Category-based User Interaction with Online User-Generated Videos: Workload Characterization

Shaiful Alam Chowdhury and Dwight Makaroff

Department of Computer Science, University of Saskatchewan
Saskatoon, SK, CANADA, S7N 5C9
sbc882@mail.usask.ca, makaroff@cs.usask.ca

Abstract

Caching video content in content distribution networks takes advantage of repeated requests amongst a population of users. The efficiency of replication and cache placement policies depends on which user requests which video object.

We develop a methodology for analyzing user interactions, based on video category. We then analyze a trace file of YouTube requests captured at a university campus from the client side of the network. Though the dataset has suffered from content deletion, enough information is available to give some insight into users' viewing patterns. Differentiation in the relative amount of repeated viewing suggests that category-specific distribution and caching policies may provide more efficient use of computational resources. Initial analysis suggests that user access patterns can be identified to a better extent when video category is considered.

1 Introduction

Over the past decade, video-on-demand (VoD) streaming applications and websites, especially user-generated content sites like YouTube, have been the most popular classes of multimedia applications on the Internet [5]. Client/server video distribution has been vulnerable to service bottlenecks due to inadequate bandwidth resources. To improve end-user experience, adoption of Content delivery networks (CDN) and multi-layer caching have been implemented by YouTube [1]. Analysis of the YouTube delivery protocol [2] claims that application level flow control may have undesirable interactions with the network transport layer across the entire delivery path. These interactions are minimized if the video can be stored closer to the users to reduce potential network bottleneck links.

Each local network would have its own request patterns related to the type of user and the usage environment. A coffee shop wireless access point may experience differing request patterns; different users may be given the same IP address via DHCP over time, or the same user returning on consecutive days may or may not be assigned the same IP address. University networks would likely experience different request patterns than residential networks. Moreover, these patterns may change slowly or rapidly over time. Short-term fluctuations in demand could occur and violate the assumptions on which the cache allocations/policies were made, so caching systems must be adaptive.

Our initial results suggest that

- Workload segmentation by category can improve caching performance, even if video uploaders exploit category definition, and
- Identifying user patterns can help cache policy decisions such as replacement policy and dynamic cache sizing.

2 Related Work

Zink *et al.* [6] examined YouTube traffic between YouTube servers and the University of Massachusetts campus network. There were gains to be made by caching the frequently requested objects. The authors determined that proxy caching

was an effective low-cost solution, outperforming local caching and P2P caching where only a single copy was kept in the entire local network.

Recent work examined information available from YouTube partner organizations through Insight analytics [3]. A small number of popular YouTube channels are examined to characterize request behaviour with respect to behaviour of specific users. Different viewing behaviours in different categories was seen, since the channels were associated with unique types of videos. The main contribution regarding viewing patterns based on channels is that partners can influence the viewing on channels by appropriate referral and promotion.

3 Methodology

The accuracy of predicting which videos will be requested multiple times by the same (set of) users may be enhanced if there are relationships between video category and repeated view distributions. There are several workload characterization measurements that can determine the potential caching benefit and guide cache policy design. Combinations of these statistics for a set of categories can inform system designers regarding which cache replacement algorithm to use and how to apportion cache space to different categories.

First of all, the request pattern for the duration of the test period can determine if there is potential for caching. If a power-law distribution for the number of requests exists, then a proxy cache could serve repeated requests by keeping the most frequently used objects. The number of videos at the head of the distribution (most requested) and the rate of decay can determine how much proxy cache space should be allocated per category.

Without more information, we cannot tell if videos become popular due to many users viewing once or a small number of times, or due to attracting the same set of users multiple times. A measure of the density of the request pattern is the average number of views per user for each video. Due to the large number of videos and the large user population, it is likely that this value will be very small, but we expect that there will be videos that some users watch a large number of times. Thus, we capture the median and the maximum number of average views per user to see if there are outlier effects that can capture detailed dynamics of requests.

IP address is used as a proxy for a user, acknowledging that this could correspond to a lab, kiosk or laptop computer using DHCP, or a NAT box. Anomalously high access IP addresses could also represent programs that crawl the video database or target certain videos to increase their popularity.

