What should you Cache? A Global Analysis on YouTube Related Video Caching

Dilip Kumar Krishnappa
and Michael Zink
University of Massachusetts Amherst
{krishnappa, zink}@ecs.umass.edu

Carsten Griwodz
Simula Research Laboratory
and University of Oslo
griff@ifi.uio.no

ABSTRACT

Following advice from the YouTube recommendation system is one of the ways users browse through the videos offered by YouTube. The system presents related videos based on several factors depending on the current video requested. This related videos list can be used by caching infrastructure to reduce network bandwidth consumption. In this paper, we analyze the differences between user-specific recommendation lists. We perform this analysis on 100s of user nodes from all around the world divided into 4 geographical regions using PlanetLab. Based on our analysis, we find that the related videos differ less in the top half (1-10) of the related video list offered by YouTube compared to the bottom half (11-20). Based on our analysis, we suggest that, caching or prefetching of the Top 10 of the related videos is advantageous over a period of time than caching the whole list offered by YouTube.

1. INTRODUCTION

YouTube has become the world's most popular Internet service that hosts user-generated videos. Viewers can choose from hundreds of millions of videos and over 4 billion hours of videos are watched each month on YouTube by about 800 million unique users. To satisfy the demands of serving this vast amount of content to viewers, Google (the owner of YouTube) uses a network of caches that are globally distributed.

Compared to other video streaming services such as Netflix and Hulu, effective caching is harder in case of YouTube. This is because, YouTube offers a huge collection of videos (> 120 million) compared to the few thousands of titles offered by Netflix or Hulu. Also, the providers of purely professionally produced content determine when new content will be made available and, thus, can much better schedule the distribution of content into caches. In the case of YouTube, the parameters in terms of viewers, availability and newly added videos as well as the popularity development of such new videos are very different and unpredictable, which increases the importance of an efficient and tailored caching approach.

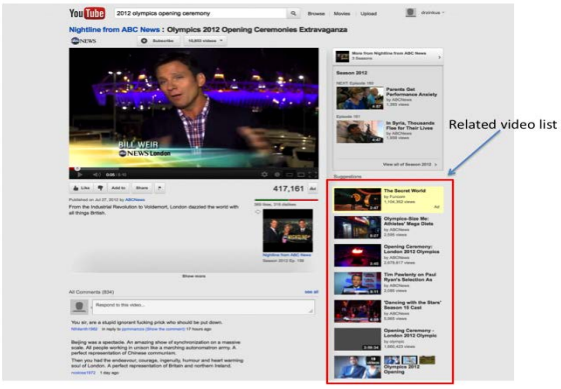


Figure 1: This screenshot shows a YouTube video page and the location of the related video list on that page.

Each YouTube video web page offers a list of 20 recommended videos, next to the video a viewer has selected (see Figure 1). In YouTube's own terms, this is described as the *related video list*. YouTube suggests the recommended videos based on factors such as title resemblance, tags, topics, view count of the videos, and other parameters. Researchers have already investigated the feasibility of caching and prefetching related videos [4, 2], and have shown that viewers make significant use of the related video list. I.e., after watching the initially selected video, the viewer chooses to watch a video that is offered on the related video list next. This information can be used to improve the streaming quality of YouTube through caching and prefetching. To achieve effective related video caching and prefetching, the related video list[1] provided to users served from the same cache should be identical.

In this paper, we investigate the related list differences for the same set of videos recommended by YouTube for different users from across the world. We use PlanetLab to generate YouTube video requests and obtain the related lists recommended by YouTube for those video requests. We use 100s of PlanetLab nodes from all around the world divided into 4 regions, USA (US), Europe (EU), Asia (AS)

[1]From here on, we refer to the "related video list" as "related list".

