基于流特征的P2P流量识别方法研究

黄烟波，周磊戈
（中南大学 信息科学与工程学院，湖南 长沙 410083）

摘 要：P2P业务流量在对互联网应用起巨大推动作用的同时，也带来了因资源过度占用而引起的网络拥塞以及安全隐患等问题，妨碍了正常的网络业务的开展。为了保证网络能正常有序的运行，有必要对P2P流量进行控制。但是，随着动态端口和数据加密技术的出现，传统的流量识别方法面临着巨大的挑战。简要介绍了三种主要的流量识别技术，并且比较了它们的优缺点。着重对基于流特征的流量识别方法效率低下的原因做了详细的分析，分别指出了引起误报和漏报的原因，并且给出了相应的解决方案。实验证明：文中方法能够有效提高P2P流量识别效率。

关键词：P2P；流特征分析；流量识别
中图分类号：TP393 **文献标识码**：A **文章编号**：1673-629X(2009)09-0046-03

Research of Identifying P2P Protocols Based on P2P Traffic Characteristics

HUANG Yan-bo, ZHOU Lei-ge
(School of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract: The P2P traffic played a huge role in the Internet promoting. At the same time it has also brought occupation, network congestion and safety problems. It is necessary to have a control over the P2P traffic. However, with the emergence of dynamic port and SSL, the P2P traffic identification is becoming more and more challenging. This article briefly introduced the main technologies of P2P traffic identification including the routing way such as port scan, deep packet inspection and transport layer identification, and makes a heavy weight on the analysis about the reasons of low efficiency of the technique based upon P2P traffic characteristics. Finally, this method shows high efficiency in the test experiment.

Key words: peer-to-peer; traffic characterization; traffic models

0 引 言

随着P2P技术的广泛应用，P2P应用占用了大量的网路带宽，加重了网络的负担。据统计，P2P应用已占ISP业务总量的60%到80%，成为网络带宽最大的消费者[1]。因此，解决带宽拥塞的关键问题是当网络资源紧张的时候限制那些使用量大的用户，保障那些使用量小的用户，反之，当网络资源有较大空闲时则取消这些限制，让每个用户都能高效地利用线路。此时如何对网络资源进行有效控制，如何对P2P流量进行有效控制就显得相当的重要了。而如何识别P2P协议，是对P2P流量进行有效控制的关键技术。

目前P2P流量检测技术可归结为三类[2]，即常规方法，深层数据包检测技术(Deep Packet Inspection)[3]和基于传输层流量特征检测技术(Transport Layer Identification)[4]。

基于传输层流量特征的检测技术利用网络流量的流量特征如IP地址、端口数量、报文长度等信息而非payload特征来检测P2P流量的方法。与基于payload特征的方法相比，基于流量特征的方法易于检测对payload进行加密的流量以及payload特征未知的P2P流量。而且，在大规模流量环境或安全网络中，出于对安全或网络性能的考虑，一般不允许布署检测payload的网络设备，从而使得基于payload的方法不适用于绝大部分这种网络，而基于流量特征的方法不会有这种障碍。从目前国内外研究的的成果来看基于流特征的P2P流量识别方法引起了重视并取得了初步的研究成果，但是误报率和漏报率仍然较大，有待进一步研究。

收稿日期：2009-01-09；**修回日期**：2009-04-06
基金项目：国家自然科学基金项目(60673164)
作者简介：周磊戈(1984-)，男，湖南长沙人，硕士研究生，研究方向为P2P流量识别；黄烟波，教授，研究方向为网络安全、网络管理。

