

## 基于用户行为分析的 PPTV 点播系统测量研究

胡 鑫, 陈兴蜀, 王海舟, 刘 磊

(四川大学 计算机学院, 四川 成都 610064)

**摘 要:** 利用基于协议分析和逆向工程的主动测量方法对 PPTV 点播系统进行了研究, 通过对 PPTV 点播协议进行分析, 获悉该协议的通信格式和语义信息, 总结出了 PPTV 点播系统的工作原理, 在此基础上设计并实现了基于分布式网络爬虫的 PPTV 点播系统主动测量平台, 并对该平台获取的用户数据进行统计分析, 获得了 PPTV 点播系统部分用户行为特征. 研究结果对 P2P 点播系统的监控及优化提供了研究方法.

**关键词:** P2P 点播; 协议分析; 爬虫; 主动测量; 行为特征

**中图分类号:** TP393

**文献标识码:** A

**文章编号:** 1000-7180(2011)08-0097-05

## A Measurement Study of PPTV VOD System Based on the Analysis of Client Behavior

HU Xin, CHENG Xing-shu, WANG Hai-zhou, LIU Lei

(College of Computer Science, Sichuan University, Beijing 610064, China)

**Abstract:** This paper conducts a research to PPTV-VOD system with active measurement method based on the reverse engineering and protocol analysis. Firstly, we analyse the protocol of PPTV-VOD to learn its communications data format and semantics. With those knowledge, we can know the working principle of PPTV-VOD system. Secondly, we design and implement a measurement platform based on distribute crawler, exploring the user data in the PPTV-VOD overlay network. In the meantime, we analyse the user data and find out some user characteristics. The research results provide some ideas and methods to monitor and optimize P2P-VOD system.

**Key words:** P2P-VOD; protocol analysis; crawler; active measurement; user characteristic

### 1 引言

随着主动测量技术的不断成熟, 在对 P2P 网络的测量研究中越来越多地采用了主动测量技术. P2P 网络的主动测量方法是使用网络爬虫(Crawler)主动加入 P2P 网络, 以获取相关的网络特性和节点属性. 主动测量方法能够直接探测获取 P2P 网络信息的扩散行为, 具有可信度高、准确性好的特点, 是目前 P2P 网络测量的主要研究方法. 2006 年 Hei 等人<sup>[1]</sup>利用主动测量技术对非常流行的 PPLive 网络电视的直播系统进行了研究, 分析了该系统的用户行为、流量峰值和冗余以及直播延迟等特性, 指出了基于 P2P 技术的 IPTV 系统的设计原则. 2008 年

李代玲等人<sup>[2]</sup>对另一款非常流行的 P2P 网络电视 PPStream 的点播系统进行了测量研究, 分析了系统中用户的不同行为对系统的影响并揭示了 PPStream 系统的缓存特点. 2008 年樊鹏翼<sup>[3]</sup>利用主动测量技术针对 PPLive 直播系统提出了 P2P PPTV 的前身就是 PPLive, 它是国内知名度最高, 节目数量最多, 用户数最多的 P2P 网络电视之一. 由于 PPTV 点播系统与直播系统存在巨大差异, 而前人的研究工作主要集中在 PPTV 直播系统, 因此对 PPTV 点播系统的研究没有可借鉴的成果, 并且 PPTV 点播系统采用私有通信协议, 研究它将面临巨大的挑战. 本文采用基于协议分析和逆向工程的主动测量方法<sup>[4]</sup>, 首先利用 WireShark<sup>[5]</sup>对 PPTV

收稿日期: 2011-05-09; 收稿日期: 2011-06-25

基金项目: 国家 242 信息安全计划((242)2009A82); 国家“九七三”基础研究项目(JG2008031)



已有的针对 P2P IPTV 的爬虫系统主要采用两种策略来控制爬行时长. Hei<sup>[1]</sup> 等人设计的爬行器采用固定时长的爬行策略,爬行时间固定为 15 s. 但是随着 P2P IPTV 系统的发展,不同节目或者不同时间段的用户数差异很大. 固定爬行时长的爬行策略准确性和完整性都不好. Vu 等人<sup>[8]</sup> 以及姜志宏等人<sup>[9]</sup> 设计的爬行器采用动态爬行时长的策略,根据节点或边的增长速度确定爬行时间,如果节点或边的增长速度小于某个阈值则停止爬行. 这种策略的缺点是当爬虫从一个节点获得的节点数或边数小于阈值时并不能保证从下一个节点获得的节点数或边数也小于这个阈值. 这种基于节点或边的增长速度的爬行策略可能导致某些节点的缺失.

