

基于双层特征的 P2P 流量检测

王春枝, 李 涛

(湖北工业大学 计算机学院, 湖北 武汉 430068)

摘 要: P2P 应用的流行和泛滥, 占用带宽, 吞噬大量的网络资源。为了更好地识别和控制 P2P 流量, 介绍了目前常用的 P2P 流量检测技术, 提出了一个基于应用层有效载荷特征和传输层流量特征的双层特征的高效混合检测方法, 介绍了选取的流量特征并总结出相应的数学公式, 详细介绍了系统流程以及端口匹配模块、流量特征匹配模块和 payload 特征匹配模块的实现。通过实验室环境下测试出该方法和单独采用深度数据包扫描方式在虚警率和误报率两方面的数据表明该方法拥有更高的识别率和准确率。

关键词: P2P; 流量检测; 有效载荷; 流量特征

中图分类号: TP31

文献标识码: A

文章编号: 1673-629X(2009)07-0238-04

P2P Traffic Identification Based on Double Layer Characteristics

WANG Chun-zhi, LI Tao

(Computer School, Hubei University of Technology, Wuhan 430068, China)

Abstract: Increasing amount of P2P services not only consume a lot of bandwidth but also influence the performance of other business. In order to identify and control P2P flow better, introduced the common method of P2P traffic identification and proposed a high efficient and mixed method to identify P2P traffic which based on both payload character of application layer and traffic character of transport layer, and then introduce the character choose and summarized as some formulae, introduce system flow chart, the implementation of the port matching module, the flow character matching module and the payload character matching module minutely. In the end, the false alarm rate and false positives rate tested in lab indicating that P2P traffic identification based on double layer characteristics had higher identify rate and more exact.

Key words: peer to peer; flow identification; payload; traffic character

0 引 言

对等计算 (Peer to Peer, 简称 P2P) 广泛应用于文件共享、即时通讯、协同工作等方面。P2P 技术允许终端用户利用 Internet 架构动态、匿名、分布式网络相互传递信息, 颠覆了传统的 C/S 信息服务模式, 允许终端用户相互直接搜索和共享信息资源。P2P 应用所产生的流量具有分布非均衡、上下行流量对称、流量隐蔽、数据集中等特性, P2P 业务的不断增加, 给网络带宽造成巨大消耗, 引起网络拥塞, 降低了其它业务的性能。对于企业用户来说, 内部员工使用企业网下载造成了巨大的资源浪费, 而通过不安全的网络环境获得的应用程序和 P2P 协议, 将为企业安全打开一扇后窗, 病毒和恶意代码可能潜入企业内部网络。因此只

有准确、快速地检测出 P2P 流量, 才能对 P2P 应用进行恰当地控制^[1]。

1 P2P 流量检测技术现状及研究

P2P 流量检测技术大致有以下三类: 基于端口的检测技术, 深层数据包检测技术 (DPI, Deep Packet Inspection) 和基于流量特征的检测技术 (Transport Layer Identification)。

1.1 基于端口的检测技术

第一代集中式 P2P 应用都是使用固定的端口号进行通信 (如表 1 所示)。使用基于端口的检测技术可以直接识别出是何种具体的 P2P 协议 (如 eMule、Bit Torrent 等), 也可以直接排除知名的非 P2P 应用 (如 FTP、E-Mail 等)。

随着 P2P 技术的演化和改进, 许多新的反监测、反封锁技术被引入, 目前大部分的 P2P 系统已经开始支持用户自定义端口号、随机动态端口号、端口跳变技术或端口伪装技术, 使基于端口的 P2P 流量检测技术

收稿日期: 2008-11-25; 修回日期: 2009-01-20

基金项目: 湖北省教育重点项目 (D20081405, D20081402)

作者简介: 王春枝 (1963-), 女, 教授, 硕士生导师, 研究方向为计算机网络和网络安全。

的识别率和准确率都大大下降。

表 1 部分 P2P 应用的特征端口号

P2P 应用名称	TCP 特征端口号	UDP 特征端口号
gnutella	6346, 6348	6346, 6348, 6349
winMX	6699	6257
Bit Torrent	6881, 6882, 6884, 6886	
eDonkey	4662	4672
Soribada	22322, 7675, 7676, 7677	22321, 7674
Sharesware	6399	6388, 6733, 6777

1.2 深度数据包检测技术

P2P 软件的控制报文,尤其是握手报文,一般会遵循一定的格式,所以,在报文的应用层的载荷内会有一些特定的字符(如 Bit Torrent 协议的通过三次握手建立连接时,应用层用户数据部分会包含的 16 进制特征字段为:“1342 6974 546f 7272 656e 7420 7072 6f74 6f63 6f6e”,翻译成 ASC II 码,含义为:“19Bit Torrent-protocol”)[2]。那么,检测工具就是通过检测报文载荷内的这些特定字符来判断该报文是否属于某种 P2P 文件共享软件所收发的报文,进而进行控制。

