浙江科技学院学报,第 30 卷第 3 期,2018 年 6 月
Journal of Zhejiang University of Science and Technology
Vol. 30 No. 3, Jun. 2018

doi: 10.3969/j.issn.1671-8798.2018.03.009

基于 YouTube 视频分享系统的信息挖掘

卢红波,钱亚冠,马　骏
(浙江科技学院 曙光大数据学院,杭州 310023)

摘　要: 研究 YouTube 的服务器数据和网络流量数据,挖掘发现其中的隐藏信息,对设计高效的视频内容分发网络具有积极意义。通过比较服务器获得的数据和网络中获取的视频流量,应用统计建模方法,对用户行为、视频传输的静态和动态特征进行挖掘分析以研究 YouTube 的视频传输机制。结果表明,视频服务器上的静态信息和网络中获得的动态信息存在显著的相异性,这种差异性有助于研究人员在设计服务器端的算法和视频传输网络的缓存算法时采用不同的优化策略。

关键词: 视频分享;社交网络;用户行为;传输机制
中图分类号: TP391.4　　**文献标志码**: A　　**文章编号**: 1671-8798(2018)03-0230-05

Mining the hiding information of YouTube sharing video system

LU Hongbo, QIAN Yaguan, MA Jun
(School of Sugon Big Data Science, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: The server data and network traffic data of YouTube video were targeted to mine some valuable hiding information for designing effective video content distribution networks. The statistical modeling method was used to analyze user behavior patterns, static and dynamic properties by comparing the server data with network video traffic, in pursuit of exploring transmission mechanisms of YouTube. The results show that the static information derived from the video server data is strikingly different from the dynamic information of the network. This difference will help researchers to employ different optimized strategies in designing algorithms for video servers and caching algorithms for transmission networks.

Keywords: video sharing; social networks; user behaviors; transmission mechanisms

收稿日期: 2017-10-27
基金项目: 浙江省自然科学基金项目(LY17F020011)

通信作者: 钱亚冠(1976—　),男,浙江省嵊州人,副教授,博士,主要从事机器学习、大数据处理及人工智能研究。E-mail:qianyg@yeah.net。

随着 Web2.0 技术的成熟以及智能手机的普及，用户可以通过视频分享网站观看、创建、共享和发布自己拍摄的视频文件[1]。据统计，目前互联网流量中大概有 25%～40%的成分是视频流[2]。其中全球著名的视频分享网站 YouTube 在视频流量中占有很大的比例，因此也成为了网络界重点研究和关注的对象。像 YouTube 这样的视频分享服务提供商与传统上的 VoD(video on demand)系统存在着很大的区别，最显著的不同是 YouTube 具有在线社交网络的特点[3]。在 YouTube 上，用户可以自由地上传自己拍摄的视频文件，并分享给其他用户；而其他用户在看完视频后，可以对其发表评论和评分。这种用户之间的交互性是传统 VoD 系统所不具备的。除了在用户的交互性上存在差异外，两者在传输协议上也存在差异[4]。YouTube 采用 HTTP/TCP 协议从 Web 服务器将视频流信息传送到客户端；而 VoD 系统则采用 RTP/UDP 协议，从专用的视频服务器上将视频流信息传送到客户端。YouTube在传输层采用面向连接的 TCP 协议，而在应用层则采用传统上用于传输 HTML 文件的HTTP协议封装视频流数据，因此其行为特征和流量模式必然有别于普通的 HTTP 流量。由于商业上的保密性，YouTube 等视频分享服务商并没有公开具体的技术细节[5]。因此，研究人员只能通过网络爬虫的方式收集分享视频系统的元数据或通过被动测量的方式从网络节点上获取视频数据。以往很多研究工作集中在服务器元数据的统计特性上，如视频的长度、观看的数量等信息或从流量中抽取的信息，如视频流的持续长度等[6-9]。但我们的研究发现，从 YouTube 内容服务器上挖掘出的信息和从网络节点中挖掘出的信息并不完全一致，视频长度和观看时间，编码率和下载速率等均存在不一致性，服务器上的静态信息不能完整地反映用户的动态行为特征。而用户的动态行为分析对提高视频传输性能，设计视频缓存算法和内容分发网络具有重要的意义。因此，本文综合这两方面的数据来进行对比分析。

