P2P流量检测与分析

蒋海明,张剑英,王青青,彭 娟
(中国矿业大学 信电学院,江苏 徐州 221008)

摘 要:P2P流量逐渐成为互联网流量的重要组成部分,精确识别P2P流量对于有效地管理网络和合理地利用网络资源都具有重要意义。对P2P应用作了简要介绍,分析了P2P和传统C/S网络的不同,并介绍了目前主流的P2P流量检测技术,分析了这些技术的优缺点,然后结合P2P流量的payload特征,设计了一种基于深度包检测的P2P流量检测方法,并通过实验证明了此方法的可行性。结果显示该方法具有非常高的检测精度和令人满意的性能。

关键词:网络流量;P2P;流量检测;深度包检测
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2008)07-0074-03

Identification and Analysis of P2P Traffic

JIANG Hai-ming, ZHANG Jian-ying, WANG Qing-qing, PENG Juan
(Coll. of Information and Electric Eng., China Univ. of Mining Techn., Xuzhou 221008, China)

Abstract: P2P traffic has become one of the most significant portions of the network traffic. Accurate identification of P2P traffic makes great sense for efficient network management and reasonable utility of network resources. The application of P2P is simply presented, then the difference between P2P and traditional C/S network was analyzed. Introduces the advantage and disadvantage of the technology of P2P traffic identification, and then designed a method to identify P2P traffic based on the deep packet inspection combining the payload characteristic of the P2P traffic. Also proved this possibility of the method through an experiment. Experimental results show that the method has proper accuracy and low cost.

Key words: networked traffic; peer-to-peer; traffic identification; deep packet inspection

0 引 言

P2P流量已经成为互联网流量的主要部分,根据英国ISP网络服务公司CacheLogic[1]调查报告,60%的互联网流量是P2P流量。互联网大量的带宽被P2P应用占据,对其它应用的服务质量形成了威胁,也损坏了ISP的利益,所以如何限制和管理P2P流量成为了人们研究的热点。目前主要的P2P流量检测方法可分为三类,第一类是基于端口的流量检测方法,但由于许多P2P应用开始使用非标准端口以隐蔽其行踪,这种方法已不再有效;第二类是深度包检测技术(DPI, Deep Packet Inspection),即基于payload的流量检测方法,通过识别报文中的应用层特征串来识别P2P,该方法简单、可靠,可对流量进行应用分类;第三类是基于流量特征的流量检测方法,指利用网络流量的流量特征检测P2P的方法,该方法有检测加密P2P应用的能力,但对P2P应用分类的能力较弱。对这三类检测方法的性能和复杂程度综合分析对比后,文中利用第二种方案设计了一种对BT流量进行检测的方法。

收稿日期:2007-10-27
基金项目:国家自然科学基金(70533050)
作者简介:蒋海明(1983-),男,湖北天门人,硕士研究生,研究方向为网络安全;张剑英,教授,主要从事信号与系统的教学与研究。

1 课题相关研究

1.1 P2P技术概述

P2P最早由Steve Crocker于1969年提出,是一种分布式网络,网络的参与者共享他们所拥有的一部分硬件资源(处理能力、存储能力、网络连接能力、打印机等),这些共享资源需要由网络提供服务和内容,能被其它对等节点(Peer)直接访问而无需经过中间实体。在此网络中的参与者既是资源(服务和内容)提供者(Server),又是资源(服务和内容)获取者(Client)。与传统C/S网络不同的是,网络中的每个结点的地位都是对等的。每个结点既充当服务器,为其他结点提供服务,同时也享用其他结点提供的服务。P2P与C/S模式的对比如图1所示。

1.2 BT下载原理

BT全名为Bit Torrent[2],是一个P2P软件,传统

的 FTP、HTTP 是把文件由服务器端传送到客户端,这样会出现一些问题:用户数量的增多要求高带宽和服务器的高性能,也会影响到服务器的稳定性,而 BT 采用的是一种类似传销的方式来达到共享,BT 首先在上传者端把一个文件分成了 Z 个部分,甲在服务器随机下载了第 N 个部分,乙在服务器随机下载了第 M 个部分,这样甲的 BT 就会根据情况到乙的电脑上 去拿乙已经下载好的 M 部分,乙的 BT 就会根据情况去到甲的电脑上去拿甲已经下载好的 N 部分,这样就不但减轻了服务器端的负荷,也加快了用户方(甲乙)的下载速度,效率也提高了,更同样减少了地域之间的限制,使用非常方便,其特点简单地说就是:下载的人越多,速度越快。

