

# BT 小包流量识别技术研究

曾 燕, 李春蔚

(四川理工学院计算机学院, 四川 自贡 643000)

**摘 要:** 由于网络中 BT 流量给带宽资源带来的沉重负担, 识别 BT 流并对其进行控制显得尤为重  
要。通过分析 BT 协议和比较 BT、迅雷等实际传输数据, 找出 BT 传输中的流量特征来识别 BT 流。通过  
实验得出数据传输中信息包有效载荷大小 5 个字节的数据包是 BT 流的流量特征, 并验证该技术对 BT  
流的识别具有一定的价值。

**关键词:** BitTorrent 流量识别; 有效载荷

**中图分类号:** TP393.1

**文献标识码:** A

## 引 言

BT 全称为 BitTorrent<sup>[1]</sup>, 是一个文件分发协议, 把上传的开销分摊到每个下载者那里, 体现的特点是下载的人越多, 文件的下载速度就越快。随着网络中 BT 流量的不断增加, 给网络带宽资源带来了沉重的负担, 迫切需要对该流量在网络中的行为特征进行分析、研究和控制, 减少其对网络的消极影响。

当前对 P2P 的研究层出不穷, 而对于 BitTorrent 则是把 BitTorrent 作为一种重要的 P2P 协议来研究。因此, 国内外针对 BitTorrent 流量识别的研究还是比较有限的, 同时也说明, 在 BitTorrent 方面进行研究具有很大的价值和意义。早期对 BitTorrent 识别采用的是分析端口<sup>[2]</sup>的识别法, 根据默认端口 6881 ~ 6889 进行识别。然而对于 BitComet、BitSpirit 等对 BT 协议进行了扩展的软件已采用随机端口进行数据传输, 因此该识别法已经非常受限。当前对 BT 流量识别的研究大量集中在应用层特征识别法方面, 在许多文献<sup>[3-5]</sup>中都提出了利用 BitTorrent 协议握手阶段的特征字段“19BitTorrent Protocol”来识别 BT 流量。但这种方法仅仅能够识别 BT 传输最开始阶段发送的数据包, 因此, 在文献<sup>[6]</sup>中分析 BT 有效载荷各阶段的数据包特征信息, 通过这些特征来更好的识别 BT 流。但这种方法要分析每个 TCP 包, 计算量大、速度慢, 并且占用大量的内存空间。

分析端口识别法和应用层特征识别法可以发现, 尽管两者的实现机理完全不同, 但是其基本思想均是基于 BT 应用的一些外在特征, 并且这些外在特征是可以隐藏的, 一旦出现上述情况, 这些识别方法就不再适用。在一些研究<sup>[7-8]</sup>中通过对传输层数据包(包括 TCP 和 UDP 数据包)的分析, 并结合 P2P 系统所表现出来的流量特征来识别某个网络流是否属于 P2P 流。那么, 对于 P2P 的典型应用之一的 BitTorrent 我们可不可以也根据 BT 传输各阶段中数据包的流量特征来识别 BT 流呢? 因此, 为了能从根本上解决这些问题, 必须分析 BT 应用与其他一些诸如 Web、P2P 等应用的根本区别, 然后利用这些本质特征将 BT 流从其他协议流中识别出来。

本文从 BT 协议、BT 源码以及实际网络流量的分析中得出: 在网络传输中, 有效载荷大小在 10 个字节内的数据包中有效载荷为 5 个字节的数据包是 BT 包。这是一个观测的结果, 所以, 仍然有小概率事件。我们将其作为经验规律, 用来对 BT 流量进行识别, 并通过实验比较证明该方法对 BT 传输数据流的识别有较高的准确性。

## 1 基于流量特征的 BT 识别技术的改进

从文献<sup>[6]</sup>所分析得到的 BT 传输各阶段信息包的有效载荷特征表 1 中注意到, 除了大小不确定的 Bitfield 包和传输数据大小通常为 1440 个字节的 piece 片之外, 理

