基于端点特征的 P2P 流媒体识别方法*

陈 伟, 兰巨龙, 张建辉
(国家数字交换系统工程技术研究中心, 郑州 450002)

摘 要: P2P 流媒体流量中的控制流与数据流,由于统计特征差异较大,致使 DFI(深度流检测)方法识别其效果不佳。借鉴 DFI 的思想,提出一种基于端点特征识别 P2P 流媒体流量的方法。该方法针对网络端点,提取了六个有效特征,并结合机器学习的方法识别 P2P 流媒体流量。实验结果表明,该方法比 DFI 识别的整体准确率要高,且可以用于 P2P 流媒体的在线识别。

关键词: P2P 流媒体; 机器学习; 流; 端点
中图分类号: TP393.04 **文献标志码**: A **文章编号**: 1001-3695(2012)07-2600-03
doi:10.3969/j.issn.1001-3695.2012.07.054

P2P streaming identification method based on endpoint features

CHEN Wei, LAN Ju-long, ZHANG Jian-hui
(*National Digital Switching System Engineering & Technological R&D Center*, *Zhengzhou* 450002, *China*)

Abstract: Due to the great differences in statistical features between control flows and data flows, the performance of DFI (deep flow inspection) in the identification of P2P streaming traffic is not so ideal. Enlightened by the idea of DFI, this paper proposed a P2P streaming identification method based on endpoint features. This method chose six features aimed at net-endpoint so as to identify P2P streaming traffic using machine learning. Experimental results show that this method performs better in overall accuracy over DFI, and it can also be used in real-time P2P streaming traffic identification.

Key words: P2P streaming; machine learning; flow; endpoint

随着 P2P 网络技术与多媒体信息处理技术相结合,P2P 流媒体已成为继 P2P 文件共享之后又一热点应用。文献[1]表明:P2P 流媒体等在线视频产生的数据流量已占全球互联网流量的 75.9%,预计到 2012 年,这一数字会上升为 81.2%。尽管互联网带宽一直在扩容,但随着 P2P 流媒体流量的增长,电子商务、电子政务等一些关键业务及其他传统业务仍受到带宽资源不足的影响。此外,P2P 流媒体应用架构的开放性,也使其在版权和安全方面产生了相应的问题。因此,对 P2P 流媒体进行合理管控势在必行,而能够准确识别网络中 P2P 流媒体的流量则是有效管控的前提。

目前,基于载荷特征检测的深度包检测法(deep packet inspection,DPI)在众多网络流量识别产品中运用广泛,但由于 P2P 流媒体应用端口使用的跳变性、载荷的加密性以及协议频繁更新等特性,致使 DPI 很难维持稳定的特征库,因此,该方法难以适用于 P2P 流媒体的识别。基于主机行为特征的识别方法[2,3],可将 IP 相关流量识别为 P2P 流媒体流量与否,然而,当在同一 IP 主机上运行多种类型应用时,流量识别结果将不可靠。基于机器学习的深度流检测法(deep flow inspection,DFI)[4]尽管不需要检测载荷特征,但由于 P2P 流媒体中同时存在统计特征差异较大的控制流与数据流,致使其识别效果并不理想。国外方面,针对 P2P 流媒体的识别研究工作主要是:对包含 P2P 流媒体流量在内的 UDP 流量,进行具体应用平台的识别[5,6],而在全网范围内针对 P2P 流媒体流量的粗识别研究较少。

为了能够识别未知 P2P 流媒体应用,对 P2P 流媒体流量进行总体管控,并保证其他业务的正常开展,有必要在全网范围内进行 P2P 流媒体流量的粗识别。本文基于前人的研究,借鉴 DFI 的思想,提出了一种基于端点特征识别 P2P 流媒体流量的方法。该方法由于从 P2P 流媒体的运行机理及流量特征出发进行研究,与 DFI 相比,更加适合于 P2P 流媒体流量的识别。

1 基于端点特征的 P2P 流媒体识别法

P2P 流媒体[7]是基于 P2P 技术的流媒体平台,当用户选定频道成为 P2P 网络的节点,便在本地开放一个 UDP 端口与服务器进行通信,查询节目源与对等节点。一旦得到其他节点的信息,便与之建立连接,依次交换节点列表,请求数据块并传送视频数据,进行资源交换。与传统 C/S 模式不同,其中存在一种 BM(BufferMap)包,用于告知其他节点自己拥有哪些数据块,以便节点请求数据。文献[8]表明,P2P 流媒体流量中,同时存在两种功能流,即控制流与数据流。控制流用于控制节点间的连接,通知 BM 信息,请求资源,单个流中包个数较少,包长较短;数据流用来传递视频数据,包含的包个数较多,且包长较大。由此可见,该两类流特征差异较大,致使 DFI 识别效果不佳。

