软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(6):1281−1298 [doi: 10.3724/SP.J.1001.2011.03995]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

P2P 流量识别*

鲁 刚[1,2+], 张宏莉[1,2], 叶 麟[1,2]

[1](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[2](哈尔滨工业大学 国家计算机信息内容安全重点实验室,黑龙江 哈尔滨 150001)

P2P Traffic Identification

LU Gang[1,2+], ZHANG Hong-Li[1,2], YE Lin[1,2]

[1](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[2](State Key Laboratory of Computer Information Content Security, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: lgang198202@126.com

Lu Gang, Zhang HL, Ye L. P2P traffic identification. *Journal of Software*, 2011,22(6):1281−1298. http://www.jos.org.cn/1000-9825/3995.htm

Abstract: The rapid increase of P2P traffic worsens the congestion of network while P2P traffic identification becomes the basic technical support for network management. The types of P2P traffic and main challenges of traffic identification are introduced first. Next, the main techniques and research progresses of P2P traffic identification are summarized. Finally, the future trend is put forward.

Key words: peer-to-peer (P2P) network; port identification; deep packet inspection; machine learning; network behavior

摘 要: P2P 流量的迅猛增长加剧了网络拥塞状况,P2P 流量识别为网络管理提供了基本的技术支持.首先介绍了 P2P 流量的类别及流量识别面临的主要困难,然后综述了 P2P 流量识别的主要技术及研究进展,最后给出下一步的主要研究方向.

关键词: 对等网络;端口识别;深层数据包检测;机器学习;网络行为

中图法分类号: TP393 **文献标识码**: A

近年来,随着 P2P 技术的不断发展,对等网络已被广泛应用于文件共享、即时通信、流媒体传输等领域.据报道,自 2004 年以来,P2P 流量已成为互联网的主宰流量,占全部流量的 60%以上[1,2].德国 ipoque 公司发现,2008~2009 年间,东欧的 P2P 流量已达到 70%[1].P2P 流量的迅猛增长一方面给网络带宽造成严重的负担,而且还以其近乎对称的流量模式加剧了网络的拥塞状况;另一方面,基于 P2P 的恶意流量也频繁出现在互联网上,大量的非法连接加快了带宽的消耗,甚至导致拒绝服务攻击[3].由此,如何识别和控制 P2P 流量,已经成为网络运营和管理者面临的挑战.

* 基金项目: 国家自然科学基金(60903166); 国家重点基础研究发展计划(973)(2007CB311101); 国家高技术研究发展计划(863)(2010AA012504); 新世纪优秀人才支持计划(NCET-07-0245)

收稿时间: 2010-09-25; 修改时间: 2010-12-23; 定稿时间: 2011-01-31

CNKI 网络优先出版: 2011-03-07 17:14, http://www.cnki.net/kcms/detail/11.2560.TP.20110307.1714.001.html

© 中国科学院软件研究所 http://www.jos.org.cn

学术界对此类问题给予了广泛关注.自 2000 年起,SIGCOMM,INFOCOM,USENIX 等会议上涌现出大量关于 P2P 测量、识别方面的研究成果.P2P 流量识别根据识别粒度可分为粗粒度识别和细粒度识别两种[4].前者主要区分 P2P 流量与非 P2P 流量,后者主要识别具体的 P2P 应用或协议.由于同一种 P2P 应用可以支持多种协议,例如国内较流行的迅雷支持 HTTP,eDonkey2000,Bittorrent 以及其自身协议等多种下载方式,且很多 P2P 协议不公开,流量负载加密,这些都给 P2P 流量识别带来巨大挑战.

根据所采用的识别方法不同,P2P 流量识别又可分为端口识别技术、深层数据包检测(deep packet inspection,简称 DPI)技术、基于机器学习的流量识别技术、基于网络行为的识别技术等.各种 P2P 流量识别技术都有其优缺点,目前还没有一种技术可以识别网络中所有 P2P 流量.表 1 针对 P2P 流量特点,列举了目前常采用的流量识别技术.

Table 1 Traffic identification techniques aiming at different P2P applications

表 1 不同类型的 P2P 应用所常用的流量识别技术

P2P traffic characteristic		Typical P2P applications	P2P traffic identification techniques
Public protocol	Non-Encryption	eDonkey2000, BitTorrent, etc.	DPI, Port-based techniques
	Encryption	Azureus Gtk-gnutella, etc.	Machine learning techniques, Traffic identification techniques based on network behavior
Private protocol	Non-Encryption	QQ voice, Pplive Ppstream, etc.	DPI
	Encryption	KaZaA, Skype, Xunlei, etc.	Machine learning techniques, Traffic identification techniques based on network behavior

P2P 流量识别面临的主要问题可归结为:(1) 如何提取稳定且准确的 P2P 特征.特征的提取和选择是确保识别准确的关键,而 P2P 网络的动态性以及大量 P2P 软件常采用规避检测技术,使得 P2P 特征并不明显;(2) 如何设计实时且准确的 P2P 流量识别算法.流量识别算法要能够实时快速地识别 P2P 流,但网速不断提高,吞吐量不断增大,这给流量识别算法提出了新的挑战.

本文详细地综述了 2004 年~2010 年来国内外主要的 P2P 流量识别技术的研究进展.第 1 节介绍 P2P 流量识别技术的主要评价指标.第 2 节阐述目前常用的 P2P 流量识别技术.第 3 节介绍 P2P 流量识别技术的研究成果.最后总结目前 P2P 流量识别的主要问题并给出下一步研究方向.

1 P2P 流量识别技术的评价指标

P2P 流量识别技术一般可以从以下 4 方面评估:

- 实时性:反映识别技术能够在线地、快速地识别 P2P 流量的能力;
- 准确性:反映识别技术能够正确地识别 P2P 流量的能力;
- 可扩展性:反映识别技术能够处理大量网络数据流的能力;
- 健壮性:反映识别技术能够克服非对称路由、包丢失和包重传等因素的影响.

目前,除了准确性以外,实时性、可扩展性和健壮性还没有一个量化的评价指标.为便于比较现有 P2P 流量识别方法的有效性,本文首先介绍误报率和漏报率、召回率和精度以及流的准确性和字节准确性评价指标.

1.1 误报率和漏报率

误报(false positive)是指非类别 C 的流量被分类成为类别 C.真阴性(true negative)是指非类别 C 的流量而被分成非类别 C.假定误报数为 FP,真阴性数为 TN,则误报率(false positive rate)为

$$\text{false positive rate} = \frac{FP}{FP+TN}.$$

漏报(false negative)是指属于类别 C 的流量而被分类成非类别 C.真阳性(true positive)是指属于类别 C 的流量而被分类成类别 C.假定漏报数为 FN,真阳性数为 TP,则漏报率(false negative rate)为

© 中国科学院软件研究所 http://www.jos.org.cn

$$\text{false negative rate} = \frac{FN}{FN+TP}.$$

1.2 召回率和精度

基于机器学习算法的流量识别技术常常用召回率(recall)和精度(precision)两项指标来评价识别结果.召回率和精度的计算方法如下:

$$\text{precision} = \frac{TP}{TP+FP},$$

$$\text{recall} = \frac{TP}{TP+FN}.$$

1.3 流的准确性和字节准确性

流通常用五元组〈源 IP 地址,目的 IP 地址,源端口号,目的端口号,协议〉来描述.在超时约束下,采用相同五元组进行通信的一组数据包的集合称之为流.

流的准确性指被正确识别的流数占网络所有流数的百分比.字节的准确性是指被正确识别的数据包承载的字节数占网络传输的总字节数的百分比.Erman 等人[5]指出,在评价流量识别的准确性时,字节的准确性是非常关键的.他们给出的数据集表明,0.1%的流占整个流量字节总数的 46%.如果流量识别算法能够识别出除了这 0.1%的流以外所有的流,那么流的准确性可以达到 99.9%,但却损失了 46%的字节准确性.因此在实际的流量识别效果评估中,在给出流的准确性的同时,也要给出字节的准确性.

P2P 流量识别的准确性已经有较完善的量化评价指标.但目前评价准确性的指标并不统一,这为客观地比较各种 P2P 流量识别技术增加了困难.此外,P2P 流量识别技术的实时性、可扩展性和健壮性常常被定性地评估,有待于进一步研究定量评估问题.

2 P2P 流量识别技术

目前的 P2P 流量识别技术主要分为端口识别技术、深层数据包检测技术、基于机器学习的流量识别技术和基于网络行为的流量识别技术.

2.1 端口识别技术

早期的 P2P 应用程序使用固定的端口号(见表 2),所以网络服务提供商(Internet service provider,简称 ISP)常利用端口号识别 P2P 流量.端口识别技术容易实现,且计算开销小,然而目前的 P2P 应用程序使用端口跳变技术和端口伪装技术来躲避流量检测.Bleul 等人[6]分析 DirectConnect 网络得出,在已观察到的端口中,70%的端口仅仅被使用了一次.可见,基于端口的 P2P 流量识别技术已不能满足当前需求.