论上来说,其余只有三种不同大小的信息包,即 5 个字节的包、9 个字节的包和 17 个字节的包。从表 1 及文献<sup>[6]</sup>分析的数据传输过程中,我们可以发现,在 BT 数据的传输过程中,10 个字节之内的信息包以大小为 5 个字节和大小为 9 个字节的包在传输中占主导地位。由于在传输块的最后一个 piece 片时很有可能出现片大小为 17 个字节或 68 个字节的包,再加上在本文中,我们只研究 10 个字节之内的信息包,所以对于大小为 17 个字节的包和 BT 传输开始阶段的 68 个字节的握手包可以被忽略。因此,从理论上来说我们可以根据 BT 流传输中大小为 5 个字节和 9 个字节的小包是 BT 包这个流量特征来识别 BT 流。值得注意的是本文中所提到的数据信息包的大小都是指数据信息包中有效载荷的大小。

表 1 BT 传输中数据包的大小

Information Packet Label	Size of Packet's Payload ( Bytes)
Request	
Have	9
Choke	5
Unchoke	5
Interest	5
Not-Interest	5
Bitfield	Uncertainty
Cancel	17
Handshake	68

为了证实以上理论结果的正确性,本文在实验室内部网络搭建简单的实验环境。将几台 PC 机连接在一起形成一个小范围的局域网网络。实验利用一台搭载了 BitTorrent 软件的 PC 通过 Internet 上传下载 BT 数据,在局域网出口路由器处监听该台 PC 机的所有 BT 业务,并用网络分析软件 Wireshark 进行抓包分析。

### 1.1 实验 1

在本实验中我们以纯 BitTorrent 流为对象,即测试的 PC 机网络中只有 BT 数据包在传输。在本文中,只研究 TCP 包,而非 TCP 包和 TCP 控制包都将被忽略。

通过 BT 流传输的抓包实验观测分析,我们得出图 1 所示情况。

图中显示了在 BT 流传输过程中 10 个字节以内的数据包分布情况。根据图中所示情况,我们可以发现,在 BT 流的传输过程中,大小为 5 个字节和大小为 9 个字节的包在 10 个字节以内的数据包中占了主要部分。这点和我们前面总结出的理论特征相一致。

经过大量的 BT 连接实验分析,我们发现大多数 BT 流的传输具备以上流量特征。然而也有一些 BT 连接不具备以上信息包大小为 5 个字节和 9 个字节的流量特征,但我们发现这些 BT 连接多数是短线连接,而且传输

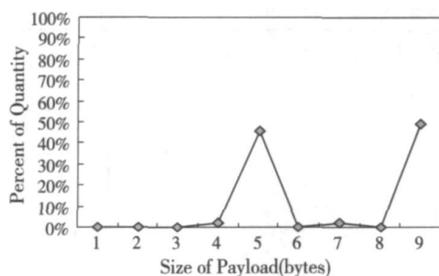


图 1 有效载荷 10 个字节之内的 BT 流信息包分布统计

的都是小包,即在 10 个字节之内。即使这些短线连接不容易被发现,但他们传输中的大多数 BT 流还是可以根据信息包在 10 个字节之内的流量特征被识别。

### 1.2 实验 2

为了进一步验证以上提出的基于小包流量特征可以识别出 BT 流,我们做了另外一些监测抓包实验,在实验测试的 PC 机网络中没有 BT 流传输,得出图 2 统计数据。

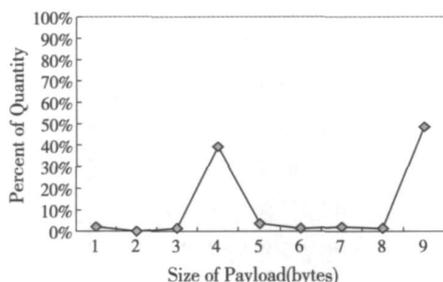


图 2 有效载荷 10 个字节之内的非 BT 流信息包分布统计

根据图 2 我们发现 10 个字节以内的数据包中大小为 5 个字节的包所占比例非常低,然而大小为 9 个字节的包所占比例却相当高。实际上,大多数协议在传输过程中 10 个字节以内的数据包主要发送的是大小为 9 个字节的包。如当前在中国非常流行的 P2P 应用软件迅雷,在传输过程中它发送的小包多数都是 9 个字节的。图 3 显示了迅雷 10 个字节以内的包的分布情况。因此,10 个字节以内的数据包中,大小为 9 个字节的包是 BT 包这个流量特征不适合用于 BT 流的识别。