1　数据集

第一个数据集是通过 YouTube 提供的 API[10]，利用网络爬虫技术遍历所有的视频文件后抽取的元数据。这些元数据包括视频 ID、上传者、视频年龄、视频类型、视频长度、观看量、编码率、评级、评论、相关视频等。我们获得 2 个不同时间跨度的流量数据集，其中一个数据子集的时间跨度为 14 d，另一个数据子集的时间跨度为 1 d。第二个数据集是从网络中的路由器上采集到的流量数据，分离出 YouTube 视频流量。我们通过五元组(源 IP、源端口、目的 IP、目的端口、协议号)来确定网络流，又通过接收到第一个 SYN 包确定该网络流的起始，接收到 FIN 包或超过某个时间间隔未收到数据包确定为网络流的结束。该数据集包含的 YouTube 视频网络流的统计属性有源 IP 地址、源端口号、目的 IP 地址、目的端口号、视频传输开始时间、结束时间、数据包数、数据包平均大小、速率等。

2　从服务器端分析用户行为模式

2.1　用户请求数模型

用户请求的到达过程是用户行为模式的主要特征，深入分析用户请求的到达过程是容量规划、用户接纳策略等的重要基础。因此，我们对两个不同时间跨度的流量数据集进行了分析，其中一个数据集的时间跨度为 14 d，共有 16 637 个用户的 611 968 个视频观看请求；而另一个数据集的时间跨度为 1 d，共有 2 377 个用户发送的 18 750 个请求。

用户的视频观看请求数量是用户的一个重要行为特征。图 1 显示了每个用户分别在 14 d 和 1 d 内发送的请求数累积概率分布(cumulative distribution function，CDF)情况，是典型的偏斜分布。由图 1 可知，80%的用户在 1 d 内发出的观看请求数少于 10，在 14 d 内少于 40。图 2 是双对数坐标下用户请求数和用户数之间的关系，由此可知这是典型的幂律关系。我们用 Pareto、Weibull 和 Log-Normal 这 3 种经典偏斜分布对上述用户请求分布建模，发现 Log-Normal 分布的拟合度最佳。其中 14 d 数据的 Log-Normal分布的参数 $\mu=2.353\ 7$，$\sigma=1.576\ 9$；1 d 数据的分布参数 $\mu=1.386\ 7$，$\sigma=1.130\ 3$。

