文章编号：1672-2892(2011)01-0117-04

P2P流快速识别技术

郑维玮，马卫东，刘作臣，马建国

(西南科技大学 信息工程学院，四川 绵阳 621010)

摘 要：P2P软件在网络中应用广泛，如何快速有效地识别P2P数据流成为十分重要的问题。传统的P2P识别方法对当前P2P动态端口以及内容加密无能为力。文章根据P2P流包括IP包数目、UDP比例以及连接数指标等动态行为特征，结合数据挖掘分类算法，提出了一种基于距离判决函数的判决算法，并对该算法进行实验验证。实验证明这种算法能对数据流进行高效的判决和预测。通过该方法，对网络中用户使用P2P软件可以进行有效的快速识别，达到对P2P监控的目的。

关键词：对等网络；流量识别；动态特征

中图分类号：TN915；TP393 **文献标识码**：A

Technology of rapid P2P stream identification

ZHENG Wei-wei，MA Wei-dong，LIU Zuo-chen，MA Jian-guo

(School of Information Engineering，Southwest University of Science and Technology，Mianyang Sichuan 621010，China)

Abstract：P2P software is widely used in networks. It is a very important issue that how to identify P2P data flows fast and effectively. The traditional P2P identification methods are not suitable to widely applied dynamic port and the content encryption. Combining with data mining classification, this paper proposes a new decision algorithm based on distance decision function according to the dynamic characteristics of P2P flows, including IP packets numbers, percent of UDP and link-numbers. The experiment results show that P2P data flows can be identified and predicated quickly and effectively using this algorithm, therefore P2P flows can be monitored.

Key words：Peer to Peer；flow identification；dynamic characteristics

目前，对等计算(Peer-to-Peer)应用极其广泛，成为当前研究的一个热点。P2P技术是一种对等网络技术，它是一种用于不同PC用户之间，不经过中间设备直接交换信息的技术，即每台PC可以直接连接到其他PC，并进行文件交换或联系，而不需要连接到服务器后再进行浏览与下载。对等计算极大地改善了网络的性能，提高了服务质量，减轻了服务器的压力，系统的可扩展性得到了显著提高。实际上，P2P网络上的所有设备都可以建立对话，每个P2P的结点既是客户机也是服务器[1]。P2P使人们在Internet上的共享以更平等、更自由、更主动的方式参与到网络活动中。然而，随着音视频、大容量文件传输等网络传输需求的增加，对基于P2P的大量音视频共享的需求呈指数增长。据统计，P2P对网络的带宽已经占用了50%~60%，极端的情况会达到80%~90%。P2P消耗了大量的带宽资源，导致ISP运营商主干链路拥塞，其他互联网业务性能显著下降。虽然运营商每年都在不断进行扩容，但是增加的带宽又会迅速地被P2P吞噬。P2P已经成为困扰世界网络带宽的一个重要问题。传统的识别方法主要有端口识别法、应用层特征识别法、流量模式识别法以及连接模式识别法。目前，P2P应用从最初的采用固定端口发展到使用可变端口甚至使用其他应用的端口进行数据传输，在传输的具体内容方面也从使用明文传输发展到对传输数据进行加密处理。本文针对传统识别方法的不足，根据目前P2P软件的动态行为特征，提出了一种基于距离函数的识别方法。该方法对P2P数据流能够做出快速、高效的识别。

1 传统P2P识别方法

1.1 端口识别法

收稿日期：2010-03-25；修回日期：2010-06-08

基金项目：国家“863”计划项目(2007AA01Z151)

在 P2P 应用兴起的早期，大多数应用使用的都是固定端口，例如，Gnutella 使用 6346-6347 端口，BitTorrent 使用 6881-6889 端口等。如果端口号与某些特定的端口号匹配，说明该数据流即为 P2P 流。这种识别方法最大的优点就是简单易行，它不需要进行复杂的处理即可准确分类。在 P2P 应用出现的初期它显得十分有效，但是随着 P2P 技术的发展，该方法逐渐变得不再适用[2]。

1.2 应用层特征识别法

这种识别方法不是通过固定端口对 P2P 识别，针对每种 P2P 数据流中携带有特定的报文信息，例如，HTTP 协议报文中会出现 GET,PUT,POST 等报文字样，与之相类似，在各种 P2P 应用协议中也具有类似的信息[3]，因此，有通过检查数据流内部携带的负载信息进行识别的方法。例如，Gnutella 的连接建立报文具有下述格式："GNUTELLA CONNECT/<protocol version string>\n\n"，而应答报文格式如下："GNUTELLA OK\n\n"。根据这些以及其他类似特征，即可确定某个流是否为 P2P 流[4]。特征码识别是国内网康、百卓等流量监控公司经常使用的手段。