1 P2P流特征分析

1.1 P2P流特征简介

基于流特征的P2P流量识别方法的前提就是选

取一个合适的 P2P 流的特征。大致可分为四种:第一,P2P 主机连接的其他主机的数量较传统主机更多;第二,P2P 主机的上下行流量基本相当,有别于传统主机的流量特征;第三,P2P 主机与传统主机不同,既作为服务器又作为客户端;第四,P2P 主机的监听端口的连接特点与传统主机不同[5]。以下的文字中将简称特征一,特征二,特征三和特征四。在文献[6]中,作者利用特征一,提出了基于 P2P 主机连接其他主机的数量比传统主机更多的特征形成了一种比较简单的算法;文献[7]中作者还利用特征二形成有效的 P2P 流量识别算法进行对比实验,文献[4]中作者利用特征四和同时使用 TCP 和 UDP 协议通信的特征形成了基于流特征的流量识别方法,但是效果不好。由此可见基于流特征的流量识别方法无论利用哪种特征作为判别算法都不可避免地存在较大的误报和漏报率(相对于深层数据包探测来说)。

1.2 P2P 流特征的进一步研究

对 P2P 流特征的研究着重于对基于流特征的 P2P 流量识别方法存在较大的误报和漏报的原因的分析上。漏报指的是把 P2P 流量误认为是非 P2P 流量,从而漏报了 P2P 流量;误报是指把非 P2P 流量误认为是 P2P 流量所引起的错误判别。

误报的原因是非 P2P 应用的行为特征类似于 P2P 应用的行为特征。笔者把引起误报的应用作了粗略的统计,容易引起误判的应用有:邮件服务,游戏,DNS,FTP 和一些 TCP 攻击。

邮件协议例如 SMTP 和 POP3 的行为方式类似于特征一,因此很容易造成错误判别;邮件服务器连接很多其他邮件服务器然后传播邮件消息,但是邮件协议有它自己的特征即邮件协议,一般使用固定的端口号例如 25(SMTP),如果一个 IP 地址的所有数据包中源端口号为 25 并且有的数据包的目的端口号也为 25,那么说明该地址是一个邮件服务器,正在和其他的邮件服务器传播邮件消息,从而可以把与该地址相关的流量判定为非 P2P 流量。

游戏的行为方式,实际上就是基于 P2P 网络架构建立起来的。从连接的方式来看,大量的 IP 地址连接固定的一些 IP 地址。少数固定的 IP 地址是游戏的服务器,而其在短时间内连接的大量目的地址就是客户端。仅从连接的方式上来说很难区分游戏与其它 P2P 应用。因此利用特征一进行流量识别的算法不能准确地区分 P2P 应用和游戏的应用。但是在线游戏有自己的特点即使用 UDP 协议通信,并且数据包大小大致相等。可以利用这个特征对在线游戏和 P2P 应用进行区分。

DNS(域名解析系统)服务器是以广播的形式来连接其他主机的,运行在 TCP 和 UDP 协议之上。这种连接方式也和 P2P 应用的连接方式类似,因此不容易区别。但是 DNS 也有其本身的特点:使用固定的端口号 53,每个数据包字节数通常比较小,每个连接时间较短。

TCP 攻击的一般攻击方式为一个源 IP 地址向一个目的 IP 地址的多个端口发送大量的 SYN 包,其端口有固定的一些端口。对于现在使用中的 P2P 而言,连接时使用的端口基本不是固定的端口,基于流特征算法误判率会显著提高,解决这个问题的办法一是过滤 TCP 攻击常用的端口,二是利用目前 P2P 连接时同时使用到 TCP 和 UDP 的方法进行过滤。从目前来看第二种方法比较合适,因为目前 P2P 客户端一般是使用随机端口,使用端口过滤有可能造成漏判的发生。基于行为特征的 P2P 流量识别方法总存在一定的漏报。原因是受时间段、网络带宽、资源稀缺度等因素的影响 P2P 应用所表现出来的行为特征会有比较大的差别。从而出现把 P2P 应用判定为非 P2P 应用的错误判定。笔者针对特征一,P2P 主机连接的其他主机的数量较传统主机更多的特征,在不同的时间段,不同的网络带宽,不同的资源的条件下做了对比实验得出结论,即 P2P 应用的特征在不同情况下有显著不同。文献[8]中 Raffaele Bolla, Marco Canini 针对 P2P 应用的平均连接时间比较长的特征在大量实验的基础上提出 P2P 流量应分为 signaling 和 downloading 阶段,在这两个阶段所展现的 P2P 流特征有显著的不同,表 1 简要概括了他们的研究成果[8]。