大量针对 P2P 流媒体的流量测量[8~10]表明:P2P 流媒体稳定运行时,流量主要由 UDP 协议产生,且集中在客户机少数一个或两个端点之上。由于端点同时汇聚了控制流与数据流,

收稿日期: 2011-11-07; **修回日期**: 2011-12-18 **基金项目**: 国家科技支撑计划课题(2011BAH19B01)

作者简介:陈伟(1987-),男,江苏盐城人,硕士研究生,主要研究方向为互联网流量识别(15939045545@163.com);兰巨龙(1962-),男,教授,博士,主要研究方向为高速宽带信息网络;张建辉(1977-),男,副教授,博士,主要研究方向为互联网路由、网络测量.

涵盖 P2P 流媒体应用的总体特性,因此,针对 P2P 流媒体端点进行识别,即可识别出 P2P 流媒体流量。本文借鉴 DFI 的思想,针对端点提出 P2P 流媒体区别于其他应用的有效特征,利用机器学习的方法来识别 P2P 流媒体流量。该方法的流程如图 1 所示。

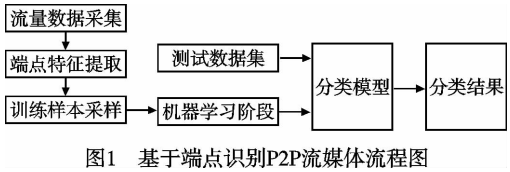

图1　基于端点识别P2P流媒体流程图

在识别流程中,该方法首先对流量数据进行端点特征提取,形成样本集,并对样本集进行采样得到训练集;其次,选择合适的机器学习算法进行训练并得到识别模型;最后,将识别模型用于测试集进行识别效果检验。由于本文不针对机器学习算法进行研究,因此,该方法与 DFI 的不同之处主要在于特征选取,选取的端点特征如表 1 所示。

表 1　端点特征的缩写及描述

编号	缩写	简单描述
1	EP_degree	端点连接度
2	D_rate	流量下载速率
3	D_stability	下载速率稳定度
4	BM_ratio	BM 包信息比
5	BM_ave_len	BM 平均包长
6	Data_ave_len	平均数据包长

表 1 的六个特征中,前两个特征主要用于识别网络中 P2P 文件共享及 P2P 流媒体端点,其余四个特征则用于区分 P2P 文件共享和 P2P 流媒体。关于该六个特征的选取,具体描述如下:

a) EP_degree。客户机上运行传统业务时,同一端点对外进行连接的次数不会超过 5 次,因此一段时间 T 内,若端点对外连接数超过 5 次以上,表明该端点可能具有 P2P 特性[10]。本文定义时间 T 内端点对外连接数为端点连接度。

b) D_rate。对于 DNS 服务器等应用端点,尽管其连接度较高,但其流量下载速率远小于 P2P 应用端点[6]。因此,端点的流量下载速率可以区别 DNS 等服务器应用。

c) D_stability。P2P 流媒体的实时播放性及不对视频数据保存等特点,要求其下载速率与视频播放的码率相当,较为稳定;而 P2P 文件共享则根据实时带宽的大小,即时调整下载速率,因此,P2P 流媒体具有更高的下载速率稳定性[11]。本文定义流量下载速率的标准差为下载速率稳定度。

d) BM_ratio。P2P 流媒体的实时播放性,要求其下载相同数据包个数时需要更多的 BM 包用于控制与调度,因此,其 BM 包个数比数据包个数的值比 P2P 文件共享要大[11]。本文定义 BM 包个数/数据包个数为 BM 包信息比。

e) BM_ave_len。P2P 流媒体不对视频数据进行保存,具有较小的缓存器,其 BM 包的平均包长比 P2P 文件共享要小[11]。

f) Data_ave_len。P2P 流媒体的实时播放性,要求 P2P 流媒体最小传输文件(chunk)比 P2P 文件共享要小,反映在数据包中便是,P2P 流媒体的平均数据包长比 P2P 文件共享要小[11]。