DPI 技术拥有检测精度高、实现简单的优点。试验表明,其准确度可以达到 95%。但是,随着 P2P 应用的不断发展,其弊端也逐渐暴露出来:(1)需访问数据包载荷中的内容,会引发隐私方面的法律问题;(2)非开源 P2P 应用程序,抽取其特征字段困难;(3)对 IP 报文进行特征串的正则表达式的匹配,其检测效率与 Payload 特征的复杂度有关,算法性能影响大;(4)无法识别基于 SSL 等加密的 P2P 应用。

1.3 基于流量特征的检测技术

P2P 应用作为一种充分利用客户端资源的新型应用,它在传输层表现出来的流量特征相对于 HTTP、FTP、DNS 等应用有许多不同的地方。基于流量特征的检测技术就是利用传输层流量特征如 IP 地址、端口数量、报文长度、上下行流量等信息进行 P2P 流量检测的技术,该技术借用了统计学领域通用的一些概念,不需要任何关于应用层协议的信息,主要解决了端口识别和深度数据包检测的不足[3]。

其优点是突出的:(1)无需访问数据包的应用层载荷,从而避免了前面提到的法律问题和 technical 问题;(2)基于流量特征的方法更易于检测对净载荷进行加密的流以及净载荷特征未知的 P2P 流。

缺点有:(1)对 P2P 应用分类的能力较弱;(2)有可能对 P2P 流量检测的精确度造成影响,很多流量特征都不是 P2P 流量唯一的,其它应用也有可能表现出这种流量特征,需要结合其它技术,如端口检测来排除。

2 基于双重特征的流量检测方法

2.1 方法设计

在充分认识以上三种技术的优缺点之后,文中设计了一种既可以有效检测加密 P2P 应用和未知的 P2P 应用,又拥有较低检测代价、较高检测精度的基于双重特征的 P2P 流量检测法,该方法的流程图如图 1 所示。

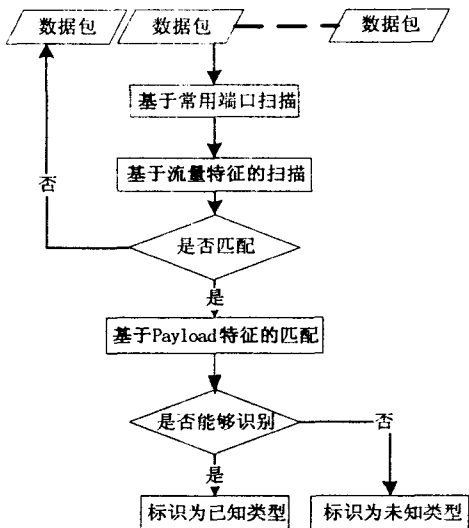


图 1 基于双重特征的流量检测流程图

该方法由三个主要步骤组成:基于常用端口匹配,基于流量特征匹配和基于 Payload 特征的匹配,其步骤如下:(1)获得一个待检测报文后,先使用常用端口过滤规则去掉使用常用端口的流量;(2)通过流特征检测模块匹配该数据流是否属于 P2P 流量;(3)对属于 P2P 流量的报文进行 Payload 特征字符串的匹配以确定其属于何种具体的 P2P 应用,如果匹配成功则标识为已知类型的 P2P 流,如果匹配失败,则标识为未知类型的 P2P 数据流,以后所有该连接的报文需要继续检测[4]。

采用这种新型的检测方法,未知和加密的 P2P 应用可以在流量特征匹配阶段进行有效的检测,从而克服了深层数据包检测技术在这方面固有的缺陷,又克服了基于流量检测技术分类能力弱的缺陷。特别是,当确定某条连接符合某种流量特征后,对该连接的报文进行的 Payload 特征串匹配并不是匹配所有的特征串,而是匹配那些符合这种流量特征的 P2P 应用的特征串;另外,由于在特征串匹配阶段该报文已经确定为 P2P 且只适合于某种 P2P 流量特征,适当地缩短或简化特征串不会影响检测精度。这两个方面使得基于双重特征检测的特征串匹配比基于 Payload 特征的方法更加高效。

2.2 流量特征的选取

选取简单、合理的 P2P 流量特征能够减少流量特

征匹配阶段的计算量,有效提高检测效率。通过对当前流行的多个 P2P 应用软件进行分析,用 ethereal 抓包工具抓取 P2P 流量,并借助于 ethereal 自带的统计功能对 P2P 流量和 C/S 流量进行研究和比较,总结出 P2P 流量一些特征^[5,6]。

