综述与评论 Review and Comment

基于网络流内在特征的 P2P 业务识别技术研究*

李燕霞,周井泉
(南京邮电大学 电子科学与工程学院,江苏 南京 210003)

摘 要:分析了几类主要的 P2P 业务识别方法,重点分析了基于流的内在特征的各种识别方法,并对其优缺点作出评价,指出了 P2P 识别技术进一步的发展方向。

关键词:P2P;识别技术;网络流的内在特征
中图分类号:TP393 **文献标识码**:A

Research of P2P traffic identification methods based on the inherent features of network flows

LI Yan-Xia, ZHOU Jing-Quan
(College of Electronic Science and Engineering, Nanjing University of Posts & Telecommunications, Nanjing 210003, China)

Abstract: This paper analyzes some major P2P identification methods, specially strength the methods based on the inherent features of network flows. It evaluates the methods's advantage and shortcoming, at last it gives the future research direction of P2P traffic identification.

Key words: P2P; identification methods; inherent features of network flows

当前 P2P 业务已占到网络业务的 60%~70%,成为网络带宽主要的使用者,而且 P2P 应用呈现快速增长的趋势。P2P 的飞速发展,一方面给用户带来方便,提供了多种类的业务,但另一方面也带来了许多负面的问题。如:P2P 文件共享过程中的版权问题;应用中大量占用网络带宽的问题;流量模式对传统网络设计带来的挑战等。因此,如何有效地识别 P2P 业务并进一步对其进行管理,对互联网业务提供者或运营商(ISP)及企业用户等已成为迫切需要解决的问题。

P2P(Peer-to-Peer)即点对点技术。该技术与传统网络技术显著的不同点在于其结构模式的不同。传统技术采用客户机/服务器(C/S)结构,而 P2P 从总体上来说是一种分布式网络模型。C/S 模式中,1 台计算机想获取服务,需向存储大量数据的服务器提出请求才能获得服务,这种模式存在着明显的主从关系,当服务器过载时,服务器就变成了网络中的瓶颈。而在 P2P 中,成千上万台彼此连接的计算机都处于对等地位,整个网络一般不依赖于专用中央服务器,每个结点既是资源提供者(Server),又是资源获取者(Client)[1]。正是这种结构促使 P2P 业务能充分利用各台主机的资源,包括带宽、存储空间和计算能力而得到了快速发展。

目前 P2P 的应用主要有如下几个方面[2]:文件共享类、视频点播类和即时通信类等。文件共享类包括迅雷,BT、电驴等,视频点播类包括 PPlive 等,即时通信类包括 QQ、MSN、SKYPE 等。此外,P2P 还应用于分布式计算、协同工作等领域。

P2P 业务在给网络用户带来便利的同时也产生了一些问题。据估计,在现今互联网带宽中,P2P 业务已占据大部分,且 P2P 用户经常是长时间占用带宽。因为大量占用带宽,从而影响了用户正常的网络使用业务,如浏览网页、收发邮件等。同时还带来了一些其他的问题,如安全、版权问题等,同时 P2P 的结构模式对现有网络设计规划也带来影响。因此,就需要对 P2P 业务进行有效的管理,而识别 P2P 业务是进行管理的前提。

1 早期 P2P 业务识别技术

1.1 端口法

最早出现的 P2P 业务具有固定的默认端口,因此用

* 基金项目:国家 863 项目(编号 2009AA01Z202),江苏省科技支撑项目(编号 BE2008134)

端口识别法能很容易地将P2P业务识别出来,但不能适应业务发展的要求,准确率也低。

1.2 签名匹配法[3]

这种方法的出发点是在P2P的各种协议中具有特定的报文信息,可根据这些特定的报文信息来识别每种具体的P2P应用。其识别准确率高,尤其是能识别具体的应用。但只能对已有的业务进行识别,对加密的业务不能识别,会涉及到隐私问题,效率不高。

以上2种识别方法对新业务的适应性都不是很好,若能提取P2P业务内在的固有的特征,则有可能从根本上解决问题,这也是目前研究的方向。

2 基于各种内在特征的P2P业务识别方法

2.1 利用传输层信息

传输层连接模式特征识别法[4]是基于P2P网络的连接模式,而不是基于分组内容。因此,不依赖于P2P协议具体的数据内容就能进行识别,而且也能识别出以前未知的P2P协议。此法只统计用户分组的首部信息,即源IP地址、目的IP地址、协议类型、源端口、目的端口,这是因为P2P业务所用的传输协议及其连接的{IP,Port}对其独有的、有别于其他业务的特征可以来识别P2P业务。以此为基础,Thomas Karagiannis等人给出了如下两种启发式方法:

(1)TCP/UDP混用

识别出那些同时使用TCP和UDP进行数据传输的应用。P2P节点一般在初始连接阶段采用UDP来发送控制信息,而采用TCP来传输数据。通常的应用极少同时使用UDP和TCP。因此,可以利用这个特征来识别P2P流。研究表明,大约2/3的P2P协议同时使用TCP和UDP协议,而其他应用中,同时使用2种协议的只有NetBIOS、游戏、视频等少量应用,而这些应用通常有固定的使用端口,因此能很容易将其区分出来。如果某个应用同时使用TCP和UDP作为传输协议,那么可以认为这个应用很有可能就是P2P应用。

(2){IP,Port}对的连接模式

P2P网络是分布式或混合式的,根据不同的网络,P2P节点可能会存储网络中其他节点、服务器或超级节点的IP地址。这一方法的基本依据是:在混合式P2P结构中,存在着超级节点,这种节点存储了其所在组内的其他节点的信息,既防止了组内某网络节点突然断开时造成的信息丢失,也方便了新的节点的加入。当1个新的主机A加入P2P系统后,它将通知超级节点其IP地址以及接受连接的端口号Port。超级节点查询缓存中满足A请求的节点,并将满足条件的节点信息返回给A节点,A节点收到返回的信息后将直接与满足条件的节点建立连接、传送数据。这样,对端口Port而言,与其建立连接的IP地址数目就等于与其建立连接的不同端口数目(因为不同主机选择同一端口与主机A建立连接的可能性是很低的,完全可以忽略不计)。而其他一些应用如Web,1个主机通常使用多个端口并行接收对象,这样建立连接的IP地址数目将远小于端口数目。但是另外一些应用,如MAIL、DNS等,也具有类似的属性,因此,使用这种方法在实际识别过程中需要将它们区分出来。

2.2 根据P2P流统计特征

2.2.1 基于连接流量特征的识别方法[5]

此方法的出发点是在P2P网络中节点既可作为客户机也可作为服务器,即既能接受其他节点的服务也能向其他节点提供服务,以此过程中表现出的特征作为识别的依据。本文给出了2个启发式算法:

(1)连出连接数与连入连接数之比。参与P2P的节点既有大量的连入连接也有大量的连出连接,而非P2P节点要么有大量的连出连接,要么有大量连入连接,它们的出入连接必定是不平衡的。因此,统计在某段时间内某节点连入连接与连出连接之比,并与使用传统网络应用的主机的经验观测值作比较,就可以判断此节点是否参与P2P。

(2)上行流量与下行流量之比。P2P节点不仅从其他节点处获得数据,同时也为其他节点提供数据,因此,它的节点流量更多体现为上行下行基本对称的特点,而且对于2个节点A、B之间建立的同一个连接,可能A既在下载B的数据,也在给B上传数据。在整个过程中,每个节点的上行流量与下行流量都是大体对称的。而对于传统的网络应用,如HTTP等,一般都是客户发送1个请求(几千字节到几百万字节),然后服务器返回客户机所需要的数据(几千字节Kb、几兆字节或更多)。在这种网络结构中,上行流量与下行流量是不平衡的。但是一些服务器主机,如FTP服务器,同时提供用户的上传与下载,其流量特征与P2P主机类似,因而要依据端口将这些特殊的服务器排除。该方法实现简单,只是不同的阈值选取识别结果不同,需根据具体情况选取。

2.2.2 根据P2P流的内在统计特征进行识别的方法

流量模式识别法,这是在Caspian路由器中实现的一种功能。该路由器记录经过它的每条流的信息,因此可以实现基于流的流量识别和控制功能,以一种新的方式对P2P流量进行识别和控制。几种常见IP服务的流量特征如表1所示。

表1 几种常见IP服务的流量特征

服务	持续时间	平均速率	传输字节数
HTTP	短	高	中-高
VPN	长	低	高
Games	长	低	高
Streaming	长	中	高
Telnet	长	低	中
Fileshare/P2P	长	中-高	高