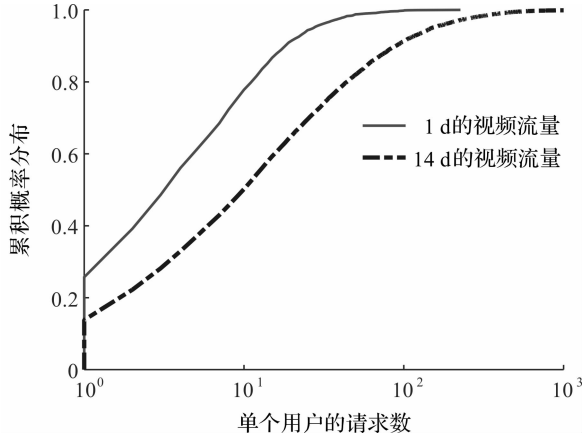

图 1 单个用户的请求数的累积概率分布

Fig. 1 CDF of request per client

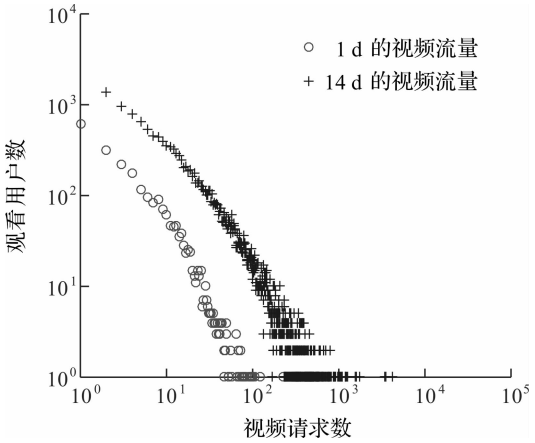

图 2 用户数和请求数之间的幂律关系

Fig. 2 Scatter graph of requests and clients

2.2 用户请求的到达过程

由于服务器的服务策略并不区分具体的用户，它关注的仅是用户请求的到达过程，因此对请求到达过程的建模对认识用户的行为特征及构建高效的服务策略具有重要的意义。图 3 显示了用户请求在内容服务器端到达间隔的累积概率分布。通过拟合度最优检验，Log-Normal 分布最适合建模请求到达间隔，其中 14 d 数据集上的模型参数 $\mu=-0.020\ 37$，$\sigma=1.195\ 2$；1 d数据集上的模型参数 $\mu=0.859\ 39$，$\sigma=1.169\ 6$。上述概率模型的分析推翻了传统上认为请求到达间隔服从指数分布的假设，即认为到达过程是 Poisson 过程，例如文献[11]把 Web 会话的到达建立为 Poisson 过程，即认为请求到达过程是长相关的（long-range dependent，LRD）。但是我们的分析发现，尽管 YouTube等视频分享网站的采用了 HTTP 协议，但其用户行为却具有自身特点，与普通的 Web 用户并不一致，因此不适合用以往研究 Web 的结论和假设来研究视频分享服务。

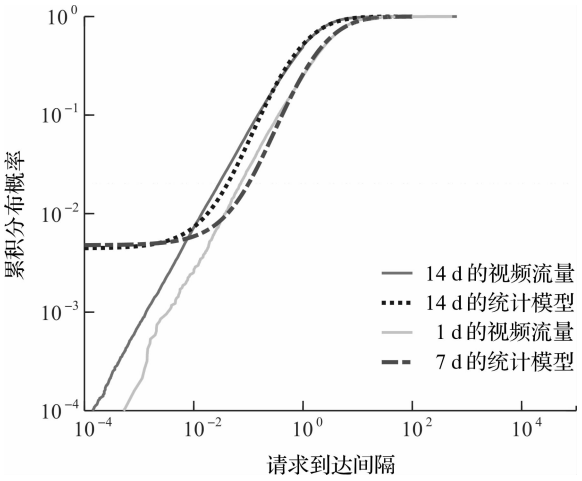

图 3 请求到达间隔的累积概率分布

Fig. 3 CDF of request inter-arrival time

2.3 连续请求之间的空闲期

用户在整个视频观看会话过程中大致可以分为搜索视频、发出观看请求、观看视频、思考、评论视频或评分等行为，这几个阶段可能是反复进行的。本文把观看视频的时间段称为激活期，而把连续 2 次观看期间的搜索、思考、评论等称为空闲期。在研究 Web 用户行为的论文中把前者称为 on 阶段，后者为 off 阶段，采用 on/off 模型来描述这种过程的变换[12]。由于视频分享系统具有社交网络的特点，用户除了搜索、浏览信息外，还会有主动的评论、评分等交互过程，行为比普通的 Web 浏览更加丰富。因此，我们认为在空闲期更能反映视频分享系统的用户行为特征。

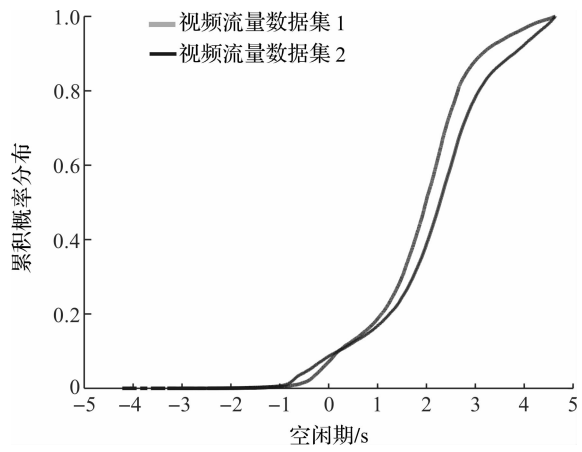

图 4 空闲期的累积概率分布

Fig. 4 CDF of idle time between viewing actions of per user

图 4 给出了空闲期的累积概率分布。假设空闲期超过 1 h 的不属于 2 个连续请求之间的会话，这类情况

不计入本文提出的有研究价值的空闲期。空闲期少于 10 s 的被认为没有评论或评分行为发生，即用户观看完一个视频后，紧接着观看另一个视频。从图 4 可知，这种情况只占到 15%左右。发生在空闲期的典型行为是搜索新视频，假设这类行为的空闲期区间为[10，400] s，从上述统计分析发现约占 55%～65%。而发表评论等行为往往需要更多的时间，大约占 10%～15%的比例。由此可以推断，大多数的用户在空闲期的主要行为是搜索新视频，而发表评论的只占少数。