根据 PPTV 点播系统的节点分发协议,PPTV 客户端的邻居节点全部由 Tracker 服务器提供,因此本文设计的爬虫客户端只向 Tracker 服务器请求节点列表. PPTV 点播系统中 Tracker 服务器是固定的,从每个 Tracker 服务器获取节点的准确性和完整性就成为影响爬虫系统性能的关键. 设一次爬行中连续向每个 Tracker 服务器请求节点列表的次数  $T$ ,为了能够在保证节点完整性的基础上在较短的时间内完成爬行,就必须确定一个合适的  $T$ . 如图 2 所示是 A、B、C 三个爬虫客户端同时针对同一节目的节点爬行示意图, $T$  分别设置为 1、2 和 3.

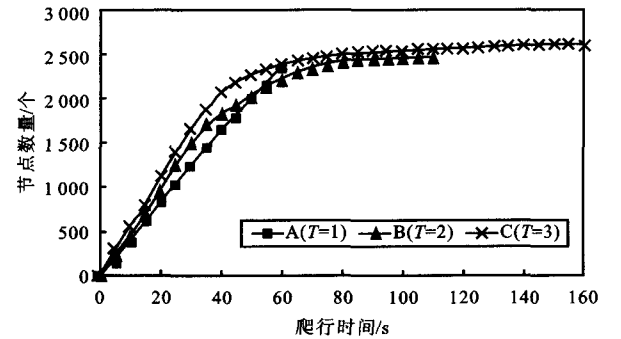


图 2 不同请求次数采样对比图

如图 2 所示, $T$  设置为 1,爬虫在 60 s 内获得超过 2 400 个节点; $T$  设置为 2,爬虫在 110 s 内获得接近 2 500 个节点; $T$  设置为 3,在 160 s 内获得约 2 600 个节点. 综上可以看出,虽然  $T$  越大,获得的节点越多,但同时爬行时间也越长,过长的爬行时间会降低爬虫系统的准确性和时效性,所以将  $T$  设置为 1 可以使爬虫的性能最优.

3.3 数据采集

在观看点播节目时,用户可以通过拖拽进度条跳过某些视频片段,这种行为称为“跳跃观看”,反之

为“连续观看”. 本文使用 MVR(Movie Viewing Record)描述用户的“连续观看”行为. 一条 MVR 表示一次连续观看行为,MVR 是由标识资源的 ResourceID,用户 IP、端口,连续观看的第一个视频文件的序号(SN),连续观看的最后一个视频文件的序号(EN)构成的四元组.

如表 1 所示,用户 A 拥有对点播视频节目 M 的三条观看记录,分析可知用户的观看行为:用户 A 连续观看了视频片段 1 和 2,在观看视频片段 3 时直接跳到视频片段 5,之后连续观看视频片段 6~8,在观看视频片段 9 时跳跃到视频片段 10,然后连续观看了视频片段 11~13,在视频片段 14 的某一个时刻或节目结束时退出.

表 1 用户 A

| ResourceID | IP、Port | SN | EN |
|------------|---------|----|----|
| M          | A       | 1  | 2  |
| M          | A       | 6  | 8  |
| M          | A       | 11 | 13 |

4 PPTV 点播系统用户行为特性

本文的分布式爬虫系统部署在四川大学校园网的多台服务器上,服务器配置为 1 颗 2.13 GHz 的志强 CPU,8 GB 内存,100 Mb/s 独立 IP 接入. 所有爬行数据均在 2010 年 9 月至 10 月期间采集.