The *fraction of repeated views* shows the relative number of views of a video were watched by a user more than once. This gives us the maximum hit ratio possible, if the video was cached in each local user's browser cache on the first view and never replaced. This measurement of repeated views, however, could be misleading as it is affected by outliers. For example, the fraction of repeated views is more than 85% for a video watched by 4 users only once, but 30 times by a single user.

Therefore, we also measure the *fraction of singleton views* of videos. Large fractions of singleton views shows *fetch-at-most-once* behaviour and local caching is undesirable. If a video category exhibits high values of average numbers of views accompanied by high singleton views, this indicates a very skewed distribution of requests; a threshold of requests could be used as a factor in replacement.

4 Preliminary Results

Zink *et al.* [6] collected six different datasets from the network of University of Massachusetts. Though this dataset is more than 6 years old, we believe our methodology is appropriate for any category-based dataset; conclusions and implications for cache design would only be relevant in the measured environment over the short term.

We analyze the longest trace (T5) and wrote a crawler to collect the category name for a given video_id for the remaining videos. Table 1 shows the number of videos accessed by category. Music is in the top position, followed by Entertainment and Comedy, similar to data from the University of Calgary [4]. Between 25% and 40% of the videos had more than one request across the categories, showing some distinction. Viewing rate and number of videos accessed are not closely correlated.

Figure 1 shows how well the Zipf distribution fits for News and Music. The goodness-of-fit value (R^2) of 0.95 is very high for Music, but News was comparatively worse. Other categories fit well, except for Autos and Games, which were similar to News, with a distorted head of the distribution. Video deletion may affect the observations, as videos 3 to 100 have lower request frequencies

Table 1: Category Specific Popularities

Category	Videos	> 1 view	
		Videos	Pct
Music	42223	16131	38.20
Entertainment	31484	11127	35.34
Comedy	14942	5985	40.05
People	9749	3407	34.95
Sports	9747	2944	30.20
Film	8100	2932	36.20
News	5295	2073	39.15
Animals	3066	1037	33.82
Howto	2720	941	34.60
Autos	2228	594	26.66
Travel	1737	539	31.03
Education	816	300	36.76
Tech	784	307	39.16
Games	474	121	25.53
Nonprofit	357	149	41.74
Total	133722	48587	36.33
Deleted	130748	52066	40.05

among remaining videos than the original set.

The mean average views/user shows a much higher viewing rate for News than Music (4.62 to 3.2); almost all medians are very small. If a video is watched many times by a small number of users, this distorts the average significantly. In particular, one News video with 513 mean views skews the distribution. The median for Travel videos is 4; users interested in this category watch the same videos multiple times. As the mean is very close to the median, outlier influence is limited.

Figure 2 shows the fractions of videos in selected categories that experience repeated views. The fraction of repeated views is plotted on the X-axis, and the Y-axis shows the complementary cumulative distribution function (CCDF) of the number of videos with the designated characteristic. As further confirmation of the general density of the Travel category request patterns, there are a significant number of repeated views for most videos. Equivalence classes exist for categories. News and Sports are the same for 75% of the videos (low fraction of repeated views), while Music and Film are significantly different for the 60% of low repeated views, but almost identical for the remaining 40%.

The denser view distribution for Music videos than News videos may indicate that the reason for longer-term popularity compared with News is repeated viewings. Film videos are similar to Music videos and thus could have the same reason for longer lifespans. This property of Music videos

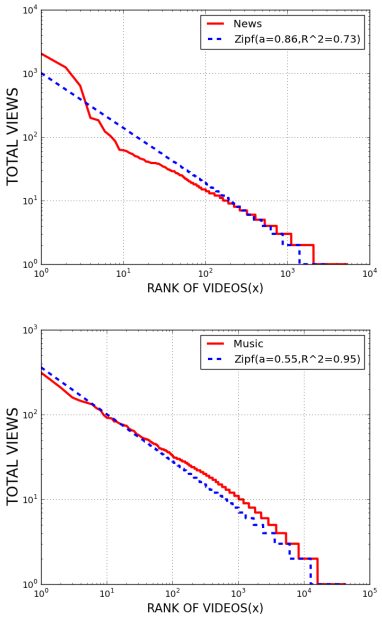

Figure 1: Views vs. rank: selected categories

was confirmed in the analysis of YouTube Channels [3].