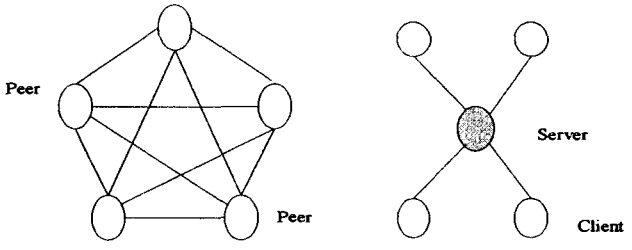

图 1　P2P 与传统 C/S 网络区别

2　P2P 流量的检测

2.1　深层数据报(DPI)检测原理

虽然大部分 P2P 流量传输使用 TCP/UDP/IP 协议,但每种 P2P 应用在自己定义的协议中都定义了一些特征头文件信息。深层数据包检测技术通过对数据包应用层协议的检测发现 P2P 应用。这种技术使用一个 payload 特征库存储 payload 特征信息,符合 payload 特征的数据包即视为 P2P 数据包。这种检测方法易于理解、升级方便、维护简单,是目前运用最普遍的。表 1 是现在较流行的 P2P 应用软件特征信息总结。

表 1　流行 P2P 协议的头特征格式总结[3]

P2P Protocol	String	Trans. prot.	Def. ports
eDonkey2000	0xe319010000	TCP/UDP	4661 - 4665
Fasttrack	"Get/hash"	TCP	1214
	0x270000002980	UDP	
BitTorrent	"0x13Bit"	TCP	6881 - 6880
Gnutella	"GNUT","GIV"	TCP	6346 - 6347
	"GND"	UDP	
MP2P	GOII,MD5,SIZ0x20	TCP	41170 UDP
Direct Connect	"$ MyN","$ Dir"	TCP	411 - 412
Ares	"GET hash"	TCP	

第一行是 P2P 协议的名称,第二行是对应协议具体的 payload 特征字符串,三、四行是数据传输时所使用的传输协议和端口号。深度包检测(DPI)[4]是一种严格的检测方案,通过深度分析 IP 包所携带的 4～7 层协议的特征进行检测,各种 P2P 应用软件在设计时定义的一些固定字段在网络上呈现出来就是数据包特征字。这样即使改变 4 层的端口,也无法躲避检测。这是一种命中率很高的检测方案,检测的关键在于,它要不断地在格式不定的数据包中判断出各种特征字,实现这一过程的基础技术就是模式匹配(Pattern - Matching)。通俗地讲,就是字符串匹配,即从数据中搜索是否存在目标字符串。

2.2　流量检测方案

由上述的原理设计出一个 P2P 流量检测软件,采用 Visual C++ 6.0 和 WinCAP 开源库实现。该软件捕捉 TCP 数据包后采用 BT 协议字符串模式匹配算法进行检测, TCP 是 BT 握手消息时使用的协议[5],一旦检验发现数据包符合 BT 协议特征时,记录该数据包主机的 IP、TCP 端口号,检测流程如图 2 所示;流量测试流程如图 3 所示,一旦确定该数据包为 BT 数据包后,记录该包的大小和到达时间,分别用 d 和 t 表示,以供下一步计算 BT 连接特定端口的流速。