P2P 应用中,每个节点 S 既担任服务器角色,也担任客户端角色,故在某段时间内 S 连入与连出的连接数之比符合某个阈值 a 可作为流量特征标准, S_{in} 表示连入连接数, S_{out} 表示连出连接数。

$$\frac{1}{a} < \frac{S_{in}}{S_{out}} < a \quad (1)$$

P2P 应用不同于传统网络应用的一点是,其节点在下载的同时,也在提供上传。而且对于两个节点 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。因此,比较某个连接的入流量和出流量之比符合某个阈值 b 可作为流量特征标准,用 T_{up} 表示出流量, T_{down} 表示入流量, T_{min} 表示流量下限。

$$\frac{1}{b} < \frac{T_{up}}{T_{down}} < b, T_{up} < T_{min}, T_{down} < T_{min} \quad (2)$$

统计某主机上所有上行流量与下行流量,当二者之比在某个范围之间,则判定该主机应用了 P2P 软件, $\sum T_{up}$ 表示出流量总和, $\sum T_{down}$ 表示入流量总和, T_{min} 表示流量总和下限。

$$\frac{1}{c} < \frac{\sum T_{up}}{\sum T_{down}} < c, \sum T_{up} < T_{min}, \sum T_{down} < T_{min} \quad (3)$$

综合选取的流量特征信息,给出如下识别方案:观测某主机一段时间内连入连接和连出连接之比,找出可能参与 P2P 的主机;在这些可疑主机的每个端口上观测上行与下行流量之比,若该端口上下行流量对称,则该端口是该主机的 P2P 端口,该端口上的所有连接都是 P2P 连接,其连接对端主机也参与 P2P,且该连接是 P2P 连接。

3 系统实现及性能测试

3.1 系统实现

系统由三个模块组成,分别为常用端口匹配模块、流量特征匹配模块和 Payload 匹配模块。

3.1.1 常用端口匹配模块

由于 P2P 应用有可能伪装成 80 端口,但使用其他低于 1024 号端口的情况并不多见,故只要端口号低于 1024 且不为 80,即认为该报文使用了常用端口,不是 P2P 报文。

3.1.2 流量特征匹配模块

具体算法流程如下:

首先针对网络中每个数据包,在列表 DataTable 中记录它的源地址、目的地址、源端口号、目的端口号、上行流量、下行流量;然后针对每个主机,计算其连入连接数和连出连接数之比,找出符合公式(1)的可疑主机;其次扫描该主机的每个端口,计算其上下行流量之比,找出符合公式(2)的端口;再次计算该主机所有符合公式(2)的端口的上下行流量的总和之比,判断该主机是否符合公式(3)。若符合,即为 P2P 节点主机,该主机符合条件(2)端口所对应的连接即为 P2P 连接^[7]。

DataTable 数据结构:

Struct DataTable

```
{
    Unsigned int sourceIP; //源 IP
    Unsigned int destIP; //目的 IP
    Unsigned long UpTraffic; //上行流量
    Unsigned long DownTraffic; //下行流量
    Ushort sport; //源端口
    Ushort dport; //目的端口
}
```

DataTable;

实现伪代码:

For each packet:

If new TCP Data (len! = 0)

Insert sourceIP, destIP, sport, dport, into DataTable; count +

+

else if UDP

Insert sourceIP, destIP, sport, dport into DataTable; count +

+

If(DataTable[i].sourceIP = DataTable[j].destIP)

&&(DataTable[i].destIP = DataTable[j].sourceIP)

If(DataTable[i].count / DataTable[j].count in the threshold)

Insert sourceIP, destIP, sport, dport into P2PTable;

If((sourceIP in P2PTable.destIP) || (destIP in P2PTable.sourceIP))

It is a P2P peer;

3.1.3 Payload 匹配模块

Payload 特征匹配模块的核心就是选取优秀的字符串匹配算法和合理的 Payload 特征(如表 2 所示)^[8]。

因为 DPI 方法中要进行匹配的特征字符串只有特定的几种,需要在大量网络数据包中进行查找,而 Karp—Rabin 算法的主要思想就是通过对字符串进行哈希运算,使得算法可以更容易地排除大量不相同的字符串,常用于单个模式字符串在多个目标字符串的同时查找,以及模糊匹配。此算法的预处理时间为 $O(m)$,它在最坏情况下的时间复杂度为 $O((2n - m + l)m)$,平均复杂度接近 $O(m + n)$,很适合作为 DPI 中的字符串匹配算法。