1.3 流量模式识别法

这种流量模式的识别法在一些路由器中已经实现，通过记录经过它的每条流的信息，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。表 1 列举了几种比较常见的网络服务流量特征。

从表 1 可以看出，P2P 应用的特点是持续时间长，平均速率较高，总的传输字节数高。这一方法不需要对数据流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。目前国内较为知名的网管软件——聚生网管系统就是采用这种方法，从而使得封堵 P2P 软件较其他网管软件有明显的优势。但是，由于需要记录每条流的信息，这种方法对内存空间以及处理速度都提出了比较高的要求。

表 1 不同协议的流量特征

Table1 Traffic features of different protocols

protocol	duration	average speed	bytes transmitted
HTTP	short	high	middle-high
Games	long	low	high
Telnet	long	low	middle
Fileshare/P2P	long	middle-high	high

1.4 连接模式识别法

该方法的基本思想是：基于数据源和目的 IP 地址的连接模式的识别。一些模式是 P2P 所独有的，因此，可以直接将 P2P 流量识别出来；另外一些模式由 P2P 和其它少数应用所共有，这时可以根据对应 IP 地址的流历史以及其他特征来减少误判概率。有 2 种基本方法：1) 识别出那些同时使用 TCP 和 UDP 进行数据传输的源—目的 IP 地址对。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、视频等少数应用。2) 基于监测{IP，端口}对的连接模式。例如浏览 Web，一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[5]。

2 基于动态行为特征识别法

基于端口和应用层特征的 P2P 识别方式不够灵活，已经不能适应当前 P2P 识别需求；文中识别法本质上属于连接模式的识别。P2P 流识别问题本质上是一个流分类问题，本文基于动态特征的 P2P 流的快速识别，不同于传统的识别方式，具有数据量小、精确度高、易于实现、有利于工程实现的特点。

2.1 实验网络拓扑结构

实验中网络拓扑结构如图 1 所示。若干终端主机通过交换机与外网联系，通过对交换机做端口映射，将用户流量数据映射到 P2P 判决器上，通过 P2P 判决器的判断对数据流做出判断。

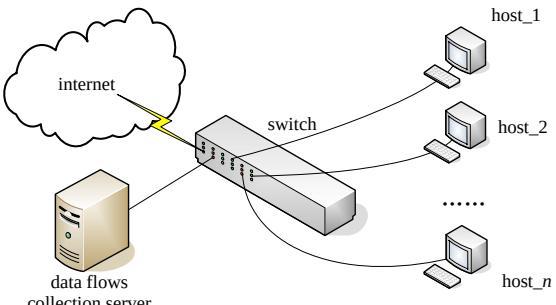


Fig.1 Network topology

图 1 网络拓扑图

2.2 数据采集与预处理

使用 wireshark 流量采集工具对 SWITCH 做端口映射，对 HOST_1…HOST_*N* 进行流量采集。

形成 IP 分组向量空间 $\boldsymbol{S}_i$=(S_IP$_i$, D_IP$_i$, Type$_i$, Class$_i$)，i=1,2,⋯,n，其中 S_IP$_i$ 为源 IP 地址，D_IP 是目的 IP$_i$ 地址，Type$_i$ 是传输层协议类型，Class$_i$ 是流量类型。

在 30 s 内，统计具有相同 S_IP/D_IP 的 $\boldsymbol{S}_i$，形成流向量空间 $\boldsymbol{L}_i=(X_i,Y_i,Z_i,P_i,P/N)$，$i$=1,2,⋯,24。$X$ 表示数据包总数，Y 表示 UDP 所占百分比，Z 表示具有相同源 IP、不相同目的 IP 特征的链接数，P 表示 P2P，N 表示非 P2P，i 表示第 i 个 IP 的向量。

2.3 基于动态行为判决模型

定义 1　特征距离：

$$D=|X-\bar{X}|+|Y-\bar{Y}|+|Z-\bar{Z}| \tag{1}$$

式中：X,Y,Z 分别为各特征属性；$\bar{X},\bar{Y},\bar{Z}$为X,Y,Z的平均值；D 为各特征属性到均值额距离之和，其物理意义是流量与流量均值的差异性。

定义 2　判决函数：

$$F(X,Y,Z)=\left(\frac{D_{\mathrm{p}}}{D_{\mathrm{n}}}\right)^2=\left(\frac{|X-\bar{X}_{\mathrm{p}}|+|Y-\bar{Y}_{\mathrm{p}}|+|Z-\bar{Z}_{\mathrm{p}}|}{|X-\bar{X}_{\mathrm{n}}|+|Y-\bar{Y}_{\mathrm{n}}|+|Z-\bar{Z}_{\mathrm{n}}|}\right)^2 \tag{2}$$