Table 2 Port features of P2P application software

表 2 P2P 应用程序的端口号特征

P2P applications	TCP/UDP port number
Gnutella	TCP 6346~6347
eDonkey	TCP 4661~4665
BitTorrent	TCP 6881~6889
DirectConnect	TCP 411~412
Kazaa	TCP 1214
Fasttrack	TCP 1214, 1215, 1331, 1337, 4329
MP2P	TCP 41170, 10240~20480, 22321

2.2 深层数据包检测(DPI)技术

DPI 技术常利用模式匹配算法搜索流量载荷中 P2P 协议的特征值,进而判断是否属于 P2P 流量.应用层负载特征的提取是确保 DPI 技术识别准确率的关键,而模式匹配算法是确保 DPI 技术性能的关键.本节从应用层负载特征和模式匹配算法两个方面阐述 DPI 技术.

© 中国科学院软件研究所 http://www.jos.org.cn

2.2.1 应用层负载特征提取

应用层负载中有能够唯一标识 P2P 流量的特征串.表 3 列出了部分 P2P 应用的负载特征.

Table 3 Payload features of P2P application

表 3 P2P 应用的负载特征

P2P applications	Application-Layer payload features
Gnutella	‘Gnutella’
eDonkey	‘0xe319010000\’
BitTorrent	‘0x13BitTorrent protocol’
DirectConnect	‘$Sen/$Get/$Fit’
Kazaa	‘*X*-Kazaa’
Fasttrack	‘Get ./ hash’, ‘GIVE’
MP2P	‘GO!!, MD5, SIZ0x20’
Ares	‘Get hash:’, ‘PUSH’
QQ voice	‘SIP/user-agent: Tencent-VQQ’

应用层负载特征的提取主要有手工和自动两种方式.手工提取特征的方法是:若协议公开,先对数据报文进行协议分析,通过查阅协议的相关文档,提取其数据报文负载特征,对不公开的协议可以采用黑盒和逆向工程两种方式进行分析.Sen 等人[7]查阅大量的 P2P 协议相关文档,提取出 Gnutella,KazaA,DirectConnect,BitTorrent,eDonkey 等 5 种 P2P 文件共享流量特征,识别准确率在 90.1%~100%.手工方式提取特征比较耗时,对于协议文档不公开或者加密的流量,获取特征更加困难.

自动提取特征方式弥补了手工方式的不足,它主要采用机器学习方法.Haffner 等人[8]采用 Naïve Bayes,AdaBoost 和 Regularized Maximum Entropy 等 3 种机器学习算法提取 7 种应用流量负载特征.利用这些特征进行分类,精度高达 99%~100%,召回率为 86.6%~99.9%.刘兴斌等人[9]采用 Apriori 算法自动提取协议特征,分类 9 种应用流量,除了加密的 emule 流量识别率较低以外,其他应用流量的字节识别率高达 97%以上.Park 等人[10]提出了 LASER 算法,LASER 算法搜索应用层负载的最长公共子序列来作为应用层负载特征.他们利用 LASER 算法提取的特征识别 LimeWire 流量的误报率为 0,漏报率为 8.42%;识别 Bittorrent 流量的误报率为 0,漏报率为 10.4%.

2.2.2 模式匹配算法

模式匹配算法很早就被应用到流量识别领域中[7,11−15].目前,模式匹配算法面临的主要挑战可归结为两点:(1) 更高的吞吐量要求.中国互联网统计报告[16]指出,截止到 2009 年 6 月,中国网络国际出口带宽达到 747541Mbps,较 2008 年增长 16.8%.这要求模式匹配算法有更高的吞吐量以适应网络流量的迅猛增长;(2) 更高的健壮性要求.网络应用流量是多变的,每一种模式匹配算法不可能在任何一种情况下都是最优的,这要求模式匹配算法能够进行动态调整以适应网络数据的动态变化.

模式匹配算法大致可分为单模式匹配和多模式匹配两种,常见的算法有 KMP,KR,BM,AC 及一些改进的算法等.文献[17]综合分析了单模式匹配算法 BM,BHM 以及 QS 算法的优点,并设计了新的用于 P2P 网络流量识别的改进算法.文献[18]比较了 AC 多模式匹配算法和 Wu-Manber 多模式匹配算法的时空复杂度,并提出了多模式匹配的硬件实现方法.鉴于文献[17−21]的研究结果,表 4 列举了目前流量识别领域中常用模式匹配算法的时间复杂度.其中,n 表示文本串长度,m 表示模式串长度.

Table 4 Time complexity among the algorithms of pattern matching

表 4 模式匹配算法的时间复杂度

Pattern matching algorithms	The worst-case time complexity
BF	$O(n\times m)$
KMP	$O(n+m)$
BM	$O(m\times n)$
QS	$O(n/(m+1))$
AC multi-pattern matching	$O(n)$
Rabin-Karp	$O((n-m+1)m)$

© 中国科学院软件研究所 http://www.jos.org.cn

由表 4 可见,模式匹配算法的时间复杂度与特征串长度相关,而 DPI 技术常常用规则表达式描述特征串.确定有限状态机(deterministic finite-state automata,简称 DFA)计算速度快,但存储开销大;不确定有限状态机(nondeterministic finite-state automata,简称 NFA)存储开销小,但计算速度慢.因此,它们都不适用于高速网络环境下描述特征[22].文献[22]提出了一种基于可扩展有限状态机的 DPI 算法,解决了 DFA 存储开销大的问题.

目前,基于 DPI 技术的 P2P 流量识别研究主要通过改进模式匹配算法以提高 DPI 技术的吞吐量.Sen 等人设计了一个基于模式匹配算法的在线分类器识别 P2P 流量,并评估了 SR(standard regex)算法、AR(AST regex)算法和 KR(Karp-Rabin)算法的流量识别性能,其吞吐量分别为 0.21%~2.39%,8.7%~77.60%和 0.07%~0.9%.可见,AR 算法的性能相对最好.文献[23]在入侵检测系统 Snort 平台上实现了对 Bittorrent 流量的识别.早期的 Snort 使用单模式串匹配 Boyer-Moore 算法,Fisk 等人将 SBMH 多模式匹配算法应用于 Snort 时发现,Snort 系统的平均性能提高了 50%.此外,为提高检测性能,基于负载特征匹配的流量分类工具 L7-filter 在默认情况下仅检测一个流开始的前 10 个数据包.如果它能够在这 10 个包内识别出相应的 P2P 协议,那么这个流的所有数据包都属于该 P2P 协议.文献[15]研究了 L7-filter 对于每个会话的搜索深度指出,72%的正则匹配是在每个会话的第 1 个数据包上进行的,且大多数的模式串出现在负载的前 32 字节中.Xu 等人[24]利用 Rabin 字符串匹配算法搜索主机上传流量和下载流量中是否存在相同的负载内容,如果存在相同的负载内容,则认为该主机为 P2P 主机.

实际上,为了保证 DPI 健壮性,模式匹配算法常常要结合其他技术,例如流状态跟踪、协议状态检测机制等.然而 Risso 等人[25]发现,在几千兆级网络环境下,TCP 会话状态表会有几百万条,这已经超出了目前硬件平台的处理能力,极大地限制了 DPI 技术的检测性能.

综上所述,在大多数情况下,DPI 技术准确性高、可靠性好,且能够细粒度地识别流量,主要适合于非加密流量的识别,其识别的准确性依赖于特征库的更新.目前,AceNet,Qosmos 等商业产品主要采用该技术识别流量,而学术界也常以该技术作为新流量识别方法的比较基准.L7-filter 能够准确识别 128 种协议流量,但对负载加密的 skype 流量和迅雷流量识别能力有限.文献[9]识别负载加密的 emule 流量,其准确性仅在 30%~70%之间.此外,在实际应用中,由于 DPI 技术侵犯个人隐私,其应用面受到限制.

2.3 基于机器学习的流量识别技术

基于机器学习的流量识别技术一般不依赖于应用层负载信息,它利用流量统计特征建立机器学习分类模型识别 P2P 流量.

2.3.1 P2P 流量的统计特征提取

P2P 流量的统计特征可以从数据包级和数据流级提取.

(1) 数据包特征

数据包特征主要统计单个流内数据包大小、数据包到达的间隔时间、数据包比率(单位时间内传输数据包的个数)等.表 5 列出统计分析部分数据包特征的时间开销和空间开销,其中,n 为统计的数据包个数.

Table 5 Memory overhead and computational complexity of extracting packet-level features

表 5 提取数据包特征的时间开销和空间开销

Packet-Level features	Time complexity	Space complexity
Single packet size	$O(1)$	$O(1)$
Mean value of packet sizes	$O(n)$	$O(1)$
Variance of packet sizes	$O(n)$	$O(n)$
Minimum/maximum of packet sizes	$O(n)$	$O(1)$
Median of packet sizes	$O(n\times\log_2 n)$	$O(n)$
Mean value of packet inter-arrival time	$O(n)$	$O(1)$
Variance of packet inter-arrival time	$O(n)$	$O(n)$
Packet rate	$O(n)$	O(1)

不同类别的 P2P 流量,其数据包特征有所差异.Bleul 等人比较分析 Bittorrent,DirectConnect,eDonkey,Gnutella 以及 FastTrack 这 5 种 P2P 流量发现,它们之间的平均数据包长差异较大.除了 eDonkey 协议外,其他 4 种频繁出现长度是小于 200 字节的数据包.Teufl 等人[26]指出,音频流的包到达间隔时间非常相似.Marcell 等

© 中国科学院软件研究所 http://www.jos.org.cn

人[27]对 Skype 呼叫流量进行实验分析发现,平均语音数据包大小在 40 字节~320 字节之间变化,单向讲话流的带宽在 20Kbit/s~80Kbit/s 之间变化,而 skype 语音数据包到达的时间间隔是 30ms 或者 60ms,相应的数据包比率分别是 33 个数据包/s 和 16 个数据包/s.他们利用这些特征将 Skype 流量与其他的 VOIP 流量(MSN,Yahoo Messenger,AOL Messenger,Gtalk)区分开.Bonfiglio 等人[28]对 skype 流量进行实验分析发现,在 Skype 呼叫连接的前 30s 内,Skype 客户端发送的数据包大小大约是以后发送数据包大小的 2 倍,平均数据包到达时间间隔是 20ms,30ms 或者 60ms.他们对 skype 流量识别的误报率为 0~0.01%,漏报率为 9.82%~29.98%.Yang 等人[29]统计包长度、包到达时间间隔和包的字节数等特征,对 Bittorrent 流量、pplive 流量、skype 流量和 MSN 流量的识别准确性在 91%~95%.