2　实验设置

2.1　实验数据集

为了评估基于端点特征识别 P2P 流媒体方法的效果,本文选用两个数据集,即 Napa-Wine 和 NDSC 数据集。Napa-Wine 是国外针对 P2P 流媒体进行识别研究的最大组织,它提供的数据集由 Napa-IT 和 Napa-PL 组成,于 2008-04-04 分别采集于 Politecnico Di Torino 和 Warsaw 科技大学。NDSC 数据集则是通过 Wireshark 抓包软件抓取的真实局域网网络流量,由于实现过程可控,该数据集抓取了比 Napa-Wine 数据集更广泛的应用流量,采集时间为 2011-05-12。由于 P2P 流媒体的流量主要由 UDP 协议产生,因此,这两个数据集都是针对 UDP 流量的采集。其中关于 UDP 的流可定义为:两个端点(endpoint)之间连续到达且在一定超时时间(60 s)内的 UDP 包序列。数据集的具体描述如表 2、3 所示。

表 2　Napa-Wine 数据集统计信息

大类别	小类别	流数	字节数	端点数
P2P_streaming	PPlive	1.0 M	24.5 G	46 k
	Tvants	35 k	1.1 G	7.1 k
	Sopcast	201 k	30.4 G	12 k
non_P2P_streaming	Emule	183 k	55 M	26 k
	Skype	4.1 M	2.4 G	860 k
	BitTorrent	33 k	14 M	2.5 k

表 3　NDSC 数据集统计信息

大类别	小类别	流数/k	字节数	端点数/k
P2P_streaming	PPLive	300	7.2 G	19
	PPStream	530	10.8 G	25
	QQlive	89	2.5 G	6.3
	UUSee	356	16.4 G	16
	Thunder	185	7.3 G	8.3
non_P2P_streaming	Emule	63	95 M	6.4
	BitTorrent	80	234 M	3.5
	Sype	35	350 M	2.4
	QQ	52	690 M	5
	DNS	74	19 M	10
	NTP	85	37 M	23

2.2　分析工具与平台

本文主要对比 DFI 方法和基于端点特征识别法的效果,不针对识别算法进行研究,因此,提取数据集两种层面的特征后,主要采用数据挖掘工具 Weka 3.5.6 进行研究。该工具是新西兰怀卡托大学 Witten 教授等人开发的开源工作平台,平台利用 Java 语言实现了决策树、朴素贝叶斯等多种机器学习法。本文所用的实验分析平台是一台普通 PC 机,其 CPU 为 Intel Pentium 1.6 GHz,内存为 DDR-667 4 GB,运行 Windows XP 操作系统。

2.3　评价指标

本文对全网 UDP 流量进行粗识别分类,分为 P2P 流媒体流量和非 P2P 流媒体流量,因此,分类器的混淆矩阵(confusion matrix)如表 4 所示。

表 4　分类器的混淆矩阵

	P2P_streaming	non_P2P_streaming
P2P_streaming	TP	FN
non_P2P_streaming	FP	TN

针对 DFI 和基于端点特征识别法的效果进行评估,主要从整体准确率这一指标来衡量。整体准确率,即 overall acuracy = (TP + TN)/(TP + FN + FP + TN)。在对两种方法的识别效果进行对比分析时,流整体准确率和字节整体准确率是可以衡量的共同指标。

3　DFI 和基于端点特征识别法的性能比对

为了对比两种识别法的效果,本文首先对 DFI 方法进行评估,选取训练集的样本数为 N_flow,其中对两大类别分别采样

N_flow/2,对小类别则进行均匀采样。通过实验分析,最终得到合适的 N_flow 以及分类的整体准确率。对基于端点特征识别法进行研究时,利用同样的方法得到合适的时间间隔 T 以及合适的训练集样本数 N_endpoint。最后,本文根据选定的N_flow、T、N_endpoint 重新训练两种识别模型,并针对相同测试集进行识别效果对比。本文选用 Weka 工具中常用的三种机器学习算法:J48 决策树(J48)、朴素贝叶斯(naive Bayes,NB)、贝叶斯网络(BayesNet,BN),进行研究,对比结果不失一般性。

3.1 DFI 识别

尽管 Moore 等人在文献[12]中针对网络流提出多达 248 个特征,但文献[2]表明:运用不同机器学习法在不同数据集上进行网络流的识别,通常只用到常用的共同特征,主要包括平均包长、包长方差、平均包到达间隔、间隔方差、包个数、流持续时间等。其他特征或是增加了特征集的冗余性,或是对识别效果不明显。本文直接选用路遥[13]的特征集用于识别研究,由于该特征集包含以上共同特征,对 DFI 的识别效果具有一般代表性。流特征集的描述如表 5 所示。

表5 流特征缩写及描述

编号	缩写	简单描述
1	C_info_ratio	控制信息比
2	P_change_frequency	包大小变换频率
3	P_ave_len	平均包长
4	P_variance	包长方差
5	P_ave_interval	平均包到达间隔
6	P_interval_variance	间隔方差
7	P_num	包个数
8	F_duration	流持续时间
9	F_rate	流带宽