3　从网络端分析分享视频系统的特点

3.1　视频长度和观看时间

视频长度即完整播放该视频需要的时间，在上传视频到内容服务器时往往作为一个描述该视频的元数据保存在服务器上。以往的研究多把视频长度作为一个静态的特征进行研究，而忽略了用户实际观看行为的模式。大多数用户并不会从头到尾完整地观看整个视频，有的用户仅观看开头的一部分，而有的用户则喜欢以快进的方式快速浏览。因此，从服务器端获得的关于视频长度的信息并不能衡量用户的实际观看时间。

为了分析用户的实际观看时间，我们根据网络节点中获得的 YouTube 视频流量数据，用数据流的持续时间来衡量用户的实际观看时间。图 5 显示了从网络中获得的 YouTube 视频流持续时间和视频内容服务器上获得的视频长度的累积概率分布，从中可知，网络中获得的 2 个数据集的持续时间在概率分布上非常相似，而与视频内容服务器上获得的视频长度的分布却存在很大的差异。我们进一步发现，大多数视频并没有被用户看完，如 80%的观看时间少于 100 s，而 80%的视频长度却超过 300 s。这一发现进一步印证了大多数用户没有完整看视频的行为模式。

图 6 给出了在双对数坐标系下观看时间和视频长度的互补累积概率分布，由此可知，对于长度超过 10 min的视频，用户的观看时间与视频长度具有相似的概率分布特性。这意味着当视频长度超过 10 min 时，用户更愿意看完视频内容，即长度超过 10 min 的视频更能吸引用户看完整个视频。这个发现有助于更好地设计高效的视频缓存系统和调度策略。

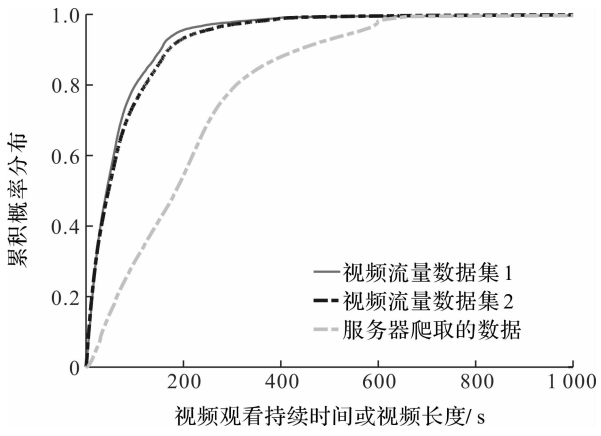

图 5　视频长度和实际观看时间的累积概率分布

Fig. 5　CDF of viewing duration and video length

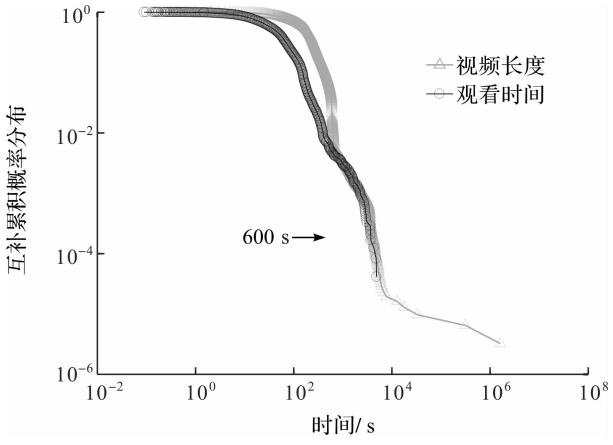

图 6　视频长度和实际观看时间的互补累积概率分布

Fig. 6　CCDF of viewing duration and video length

3.2　编码率和下载速率

为了节省存储空间和带宽，YouTube 上的视频都采用 H. 264 编码器进行了压缩。编码率和图像的质量直接相关，更低的编码率可以获得更高的压缩比，但代价是牺牲图像的画面质量。反之，更高的编码率需要更高的网络带宽来传输视频。为此，YouTube 必须在图像的分辨率和网络的可用带宽之间做出合理的平衡。图 7 给出了 YouTube 视频编码率和下载速率的累积概率分布。由此可知，大约 80%的视频其编码率在很窄的范围：[285，350]kB/s。这意味着YouTube采用的是中等编码率来平衡画面质量和传输带宽。