Este 等人[30]研究了数据包特征的时空稳定性,发现数据包大小受到网络时空环境变化的影响相对最小,而且每个 TCP 连接成功后的第 1 个数据包大小对分类的贡献最大.她们仅分析了 TCP 协议下的数据包特征稳定性,对于 UDP 协议下的特征稳定性还可以进一步深入研究.文献[31]利用数据包大小和数据包方向(客户端发送的数据包为正,服务器发送的数据包为负)分类网络流,对 Bittorrent 的识别准确率为 96.8%.

此外,Roughan 等人[32]研究,仅统计数据包特征还不足以区分大数据块流和流媒体,也不能将 FTP 流与 WWW 流区分开,因此还需要在数据流级获取更多的统计特征.

(2) 数据流特征

数据流特征主要包括流的源/目的端口号、流大小、流持续时间以及标识位(FIN,SYN,RST,PUSH,ACK,URG)被设置的 TCP 数据包数目等等.流大小是指同属于一个数据流的所有数据包字节数总和.流持续时间由一个流的结束时刻减去流开始时刻得到.一般而言,TCP 流的开始时刻是其 SYN 数据包到达时刻,TCP 流的结束时刻是其 FIN 或 RST 数据包到达时刻.UDP 流的开始时刻和结束时刻还没有明确定义,目前,Cisco Netflow 将流的超时值设置为 60s.即,连续两个 UDP 数据包到达时间间隔超过 60s 则认为是两个流.

目前,对于数据流特征提取,国内外学术界已有大量工作.文献[33,34]对 P2P 数据流和 Web 数据流的统计特征进行了比较分析发现,P2P 流大小的均值比 Web 流大小的均值大,P2P 流的平均持续时间要比 Web 流的平均持续时间长.陈庆章等人[35]指出 FTP 流量和 P2P 流量各自的数据流特征发现,P2P 流的数据包大小变化幅度更大,流的持续时间更长,流的总长度更大.Moore 等人[36]提取 249 种 TCP 数据流特征,将网络流量粗略分成 10 种类别,识别 Web 流量的准确性高达 99.27%,而对 P2P 文件共享流量(KazaA,Bittorrent,Gnutella)识别准确性仅达到 36.45%.由于 249 维特征向量有较大的计算开销和存储开销,Li[37]利用基于相关的快速特征选择算法(fast correlation-based filter,简称 FCBF)从 249 种数据流特征中选择出 12 种 TCP 流特征.此外,Li 还提取了 9 种 UDP 流特征.Erman 等人[38]用向后贪婪特征选择算法从 25 种 TCP 数据流特征中选择 11 种流特征.Chhabra[39]等人提出 PISA 算法,自动提取流量统计特征.鉴于以往研究,表 6 列出统计常用数据流特征的时间开销和空间开销,其中,n 为统计的数据包个数.

Table 6 Memory overhead and computational complexity of extracting flow-level features

表 6 提取数据包特征的时间开销和空间开销

Flow-Level features	TCP flow or UDP flow	Time complexity	Space complexity
Client port	TCP or UDP	$O(1)$	$O(1)$
Server port	TCP or UDP	$O(1)$	$O(1)$
Count of all packets with flag bit set in TCP header	TCP	$O(n)$	$O(1)$
The total number of bytes sent in initial window	TCP	$O(1)$	$O(1)$
Count of packets with at least 1 byte of TCP data payload	TCP	$O(n)$	$O(1)$
Count of packets in a flow	TCP or UDP	$O(n)$	$O(1)$
Flow size	TCP or UDP	$O(n)$	$O(1)$
Count of all packets with 0 byte of data payload	TCP or UDP	$O(n)$	$O(1)$
Flow duration	TCP or UDP	$O(n)$	$O(1)$

数据流特征常与机器学习算法结合使用,各个数据流特征结合在一起形成特征向量,作为机器学习算法训练和测试的样本.目前,利用数据流特征分类网络流的研究工作已有很多.文献[32]利用流的持续时间和平均数据包大小分类包含 Kazaa 在内的 7 种业务流,总体识别错误率为 9.7%.文献[40]利用平均流大小、平均流持续

© 中国科学院软件研究所 http://www.jos.org.cn

时间等特征分类P2P流量和Web流量,流准确性达到95%,字节准确性达到80%.Jiang等人[41]利用Cisco Netflow得到数据流级的统计信息进行网络流分类,平均准确性达到 88.3%.

P2P 流量统计特征不依赖于应用层负载内容,但流量统计特征受网络环境的影响较大,其稳定性相对于应用层负载特征较差.文献[34]指出,Web 流持续时间服从双模的 Pareto 分布,P2P 流持续时间服从 Weibull-Pareto 分布,而文献[33]指出,Web 流持续时间和 P2P 流持续时间近似服从对数正态分布.由于 P2P 流量在地域分布上的差异性,所以不同的网络实验环境下得到的实验数据不同,流量模型也有所差异.这意味着,利用数据流特征识别 P2P 流量受网络时空环境的影响较大.文献[42]利用数据流特征建立分类器,并在不同地点采集的数据集上交叉测试,发现分类整体准确性下降.

2.3.2 机器学习算法

本节先从无监督学习、监督学习、半监督学习 3 个部分阐述机器学习算法在 P2P 流量识别领域的应用,再评估与比较各种机器学习算法的性能.

(1) 无监督学习

目前,国内外基于无监督学习的流量识别研究主要使用了 EM(expectation-maximization,期望最大化)算法、AutoClass 算法、*K* 均值聚类算法、DBSCAN 聚类算法、GMM(高斯混合模型)聚类和 HMM(隐马尔可夫模型)聚类算法等.

EM 算法是一种基于概率的聚类算法,它将样本以一定概率指派到簇中.McGregor 等人[43]首次将 EM 算法应用到流量分类中.EM 算法将流量样本聚类成若干个簇,从这些簇中选择对分类贡献最大的流量统计特征.McGregor 等人没有考虑 P2P 流量的识别,而是粗略地将网络流量分成大数据块传输流、交互流等.EM 算法简单且容易实现.在实际流量分类时,EM 算法收敛速度较快,但可能达不到全局最优.其计算复杂度线性于流量统计特征数、数据流数和算法迭代次数.

AutoClass 算法是一种无监督贝叶斯聚类算法,是 EM 算法的一种拓展.它反复使用 EM 算法以便找到全局最优解.Zander 等人[44]用 AutoClass 算法分类 8 种协议流量,其中包括 Napster 的 P2P 应用流量.他们对 Napster 的识别准确性不稳定,最差时为 0,最好时大约为 90%.

K 均值算法以 *k* 作为输入参数,把样本集分成 *k* 个簇,使得结果簇内相似度高,而簇间相似度低.文献[45]使用 *K* 均值聚类算法分类 10 种协议流量,其中,P2P 流量包括 eDonkey 和 kazaa.他们将流量样本集分成 50 个簇,识别 eDonkey 的准确性为 84.2%,识别 kazaa 的准确性为 95.24%.*k* 均值算法简单且易实现,但该算法有两个局限:一是 *K* 均值算法事先随机地选择 *k* 个样本作为簇的初始中心,如果初始中心选择不好,*K* 均值算法将要收敛到次优解;二是结果簇总是凸球状的(spherical).在实际的流量分类中,由于数据流的统计特征对分类的贡献不同,结果簇也不应该都是凸球状的.

文献[46]评估了 *K* 均值聚类、GMM 和 HMM 等 3 种聚类方法分类 10 种 TCP 应用流量的性能,发现 *K* 均值分类整体准确性为 95%左右,而 GMM 和 HMM 的分类整体准确性为 99%左右.GMM 对 eDonkey 识别的准确率 94.1%,对 kazaa 识别准确率 88.9%;而 HMM 对 eDonkey 识别的准确率为 71.4%,对 kazaa 识别的准确率为 67.7%.苏欣等人[47]比较分析了流量识别中常用的 3 种聚类算法 *K* 均值、DBSCAN 和 *k*-medoids,发现 *K*-medoids 聚类算法的字节识别率最高.*K*-medoids 聚类算法识别率在 91.4%~93.6%之间,计算复杂度为 $O(nkt)$(n 是对象的总数,k 是簇的个数,t 是迭代的次数).文献[48]比较了 *K* 均值、DBSCAN 和 AutoClass 等 3 种聚类算法,发现使用 *K* 均值和 DBSCAN 建立分类模型的时间要小于 AutoClass 的建模时间,DBSCAN 分类流量的准确性要低于 *K* 均值聚类算法,但 DBSCAN 算法的分类精度是最高的.