表 1　eDonkey 和 BT 连接持续时间的均值和标准差

连接类型	均值	标准差
eDonkey - download	276.76	784.79
eDonkey - signaling	51.64	490.36
eDonkey - all	64.56	514.53
BitTorrent - download	749.43	2076.03
BitTorrent - signaling	141.10	1513.28
BitTorrent - all	293.85	1693.15

综上所述,引起基于行为特征的 P2P 流量判别算法的误判和漏判的原因是:存在类似于 P2P 应用的非 P2P 应用,和 P2P 应用行为特征的不稳定性。所以在基于行为特征设计 P2P 流量识别算法的时候应该充分考虑以上的几种情况。

2 流量识别算法的设计与性能分析

2.1 算法的基本思想

文中提出的基于流特征的识别算法是以 1.1 节提

出的特征一和特征四为基础，结合 1.2 节分析的影响识别算法效率的因素的一种综合解决方案。P2P 主机连接的其他主机的数量较传统主机更多，由于 P2P 协议自身的特点，它会与众多的用户连接并交换信息，其连接用户的数量远远多于普通用户。P2P 主机的监听端口的连接特点与传统主机不同。若有一个主机(Host A)加入 P2P 网络的时候，它告诉超级节点它的 IP 地址和监听端口号(Host A. ip, Host A. port)，超级节点必须传播这个消息(Host A. ip, Host A. port)给其他主机(Host others)，以便 Host others 连接 Host A。如果有 10 个客户端连接 Host A，那么这 10 客户端都连接 Host A 的 Host A. port，而这 10 个客户端都随机选择一个端口，即有 10 个不同的 IP 地址与 10 个不同的端口与 Host A. port 进行连接。这两个相同的值是 P2P 主机的特征，可以依据这两个值来判定 P2P 用户。然后再结合 1.2 节所分析的干扰因素，排除可能产生误报和漏报的网络流量。算法的基本思路如下：

1) 通过排除法把源地址相同的情况下目标地址、源端口相同的连接排除[9]。

2) 按照不同的源 IP 地址，把 T 分成多个子集，每个子集中元素的源 IP 地址相同，统计每个子集中不同的目的地址数记为 Ti 和不同的目的端口数记为 A。如果 Ti 小于 n1 则判定为非 P2P，如果 Ti 大于 n2 则判定为 P2P，如果 Ti 大于 n1 小于 n2 则进入第三步。

3) A 和 Ti 的商小于 n3 则将该主机判定为 P2P 主机，如果 A 和 Ti 的商大于 n3 则进一步分析该连接的其他的特征，判断是否符合上一节提出的 DNS，邮件，游戏等应用的行为特征，如果符合这些特征则将它归类为非 P2P，反之归类为 P2P。

2.2 算法中阈值的选择

上述方法涉及三个阈值选择 n1，n2，n3。n1，n2 是一个节点所同时连接的 IP 数，n1 是非 P2P 节点平均连接的 IP 数，n2 是 P2P 节点平均连接的 IP 数。通常情况下 n1 小于 n2。n3 是一个节点所产生的所有目的端口数目与所有连接 IP 数的商。n3 必大于 1，因为一个节点 A 和另外一个节点 B 通信，节点 A 至少有一个目的端口。为了获得比较准确的阈值，在实验室的环境下进行了跟踪观察，观察的结果如表 2 所示。

表 2 P2P 和非 P2P 平均连接 IP 数和平均连接数

	观察时间	平均连接 IP 数	平均连接数
BitTorrent	2 hour	78.3	1.02
eMule	2 hour	69.7	1.14
pplive	2 hour	235.5	1.04
Non－p2p	2 hour	3.8	5.13