YouTube 上的视频文件是典型的流式媒体，采用 FLV(Adobe flash video)格式可使用户在浏览器上方便地观看视频。为了使用户在视频还没有完全下载完之前就可观看，YouTube 采用了累进下载技术，即只要在缓冲区装入足够的视频就可以启动视频播放，余下的视频内容可以边播放边下载[13-14]。因此，只要保证下载速率略大于编码率就可以保证播放过程不出现停顿。从图 7 可知，99%的视频流下载速率大于编码率。这一方面反映出传输视频的网络状况，另一方面反映出 HTTP/TCP 流控机制。由前面的分析可知，大多数用户并不看完完整的视频，缓存中的视频不会再被观看。因此，过高的下载速率反而会浪费大量带宽。为此，除了采用 TCP 进行下载速率的流控机制外，推断 YouTube 还有自己的速率控制机制。在图 7 中，下载速率的 CDF 曲线有 2 个接近垂直的陡峭上升部分，其中一个表明大约有 20%的视频流速率在 500 kB/s，另一个则表明 48%的视频流速率在 1 250 kB/s，即大约有 70%的视频流下载速率控制在 500 kB/s 或 1 250 kB/s。因此我们可以推断 YouTube 除了采用 TCP 流控机制外，在服务器端还有专门的下载速率控制机制。由此我们可以得出如下结论：在保证图像质量的前提下，合理的下载速率控制可以节省大量的网络带宽资源，而过快的速率对播放视频并没有额外的好处。

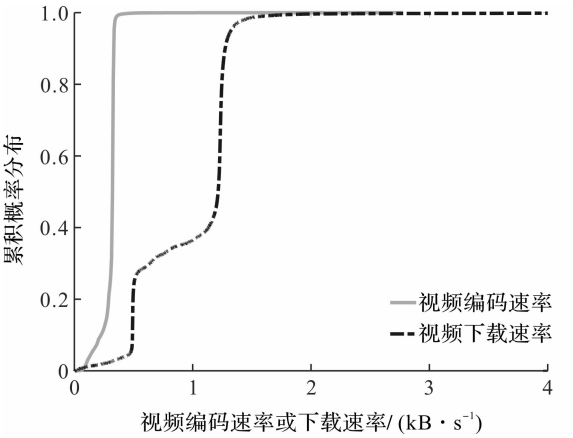

图 7 视频编码率和下载速率的累积概率分布

Fig. 7 CDF of video coding rate and download rate

4 结 论

我们通过 YouTube 视频服务器的数据和网络流量数据来研究用户行为和视频服务器的机制，通过比较研究挖掘出有价值的信息。结果表明，从 YouTube 内容服务器上挖掘出的信息和从网络节点中挖掘出的信息并不完全一致，因此，本文综合这两方面的数据来进行对比分析。视频长度和观看时间，编码率和下载速率等均存在不一致性，服务器上的静态信息不能完整地反映用户的动态行为特征。而用户的动态行为分析对提高视频传输性能，设计视频缓存算法和内容分发网络具有重要的意义，通过此研究可为提高视频传输性能，如视频缓存和内容分发网络的设计提供参考。

参考文献：

[1] JIANG J, SEKAR V, ZHANG H. Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive[C]//Proceedings of the 8th international conference on Emerging networking experiments and technologies. Nice: ACM, 2012:97.

[2] MAIER G, FELDMANN A, PAXSON V, et al. On dominant characteristics of residential broadband internet traffic [C]//Proceedings of the 9th ACM SIGCOMM conference on Internet measurement. Chicago: ACM, 2009:90.

[3] SPART Z, JAMES T, SU L, et al. YouTube, social norms and perceived salience of climate change in the American mind[J]. Environmental Communication, 2017,11(1):1.

[4] AMEUR C B, MORY E, COUSIN B, et al. TcpHas: TCP for HTTP adaptive streaming[C]//2017 IEEE International Conference on Communications (ICC). Paris: IEEE, 2017:1.

[5] CHENG X, MEHRDAD F, MA X Q, et al. Understanding the YouTube partners and their data: measurement and analysis[J]. China Communications, 2014, 11(12):26.

[6] CHENG X, DALE C, LIU J. Statistics and social network of youtube videos [C]//16th IEEE International Workshop on Quality of Service. Enschede: IEEE, 2008:229.

[7] ADHIKARI V, JAIN S, CHEN Y, et al. Vivisecting youtube: an active measurement study[C]//The 31st Annual IEEE International Conference on Computer Communications. Orlando: IEEE, 2012:2521.

(下转第 250 页)