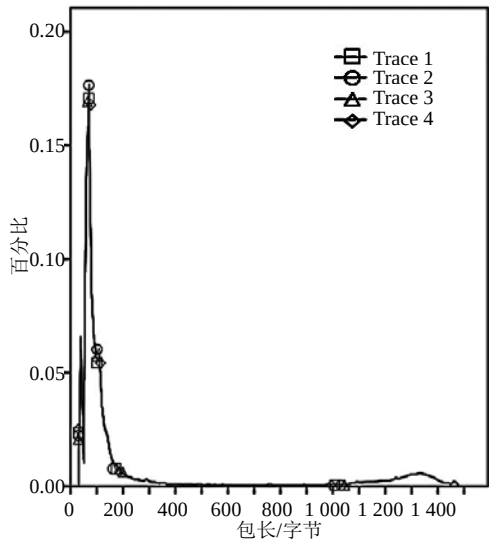

图10 迅雷上行包长分布

从图中容易看出，各业务的包长分布曲线都具有较为明显的特征。具体来说，BitTorrent下行包长度集中分布在(50,200)与(300,400)两个区间，上行包长集中分布在(100,200)之间，但在(300,400)之间也存在峰值。HTTP的下行包长多分布在300字节以下，但较为分散，且存在较多1 400字节以上的大报文分布，相比之下，HTTP的上行流量中则几乎不存在大于700字节的大报文。PPStream的上下行包长分布较为一致，大多分布在200字节以下，且在1 000～1 100字节之间存在峰值。QQ的上下行包长都集中分布在100字节左右，且上行包长在500～600字节之间存在峰值。迅雷的上下行包长同样集中分布在小于200字节的小报文区间。

不仅如此，图中不同时间段数据的各条曲线形态与基准曲线十分吻合。为了分析各分布曲线间的相似程度，将报文长度值域以10字节为一个单位长度进行划分，从0～1 510字节，共划分为151个区间。统计数据集在各个区间上的样本个数。将数据集i在区间j上的样本数记做 X_{ij}，数据集i的样本总数记做T_i，为了消除数据集间样本数的差异，对各区间上的样本数进行归一化，并记区间j上的归一化值为 $R_{ij}=X_{ij}/T_i$。定义数据集a和b之间归一化后的距离为 $D(a,b)=\sqrt{\sum_k R_{ak}R_{bk}}$，这两个数据集$a$和$b$之间的距离值域为$[0,\sqrt{2}]$。最小值取在$a$和$b$在各个区间上的分布百分比数相同；最大值取在$a$和$b$各仅在一个区间上分布，即其分布的百分比为100%，且a和b不分布在同一区间上。表2给出不同时间段的3组数据与完整数据集之间距离。

表2 不同时间段分布与基准偏离

		BitTorrent	HTTP	PPStream	QQ	迅雷
Trace 1	下行	0.378	0.291	0.606	0.751	0.961
	上行	0.428	0.361	0.769	0.773	0.638
Trace 2	下行	0.463	0.316	0.183	0.251	0.995
	上行	0.123	0.138	0.140	0.159	0.301
Trace 3	下行	0.364	0.317	0.545	0.494	0.906
	上行	0.350	0.301	0.541	0.555	0.439

从表中可看出，分布最为平稳的是HTTP业务，最大的差异仅0.3%左右。而分布最不平稳的是迅雷业务，尤其是下行流量，偏差均在0.9%以上。但总体来说，各时间段分布与基准的偏差都不大，均不到百分之一。这说明各业务的包长分布随时间不同仅存在微小变化。

2.2 不同采样比数据的包长分布规律

随着样本数的不同，总体所呈现的分布规律会发生变化。通过对完整数据集按不同采样比进行随机采样，获得4组新的数据，其详细信息如表3所示。

表3 不同采样比数据详细信息

数据	流记录数/个	采样比	各业务流记录数分布	
			业务	流记录数/个
Trace 5	22 543 566	10:1	BitTorrent	1 097 160
			HTTP	10 287 533
			PPStream	1 380 964
			QQ	637 766
			Thunder	517 398
Trace 6	4 509 116	50:1	BitTorrent	218 982
			HTTP	2 058 251
			PPStream	275 604
			QQ	127 502
			Thunder	102 566
Trace 7	902 150	250:1	BitTorrent	43 493
			HTTP	412 075
			PPStream	54 922
			QQ	25 623
			Thunder	20 573
Trace 8	225 186	1 000:1	BitTorrent	10 979
			HTTP	102 972
			PPStream	13 769
			QQ	6 308
			Thunder	5 136

根据定义的数据集之间的距离，同样将报文长度值域以10字节为一个单位长度进行划分，比较不同采样比数据与基准之间的差异，详细结果如表4所示。

表4 不同采样比数据与基准偏离

		BitTorrent	HTTP	PPStream	QQ	迅雷
Trace 5	下行	0.268	0.056	0.274	0.291	0.308
	上行	0.075	0.054	0.103	0.116	0.120
Trace 6	下行	0.461	0.110	0.583	0.570	0.870
	上行	0.240	0.098	0.179	0.240	0.283
Trace 7	下行	1.018	0.225	1.139	1.263	1.605
	上行	0.385	0.250	0.421	0.489	0.623
Trace 8	下行	1.953	0.465	2.887	1.980	3.602
	上行	0.756	0.301	0.736	1.454	1.017