很明显，P2P 应用平均连接 IP 数普遍很大，不同的 P2P 应用其平均 IP 数可能相差很大，但不同 P2P 应用的平均连接数比较相近，都接近 1，这充分说明利用特征一和特征四来区分 P2P 流量和非 P2P 流量的可行性。由上表可以看出非 P2P 应用的平均连接 IP 数是 3.8，因此 n1 取值 4，n2 代表的是 P2P 节点所同时连接的 IP 数，多次试验表明当 n2 取值为 10 时有比较好的区分度。n3 代表一个节点所产生的所有目的端口数目与所有连接 IP 数的商，n3 取值 1.2 效果比较好。

2.3 算法性能分析

由于很难确定校园网内使用 P2P 的主机，而且即使能够确定使用 P2P 的结点，也很难确定该结点的流量有多少是 P2P 应用产生的，多少是非 P2P 应用产生的，所以采用实验机器产生 P2P 流量。共有 6 台实验机器用作 P2P 对等点以产生足够大的 P2P 流量以检测这两种方法的鲁棒性和检测精度。这 6 台机器在不同的时段产生不同类型的 P2P 流量。实验从 2008 年 12 月 31 日晚上八点开始，每 12 小时换一种 P2P 应用，连续测试了 3 天。根据 2.2 节分析得出的阈值，将 n1 设置为 3.3，n2 设置为 10，n3 设置为 1.2。而非 P2P 的数据则是真实的用户数据，通过 IP 地址找到了 40 台不可能使用 P2P 的用户机器进行测试。实验结果如表 3 所示。

表 3 文中方法与 payload 特征法效率对比

应用程序	Payload 特征法漏报率	Payload 特征法误报率	文中方法漏报率	文中方法误报率
BT	2.45%	0	5.12%	2.34%
eMule	56.45%	0	5.32%	3.82%
pplive	2.57%	0	4.45%	4.43%
kugoo	100%	0	4.12%	3.12%
Non－p2p	0	0	4.63%	4.14%

payload 特征法在测试中误报率总为 0，说明该方法能够有效地防止非 P2P 数据被识别为 P2P，在漏报率的对比中，payload 特征法和文中方法相差不大，但 payload 特征法应用到 BitTorrent 却产生了很大的漏报率，说明 BitTorrent 特征串已经过时，Kugoo 是一种新的 P2P 应用，payload 特征法没有它的特征，所以检测不到它的流量，而双重特征法能够检测到。虽然文中方法仍然存在一定的漏报和误报，但该漏报和误报控制在可接受的水平。

3 结束语

随着 P2P 技术的不断进步，P2P 客户端的发展经历了使用固定端口、随机端口、加密报文、隧道方式等

(下转第 53 页)

所以随着实验用例中页面数的增加,时间复杂度也迅速增加。而 DBSCAN 算法的时间复杂度与实验用例的大小变化并不明显,只是在一定的水平上下波动。主要原因是 DBSCAN 算法是一种通过对局部密度分析,将相邻点聚集在一起的聚类算法[7]。在整个算法进行过程中,它只对数据库进行一次扫描。因此只要 DBMS 对相邻点的查询效率很高,DBSCAN 的效率将会很高。而 DBMS 的查询效率目前已经完全满足这一点[11]。

4　结束语

采用有向带权图表示用户会话,使得挖掘的结果包含更多的信息。运用不同的聚类算法实现页面聚类,可以为网站结构的离线优化提供更多的研究依据。实验结果表明,COBWEB 算法在准确率和时间性能上总体优于 K-means 算法和 DBSCAN 算法。但是实验数据并不具有权威性。同时,文中的有向带权图模型是否适用于大数据量,是否能应付异常数据等还需要继续研究。

参考文献:

[1] 周　军,余智学,姜元春. 基于有向带权图的 Web 用户浏览行为模型[J]. 情报理论与实践,2008,31(5):795-798.