(2) 监督学习

监督学习对训练样本集中的每个输入样本提供类别标记和分类代价,并寻找能降低总体代价的方向.Roughan 等人最早使用监督学习方法分类网络流.他们利用 *K* 最近邻法分类 7 种应用流量.*K* 最近邻算法可以简单描述为:取未知样本 x 的 k 个近邻,判断这 k 个近邻中多数属于哪一类,就把 x 归为哪一类.他们的实验结果发现:在将所有流量分成 3 类时,错误率在 2.5%~3.4%之间;在将所有流量分成 7 类时,错误率在 9.4%~11.4%之间.

© 中国科学院软件研究所 http://www.jos.org.cn

可见,利用 K 最近邻方法识别流量时,流量划分得越细,识别的准确率越低.

Moore 等人使用手工方式标记数据流量形成实验数据集(下文简称 Moore 数据集),他们先用简单的朴素贝叶斯分类算法分类网络流量,分类的总体流准确性较低,仅有 65%.这是因为朴素贝叶斯分类算法假定训练集的样本分布为正态分布,但实际上,训练集中的样本分布常常是未知的.贝叶斯核估计算法是用核函数来逼近原有数据集的样本分布,Moore 等人在应用贝叶斯核估计分类算法时,分类的流准确性提高到 95%以上.尽管贝叶斯核估计算法准确性提高了,但该算法仍旧依赖于各类别样本所占比例.在实际的网络环境中,不同类型的网络流比例是动态变化的,这会影响贝叶斯核估计分类算法的稳定性.

徐鹏[49]在 Moore 数据集上使用支持向量机分类流量,对 P2P 类别的样本识别准确率为 86.73%.支持向量机算法将实际问题通过非线性变换到高维的特征空间,并在高维空间中构造线性判别函数来实现原空间中的非线性判别函数,这种变换策略可以有效降低冗余属性和无关属性对分类性能的影响.由于采用二次寻优方法,即使在各流量类别的先验概率不足的情况下,它的分类准确性和稳定性也要比贝叶斯分类算法好.

C4.5 分类算法也不完全依赖于样本分布,文献[37,50]使用 C4.5 算法分类流量.该算法的准确性比贝叶斯分类算法好,但建模时间较长.文献[51]评估 Bayes 网络、决策树和多层感知机(multi-layer perceptron)等 3 种机器学习算法.他们的实验结果表明,Bayes 网络和决策树更适合于高速网络下的流量分类.Williams[52]对贝叶斯网络、C4.5 决策树、朴素贝叶斯和朴素贝叶斯分类树等 4 种算法进行了比较,发现这 4 种算法流量识别的准确性都在 95%以上.但是 C4.5 决策树算法测试时间最短,更适合实时流量识别.

基于监督学习的流量分类主要面临两个问题:一是标记的流样本稀缺或难以获取.传统的监督学习算法使用少量已标记的样本建立的分类器,往往不能准确识别训练样本集中没有出现的流类型;二是当网络应用行为发生变化时,传统监督学习算法建立的流量分类器要重新训练.

(3) 半监督学习

半监督学习是指用大量的未标记的样本和少量已标记的样本建立分类器[53].文献[38]首次使用半监督学习方法分类网络流量.他们先用 K 均值算法将大量未标记样本和少量标记样本混合的训练集聚类成若干个不相交的簇,然后使用标记的样本完成簇与类别之间的映射,选择簇中已标记的样本比例最大的类别作为该簇的类别.他们分类 8 种应用流量,分类的字节准确性在 70%~90%之间.

Qian 等人[54]提出基于 GMM 的半监督流量分类系统,所用的实验数据集是 Moore 等人[36]的实验数据集.他们在 Moore 数据集上,每 n 个样本中选择一个样本作为有标记样本,其余 n–1 个样本作为未标记样本.实验结果表明,当 n 的取值越大,分类错误率越高.实际上,n 值的增大意味着有标记的样本数减少,未标记样本数增多,所以分类错误率会增大.可见,在实际的流量分类中,为提高分类的准确率,也要保证有标记的样本数.

基于机器学习的流量识别面临的最大问题就是概念飘移(concept drift),即在时刻 t 得到的最佳分类模型 y_t,与前一时刻 t–1 得到的最佳分类模型 y_{t-1} 不一致.导致这种现象的原因是 P2P 网络的动态性.Williams[52]也指出,目前机器学习算法识别流量的速度跟不上网络的限速.如何在 P2P 流量识别中解决概念飘移的问题,可作为未来研究方向.

(4) 机器学习算法的评估与比较

在流量识别研究中,评估与比较不同的机器学习算法所面临的主要困难在于两点:一是缺少一个可信的评估数据集.基于机器学习的流量分类器在训练的过程中应该采用纯净的数据集,即训练集中的每个样本唯一地属于某个特定的应用(或协议)类别.但是目前已公开的数据集,例如 Moore 数据集,只是粗略地给出 P2P 文件共享流量.为此,我们曾开发了一个基于进程的包捕获器[55]以获取纯净的数据样本集,并尝试用单分类支持向量机过滤掉噪音流量[56].但由于一个应用进程可以支持多种协议的通信方式,该方法也仅适用于分类应用,而不适用分类协议;二是目前分类的流量不同,且机器学习算法使用的参数尚未公布,这给客观地比较机器学习算法的优劣增加了困难.表 7 比较了目前用于流量识别的机器学习算法.

© 中国科学院软件研究所 http://www.jos.org.cn

Table 7 Comparisons of machine learning algorithms for traffic identification

表 7 用于流量识别的机器学习算法比较

References	Traffic category	Machine learning algorithms	Evaluation results
Moore, *et al*.[36]	WWW, P2P, Mail, etc.	Naïve Bayes Naïve Bayes kernel	The overall accuracy of Naïve Bayes is 65%, the overall accuracy of Naïve Bayes kernel estimator is 95%
Bernaille, *et al*.[46]	HTTP, EDonkey, kazaa, FTP, etc.	*K*-means, GMM, HMM	GMM and HMM are better than K-means, their accuracies achieve about 99%
Shu, *et al*.[47]	BitTorrent, POCO, EDonkey	*K*-means, DBSCAN *K*-medoids	The byte accuracy of *K*-medoids is the highest among the three cluster algorithms. It is about 91.4%~93.6%
Erman, *et al*.[48]	HTTP, P2P, SMTP, etc.	*K*-means, DBSCAN AutoClass	The model time of *K* means and DBSCAN is shorter and the classification precision of DBSCAN is the highest
Xu, *et al*.[49]	WWW, P2P, Mail, etc.	Naïve Bayes Naïve Bayes kernel Support vector machine (SVM)	The classification accuracy of SVM is the best and achieves about 94.02%
Soysal, *et al*.[51]	HTTP, P2P, FTP, etc.	Bayes networks Decision tree Multi-Layer perceptron	Decision tree achieves the highest accuracy, Bayes networks and decision trees are more suitable for Internet traffic flow classification at a high speed
Williams, *et al*.[52]	HTTP, FTP, SMTP P2P, etc.	Bayes networks C4.5 decision trees Naïve Bayes Naive Bayes trees	The accuracy of the four algorithms is similar, but the test time of C4.5 is shorter and C4.5 is more suitable for real-time traffic identification

2.4 基于P2P网络行为特征的流量识别技术

2.4.1 P2P 网络行为特征提取

P2P 网络的每个对等体(peer)都承担着两种功能角色,它们既是服务的提供者也是服务的使用者,资源的所有权和控制权分散到网络的每一个结点中.P2P 网络行为特征主要包括对等体的连接模式、流行度、扰动性.

(1) P2P 连接模式

Karagiannis 等人[57]发现,P2P 网络传输层连接的两个特征:一是大约 2/3 的 P2P 应用同时使用 TCP 和 UDP 协议,而其他少数应用中同时使用两种协议的仅仅包括 NetBIOS,DNS,游戏等,这些少数应用大多使用固定端口进行通信,例如 NetBIOS 使用 135,137,139 和 445 端口,通过端口号可排除掉这些非 P2P 应用;二是在 P2P 文件共享网络中,对等体之间通常仅使用一条 TCP 连接进行文件传输;而对于 Web 等非 P2P 应用,客户端和服务器之间通常存在多条并发的 TCP 连接.Karagiannis 利用这两个特征识别 P2P 流量,其具体的实现算法详见第 2.4.2 节 PTP 算法所述.该方法识别 P2P 流量的误报率在 8%~12%之间.

Constantinou 等人[58]研究指出,与其他网络所形成的逻辑拓扑图相比,P2P 网络具有更大的直径.他们通过记录每个节点与其他节点建立连接的情况而得到 P2P 网络的逻辑连接拓扑图,并计算其网络直径.若某个网络的直径大于规定的最大直径阈值,并且网络中的既是服务器又是客户端的结点数超过特定的阈值,则该网络是 P2P 网络.该方法识别 P2P 流量,平均漏报率约为 10%.