为了研究 DFI 的识别效果,本文针对 Napa-Wine 和 NDSC 数据集分别进行流采样得到训练集,采样数 N_flow 从 500 递增至5 000,建立好训练模型后,利用 N_flow 相同数目的未被抽样流进行测试。该实验在两个数据集上的测试结果分别如图 2、3 所示。

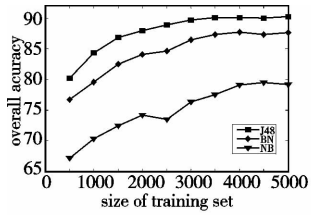

图2 Napa-Wine数据集在不同规模训练集下的整体准确率

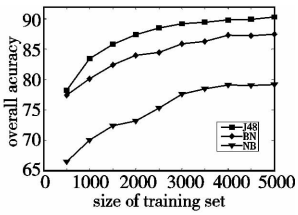

图3 NDSC数据集在不同规模训练集下的整体准确率

在图 2 和 3 中,横轴表示训练集采样数 N_flow,纵轴为流整体准确率。从图中可以看出,在相同情况下,随着 N_flow 的增大,三种算法的流整体准确率增大,并且在 N_flow 为4 000时均趋于平稳。在整体准确率方面,NB 算法总是最低,不到 80%,且在样本增大过程中,波动较大,会出现训练集增大整体准确率反而降低的情形;J48 算法的整体准确率最高且最平稳,可以达到 90%;BN 算法的整体准确率位于 NB 算法和 J48 算法之间,达到 87.6%,其整体准确率的波动不如 NB 算法明显。因此,当训练集规模 N_flow 为4 000时,利用 J48 决策树的识别法能够取得较好的流识别效果。

3.2 基于端点特征的识别

为了研究基于端点特征识别法的效果,首先令时间间隔 T 从 1 s 递增至 10 s,取 N_endpoint 为 5 000,忽略训练集规模的影响,建立训练模型并进行测试。该实验在两个数据集上的测试结果分别如图 4、5 所示。

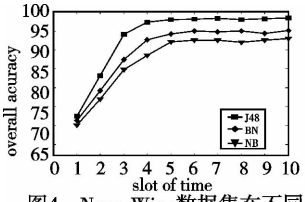

图4 Napa-Wine数据集在不同时间间隔T下的整体准确率

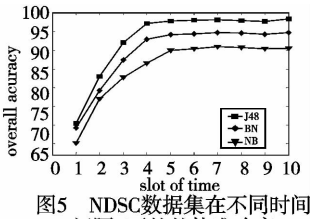

图5 NDSC数据集在不同时间间隔T下的整体准确率

在图 4 和 5 中,横轴表示时间间隔 T,纵轴表示端点整体准确率。可以看出,在相同情况下,随着 T 的增大,三种算法的端点整体准确率均增大,且在 T 为 5 s 时趋于平稳,因此,令 T 为 5 s 可取得较好的识别效果。

为了研究训练集规模对识别效果的影响,取 T 为5 s,令样本采样数 N_endpoint 从 500 递增至 5000,建立训练模型并以 N_endpoint规模的测试集进行测试。该实验在两个数据集上的测试结果分别如图 6、7 所示。

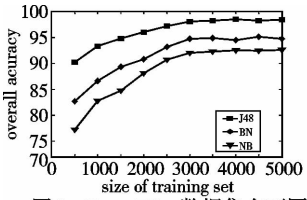

图6 Napa-Wine数据集在不同规模训练集下的整体准确率

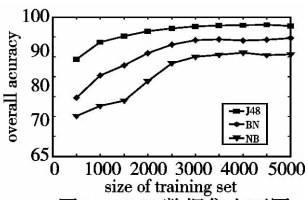

图7 NDSC数据集在不同规模训练集下的整体准确率

在图 6 和 7 中,横轴表示训练集采样数 N_endpoint,纵轴为端点整体准确率。可以看出,在相同情况下,随着 N_endpoint 的增大,三种算法的端点整体准确率均增大,且在 N_endpoint 为 3 000 时趋于平稳,因此,选取训练集数目 N_endpoint 为 3 000 可取得较好的识别效果。

在端点整体准确率方面,从图 4 ~ 7 中均可以看出,三种算法的识别率普遍较高,其中 J48 算法最高,达到 98%;NB 算法最低,达到 90%;BN 算法居于两者之间,可以达到 94%。因此,当时间间隔 T 为 5 s,训练集采样数 N_endpoint 为 3 000 时,利用 J48 决策树的识别算法能够取得较好的端点识别效果。