式中：$\bar{X}_{\mathrm{p}},\bar{Y}_{\mathrm{p}},\bar{Z}_{\mathrm{p}}$ 分别为 P2P 的 X,Y,Z 的平均值；$\bar{X}_{\mathrm{n}},\bar{Y}_{\mathrm{n}},\bar{Z}_{\mathrm{n}}$ 分别为非 P2P 的 X,Y,Z 的平均值；D_{p} 与 D_{n} 分别为 P2P 与非 P2P 特征距离。

$F(X,Y,Z)$）由 D_{p} 与 D_{n} 之比确定，D_{p} 和 D_{n} 分别表示了数据流与 P2P 流均值和非 P2P 流均值的相似度。然后对两者求比例后平方，进一步放大了两者的相似度的差异。根据判决函数对 P2P 流做出识别，基于动态行为特征判决模型框架如图 2 所示。

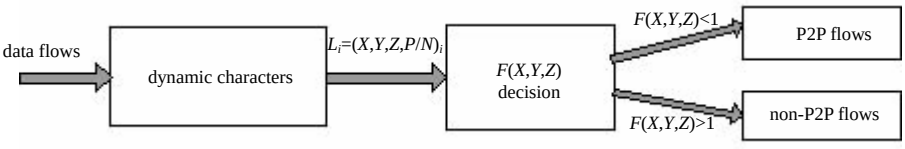
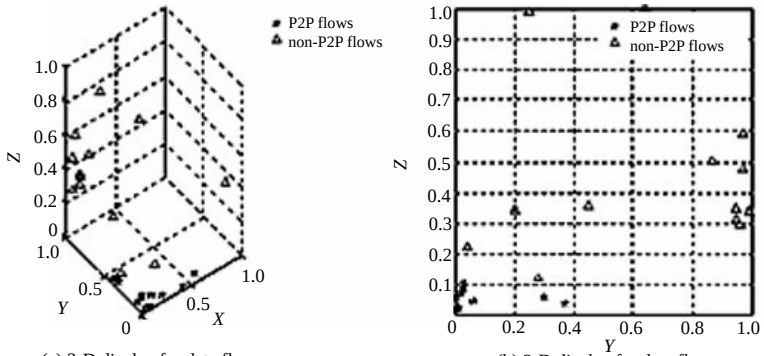


Fig.2 Model of dynamic-character decision
图 2 基于动态特征判决模型

数据流通过特征提取器后相同的源 IP 数据包形成向量空间，经过判决函数 $F(X,Y,Z)$ 判决后，如果 $F(X,Y,Z)<1$，说明数据流与 P2P 流的差异小于非 P2P 流的差异，也就是该流量更接近 P2P，所以判决为 P2P 流；反之 $F(X,Y,Z)>1$，说明数据流与 P2P 流的差异大于非 P2P 流的差异，所以判决为非 P2P 流量。



(a) 3-D display for data flows　　(b) 2-D display for data flows

Fig.3 Display of data flows
图 3 数据流的显示

3　实验验证

根据文献[6]对属性值标准化，如：$X_i / X_{\max}$，Y 用小数表示百分数，Z 的处理与 X 类似。使用 Matlab 将向量空间在三维坐标上绘出点(如图 3(a)所示)；在 weka 中进行相关性分析，X 属性与判决的相关性较小，删除 X 属性，从而将三维的流分布图降为二维表示(如图 3(b)所示)。消除个别离群点的影响，选取特征最不明显的 3 个 P2P 流，对每个属性求均值，得到具有一般特点的 P2P 流属性：$\bar{L}_{\mathrm{p}}=(\bar{Y}_{\mathrm{p}},\bar{Z}_{\mathrm{p}},P)$。同理，求得具有一般特点的非 P2P 流属性：$\bar{L}_{\mathrm{n}}=(\bar{Y}_{\mathrm{n}},\bar{Z}_{\mathrm{n}},N)$。

将向量空间 $\boldsymbol{L}_i=(Y,Z,P/N)_i$ (i=1,2,⋯,24)代入式(2)，求出 $F(X,Y,Z)$ 值，对捕获数据流进行判决。如图 4 所示，圆点表示判决函数值大于 1，判决为非 P2P 流；方块表示判决函数值小于 1，判决为 P2P 流(例如，横坐标“2”对应的一个圆点表示第 2 个非 P2P 流)。通过判决函数 $F(Y,Z)$，P2P 流和非 P2P 流均能进行识别。