(2) 流行度

这里的流行度是指在时间 t 内网络中与某台主机建立连接的数量.一些 P2P 应用在使用时要发起大量的连接,流行度会突然增加,这是 P2P 网络的一个行为特征.

文献[59]指出,流行度还不足以识别 P2P 流量,只能作为一种启发信息.在较短时间内,某台主机的流行度突然增加,这意味着两种可能情况:一是主机正运行 P2P 应用,二是主机正遭受恶意攻击.

(3) 扰动性

在 P2P 网络中,对等体可以随时、任意地加入或离开网络,而其频繁加入或离开称为 P2P 网络的扰动性[60,61].扰动性常用对等体的在线时间来衡量.对等体在线时间短且变化大是 P2P 网络扰动性的基本表现.在线时间一般需要通过主动测量技术得到.周丽娟[62]利用 P2P 流媒体具有节点扰动性大、资源暂存性强的特征,对 P2P 流媒体应用识别的准确性在 90%左右.

© 中国科学院软件研究所 http://www.jos.org.cn

表 8 列出提取 P2P 网络行为特征的时间开销和空间开销.其中:m 为 IP 地址数;n 为{IP,Port}列表长度;p 为在时间 t 内,使用相同{IP,Port}的连接数.

Table 8 Memory overhead and computational complexity of extracting network behavior features

表 8 提取网络行为特征的时间开销和空间开销

P2P network behavior features	Time complexity	Space complexity
P2P link pattern	$O(m+mn+n)$	$O(m+n)$
Popularity	$O(mp)$	$O(m+p)$
Churn	$O(mn)$	$O(m+n)$

2.4.2 基于 P2P 网络行为特征的流量识别算法

针对不同的网络行为特征可以设计出多种流量识别算法,本节仅详细阐述两种经典算法.

(1) PTP 算法

PTP 算法[57]首次利用 P2P 的连接模式来识别流量,该算法的设计思想在学术界已得到广泛的应用.其主要思想是,如果源主机与目的主机在预设时间 t 内既使用 TCP 又使用 UDP 协议进行通信,那么它们之间的数据流很可能是 P2P 流.PTP 算法通过端口号排除掉非 P2P 应用流,并将排除后剩余的 IP 地址和端口号记录到{IP,Port}列表中.如果列表中 IP 地址数目与端口数目的差值在某个预设的阈值内,那么该源 IP 与目的 IP 地址之间的数据流被确认为 P2P 流.

PTP 算法主要是根据国外网络环境中 P2P 应用的传输层行为特征提出的,而国内大量使用网络地址翻译技术和被动式 FTP 等技术,这使得 PTP 算法还不能够直接应用于国内网络环境.徐鹏等人[63]针对国内网络环境,提出了 3 条改进策略:① 基于非 P2P 知名端口的过滤机制;② 基于有效数据流的计数机制;③ 基于反向流的 FTP 过滤机制.他们对 P2P 流识别准确率接近 95%,对 P2P 字节识别准确率约为 99%.

(2) 应用层连接同质性(link homophily in the application layer)算法

应用层连接同质性是指运行同一种应用的 IP 主机所产生的流的倾向性.应用层连接同质性算法首次将统计关联学习和图挖掘方法应用于流量识别,为基于网络行为特征的流量识别技术提供了新的思路.Gallagher 等人[64]给出了计算应用层连接同质性的算法.其基本思想是,基于给定的网络流量建立网络踪迹图 G,图 G 的节点为 IP 主机,而 IP 主机间的流作为图的边.如果两条边有共同的节点,那么这两条边被视作邻边.给定已标记类别的边 l,其连接同质性为:与 l 有相同类别的邻边所占的比例.类别的同质性是指图 G 中所有标记为该类别的边 l 的连接同质性的和.

基于连接同质性的分类算法由两部分构成:NLC(neighboring link classifier,邻接边分类器)和 NLC+RL (neighboring link classifier with relaxation labeling,带有松弛标签的邻接边分类器).NLC 算法计算图 G 中每个无标记的边 u 属于类别 c 的连接同质性.NLC+RL 算法将 NLC 算法执行多次,选择连接同质性最大的类别作为边 u 的类别标记.该算法识别 P2P 流量的准确率在 90%以上.

基于网络行为特征的 P2P 流量识别技术不依赖于应用层负载特征,其识别的对象主要是 P2P 网络中的对等体,将对等体之间传输的数据流视为 P2P 流.该技术在实际应用中面临的主要问题:一是它仅能够粗粒度地识别 P2P 流,不能将 P2P 流细化到具体的协议,例如 eDonkey 协议、Bittorrent 协议等等;二是它需监控网络中每台主机的行为模式,但由于一台主机常运行多个应用,如 P2P,Web,E-mail 等,该技术很难从中识别出 P2P 应用的行为模式;三是由表 8 可以看出,提取 P2P 网络行为特征的时间开销和空间开销较大,此技术一般不适用于高速网络环境下的流量识别.

3 P2P 流量识别技术的研究成果

本节从实际应用角度总结了目前国内外 P2P 流量识别技术的研究成果,并比较了各种 P2P 流量识别技术的特点.

定量地比较 P2P 流量识别技术易受到两个条件的约束:一是缺少可信的评估数据集.目前已发布的数据集并未包含应用层的数据信息,这使得部分技术在这些数据集上分类不可行;二是准确性评价指标并不统一.

© 中国科学院软件研究所 http://www.jos.org.cn

因此,表 9 列出了当前主流 P2P 流量识别技术的研究成果.基于端口的流量识别技术由于仅需要检测数据包头部,不采用复杂的计算,实现简单,可用于高速网络环境下实时流量分类.但由于 P2P 应用常使用动态端口,使得该技术识别的召回率较低.

DPI 流量识别技术由于其准确性较好且可用于在线识别,已被广泛应用于商业产品,可是该技术侵犯用户隐私且对加密流量的识别召回率较低.基于机器学习的流量识别技术弥补了 DPI 技术的不足,可用于在线识别加密流量.由于流量统计特征受网络环境的影响较大,这导致该技术健壮性较差.目前,实现机器学习技术的流量分类工具只有 Tstat2.0,而 Tstat2.0 仅能够识别 Skype 流量,尚无法利用机器学习技术识别所有应用流量.基于网络行为特征的流量识别技术不检测数据包的负载信息,保护了用户隐私,也可用于识别加密流量.可是,它仅能够粗粒度地识别 P2P 流,且不适用于高速网络环境下在线实时识别流量.文献[59]提出的 BLINC 系统利用主机的行为模式识别 P2P 流量.该系统仅适合于部署到单宿主边缘网络(single-homed edge network)的边界连接处,而不适合于部署到骨干网连接处.

实际上,每一种流量识别技术都有其优缺点,各种技术的有效结合是很有必要的.混合流量识别技术是当前学术界研究的热点之一.由表 9 可以看出,目前工作常将 DPI 技术和基于网络行为特征的流量识别技术结合在一起,以求获得更好的识别效果.

Table 9 Research reviews of P2P traffic identification technology

表 9 P2P 流量识别技术的研究成果

Traffic identification techniques	Ref.	Ground truth	On-Line	P2P traffic category	Extractive features	Transport-Layer protocol	Identification results	Application mode
Port-Based techniques	[65]	Unspecified	Yes	Gnutella	6346~6347	TCP	It can not classify 30%~70% Internet traffic	Technique comparison
	[66]	Payoad-Based method	Yes	Coarse-Grained P2P	Unspecified	TCP/ UDP	Identification precision is over 80%. The recall is below 60%	
DPI techniques	[7]	Using P2P default port number	Yes	Gnutella	‘GNUTELLA’	TCP	False positive rate and false negative rate are below 5%	Applied practice
				Kazaa	‘*X*-Kazaa-SupernodeIP’ ‘*x*-Kazaa’			
				DirectConnect	The first byte in the payload is ‘$’. The last byte is ‘\|’, the string after ‘$’ ended with space			
				BitTorrent	The first byte in the payload is ‘0x13’, the next 19 bytes match the string ‘BitTorrent protocol’			
				eDonkey	The first byte in the payload is ‘0xe3’. The next 4 bytes is equal to the size of the entire packet			
	[9]	Unspecified	Yes	Bittorrent, Xunlei, pplive ppstream qqlive, emule	Automated protocol signature generation with apriori algorithm (characteristic string and packet length)	TCP	Packet accuracy is over 96%; byte accuracy is over 97%	

© 中国科学院软件研究所 http://www.jos.org.cn

Table 9 Research reviews of P2P traffic identification technology (continue)

表 9 P2P 流量识别技术的研究成果(续)