因此,从 BT 协议、BT 源码以及实际网络流量的分析,我们可以得出这样一个结论:在网络传输过程中,有效载荷的大小为 5 个字节的包是 BT 包。这是一个观测的结果,所以,仍然有小概率事件。我们将其作为经验规律,用来对 BT 流量进行识别。对于实际应用,也就是对于实际的 P2P 流量的识别,还是具有很大的现实意义。对这一现象规律的深层次解释和理论分析有待进一步的研究。

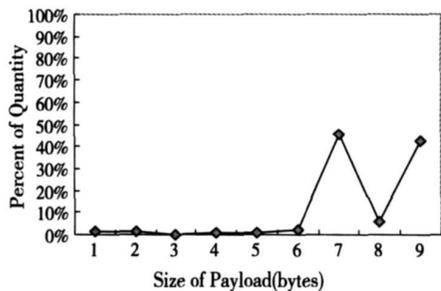


图 3 有效载荷 10 个字节之内的迅雷信息包分布统计

## 2 性能研究

为了验证能否达到研究目标,用以上实验得出的经验规律进行软件开发,并用该软件来研究 BT 小包识别技术的准确性。我们从误报率和漏报率两个指标来评价准确性。

### 2.1 误报率

为了测试误报率,从 2009 年 2 月 22 日 09 时起到 2009 年 2 月 22 日 11 时止进行抓包实验,且环境中没有运行 BT 软件,因而从理论上来说,将不会捕捉到 BT 包。表 2 列出了得到的数据。

通过表 2 可知,在测试环境中捕捉到了 18 个 BT 包,通过对这 18 个 BT 包的分析,发现没有一个包的源端 IP 或是目的地 IP 是研究环境的 IP。考虑到抓包时网卡是设置为混杂模式的,因而我们推测这 18 个 BT 包应该是局域网内其它机器之间进行 BT 传输,经过研究环境中的网卡时被捕捉到的,而非研究环境中的传输的数据。因而,可以认为本软件从研究环境中捕捉到的 BT 包为 0。不过,即便 18 个 BT 包是由于本软件将研究环境中非 BT 包误认为是 BT 包,则误报率亦只有 0.023%,可以忽略。

表 2 没有运行 BT 的捕包情况表

协议名称	捕到的包的数目	百分比
BT	18	0.023%
DNS	564	0.7%
FTP	25	0.032%
HTTP	42315	54.2%
SMTP	3	0.004%
POP3	0	0
其它协议	35098	45.0%

### 2.2 漏报率

2009 年 2 月 26 日 15:10:58 到 15:38:59 之间,在研究环境中仅仅运行 BitTorrent 软件,用本识别方式进行抓包识别。表 3 列出了得到的数据。

对 HTTP 和 DNS 类的包进行分析,可以确定这些包为非 BT 包。故假设认为研究环境中传输的其它协议类的包为 BT,则漏报率为 1.6%。

表 3 本文识别技术识别 BT

协议名称	本文识别技术捕到的包的数目
BT	33047
DNS	3
FTP	0
HTTP	58
SMTP	0
POP3	0
其它协议	541

为了进行对比,还是只运行 BitTorrent 软件,用 Wireshark 软件抓包,用基于握手包特征字符串“19BitTorrent Protocol”的 BT 识别方式进行识别,得到表 4 所示数据。

表 4 “握手”包特征字符串识别方式

协议名称	"19BitTorrent Protocol" 识别技术捕到的包的数目
BT	48348
DNS	17
FTP	0
HTTP	169
SMTP	0
POP3	0
其它协议	1112