4.1 用户数量变化趋势

如图 3、图 4、图 5 和图 6 所示为 2010 年 9 月 6 日至 12 日一周内《寰宇地理—狒狒》、《宫心计第九集》、《10 进 8 快乐男声总决赛第三场》和《新三国第四集》四路节目的用户数量变化趋势. 用 Movie1、Movie2、Movie3、Movie4 表示以上四个节目,其中 Movie1 为不流行节目,其余 3 个节目为流行节目,这四个节目代表了不同的用户群.

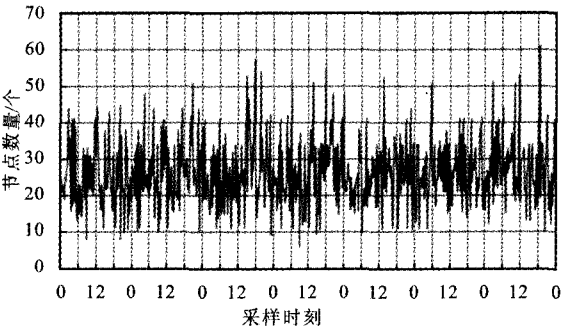


图 3 Movie1 一周用户数量变化

对各节目的一周用户数量采样可知:(1)不流行

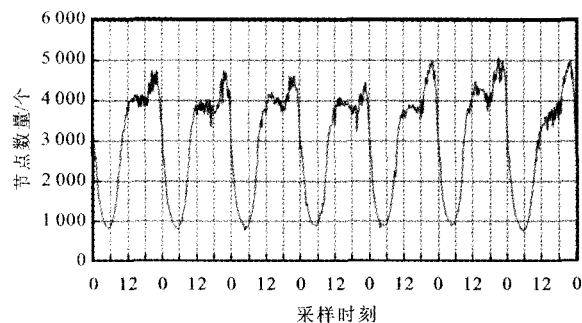


图4 Movie2一周用户数量变化

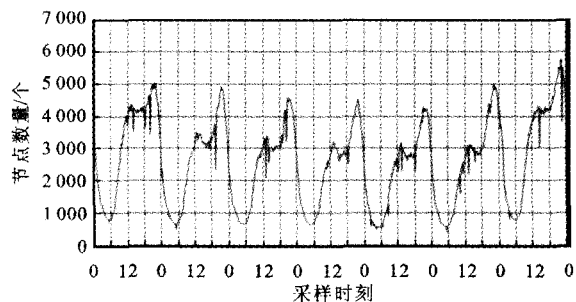


图5 Movie3一周用户数量变化

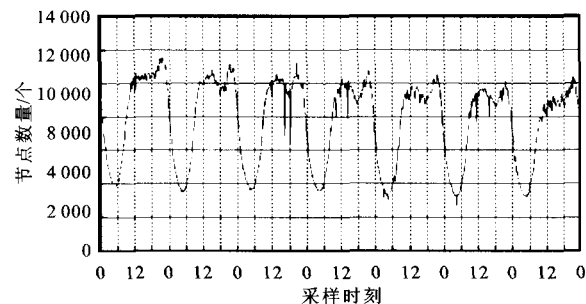


图6 Movie4一周用户数量变化趋势

节目 Movie1 用户数量较少, 数量变化趋势规律不明显, 而其余 3 个流行节目的用户数量较多且存在明显数量变化趋势; (2) 从一天的用户数量变化趋势来看, 三个流行节目在每天的凌晨 5 点左右用户数量最少, 从 5 点开始至 12 点用户数量一直呈增加趋势, 12 点至 19 点用户数量稍有下降, 从 19 点开始用户数量开始上升, 在 21 点左右达到一天的最大值, 用户数量一直减小到第二天凌晨 5 点; (3) 从一周的统计数据来看, 三个流行节目每天的用户数量变化趋势都很相似, 峰值和低谷都在相近的时间点出现。综上, 可以发现点播系统中存在一个明显的用户行为特征: PPTV 点播系统用户数量变化规律与人们的作息时间表关联明显。