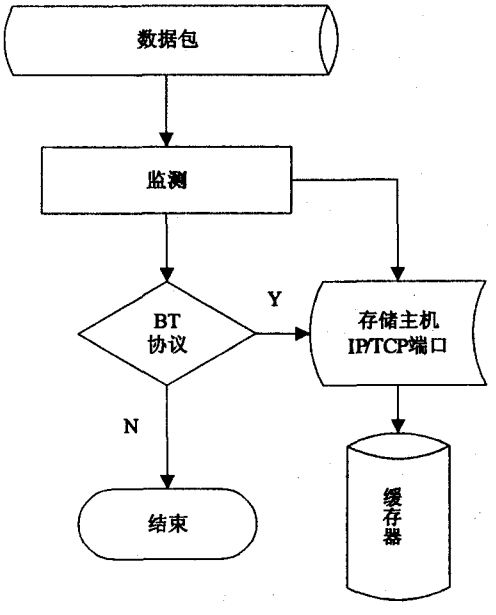

图 2　BT 流量识别流程图

在实际检测中,由于 BT 下载数据包数量多,频繁的流量计算反而影响检测的性能。与利用单个数据包依次检测流量相比,采用定时检测流量将大大减少计算的复杂程度[6],进一步提高实验结果的准确度,灵活地适用于各种网络环境。该实验利用下面等式计算特定连接的 BT 流量速率,实验中 ΔT 取为 1 秒,为了对 BT 流量有个直观的分析,利用数据库将计算得到的数据存储起来供日后对比分析。

$$r_k^i = \frac{\sum_{j \in [(k-1)\Delta T,(k\Delta T)]} d_j^i}{\Delta T} \tag{1}$$

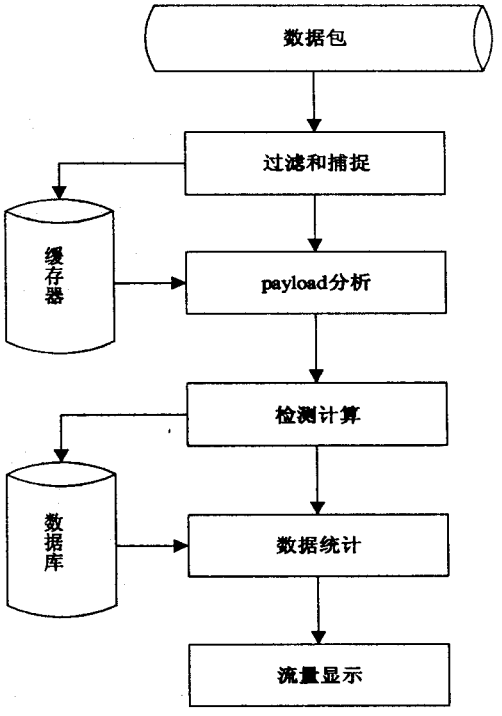

图 3 流量测试流程图

另外,对于大流量、低延迟的网络环境,需要考虑合理的硬件设计以满足严格的实时性检测要求。该实验方案采用旁路技术[7],对数据包进行"查看"而不是"停止并检查",当数据包命中时根据控制策略采取行动。如果由于网络流量太大,无法满足实时监控的要求,则可以利用硬盘缓存适当调整流量速度以期达到更好的检测效果。这种方式在一定程度上消除了检测节点对网络速度的制约,且不会因为检测节点的失效而导致网络瘫痪。

3 实验结果和分析

3.1 流量检测结果

在校园网环境下对 DPI 检测的性能进行了测试,把检测软件布署在一台 HUAWEI 路由器上,实验室局域网通过此路由器与广域网互联。检测软件测量通过路由器的 P2P 流量。

显然,此方法有效地检测了 P2P 流量,性能稳定性好。P2P 流量通常是分布式、多连接,具有较高的速率。由图 4 可知,检测到的每秒平均连接数大约为 155 次;具体一实验特定测试连接端口数据传输速率如图 5 所示大约为 130kbps。这种检测方法易于理解、升级方便、维护简单,可应用于多种网络环境中。