Traffic identification techniques	Ref.	Ground truth	On-Line	P2P traffic category	Extractive features	Transport-Layer protocol	Identification results	Application mode
DPI techniques	[23]	Unspecified	Yes	Bittorrent	The first byte is ‘0x13’, the next 19 bytes match the string ‘BitTorrent protocol’	TCP	False negative rate is 0.6%~10.5% as packet sampling. Fasle negative rate is 0.06%~1.1% as flow sampling	Applied pratice
					The length of connecting request is 16B, the first 8B is 0x41727101980	UDP		
	[67]	Using L7-filter	Yes	Gnutella	Automatic text-oriented multi-protocol signatures extraction	TCP	The recall of identifying Gnutella is below 40%	
Machine learning techniques	[28]	Payoad-based method	Yes	Skype	At the connection beginning (time [0:30]s), messages are approximately double the size of the messages in the second part of the call. The average message framing inter-arrival time takes values of 20, 30, 60ms	TCP/UDP	False positive rate is 0.0%~1.1%; False negative rate is 2.4%~29.98%	Experiment (skype identification techniques presented by ref. [28] is already used by Tstat2.0)
	[29]	Using default ports and payoad-based method	No	Bittorrent	(1) The statistical features of packet size; (2) The statistical features of packet inter-arrival time; (3) flow duration; (4) Count of packets for each flow; (5) flow size	TCP	Flow accuracy is 96.8%	
	[36]	Manual	No	Corase-Grained P2P	Extracting 249 TCP flow-level features	TCP	P2P flow accuracy is 55.18%	
	[37]	Manual	No	Coarse-Grained P2P	(1) Average segment size; (2) Server port; (3) Client port; (4) Minimum payload size; (5) Maximum payload size, etc	TCP/UDP	Identification precision is 96.46%; Identification recall is 96.59%	
	[45]	Payoad-Based method	Yes	eDonkey Kazaa	The first 5 packet size at the begin of TCP connection	TCP	The accuracy of identifying eDonkey is 84.2%; The accuracy of identifying Kazaa is 95.24%	

© 中国科学院软件研究所 http://www.jos.org.cn

Table 9 Research reviews of P2P traffic identification technology (continue)

表 9 P2P 流量识别技术的研究成果(续)

Traffic identification techniques	Ref.	Ground truth	On-Line	P2P traffic category	Extractive features	Transport-Layer protocol	Identification results	Application mode
Machine learning techniques	[50]	Manual	No	Coarse-Grained P2P	(1) The statistical features of packet size (2) The statistical features of packet inter-arrival time (3) Flow duration (4) Count of packets for each flow (5) Flow size	TCP	P2P flow accuracy is 84.13%	Experiment (skype identification techniques presented by ref. [28] is already used by Tstat2.0)
Traffic identification techniques based on network behavior	[57]	Using default ports and payoad-based method	No	Coarse-Grained P2P	(1) Non-P2P application servers provide services with known port number (2) P2P applications use TCP as well as UDP (3) If the differences of IPs and ports from {IP,PORT} are less than specified threshold value, the {IP, Port} is classified to be P2P application	TCP/UDP	False positive rate is 8%~12%	Experiment
	[59]	Using default ports and payoad-based method	No	Coarse-Grained P2P	(1) the statistical features of data flow (2) popularity of host and link behavior features (3) P2P hosts are service providers as well as requesters	TCP/UDP	Flow accuracy is about 95%	
	[63]	Payoad-Based method	No	Coarse-Grained P2P	Based on the features extracted by Ref. [57], counting the effective data flows	TCP/UDP	Flow accuracy is 95%; Byte accuracy is 99%	
	[68]	Unspecified	No	Coarse-Grained P2P	(1) Popularity (2) Link behavior	TCP/UDP	Flow accuracy is 94.85%	
Hybrid traffic identification techniques	[24]	The traces are captured from a host which only has the given application running on it.	Yes	P2P file sharing and P2P video (such as Bittorrent, Emule and PPlive etc.)	(1) Automated generating payload strings based on Rabin finger printing algorithm (2) Downloaded data from a P2P host will be uploaded to other hosts later	TCP/UDP	The accuracy of identifying Emule is 40%, others are over 90%	Applied practice
	[69]	Unspecified	Yes	Gnutella eDonkey Bittorrent Skype Kazaa	(1) Automated generating payload strings based on LASER algorithm (2) P2P network churn (3) P2P host link pattern	TCP/UDP	Flow accuracy is about 95%	

4 下一步主要研究工作

目前,P2P 流量识别技术是网络流量工程的研究热点.该技术面临的主要困难总结如下:

© 中国科学院软件研究所 http://www.jos.org.cn

(1) P2P 网络的自身复杂特性

P2P 网络最本质的特征是动态性.P2P 流量的动态性使得某些状况下,P2P 流量不具备区别于其他流量的明显特征;P2P 网络行为的动态性,使得基于机器学习的 P2P 流量识别极易出现概念飘移情况.

(2) P2P 网络自身发展特性

P2P 网络作为一种新型的网络应用也在不断的发展与完善.目前,P2P 应用软件正在不断地更新,新的 P2P 应用软件不断涌现.不同 P2P 应用软件的网络行为不同,同一种应用软件的不同版本网络行为也有所差异,所以流量识别技术也需要不断地改进以适应 P2P 网络自身发展特性.

鉴于 P2P 流量识别技术的研究现状,下一步的研究工作主要概括为以下 6 个方面:

(1) P2P 流量识别首先要进行数据采集,而目前网络速度不断提高,在内存资源有限的前提下,不可能采集所有的流量数据.结合流抽样和包抽样的 P2P 流量识别技术,国内外已经有部分研究,但是单独基于包抽样的 P2P 流量识别研究相对较少.包抽样技术可以不维护流的状态信息,有助于提高流量识别效率.此外,面向 P2P 应用的抽样技术也可以作为进一步的研究方向;

(2) 迅雷是目前国内用户使用较广泛的 P2SP(peer to server and peer)下载软件,识别迅雷流量的困难在于:一是其负载内容加密且协议文档不公开;二是迅雷支持多种协议下载方式,网络行为模式不显著.对迅雷流量的进一步识别以及建立迅雷应用的网络行为模型可以作为进一步的研究方向;

(3) Kazaa 和 Gnutella 目前采用协议伪装技术躲避流量检测.他们将自身流量伪装成 HTTP 协议流量进行文件下载传输.对于协议伪装的 P2P 流量识别,可以作为进一步研究方向;

(4) 基于机器学习的 P2P 流量识别技术经常面临概念飘移情况.引起概念漂移的情况有很多,例如网络时空环境发生变化、网络应用分布发生变化等.如何克服概念漂移、提高 P2P 流量识别的健壮性,可作为进一步研究方向;

(5) 网络流量分布状况常常是不平衡的,即一些应用的网络流占据了很大比例,而另外一些应用的网络流所占比例很小.以往基于机器学习的流量识别技术常常把比例很小的网络应用流忽略掉,这是不可取的.因为即使这些应用的流数所占比例很小,但是其字节比例可能会很大(例如 P2P 文件共享数据流).因此,利用机器学习技术分类不平衡的网络数据流可以作为进一步的研究方向;

(6) 部分网络常采用隧道技术保障用户数据的隐私,但这也隐藏了网络应用的行为.目前,对于一种 P2P 应用仅使用一个加密隧道的流量识别情况,国内外已有相关研究.但实际情况下,多个 P2P 应用可同时复用同一个加密隧道.对于这种复杂情况下的 P2P 流量识别,可以作为进一步的研究方向.

References:

[1] Mochalski K, Schulze H. Ipoque internet study 2008/2009. 2009. http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009

[2] MacManus R. Trend watch: P2P traffic much bigger than Web traffic. 2006. http://www.readwriteweb.com/archives/p2p_growth_trend_watch.php

[3] Sun X, Torres R, Rao S. Preventing DDOS attacks on Internet servers exploiting P2P systems. Computer Networks, 2010,54(15): 2756−2774. [doi: 10.1016/j.comnet.2010.05.021]

[4] CAIDA. Internet traffic classification. 2009. http://www.caida:research/traffic-analysis/classification-overview/#P-47

[5] Erman J, Mahanti A, Arlitt MF. Byte me: A case for byte accuracy in traffic classification. In: Sen S, Sahu S, eds. Proc. of the 3rd Annual ACM Workshop on Mining Network Data (MineNet 2007). New York: ACM Press, 2007. 35−37. [doi: 10.1145/1269880.1269890]

[6] Bleul H, Rathgeb EP, Zilling S. Advanced P2P multiprotocol traffic analysis based on application level signature detection. In: Proc. of the Telecommunications Network Strategy and Planning. New Delhi: IEEE Computer Society, 2006. 1−6. [doi: 10.1109/NETWKS.2006.300369]

[7] Sen S, Spatscheck O, Wang DM. Accurate, scalable in-network identification of P2P traffic using application signatures. In: Feldman S, Uretsky M, Najork M, Wills C, eds. Proc. of the 13th Int'l Conf. on World Wide Web (WWW 2004). New York: ACM Press, 2004. 512−521. [doi: 10.1145 /988672.988742]

© 中国科学院软件研究所 http://www.jos.org.cn

[8] Haffner P, Sen S, Spatscheck O, Wang DM. ACAS: Automated construction of application signatures. In: Sen S, Ji C, Saha D, McCloskey J, eds. Proc. of the 2005 ACM SIGCOMM Workshop on Mining Network Data (MineNet 2005). New York: ACM Press, 2005. 197−202. [doi: 10.1145/1080173.1080183]

[9] Liu XB, Yang JH, Xie GG, Hu Y. Automated mining of packet signatures for traffic identification at application layer with apriori algorithm. Journal on Communications, 2009,29(12):51−59 (in Chinese with English abstract).