#### 4.2 “连续观看”时间与用户停留时间

一次“连续观看”的时长可以通过一条 MVR 记录的 SN、EN 计算和表示:

$$t_{\text{mvr}} = (EN - SN + 1) \times FT \quad (1)$$

FT 表示每个视频文件的视频时长。Tracker 服

务器通常只保存拥有完整视频文件的用户列表信息, 式(1)只能计算 MVR 中完整视频片段时长。

图 7 所示为表 2 中用户 A 的第二条 MVR 的示意图, 用户 A 在视频片段 5 的 X 处开始一次“连续观看”, 由于从 X 点到视频片段 6 的开始位置之间的视频片段未能构成一个完整的视频文件, 因此将不会被 Tracker 服务器记录, 这部分长度的计算方法如下: X 满足  $0 < X < FT$ , 且 X 在  $(0, FT)$  区间上服从均匀分布, 因此 X 的密度函数  $f_{(X)} = 1/FT$ ,

$$\text{期望 } E_{(X)} = \int_0^{FT} X f_{(X)} dx.$$

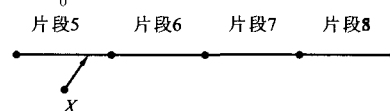


图7 MVR示意图

综上, 假设某个节目由 FN 个视频文件组成, 每个视频文件时长为 FT, 一次“连续观看”的时间:

$$t_{\text{MVR}} = (EN - SN + 1) \times FT + n \int_0^{FT} X f_{(X)} dx$$

n 表示该 MVR 中的未能构成一个完整的视频文件的视频片段的数量。

当  $SN \neq 0$  且  $EN \neq FN$  时,  $n = 2$ 。

当  $SN \neq 0, EN = FN$  或  $SN = 0, EN \neq FN$  时,  $n = 1$ 。

当  $SN = 0$  且  $EN = FN$  时,  $n = 0$ 。

设  $u = t_{\text{MVR}}/pT$ , 表示一次“连续观看”的时长占整个节目时长的比例,  $pT$  表示一个完整节目的时间。由于 Movie1 用户数量较少, 不具有代表性, 只选取了其余三个节目进行研究。如图 8 所示三个节目 u 的分布图, 可以看到, Movie2 和 Movie3 的用户在连续观看很短的时间后就会发生一次跳跃, 而 Movie4 的用户会观看较长一段时间后才发生一次跳跃。

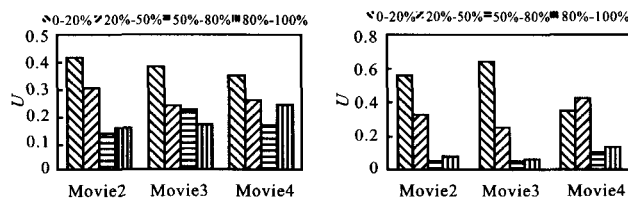


图8 U和u的分布

用户停留时间是指某用户观看某一个节目的时间, 可以通过属于该用户的所有 MVR 计算出来:

$$T = \sum_{i=1}^k t_{\text{MVR}_i}$$

$t_{\text{MVR}_i}$  表示用户的一次连续观看的时长, k 表示某用户针对该节目拥有的 MVR 的数量。设  $U =$

$T/pT$ ,表示用户停留时长占整个节目时长的比例, $pT$ 表示一个完整节目的时间.如图8所示为 $U$ 的分布图,可以看到,三个节目的测量结果都相似,一部分用户在观看一段时间后就离开了该节目,另一部分用户则会在观看完节目后离开.

综上,点播系统中存在一个明显的用户行为特征:跳跃观看者数量远远大于平滑观看者数量.当用户观看点播节目时会迅速的浏览一些节目片段直到发现他们感兴趣的节目,然后就会长时间的连续观看,如果没有感兴趣的节目就会退出系统.