由表1可以看出,P2P应用的特点是持续时间长、平均速率较高以及总的传输字节数高。这与文件传输如FTP等应用有些类似。但是该类应用可以很方便地通过端口号识别出来,而且由于这些应用与用户的交互性不如Web、视频等应用高,因此,出现一定的误判对它们的流量限制不会造成大的问题。另外,根据流所包含的字节数,可以很容易将普通Web流量同P2P文件共享流量区分开。

通过分析不同应用的流量模式,可以实现识别P2P流量的目的。而且这一方法不需要对分组内部用户数据进行检查,因此不受数据是否加密的限制,扩大了其适用范围。

除了流量模式特征外,P2P业务一些其他的内在特征,如持续时间、分组数量、平均到达间隔时间[6-7]等,也可以作为识别的依据,同时也可对这些特征进行分析,进一步提取新的特征[8]。参考文献[8]提出了1个新的参数:分组大小变换频率。在这些特征的基础上可应用各种机器学习(ML)算法等对P2P业务进行识别[9-12],算法识别的精确性很高,其中C4.5识别的精确性可达99%以上,但算法的选择、特征的选取、训练数据的数量,对识别结果的精确性、识别时间等有影响,不易实现实时识别,且不能实现具体识别。

2.3 根据P2P协议特征

2.3.1 从数据信号和信令信号的特征的不同对P2P业务进行区分

参考文献[13]更多的是从网络的角度对P2P业务进行特征提取,如对P2P业务的数据下载和信令会话阶段的行为,包括传送内容大小、会话到达间隔时间以及时长等进行进一步深入的分析,提取出不同的特征。一般可以用3个不同的流特征来描述流的长度:分组数量、有效载荷字节数量、头部和有效负载字节数,但参考文献[13]只采用有效负载字节数作为流的长度。

该方法识别的依据是:在文件下载期间,1个节点一般接收到的分组数据是很大的,会达到MTU值,对接收到的数据进行确认,这时向发送节点发送的确认分组数量就会很少,即使节点向其他节点交换1个文件的片段,在这个过程中其会话是平衡的,平均有效负载一般也会达到MSS(最大段大小)。相反,在信令会话过程中具有平稳的特征,在这个过程中交换的字节和分组数量都更加地均衡。可以用1个阈值来区分这2种不同的业务,如果1个会话CTI(内容转变索引)值超过这个阈值即为数据下载业务;反之则为信令业务。可根据这一特征对不同的网络行为进行识别。

2.3.2 根据P2P协议内容重新分配特征

每个节点既是服务器也是客户机[14],因此,节点会重新分配从其他节点收到的内容,依据这个过程中表现出来的行为特征作为识别的依据。此方法在识别PPlive方面能达到99%的精确率。

2.3.3 利用P2P协议的基本特征

P2P协议的基本特征是:较大的网络直径、绝大多数节点既作为服务器又作为客户机,利用这一特征对P2P业务进行识别[15]。

2.4 基于网络中最大流的识别[16]

基于此方法的识别思想是网络中一小部分流量产生了网络中大部分的分组和字节,将网络中这一部分流量识别出来以后再对其业务类型进行分析。参考文献[16]主要识别5种业务:P2P文件共享类业务、P2P流类业务、扫描行为、蠕虫病毒行为、DOS攻击行为。所提出的各种识别的特征依据是:如短时间内连接数或业务数据的大小、时长、节点分布等,根据不同业务具有的不同的特征可将其识别出来,如P2P文件共享类业务具有连接数量大、快速传送大量业务数据、时间周期短、端口分布规律、大量节点参与业务等特征。

这种方法可解决重量级P2P协议流量识别问题,还可以识别出杀手级用户和热点文件的特征,在解决实际问题方面有重要的意义。

2.5 BLINC方法

参考文献[17]提出的BLINC方法,按照主机的行为模式作为识别依据,对节点在社会、功能、应用等3个层面进行分析,得出P2P节点行为特征,利用这些特征,识别出P2P节点。(1)社会层面:1台主机与其他主机的互动。首先是检查这台主机的活跃性,其次识别与这台主机通信的节点。(2)功能层面:捕获主机的行为特征,分析看其在网络中扮演的角色是业务提供者还是业务接收者或者两者兼有。例如,若1台主机用1个端口与其他多台主机通信,那么这台主机在这个端口上应是一个业务提供者的角色。(3)应用层面:捕获特定主机的特定端口传输层之间的互动识别业务的发起方。