[10] Park BC, Won YJ, Kim MS, Hong JW. Towards automated application signature generation for traffic identification. In: Brunner M, Westphall CB, Granville LZ, eds. Proc. of the Network Operations and Management Symp. (NOMS). Salvador: IEEE Press, 2008. 160−167. [doi: 10.1109/NOMS.2008.4575130]

[11] AceNet. The appalachian center for economic networks. 2009. http://acenettech.com/

[12] Qosmos. Deep packet inspection and network intelligence tools. 2010. http://www.qosmos.com/

[13] L7-filter supported protocols. 2009. http://l7-filter.sourceforge.net/protocols

[14] Snort. Network intrusion prevention and detection system. 2010. http://www.snort.org

[15] Aceto G, Dainotti A, Donato W, Pescapé A. PortLoad: Taking the best of two worlds in traffic classification. In: Proc. of the INFOCOM IEEE Conf. on Computer Communications Workshops. San Diego: IEEE Press, 2010. 1−5. [doi: 10.1109/INFCOMW.2010.5466645]

[16] CNNIC. The statistical survey report on Internet developing in China. 2009 (in Chinese). http://www.cernet.com/news/chinacngi.pdf

[17] Zhao R. The research and implementation of P2P traffic identification based on feature string [MS. Thesis]. Chengdu: University of Electronic Science and Technology of China, 2009 (in Chinese with English abstract).

[18] Li WN, E YP, Ge JG, Qian HL. Multi-Pattern matching algorithms and hardware based implementation. Journal of Software, 2006, 17(12):2403−2415 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/2403.htm [doi: 10.1360/jos172403]

[19] Tan JL. String matching algorithm and application of network content analysis [Ph.D. Thesis]. Beijing: Graduate University of the Chinese Academy of Sciences, 2003 (in Chinese with English abstract).

[20] Cormen TH, Leiserson CE, Wrote; Pan JG, Gu TC, Li CF, Ye M, Trans. Introduction to Algorithms. 2nd ed., Beijing: China Machine Press, 2008. 557−568 (in Chinese).

[21] Navarro G, Raffinot M, Wrote; Network Information Security Research Center in Institute of Computing Technology, Trans. Flexible Pattern Matching in Strings. Beijing: Publishing House of Electronics Industry, 2007. 13−68 (in Chinese).

[22] Smith R, Estan C, Jha S, Kong SJ. Deflating the big bang: Fast and scalable deep packet inspection with extended finite automata. In: Bahl V, Wetherall D, Savage S, Stoica I, eds. Proc. of the ACM SIGCOMM 2008 Conf. on Data Communication (SIGCOMM 2008). New York: ACM Press, 2008. 207−218. [doi: 10.1145/1402958.1402983]

[23] Guo ZB, Qiu ZD. Identification of BitTorrent traffic for high speed network using packet sampling and application signatures. Journal of Computer Research and Development, 2008,45(2):227−236 (in Chinese with English abstract).

[24] Xu K, Zhang M, Ye MJ, Chiu DM, Wu JP. Identify P2P traffic by inspecting data transfer behavior. Journal of Computer Communications, 2010,33(10):1141−1150. [doi: 10.1016/j.comcom.2010.01.005]

[25] Risso F, Baldi M, Morandi O, Baldini A, Monclus P. Lightweight, payload-based traffic classification: An experimental evaluation. In: Sun L, Zhang JW, Wang YG, eds. Proc. of the IEEE Int'l Conf. on Communications (ICC 2008). Beijing: IEEE Press, 2008. 5869−5875. [doi: 10.1109/ICC.2008.1097]

[26] Teufl P, Payer U, Amling M, Godec M, Ruff S, Scheikl G, Walzl G. InFeCT—Network traffic classification. In: Bi J, Gyires T, eds. Proc. of the 7th Int'l Conf. on Networking (ICN). Cancun: IEEE Computer Society, 2008. 439−444. [doi: 10.1109/ICN.2008.42]

[27] Perényi M, Molnár S. Enhanced skype traffic identification. In: Proc. of the 2nd Int'l Conf. on Performance Evaluation Methodologies and Tools (Valuetools 2007). Brussels: ICST Press, 2007. 1−9.

[28] Bonfiglio D, Mellia M, Meo M, Rossi D, Tofanelli P. Revealing skype traffic: When randomness plays with you. ACM SIGCOMM Computer Communication Review, 2007,37(4):37−48.

[29] Yang AM, Jiang SY, Deng H. A P2P network traffic classification method using SVM. In: Wang GJ, Chen J, eds. Proc. of the 9th Int'l Conf. for Young Computer Scientists (ICYCS 2008). IEEE Computer Society, 2008. 398−403. [doi: 10.1109/ICYCS.2008.247]

[30] Este A, Gringoli F, Salgarelli L. On the stability of the information carried by traffic flow features at the packet level. ACM SIGCOMM Computer Communication Review, 2009,39(3):13−18. [doi: 10.1145/1568613.1568616]

[31] Este A, Gringoli F, Salgarelli L. Support vector machines for TCP traffic classification. Computer Networks, 2009,53(14): 2476−2490. [doi: 10.1016/j.comnet.2009.05.003]

© 中国科学院软件研究所 http://www.jos.org.cn

[32] Roughan M, Sen S, Spatscheck O, Duffield N. Class-of-Service mapping for QoS: A statistical signature-based approach to IP traffic classification. In: Lombardo A, Kurose J, eds. Proc. of the 4th ACM SIGCOMM Conf. on Internet Measurement. Sicily: ACM Press, 2004. 135−148. [doi: 10.1002/scj.20283]

[33] Mori T, Uchida M, Goto S. Flow analysis of Internet traffic: World Wide Web versus peer-to-peer. Journal Systems and Computers in Japan, 2005,36(11):70−81.

[34] Basher N, Mahanti A, Williamson C, Arlitt M, Mahanti A,. A comparative analysis of Web and peer-to-peer traffic. In: Huai JP, Chen R, Hon HW, eds. Proc. of the 17th Int'l Conf. on World Wide Web. New York: ACM Press, 2008. 287−296. [doi: 10.1145/1367497.1367537]

[35] Chen QZ, Shao B, Chen C. Design and implementation of P2P traffic identification system based on compound characteristics. Journal of Southest University (natural science edition), 2008,38(S1):109−113 (in Chinese with English abstract).

[36] Moore AW, Zuev D. Internet traffic classification using bayesian analysis techniques. ACM SIGMETRICS Performance Evaluation Review, 2005,33(1):50−60. [doi: 10.1145/1071690.1064220]

[37] Li W, Canini M, Moore AW, Bolla R. Efficient application identification and the temporal and spatial stability of classification schema. Computer Networks, 2009,53(6):790−809. [doi: 10.1016/j.comnet.2008.11.016]

[38] Erman J, Mahanti A, Arlitt M, Cohen I, Williamson C. Offline/Realtime traffic classification using semi-supervised learning. Performance Evaluation, 2007,64(9-12):1194−1213. [doi: 10.1016/j.peva.2007.06.014]

[39] Chhabra P, John A, Saran H. PISA: Automatic extraction of traffic signatures. In: Boutaba R, Almeroth K, Puigjaner R, Shen S, eds. Proc. of the 4th Int'l IFIP-TC6 Networking Conf. Ontario, Heidelberg: Springer-Verlag, 2005. [doi: 10.1007/11422778_59]

[40] Erman J, Mahanti A, Arlitt M, Williamson C. Identifying and discriminating between Web and peer-to-peer traffic in the network core. In: Williamson C, Zurko ME, eds. Proc. of the 16th Int'l Conf. on World Wide Web (WWW 2007). New York: ACM Press, 2007. 883−892. [doi: 10.1145/1242572.1242692]

[41] Jiang H, Moore AW, Ge ZH, Jin SD, Wang J. Lightweight application classification for network management. In: Proc. of the 2007 SIGCOMM Workshop on Internet Network Management (INM 2007). New York: ACM Press, 2007. 299−304. [doi: 10.1145/1321753.1321771]

[42] Pietrzyk M, Costeux JL, Urvoy-Keller G, En-Najjary T. Challenging statistical classification for operational usage: The ADSL case. In: Feldmann A, Mathy L, eds. Proc. of the 9th ACM SIGCOMM Conf. on Internet Measurement Conf. (IMC 2009). New York: ACM Press, 2009. 122−135. [doi: 10.1145/1644893.1644908]

[43] McGregor A, Hall M, Lorier P, Brunskill J. Flow clustering using machine learning techniques. In: Barakat C, Pratt I, eds. Proc. of the Passive and Active Network Measurement (PAM). LNCS 3015, Heidelberg: Springer-Verlag, 2004. 205−214. [doi: 10.1007/978-3-540-24668-8_21]

[44] Zander S, Nguyen T, Armitage G. Automated traffic classification and application identification using machine learning. In: Hassanein H, Waldvogel M, eds. Proc. of the IEEE Conf. on Local Computer Networks (LCN 2005). Sydney: IEEE Computer Society Press, 2005. 250−257. [doi: 10.1109/LCN.2005.35]

[45] Bernaille L, Teixeira R, Akodkenou I, Soule A, Salamatian K. Traffic classification on the fly. ACM SIGCOMM Computer Communication Review, 2006, 36(2):23−26. [doi: 10.1145/1129582.1129589]

[46] Bernaille L, Teixeira R, Salamatian K. Early application identification. In: Diot C, Ammar M, eds. Proc. of the 2006 ACM CoNEXT Conf. New York: ACM Press, 2006. 1−12. [doi: 10.1145/1368436.1368445]

[47] Shu X, Yang JH, Zhang DF, Xie GG. Compare and analysis of clustering algorithms oriented traffic identification system. Computing Technology and Automation, 2008,27(3):1−6 (in Chinese with English abstract).