[2] 杨怡玲,管旭东,尤晋元. 基于页面内容和站点结构的页面聚类挖掘算法[J]. 软件学报,2002,13(3):467-469.

[3] Chen M S, Park J S, Yu P S. Data nining fo r path traversalpatterns in a Web environment[J]. In: Proc of the 16th Int'l Conf on Distributed Computing Systems. Hong Kong: [s.n.], 1996:385-392.

[4] 宋擒豹,沈钧毅. Web 日志的高效多能挖掘算法[J]. 计算机研究与发展,2001,38(3):328-333.

[5] 孙吉贵,刘　杰,赵连宇. 聚类算法研究[J]. Journal of Software,2008,19(1):48-61.

[6] Ester M, Kriegal H P, Sander J. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// In: Simoudis, Evangelos, Han Jia-wei, Fayyad U M. KDD'96 Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. [s.l.]: AAAI Press, 1996.

[7] 王　庆,周俊梅,吴红伟,等. XML 文档及其函数依赖到关系的映射[J]. 软件学报, 2003(7):1275-1281.

[8] 黄　松,刘晓明,宋自林. 基于归纳化会话的网络用户的聚类[J]. 计算机研究与发展,2001,38(10):1224-1228.

[9] 郭　岩. 用户兴趣空间的 Web 页面聚类[J]. 微电子与计算机,2003(8):10-14.

[10] 宋摘豹,沈钧毅. Web 页面和客户群体的模糊聚类算法[J]. 小型微型计算机系统,2001,22(2):229-231.

[11] 苏　中,马少平,杨　强,等. 基于 Web-Log Mining 的 Web 文档聚类[J]. 软件学报,2002,13(1):99-104.

(上接第 48 页)

阶段。人们识别基于 P2P 协议报文的困难变得越来越大,一些传统的识别方法,比如端口识别法、应用层协议内容识别法的识别效果已经不是很好了。文中提出的方法能够检测使用动态端口,加密传输的 P2P 流量,在算法的执行效率上也有一定程度的提高。挖掘更深层次的流特征,把各种流特征有机地结合起来,运用人工神经网络的思想,对 P2P 流量进行识别,是最新的研究方向。

参考文献:

[1] 李江涛,姜永玲. P2P 流量识别与管理技术[J]. 电信科学,2006(7):57-60.

[2] 石　萍,陈贞翔,荆　山. 基于对等特征的 P2P 流量识别方法[J]. 中国教育网络, 2007(2):36-38.

[3] Sen S, Spatscheck O, Wang Dongmei. Accurate, scalable in-network identification of p2p traffic using application signatures[C]// In WWW '04: Proceedings of the 13th international conference on World WideWeb. New York, NY, USA: ACMPress, 2004:512-521.

[4] Karagiannis T, Broido A, Faloutsos M. Transport layer identification of p2p traffic[C]// In Proceedings of the 4th ACM SIGCOMM conference on Intrnet measurement. New York, USA: ACM Press, 2004:121-134.

[5] 蒋海明,张剑英,王青青,等. P2P 流量检测与分析[J]. 计算机技术与发展, 2008,18(7):116-119.

[6] 温　超,郑雪峰. 基于流量分析的 P2P 协议识别方法的研究[J]. 微计算机应用,2007(7):714-717.

[7] Constantinou F, Mavrommatis P. Identifying Known and Unknown Peer-to-Peer Traffic[C]// Network Computing and Applications, 2006. NCA 2006. Fifth IEEE International Symposium onVolume, 2006:93-102.

[8] Bolla R, Canini M, Rapuzzi R. On the Double-Faced Nature of P2P Traffic Department of Communication[C]// In Proceedings of the Sixteenth Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP'08). [s.l.]:[s.n.],2008:524-530.

[9] 柳　斌,李之棠. 一种基于流特征的 P2P 流量识别实时识别方法[J]. 厦门大学学报,2007(11):56-60.