3.2 检测性能分析

通过对以上的实验过程和结果进行分析可以得到,基于深度包检测(DPI)的 P2P 流量检测具有以下

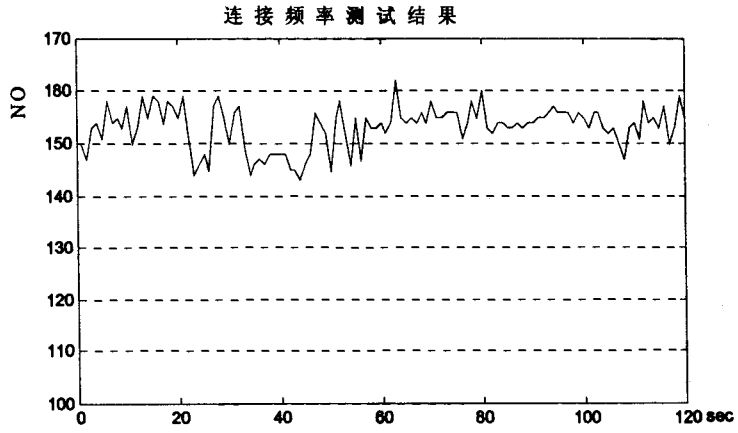

图 4 P2P 流量检测性能

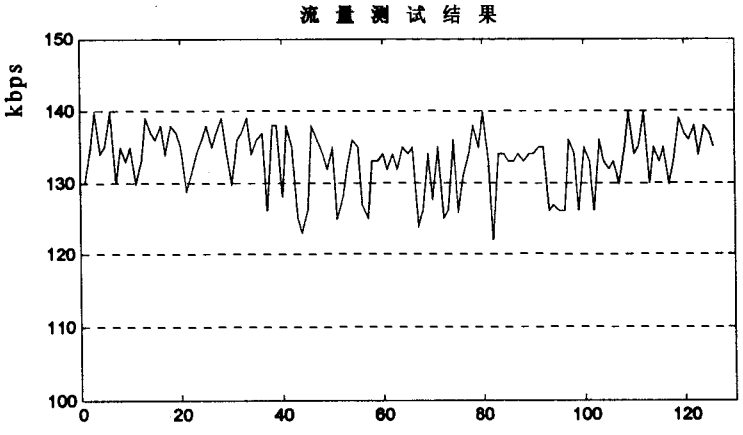

图 5 P2P 流量测试性能

优点:

第一,通过使用各种优化措施和发现更多的 payload 特征,DPI 技术可以达到非常高的检测精度和令人满意的性能。

第二,目前主流的 P2P 协议都是非加密传输的,破解相对容易,使用 DPI 技术就能满足目前运营商限制 P2P 流量的需求。

第三,对于新出现的 P2P 协议,只需在破译后升级 payload 特征库就可以实现对新出现的 P2P 协议的监控,后期维护简单。

4 结束语

设计了一种基于深度数据包检测 P2P 流量的方法,并利用程序实现了这一方案,以实验直观地证明了此方法的可行性。经过对 BT 流量的检测体现了此方法具有非常高的精度,有很强的实用性,可广泛应用于校园、企业网络管理,实施相应的 QoS 技术,保障用户的业务质量。

P2P 流量的监控和管理是一个不断变化发展的过程[8]。P2P 应用由简单的基于固定的端口号,容易被

(下转第 79 页)

用的信息,比如广告,站点导航,赞助商,用于友情交换的链接[17,18],对于链接分析不仅没有帮助,而且还影响结果。如何有效去除这些无关链接,也是算法的一个关键点。

(4)查询的分类。每种算法都有自身的适用情况,对于不同的查询,应该采用不同的算法,以求获得最好的结果。因此,对于查询的分类也显得非常重要。

文中对两种算法进行了深入探讨和比较,但在这几个方面需要继续做深入的研究,相信在不久的将来会有更多的有价值的成果出现。就目前的研究来看,值得关注的是,已有学者对这两种算法相结合的可能性作了理论上的探讨。

参考文献:

[1] 戚华春,黄德才,郑月峰. 具有时间反馈的PageRank改进算法[J]. 浙江工业大学学报, 2005,33(3):272-275.

[2] Lempel R, Moran S. The Stochastic Approach for Link-Structure Analysis(SALSA) and the TKC effect[J]. Computer Networks,2000,19(2):33-36.

[3] Brin S, Page L. Anatomy of a Large-Scale Hypertextual-Web Search Engine[C]//Proc. 7th International World Wide Web Conference. Stanford: Stanford University,1998.

[4] 赖茂生. 计算机情报检索[M]. 北京:北京大学出版社, 1993:89-112.