从表中可以看出，分布最为平稳的仍然是HTTP业务，在采样比为1 000∶1的情况下，下行流量的差异也仅有0.4%左右；而分布最不平稳的仍然是迅雷业务，尤其是在采样比为1000∶1的情况下，下行流量与基准距离在3.6%以上。总体来看，各业务上行流量与基准的距离比下行流量的小。这说明各业务下行流量的包长特征比上行流量更富于变化，从文献[13]中也得到了证实，业务的下行流量比上行流量携带更多的信息。从样本数量上看，由于数据集中HTTP占了大多数，HTTP协议经过较高比例采样后，获得的样本数也较多，从数据集Trace 8中可以看出，HTTP的样本数是迅雷样本数的约20倍。所以，采样比的增加对HTTP包长分布的影响比其他业务要小得多。容易看出，当样本数为十万数量级时，各业务与基准的距离都低于1%，而当样本数为几万甚至更少时，各业务下行流量包长分布的偏离超过了1%，并且与基准的距离随着采样比的增加而增大。

根据以上分析可以得出，由于各业务在互联网中所占比例差异较大，对各业务进行随机采样后，所获得样本集合规模差异明显。而总体样本数的多少，决定了其分布形态与基准之间差异的大小。为了在各业务同等样本数规模的条件下进行包长分布形态的分析，下面将在对各业务样本进行采样的同时，控制样本个数，使得各业务样本在同等样本个数的情况下，进行分布形态分析。

2.3　不同采样比同等样本数的包长分布规律

为使各业务样本数一致，在对各业务按照不同比例采样的同时，将各业务多余样本舍去，构成新的5组数据集。由于流记录中，存在流量仅集中在上行或下行一个方向的记录。所以，为了更精确地分析，采集和统计将分别对上下行两个方向进行。数据详细信息如表5所示。

表5　不同采样比同等样本数各数据详细信息

数据集	采样比	各业务下行样本数	各业务上行样本数	总下行样本数	总上行样本数
Trace 9	10∶1	65 000	130 000	6 379 645	15 425 616
Trace 10	50∶1	14 000	25 000	1 276 640	3 087 408
Trace 11	250∶1	2500	5000	255 043	617 485
Trace 12	1 000∶1	500	1000	63 171	153 991
Trace 13	5 000∶1	100	200	12 565	30 654

根据定义的数据集之间的距离，同样将报文长度值域以10字节为一个单位长度进行划分，比较不同采样比数据与基准之间的差异，详细结果如表6所示。

表6　不同采样比同等样本数各数据与基准偏离

		BitTorrent	HTTP	PPStream	QQ	迅雷
Trace 9	下行	0.489	0.722	0.269	0.749	0.529
	上行	0.289	0.689	0.595	0.515	0.266
Trace 10	下行	0.763	1.027	0.429	0.927	0.511
	上行	0.692	0.907	0.652	0.680	0.456
Trace 11	下行	2.577	2.133	1.593	1.300	1.864
	上行	1.467	1.353	1.460	1.357	1.687
Trace 12	下行	3.375	4.657	3.493	4.289	5.057
	上行	3.102	2.717	3.668	4.348	2.966
Trace 13	下行	6.947	8.109	8.221	17.575	6.672
	上行	4.498	7.235	7.187	6.671	6.943

从表中可以看出，在同等样本数的前提下，分布最平稳的不再是HTTP业务，并且在下行样本数为14 000时，只有HTTP下行流量与基准的偏离超过了1%。可见，当HTTP与其他业务具有相同样本数时，其包长分布的平稳性相对较差。总体来看，当下行样本数为14 000，上行样本数为25 000时，几乎所有业务与基准的偏离都不超过1%，除了HTTP的下行流量，但也仅超过0.027%。而对于样本数更少的Trace11～Trace13，各业务的分布与基准的距离普遍超过1%。而对于Trace13，仅有极少的上下行样本数，各业务包长分布与基准的距离，除了最大值QQ的下行流量之外，其余的都在8%左右。这说明了各业务随样本总数的递减，包长分布曲线有所变化，但整体形态和趋势并没有显著改变，即使在样本数很少的极限情况，其分布也与基准较为吻合。

3　结 束 语

本文基于互联网现网采集的流记录数据，针对性地选择互联网典型业务进行分析，统计各业务包长分布规律。通过对不同时间段采集数据的分析结果进行对比分析，说明了业务的包长分布随时间不同仅存在微小变化。之后，对数据集按不同采样比分别进行等样本数和不等样本数采样，分析和统计的结果表明：随着样本总数的递减，包长分布曲线有所变化，但整体形态和趋势并没有显著改变。本文对于研究互联网各业务流量特点，分析和掌握网络流量规律等方面具有重要价值。

参 考 文 献

[1] CROTTI M, DUSI M, GRINGOLI F, et al. Traffic classification through simple statistical fingerprinting[J]. ACM SIGCOMM Computer Communication Review, 2007, 37(1): 7-16.

(下转第256页)