由表 4 进行计算得到漏报率为 2.3%。所以,从以上数据显示,本文识别方式更有效率,准确性更高。

## 3 结束语

本文从 BT 协议、BT 源码以及实际网络流量的分析中得出 BT 小包流量识别技术,可以看出,基于有效载荷大小为 5 个字节的小包识别技术对 BitTorrent 传输的数据有比较高的准确性,而对 BitComet 和 BitSpirit 传输数据的识别,则在漏报率上稍差。由于环境所限,在可扩展性方面,本识别技术则表现较差,有待进一步的改进。另外,也是由于环境所限,对在高速网络上的性能没有进行研究,这也有待下一阶段的工作。

## 参考文献

- [1] Cohen B BitTorrent Protocol [OL]. <http://www.bittorrent.org/protocol.htm> 2006-3-1.
- [2] Saroiu S, Gummadi K P, Dunn R J et al. An Analysis of Internet Content Delivery Systems [C]. In Proceedings of the 5th Symposium on Operating Systems Design and Implementation New York, USA: ACM Press, 2002 315-328.
- [3] 杨凌霄, 阎卫杰, 张镨. P2P 流量监控技术分析 [J]. 数据通信, 2007, 5(5): 61-63.
- [4] 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): 74-79.

(下转第 685 页)

- 法研究 [J]. 计算机应用研究, 2005, 22(4): 52-53
- [9] 宋艳, 梁静国. 基于模糊聚类的客户分类应用研究 [J]. 物流科技, 2005, 28(1): 26-28
- [10] [http://www.zgrtu.com/adm\\_inuser/htglxt.php](http://www.zgrtu.com/adm_inuser/htglxt.php) 2010-3-16
- [11] 别荣芳, 尹静, 邓六爱, 译. 数据挖掘技术 [M]. 北京: 机械工业出版社, 2006
- [12] 邝祝芳, 焦贤龙, 高升, 译. 数据挖掘原理与应用—SQL Server2005数据库 [M]. 北京: 清华大学出版社, 2007.
- [13] 王慧敏, 陈泽宇, 王敏娟, 等. 移动学习情境中教育智能应用探究 [J]. 中国远程教育, 2010(1): 68-71.

## Empirical Research on Student of Internet Based on Clustering

MAO Bu<sup>1</sup>, TIAN Lin<sup>2</sup>, XIE Wen<sup>3</sup>

(1. Sichuan Broadcast and Television College, Zigong 643000, China

2. Chuxiong Teachers College, Chuxiong 675000, China; 3. Sichuan University, Chengdu 61000, China)

**Abstract** Online unattended learning is of key importance in distance education, therefore making the establishment of a knowledge resource pool for the learners in a more efficient and faster manner a hot field for our research. This essay tries to venture an effective analysis with data and study with the basic information and study segmentation of almost a hundred students of the year 2009 as sampling data and based on that proposes solutions to create resource pool in keeping with distance education.

**Keywords** dynamic clustering; study on Internet; student segmentation; data study

(上接第 681 页)

- [5] 陆庆, 周世杰, 秦志光, 等. 对等网络流量检测技术 [J]. 电子科技大学学报, 2007, 36(6): 133-137.
- [6] 李明伟, 张大方. 基于有效载荷分析的 BT 流量识别技术 [J]. 计算机应用, 2007, 27(9): 230-232.
- [7] 侯自强. P2P: 让互联网无处不在. <http://www.dchljnpit.net.cn>, 2005.
- [8] 陆晓雯. P2P 流量识别与控制系统的的设计研究 [D]. 南京理工大学, 2008.

## Technology of BT Traffic Identification Based on Small Data Package

ZENG Yan, LI Chunwei

(School of Computer Science, Sichuan University of Science & Engineering, Zigong 643000, China)

**Abstract** Due to the heavy burden of bandwidth resources caused by the BT traffic, traffic identification and control is very important. Through the BitTorrent protocol analysis and comparing the actual transmission data of BitTorrent, Thunderbolt etc, BT traffic was identified by finding its characteristics of transmission. The experiment shows that BT traffic was characterized by data packages payload size of 5 bytes on the data transmission process and it has a certain value to identify BT traffic.

**Keywords** BitTorrent; traffic identification; payload