[48] Erman J, Arlitt M, Mahanti A. Traffic classification using clustering algorithms. In: Proc. of the 2006 SIGCOMM Workshop on Mining Network Data (MineNet 2006). New York: ACM Press, 2006. 281−286. [doi: 10.1145/1162678.1162679]

[49] Xu P, Liu Q, Lin S. Internet traffic classification using support vector machine. Journal of Computer Research and Development, 2009,46(3):407−414 (in Chinese with English abstract).

[50] Xu P, Lin S. Internet traffic classification using C4.5 decision tree. Journal of Software, 2009,20(10):2692−2704 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/20/2692.htm [doi:10.3724/SP.J.1001.2009.03444]

[51] Soysal M, Schmidt EG. Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. Performance Evaluation, 2010,67(6):451−467. [doi: 10.1016/j.peva.2010.01.001]

[52] Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. ACM SIGCOMM Computer Communication Review, 2006,36(5):5−15. [doi: 10.1145/1163593.1163596]

[53] Zhu XJ. Semi-Supervised learning literature survey. 2008. http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey_7_19_2008.pdf

© 中国科学院软件研究所 http://www.jos.org.cn

[54] Qian F, Hu GM, Yao XM. Semi-Supervised Internet network traffic classification using a Gaussian mixture model. AEU—Int'l Journal of Electronics and Communications, 2008,62(7):557−564. [doi: 10.1016/j.aeue.2007.07.006]

[55] Peng LZ, Zhang HL, Yang B, Chen YH, Qassrawi MT, Lu G. Traffic identification using flexible neural trees. In: Proc. of the 18th Int'l Workshop on Quality of Service (IWQOS). Beijing, 2010. 1−5. [doi: 10.1109/IWQoS.2010.5542729]

[56] Lu G, Zhang HL, Sha XF, Chen C, Peng LZ. TCFOM: A robust traffic classification framework based on OC-SVM combined MC-SVM. In: E.Guerrero J, ed. Proc. of the Int'l Conf. on Communications and Intelligence Information Security (ICCIIS). Nanning: IEEE Computer Society, 2010. 180−186. [doi: 10.1109/ICCIIS.2010.57]

[57] Karagiannis T, Broido A, Faloutsos M, claffy K. Transport layer identification of P2P traffic. In: Lombardo A, Kurose J, eds. Proc. of the 4th ACM SIGCOMM Conf. on Internet Measurement. New York: ACM Press, 2004. [doi: 10.1145/1028788.1028804]

[58] Constantinou F, Mavrommatis PBI. Identifying known and unknown peer-to-peer traffic. In: Bilof R, ed. Proc. of the 5th IEEE Int'l Symp. on Network Computing and Applications. New York: IEEE Computer Society, 2006. 93−102. [doi: 10.1109/NCA. 2006.34]

[59] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: Multilevel traffic classification in the dark. In: Guerin R, Govindan R, Minshall G, eds. Proc. of the 2005 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM 2005). New York: ACM Press, 2005. [doi: 10.1145/1080091.1080119]

[60] Stutzbach D, Rejaie R. Understanding Churn in peer-to-peer networks. In: Almeida J, Almeida V, Barford P, eds. Proc. of the 6th ACM SIGCOMM Conf. on Internet Measurement (IMC 2006). New York: ACM Press, 2006. 189−202. [doi: 10.1145/1177080. 1177105]

[61] Zhang YX, Yang D, Zhang HK. Research on Churn problem in P2P networks. Journal of Software, 2009,20(5):1362−1376 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/20/1362.htm [doi: 10.3724/SP.J.1001.2009.03485]

[62] Zhou LJ. The research on identification of P2P streaming traffic [Ph.D. Thesis]. Wuhan: Huazhong University of Science and Teehnology, 2008 (in Chinese with English abstract).

[63] Xu P, Liu Q, Lin S. An improved transport layer identification of peer-to-peer traffic. Journal of Computer Research and Development, 2008,45(5):794−802 (in Chinese with English abstract).

[64] Gallagher B, Iliofotou M, Eliassi-Rad T, Faloutsos M. Link homophily in the application layer and its usage in traffic classification. In: Proc. of the INFOCOM. San Diego: IEEE, 2010. 1−5. [doi: 10.1109/INFCOM.2010.5462239]

[65] Madhukar A, Williamson CL. A longitudinal study of P2P traffic classification. In: Ceballos S, ed. Proc. of the 14th IEEE Int'l Symp. on Modeling, Analysis, and Simulation. Washington: IEEE Computer Society Press, 2006. 179−188. [doi: 10.1109/MASCOTS. 2006.6]

[66] Kim H, Claffy KC, Fomenkov M, Barman D, Faloutsos M, Lee K. Internet traffic classification demystified: Myths, caveats, and the best practices. In: Azcorra A, Veciana G, eds. Proc. of the 2008 ACM CoNEXT Conf. New York: ACM Press, 2008. 1−12.

[67] Zhao Y, Yao QL, Zhang ZB, Guo L, Fang BX. TPCAD: A text-oriented multi-protocol inference approach. Journal on Communications, 2009,30(10): 28−35 (in Chinese with English abstract).

[68] Chen ZX. The research on Internet traffic identification methods with scale adatability [Ph.D. Thesis]. Ji'nan: Shandong University, 2008 (in Chinese with English abstract).

[69] Keralapura R, Nucci A, Chuah C. Self-Learning peer-to-peer traffic classifier. In: Proc. of the Int'l Conf. on Computer Communications and Networks (ICCCN). San Francisco: IEEE Press, 2009. 1−8. [doi: 10.1109/ICCCN.2009.5235313]

附中文参考文献:

[9] 刘兴彬,杨建华,谢高岗,胡明.基于 Apriori 算法的流量识别特征自动提取方法.通信学报,2009,29(12):51−59.

[16] CNNIC.中国互联网络发展状况统计报告.2009. http://www.cernet.com/news/chinacngi.pdf

[17] 赵瑞.基于特征串的 P2P 流量识别研究与实现[硕士学位论文].成都:电子科技大学,2009.

[18] 李伟男,鄂跃鹏,葛敬国,钱华林.多模式匹配算法及硬件实现.软件学报,2006,17(12):2403−2415. http://www.jos.org.cn/1000-9825/17/2403.htm [doi: 10.1360/jos172403]

[19] 谭建龙.串匹配算法及其在网络内容分析中的应用[博士学位论文].北京:中国科学院研究生院,2003.

[20] Cormen T, Leiserson CE,著;潘金贵,顾铁成,李成法,叶懋,译.算法导论.第 2 版,北京:机械工业出版社,2008.557−568.

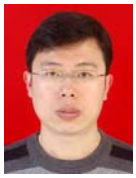
[21] Navarro G, Raffinot M,著;中科院计算所网络信息安全研究组,译.柔性字符串匹配.北京:电子工业出版社,2007.13−68.

[23] 郭振滨,裘正定.应用于高速网络的基于报文采样和应用签名的 BitTorrent 流量识别算法.计算机研究与发展,2008,45(2): 227−236.

[35] 陈庆章,邵奔,陈超.基于复合特征的 P2P 业务识别系统的研究与实现.东南大学学报(自然科学版),2008,38(S1):109−113.

© 中国科学院软件研究所 http://www.jos.org.cn

[47] 苏欣,杨建华,张大方,谢高岗.面向流量识别系统的聚类算法的比较与分析.计算技术与自动化,2008,27(3):1−6.
[49] 徐鹏,刘琼,林森.基于支持向量机的 Internet 流量分类研究.计算机研究与发展,2009,46(3):407−414.
[50] 徐鹏,林森.基于 C4.5 决策树的流量分类方法.软件学报,2009,20(10):2692−2704. http://www.jos.org.cn/1000-9825/20/2692.htm [doi:10.3724/SP.J.1001.2009.03444]
[61] 张宇翔,杨东,张宏科.P2P 网络中 Churn 问题研究.软件学报,2009,20(5):1362−1376. http://www.jos.org.cn/1000-9825/20/1362.htm [doi: 10.3724/SP.J.1001.2009.03485]
[62] 周丽娟.P2P 流媒体识别方法的研究[博士学位论文].武汉:华中科技大学,2008.
[63] 徐鹏,刘琼,林森.改进的对等网络流量传输层识别方法.计算机研究与发展,2008,45(5):794−802.
[67] 赵咏,姚秋林,张志斌,郭莉,方滨兴.TPCAD:一种文本类多协议特征自动发现方法.通信学报,2009,30(10):28−35.
[68] 陈贞翔.具有规模适应性的互联网流量识别方法研究[博士学位论文].济南:山东大学,2008.



鲁刚(1982－),男,辽宁沈阳人,博士生,主要研究领域为 P2P 技术,网络测量.

叶麟(1982－),男,博士生,主要研究领域为网络测量.

张宏莉(1973－),女,博士,教授,博士生导师,主要研究领域为网络信息安全,网络测量,并行处理.

© 中国科学院软件研究所 http://www.jos.org.cn