3.3 DFI 和基于端点特征识别法的对比

为了对比 DFI 和基于端点特征识别法的识别效果,选取 N_flow为 4 000,T 为 5 s,N_endpoint 为 3 000,建立两种识别模型,分别用 500 ~ 5000 这 10 个不同规模的测试集进行识别效果对比。识别结果可映射到流整体准确率和字节整体准确率上进行分析,对该 10 次测试结果取均值,如表 6 所示。

表6 DFI 与基于端点特征识别法(EP)效果对比

算法	J48/%		BN/%		NB/%	
	Flows	Bytes	Flows	Bytes	Flows	Bytes
DFI (Napa-Wine)	90.1	93.2	87.4	90.7	79.3	83.9
EP (Napa-Wine)	98.6	99.4	96.3	99.1	92.8	95.6
DFI (NDSC)	89.6	92.8	87.2	90.2	79.0	83.1
EP (NDSC)	98.5	99.4	95.4	98.5	91.3	94.2

从表 6 中可以看出,采用任一机器学习法及任一数据集,基于端点特征识别法的整体准确率都要高于 DFI,表明其在识别 P2P 流媒体流量时的有效性。此外,由于端点特征比流特征具有更好的区分性,其所需训练集的数目以及特征的个数均较少,并且端点特征的提取可在时间间隔 T 内完成,这对于 P2P 流媒体流量的在线识别也具有重要意义。

(下转第 2606 页)

(上接第 2602 页)

4 结束语

P2P 流媒体识别是当前流量识别领域的一个重要方面,由于 P2P 流媒体流量中存在统计特征差异较大的控制流与数据流,致使 DFI 方法识别效果不佳。流量测量表明,P2P 流媒体端点汇聚了控制流与数据流,涵盖其总体特性,因此,本文提出一种基于端点特征识别 P2P 流媒体的方法。该方法针对端点提出六个特征,并利用机器学习法进行识别。实验表明,该方法比 DFI 识别时具有更高的流整体准确率和字节整体准确率。此外,该方法在训练集规模、特征数目以及特征提取时间上的优势,对于 P2P 流媒体流量的在线识别也具有重要意义。

参考文献:

[1] 张艺濒,张志斌,赵咏,等. TCP 与 UDP 网络流量对比分析研究[J]. 计算机应用研究,2010,26(6):2192-2197.

[2] 雷蕾,沈富可. 基于连接特征的 P2P 流媒体应用的识别[J]. 计算机应用,2007,27(B12):41-43.

[3] LIU Chao-bin, YANG Yue-xiang, TAN Chuan. A classification method of unstructured P2P multicast video streaming based on SVM[C]// Proc of International Conference on Multimedia Information Networking and Security. Washington DC: IEEE Computer Society, 2009: 68-72.

[4] NGUYEN T, ARMITAGE G. A survey of techniques for Internet traffic classification using machine learning[J]. IEEE Communications Surveys & Tutorials, 2008, 10(4): 56-76.

[5] FINAMORE A, MELLIA M, MEO M, *et al*. KISS: stochastic packet inspection[C]//Proc of the 1st International Workshop on Traffic Monitoring and Analysis. Berlin: Spring-Verlag, 2009: 117-125.

[6] ROSSI D, VALENTI S. Fine-grained traffic classification with Netflow data[C]//Proc of the 6th International Wireless Communications and Mobile Computing Conference. New York: ACM, 2010: 4503-4507.

[7] 蒋海明,张剑英,刘琼,等. PPLive 协议分析及流量识别[J]. 电讯技术,2009,49(5): 21-24.

[8] ALESSANDRIA E, GALLO M, LEONARDI E, *et al*. P2P-TV systems under adverse network conditions: a measurement study[C]//Proc of INFOCOM. 2009: 100-108.

[9] SILVERSTON T, FOURMAUX O. P2P IPTV measurement: a comparison study[J]. Computing and Computers, 2006, 7(3): 53-59.

[10] PERENYI M, DANG T D, GEFFERTH A, *et al*. Identification and analysis of peer-to-peer traffic[J]. Journal of Communications, 2006, 1(7): 36-46.

[11] 刘朝斌. 非结构化 P2P 视频组播流的实时识别技术研究[D]. 长沙:国防科学技术大学,2009.

[12] MOORE A W, ZUEV D. Discriminators for use in flow-based classification[R]. Cambridge: Intel Research, 2005.

[13] 路遥. 基于复合特征的 P2P 流媒体识别技术研究[D]. 重庆:重庆大学,2010.