本识别方案在下列 3 个方面有较大改善：

1) 准确度：传统的 P2P 流识别是基于端口和特征关键字的，随着端口的跳变和加密技术的应用，传统 P2P

流量识别的效率大幅度下降。而本方案将目光聚焦在 P2P 动态行为特征上，特征属性选取鲜明，识别精确度高。

2) 实时性：经典的基于流量模式和连接模式识别观测时间长，本方法在进行流分类时，提取一个短时间数据流，提取少量特征，这样也提高了分类的实时性能。

3) 健壮性：本方案可以随着网络流量灵活地调整 X,Y,Z 均值，以提高识别效率。与其他复杂的分类算法相比，本判决函数形式简洁，便于根据流量变化自适应调整。

本方案以 P2P 流的动态行为特性作为流量特征，不但能提高整个识别系统的精确度，而且能有效地识别协议伪装流和加密数据流，达到了较好的识别效果。

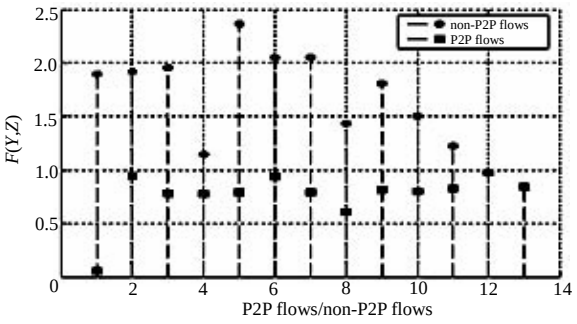


Fig.4 Decision function

图 4 判决函数

4 结论

文章提出基于动态行为特征的 P2P 流的识别算法，对于图 4 所示的共 24 组典型的 P2P 流和非 P2P 流向量，均能够正确识别。另一方面，该方法对数据量需求较小，较利于快速判别。基于统计动态行为特征的流识别技术可以很好地与传统技术结合，例如：首先对 P2P 提供预先的判断，然后再进行深度扫描，进而对数据流做出精确判决。这种根据流动态特征与距离分类的判决方法较多数识别方法流判决速度快，判决规则简单，有利于工程应用。本文识别方法可以用于核心路由器，对 P2P 做出快速判断，保证网络的传输质量。通过对实验室小数据量的测试证明本识别方式的可行性，下一步将对更大规模的数据流进行实验研究，进而建立更加适用的识别模型。

参考文献：

[1] 蒋瑜,刘嘉勇,李波. 基于信誉的 P2P 网络信任模型研究[J]. 信息与电子工程, 2007,5(6):452–453. (JIANG Yu,LIU Jiayong,LI Bo. Recommendation-based Trust Model in P2P Network[J]. Information and Electronic Engineering, 2007,5(6): 452–453.)

[2] Sen S,Jia W. Analyzing Peer-to-Peer traffic across large networks[J]. IEEE/ACM Transactions on Networking, 2004,12(2): 219–232.

[3] 邓远林,方勇,张野,等. 基于 P2P 网络的局域网监控审计模型[J]. 信息与电子工程, 2007,5(2):130–133. (DENG Yuanlin, FANG Yong,ZHANG Ye,et al. LAN-based P2P network monitoring audit model[J]. Information and Electronic Engineering, 2007,5(2):130–133.)

[4] Madhukar A,Williamson C. A Longitudinal Study of P2P Traffic Classification[C]// Proceedings of 14th IEEE International Symposium on Modeling,Analysis and Simulation of Computer and Telecommunication Systems. Monterey,California:IEEE, 2006:179–188.

[5] Karagiannis T,Broido A,Faloutsos M,et al. Transport Layer Identification of P2P Traffic[C]// IMC'04, Taormina,Italy: [s.n.], 2004:25–27.

[6] 威滕. 数据挖掘实用机器学习技术[M]. 董琳,译. 北京:机械工业出版社, 2005. (Witten l H. Data Mining Practical Machine Learning Technology[M]. Translated by Dong Lin. Beijing:Mechanical Industry Press, 2005.)

作者简介：

郑维玮(1983–)，男，福建省三明市人，在读硕士研究生，主要研究方向为网络业务流量识别. email:zheng252635@163.com.

刘作臣(1986–)，男，山东省聊城市人，在读硕士研究生，主要研究方向为网络业务流量识别.

马卫东(1968–)，男，河北省玉田县人，副研究员，博士，主要研究方向为非线性复杂网络、语义信息处理、嵌入式系统.

马建国(1957–)，男，四川省梓潼县人，博士，教授，主要研究方向为信息共享技术.