## 5 结束语

本文通过对PPTV点播系统进行协议分析,利用协议分析和逆向工程的主动测量方法,获悉了它的工作原理,并在此基础上设计实现了PPTV点播系统分布式爬虫系统,利用该爬虫完成了对PPTV点播系统覆盖网的主动测量.测量结果发现:(1)PPTV点播系统用户数量变化规律与用户的作息時間关联明显;(2)在点播系统中,跳跃观看者数量远远大于平滑观看者数量,用户会快速的浏览一些节目直到发现感兴趣的节目.本文的研究方法为P2P点播系统的测量与监控提供了参考,研究结果为P2P点播系统的监控及优化提供了依据.

### 参考文献:

- [1] Hei X, Liang C, Liang J, et al. A measurement study of a large-scale P2P IPTV system [J]. IEEE Transactions

on Multimedia, 2007, 9(8):1672-1687.

- [2] 李代玲. 基于网络测量的PPStream网络电视系统研究[D]. 北京:北京交通大学,2008.
- [3] 樊鹏翼. P2P流媒体监管关键技术研究[D]. 长沙:国防科学技术大学,2008.
- [4] 时金桥,张志斌,郭莉,等. 大规模分布式互联网应用的测量[J]. 信息技术快报, 2008, 6(6):1-14.
- [5] The WireShark Homepage [EB/OL]. [2010-09-30]. <http://www.WireShark.org/>.
- [6] 冯健. P2P流媒体关键技术研究[J]. 微电子学与计算机, 2009, 26(8):49-54.
- [7] 王海舟,陈兴蜀,王文贤. PPLive网络电视系统的测量研究[J]. 计算机应用, 2009(7):1988-1991.
- [8] Long V, Guta I, Liang J, et al. Measurement and modeling of a large-scale overlay for multimedia streaming [EB/OL]. [2009-06-07]. <http://dprg.cs.uiuc.edu/docs/pplive/qshine07-4.pdf>.
- [9] 姜志宏,王晖,樊鹏翼,等. 一个P2P IPTV多协议爬行器—TVCrawler[J]. 计算机应用, 2010, 30(3):58-60.

### 作者简介:

- 胡鑫 男,(1988-),硕士研究生. 研究方向为P2P网络.
- 陈兴蜀 女,(1968-),博士,教授,博士生导师. 研究方向为信息安全、计算机网络.
- 王海舟 男,(1986-),博士研究生. 研究方向为信息安全、P2P网络.
- 刘磊 男,(1983-),硕士研究生. 研究方向为信息安全、P2P网络.

(上接第96页)

比例上比较理想.(1)阈值滑动机制中阈值的划定(2)如何提高压缩比例及在其他领域的应用是需要进一步研究的工作.

### 参考文献:

- [1] Chen Yu, Tang Changjie, Li Rui, et al. Reduced-GEP: Improving gene expression programming by gene reduction[C]// Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2010 2nd International Conference. China: Hangzhou, 2010:176-179.
- [2] 陈华辉,施伯乐,钱江波,等. 基于小波概要的并行数据流聚类[J]. 软件学报, 2010, 21(4):644-658.
- [3] 丁超,元昌安,李桂来,等. 基于GEP的多数据流的压缩算法[J]. 计算机研究与发展, 2008, 45(S0):191-195.

- [4] 元昌安,彭昱忠,覃晓,等. 基因表达式编程算法原理与应用[M]. 北京:科学出版社,2010.
- [5] 刘陶刚,赵荣彩,姚远,等. 分块存储的滑动窗口数据重用技术[J]. 计算机应用, 2010, 30(5):1371-1375.
- [6] 王元元,曾建潮,谭瑛. 基于并行计算模型的并行微粒群算法的性能分析[J]. 微电子学与计算机, 2007, 24(9):25-28.

### 作者简介:

- 杨文 男,(1987-),硕士研究生. 研究方向为并行计算和人工智能.
- 李文敬 男,(1964-),教授. 研究方向为并行算法和演化算法.
- 罗锦坤 男,(1985-),硕士研究生. 研究方向为数据挖掘与智能计算.