实验结果表明,BLINC方法能够对网络中的80%~90%的流量进行识别,识别的准确率达95%。但这种方法难以做到实时检测,且复杂度高,易将P2P文件共享业务流与DOS攻击流混淆。

以上各种基于P2P业务内在特征的方法可直接应用,也可作为启发式算法与其他业务结合应用或作为ML算法的训练数据应用,能解决数据加密等问题,可以识别出未知的P2P业务,但不能具体识别出P2P业务。

根据对各种技术优缺点的分析,预测P2P识别技术未来的发展方向是:(1)将基于P2P内在特征的技术与以前的技术相结合,根据不同的环境、不同的要求,综合运用各种识别方法,以实现更加有效、具体的识别;(2)深入挖掘P2P业务独有的业务特征,并从已有的特征中分析出其有决定作用的核心特征,进行具体分析,发现各种主流P2P业务其各自独有的特征,以实现具体的识别。

参考文献

[1] 暴励.P2P 技术的应用与研究[J].电脑开发与应用,2009(2):67-69.

[2] 董振江.P2P 发展现状与运营方案[J].中兴通讯技术,2008(2):49-53.

[3] SUBHABRATA S, OLIVER S, DONGMEI W. Accurate, scalable in-network identification of P2P traffic using application signatures. WWW2004, NewYork, New York,ACM 1-58113-844-X/04/0005,2004:512-521.

[4] THOMAS K, ANDRE B.Transport layer identification of P2P traffic [C]. IMC'04, October 25-27, 2004, Taormina, Sicily, Italy. Copyright 2004 ACM 1-58113-821-0/04/0010 2004:121-134.

[5] 柳斌,李之棠,李佳.一种基于流特征的 P2P 流量实时识别方法 [J]. 厦门大学学报 (自然科学版),2007,46(2):132-135.

[6] CLAFFY K C, BRAUN H W. Internet traffic profiling.1-25.Caida, San diego supercomputer Center,http://www.caida.org/outreach/papers/1994/itf/,1994.

[7] LAN K C,HEIDEMANN J. A measurement study of correlations of Internet flow characteristics[J]. Computer Network, 2006,50:46-62.

[8] 周豐谷.P2P flow identidication[D].台湾科技大学,2006.

[9] ARLITT J E M, MAHANTI A.Traffic classification using clustering algorithms. in Proceedings of the 2006 SIGCOMM Workshop on Mining Network Data (MineNet06), 2006.

[10] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques. in ACM Sigmetrics, 2005:50-60.

[11] AULD A W M T, GULL S F. Bayesian neural networks for internet traffic classification. IEEE Transactions on Neural Networks,2007,18:223-239.

[12] SCHMIDT S E G, SOYSAL M. An intrusion detection based approach for the scalable detection of P2P traffic in the national academic network backbone [R]. in 7th International Symposium On Computer Networks(ISCN06), 2006.

[13] RAFFAELE B, MARCO C, RICCARDO R.Characterizing the network behavior of P2P traffic [EB]. IT-NEWS 2008, 978-1-4244-1845-9/08,IEEE,2008:14-19.

[14] LU Xing, DUAN Hai xin,LI Xing. Identification of P2P traffic based on the content redistribution Characteristic[M]. ISCIT 2007, IEEE, 1-4244-0977-2/07/:596-601.

[15] FIVOS C, PANAYIOTIS M. Identifying known and unknown Peer-to-Peer traffic [R]. Fifth IEEE International Symposium on Network Computing and Applications (NCA'06),0-7695-2640-3/06,2006.

[16] WANG Jiao, ZHOU Ya Jian, YANG Yi Xian.Classify the majority of the total bytes on the Internet [R]. 2008 International Symposiums on Information Processing, 978-0-7695-3151-9/08 IEEE,2008:68-72.

[17] KARAGIANNIS T, PAPAGIANNAKI K, FALOUTSO S M. Blinc: multilevel traffic classification in the dark. Proceeding of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications[C]. Los Angeles: ACM Press,2005:229-240.

(收稿日期:2009-07-06)