[5] 杨思洛. 搜索引擎的排序技术的研究[J]. 信息检索技术, 2005,119(1):43-45.

[6] 陈定权. Web信息检索技术最新进展[J]. 现代图书情报技术,2002(2):2-3.

[7] Henzinger R. Hyperlink Analysis for the Web[J/OL]. IEEE Internet Computing,2001:45-50. http://computer.org/internet/.

[8] 曹　军. Google的PageRank技术剖析[J]. 情报杂志,2002(10):10-12.

[9] 彭绪富,邹友宽,邓荣华. INTERNET搜索引擎探解[J]. 高等函授学报:自然科学版,2001(2):15-16.

[10] 苏宁新. 信息检索理论与技术[M]. 北京: 科学技术文献出版社,2004:231-233.

[11] Goldberg D E. Genetic Algorithms in Search, Optimization and Machine Learning[M]. New York: Addison Wesley, 1989.

[12] Mernik M, Crepinsek M, Zumer V. A metavolutionary approach in searching of the best combination of crossover operators for the TSP[C]// Proceeding of the IASTED ICNN. Pittsburgh, PennsyIvania: IASTED/ACTA Press, 2000:32-36.

[13] Kamvar S, Haveliwala T, Golub G. Adaptive methods for the computation of PageRank[C]// Linear Algebra and its Applications 386. Proc of International Conference. [s.l.]: American Pacific University,2004:51-65.

[14] Kleinberg J M. Authoritative Source in a hyperlinked Environment[J]. Journal of ACM, 1999,46(5):604-632.

[15] 肖　文,庞丽萍. 电子出版物的全文检索技术研究[J]. 计算机与数字工程,2002(4):45-48.

[16] 钟敏娟. 基于Web的文本信息检索算法研究[D]. 长沙: 湖南大学, 2004:9-19.

[17] Lee Min-Hyung, Kim Yeon-Seok, Lee Kyong-Ho. Logical structure analysis: From HTML to XML[J]. Computer Standards & Interfaces, 2007,29(1):109-124.

[18] Gupta S, Kaiser G E. Context-based content extraction of html documents[D]. Columbia: Columbia University,2006.

(上接第76页)

检测逐渐发展到动态随机端口号,近期涌现的新型P2P应用越来越具有反侦察的意识,采用一些加密的手法,伪装HTTP协议,传输分块等来逃避识别和检测,可见将来P2P软件必将走向加密通信的方向。针对现在P2P应用发展的趋势,P2P流量管理将逐步走向根据其不变传输特性建立相应的分析模型,应用更高级的机器学习和数据挖掘的方法去识别和管理流量,从而提高网络服务质量。

参考文献:

[1] CacheLogic[EB/OL]. 2006. http://www.cachelogic.com/.

[2] BitTorrent[EB/OL]. 2007. http://www.bittorrent.com/.

[3] Karagiann I T, Broidoa, Faloutsosm. Transport Layer Identification of P2P Traffic[C]// Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement. New York: ACM Press, 2004:121-134.

[4] Sen S, Wang Jia. Analyzing Peer-to-Peer Traffic across Large Networks[C]// In: IEEE/ACM Transactions on Networking. NJ: IEEE Press, 2004:219-232.

[5] Karagiann I T, Bro I I, Brownlee N, et al. Is P2P dying or just hiding[C]// Globecom. Dallas, TX, USA: [s.n.], 2004.

[6] Kim M S, Kang H J, Hong J W. Towards Peer-to-Peer Traffic Analysis Using Flows[C]// SelfManaging Distributed Systems, 14 th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management. Heidelberg, Germany:[s.n.], 2003:55-67.

[7] Kim M S, Won Y J, Hong J W K. Application-Level Traffic Monitoring and an Analysis on IP Networks[J]. ETRI Journal, 2005,27(11):22-42.

[8] P-Cube Inc. Approaches to Controlling Peer-to-Peer Traffic: A Technical Analysis[EB/OL]. 2004. http://www.pcube.com/doc_root/products/Engage/WP_Approaches_Controlling_P2P_Traffic_31403.pdf.