With respect to the maximum number of views by a user for videos, Music videos attract users to watch videos repeatedly much more than any other category. Film and Music videos have approximately the same shape, but music is about 4 to 5 times more popular with respect to this metric.

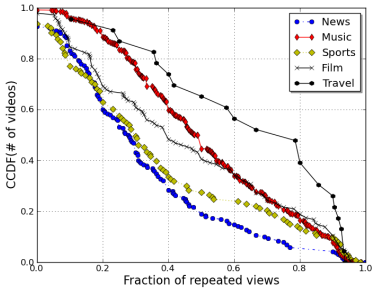

Figure 2: Repeated views of categories

The *singleton views* for our dataset are virtual mirrors of the *repeated views* with few exceptions. For the first 20% of the videos sorted by the lowest fraction of users with singleton views, Music and Sports show the same pattern, but for the last 75% of the videos, Sports and News are comparable.

5 Design Issues

Based on this analysis, we can identify some principles upon which caching policies could be based. For example, the local cache space could be divided up into two components: peer-cache and local-cache. The caches would be adaptive in size and algorithm, depending on the relative hit ratios. To determine if a video should be cached, and which cache to use, popularity thresholds could be set. If video requests exceed a threshold, then it is put in the local cache, which might use an LFU replacement policy, counting total local requests within a medium-term threshold of elapsed time.

Otherwise, it is placed in the peer-cache. Requests from peers for this object would be recorded for use in replacement policies, requiring periodic flooding of this metadata. LRU may be appropriate, since some videos would become popular over a small time period or within a particular geographic or socio-political area of interest.

Proxy caches would see requests not satisfied locally or by peers and could also have a hybrid replacement policy. They would keep request information for all objects in their sphere of responsibility. A community of users may show preference for a category of videos. Communication between local and proxy caches could help inform each other of the request patterns in the local neighbourhood; if an item is evicted from a local cache, the next higher level cache has some idea of the history at lower levels and can properly rank the video when it is requested at that level. These policies could then be extended to categories, where each request is compared to a threshold and all subsequent requests for videos of that category would have the same decisions made. Periodic aging of the requests would enable the system to operate on recent request patterns as the behaviour of uses with respect to categories may shift over time and will likely differ between local contexts. Categories with similar patterns could share a cache space. Any realistic simulation of such caching systems would have to account for the continual generation of new content over any measurement period.

6 Conclusion and Future Work

Request patterns differ between videos in terms of repeat views and singleton views per user, and that category differentiation maintains these characteristics to some degree. Our characterization methodology can be used for any dataset of video requests from a local population of users. Identifying category characteristics and basing caching decisions on these characteristics may provide better use of resources in supplying the streaming bandwidth and storage capacity for the video objects.

Our data set shows potential for category based analysis which matches intuition regarding these categories. Further exploration will be able to quantify the differences and enable accurate parameterization for caching systems. Additionally, there may be patterns in the repetition of views that could be exploited. For example, are views distributed in a regular pattern over the measurement period by a differing number of users or did it have the same burstiness in requests for each user as the overall request pattern?

References

[1] V.K. Adhikari, S. Jain, Yingying Chen, and Zhi-Li Zhang. Vivisecting YouTube: An active measurement study. In *INFOCOM 2012*, pages 2521–2525, Orlando, FL, March 2012.

[2] S. Alcock and R. Nelson. Application flow control in YouTube video streams. *SIGCOMM Comput. Commun. Rev.*, 41(2):24–30, April 2011.

[3] X. Cheng, M. Fatourechi, X. Ma, C. Zhang, L. Zhang, and J. Liu. Insight data of YouTube from a partner's view. In *ACM NOSSDAV 2014*, pages 73–78, Singapore, Singapore, 2014.

[4] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube Traffic Characterization: A View From the Edge. In *ACM IMC 2007*, pages 15–28, San Diego, CA, October 2007.

[5] X. Zhang and H. Hassanein. Video on-demand streaming on the internet - A survey. In *25th Biennial Symposium on Communications (QBSC 2010)*, pages 88–91, Kingston, Canada, May 2010.

[6] M. Zink, K. Suh, Y. Gu, and J. Kurose. Characteristics of YouTube Network Traffic at a Campus Network - Measurements, Models, and Implications. *Computer Networks*, 53(4):501–514, March 2009.