表 2 典型 P2P 应用 Payload 特征

应用类型	流量类型	Payload 特征
PPLive	TCP	连接建立后的第 1 个报文为 4 字节,内容为 0x39 00 00 00
	UDP	前两个字节是 0xe9 03
BitTorrent	TCP	包含:0x19BitTorrentprotocol
QQLive	UDP	第 1 个字节为:0xfe
eMule	TCP	第 1 个字节为:0xe3 或 0xc5 或 0xd4

3.2 系统测试

在实验室环境下以拥用 10 台电脑的局域网作为检测对象分别对 PPLive、BitTorrent、QQLive、eMule、Kugoo 和非 P2P 应用进行连续 24 小时测试。Payload 特征法和双重特征法采用的特征串中,前四种 P2P 使用表 2 中 Payload 特征作为匹配字符串,Kugoo 无特征字段。采用虚警率和误报率作为比较标准,实验结果如表 3 所示。

表 3 双重特征法与 DPI 法的虚警率
和误报率对比表

应用类型	数据量 (GB)	Payload 虚警率(%)	Payload 误报率(%)	双重特征 虚警率(%)	双重特征 误报率(%)
PPLive	30.6	2.20	0	3.41	0.14
BitTorrent	20.3	62.51	0	2.52	0.11
QQLive	32.5	0.51	0	0.51	0.09
eMule	10.9	2.82	0	1.95	0.02
Kugoo	2.5	100	0	4.17	0.17
传统应用	5	0	0	0	0.04

由表可知,基于双重特征的 P2P 流量检测法相比于单纯的 DPI 法拥有较低的误报率,特别是对于 Payload 特征字段落后的 Bit Torrent 和无 Payload 特征字段的 Kugoo,仍有较高的识别率。虽然双重特征的 P2P 流量检测法有一定的虚警率,但是仍一直维持在可接受的范围之内。

(上接第 237 页)

指令的成功分解。提出的具有统一架构的指令分解规则以及通用转换器定制的方法灵活适应了业务需求的变化,当有新的业务产生时,只要制定新的输入界面和按照架构定制新规则[0]就可以完成新的业务指令的分解工作,当规则发生变化时,只要按照架构修改旧规则即可,而无需对系统进行修改。此方法也为其他领域的数据库转换技术提供了新的思路和方法。

参考文献:

[1] 盛贤良,瞿有甜. 基于 XML 的异构构件组装技术研究[J]. 计算机技术与发展,2008,18(2):83-87.
[2] 元祥波,南琳,张福顺. 基于元数据和 XML 的信息抽取

4 结束语

基于双重特征的流量检测方法不仅继承了 Payload 特征法检测准确、误报率低的优点,而且通过选取合适的流量特征能够识别经过加密和未知的 P2P 流,拥有较高的识别率。

为了适应 P2P 应用进一步发展,进一步降低误报率和漏报率,需要发现新的更具代表性的 P2P 流量特征,以及采用机器学习、数据挖掘和硬件实现该检测算法的新方法,使之能够应用于大规模流量环境中,提高网络服务质量。

参考文献:

[1] 杨磊,姜浩. P2P 技术在网络文件管理中的应用[J]. 计算机技术与发展,2006,16(12):127-129.
[2] Sen S, Spatschecko, Wangdm. Accurate, Scalable In - Network Identification of P2P Traffic Using Application Signatures [C]//WWW2004. New York, USA:[s.n],2004.
[3] Karagiann I T, Brodo A, Faloutsosm. Transport Layer Identification of P2P Traffic[C]//Proceedings of the 4th ACM SIG - COMM Conference on Internet Measurement. New York: ACM Press,2004:121-134.
[4] 王锐,王逸欣. 一种跨层 P2P 流量检测方法[J]. 计算机应用,2006,26:30-32.
[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,2007,46:132-135.
[6] 温超,郑雪峰,戚翔,等. 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] 石萍,陈贞翔. 基于对等特征的 P2P 流量识别方法[J]. 网络测量与规划,2007(2):36-38.
[8] 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):74-76.

与集成技术研究[J]. 信息与控制,2008,37(1):52-57.
[3] Florescu D D, Kossmann D. Storing Querying XML Data using an RBMS[J]. IEEE Data Engineering Bulletin,1999,22(3):27-34.
[4] 王大刚,谢荣传,彭俊. 基于 XML Schema 的数据匹配方法的研究[J]. 计算机技术与发展,2008,18(6):28-31.
[5] 黎建辉,吴威,阎保平. 一种基于 XML 的元数据映射与转换方法[J]. 微电子学与计算机,2008,25(1):34-38.
[6] 孙国强. XML、XSLT 在应用系统集成中的功能研究[J]. 科学技术与工程,2007,7(13):3124-3128.
[7] Gardner J R, Rendon Z L. XSLT and Xpath: a guide to XML transformations[M]. [s.l.]:Prentice Hall,2002.