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
NOSSDAV'13, February 26 – March 1, 2013, Oslo, Norway
Copyright 2013 ACM 978-1-4503-1892-1/13/02 ...$15.00.

and South America (SA) to understand how the related list differs between users from one region to other. We determine two related list difference counts (Content Change and Order Change) to analyze the related video differences between nodes in a region. We look into the order change of the related list between clients because recently we have suggested a recommended list reordering approach [3], which yields better cache hit rate when the positions of the related list are not changed from the previously recommended related list. We also perform a related list difference analysis on the same node for five consecutive days as client based caching and prefetching have been shown to be effective [2]. From our analysis, we find that YouTube suggests related videos differently for different regions and also for different users in the same region. The number of changes is less in the top half (Top 10) than the bottom half (11 - 20) of the related list. From our observations, we suggest that caching or prefetching only the top half of the related video list is advantageous for two reasons. First, viewers tend to click on related videos from the top half of the related video list, and second, the related videos suggested by YouTube tend to change less from client to client in the top half of the list than the bottom half.

The outline of the paper is as follows. Section 2 presents the related work closest to the one presented in this paper. Section 3 provides details of the experiment setup and methodology used to analyze the related list differences. Section 4 presents the results from our analysis of related video differences. Section 5 discusses the impact of the results on efficiency of caching or prefetching related videos and Section 6 concludes the paper.

2. RELATED WORK

Researchers have investigated the performance of proxies and caches for YouTube videos in different contexts. In the following, we mention the ones closest to our work. In [1] and [6], trace-driven simulations were performed to investigate the effectiveness of caching for YouTube videos. Although the traces for both studies were different, the results showed that caching can reduce server and network load significantly. Both studies did not investigate the change in related lists depending on the location or the time a video is requested from.

Besides caching, YouTube's recommendation system (the related video list) has also been studied in related work. In [4], Zhou et al. analyzed two data sets (one directly crawled from YouTube and the other one a trace from a campus network) to investigate if the position of a video in the related list has significant impact on it being selected by the viewers. The results of this analysis show that a large percentage of viewers select videos they watch from the related list. In follow on work [5], Zhou et al. perform further analysis of YouTube's recommendation system based on global statistics for YouTube videos. The authors show that the click through rate is proportional to the position of the video on the related list (the higher the position of the video on the list, the higher the chance that it will be selected by the viewer). While the goal of the work presented in [2] was to show how the prefetching of prefixes from videos on YouTube's related list can improve caching, it also shows that the related list is often used by viewers to select a video. In contrast to the work we present in this paper, no analysis on the related list differences in various nodes is presented.

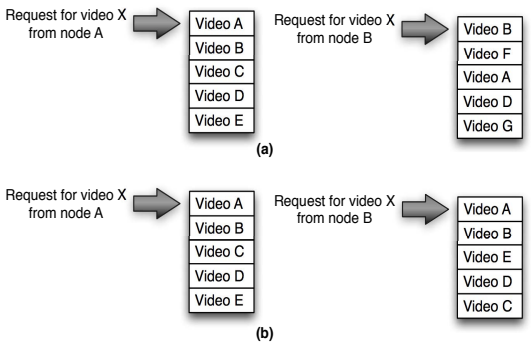


Figure 2: Recommended list changes between requests from different clients: (a) Content Change count, (b) Order Change count

To the best of our knowledge the work we present in this paper is the first that investigates how related lists provided by YouTube in their recommendation system differs from node to node and how this affects the efficiency of caching and prefetching of the related videos.

3. EXPERIMENT METHODOLOGY

In this section, we describe the experiment setup and the measures we use to analyze the related list differences for a set of videos requested from PlanetLab nodes around the globe. Such related list changes can have a significant impact on caching and prefetching performance.

The goal of this measurement is to evaluate if the related list is always identical for all client requests for a specific video or if it changes (either between requests from different clients or between requests for the same video from the same client at different points in time). If the list changes frequently, this can lead to a significant amount of traffic between servers and caches. Thus, it is important to know how much of the related videos change from request to request, which allows us to identify how much of the related videos to cache or prefetch to improve the efficiency of caching and prefetching related videos. We decided to use PlanetLab nodes for our experiment, since it allows us to obtain global and regional information about related list changes.

To investigate differences between the related list for requests that originate from different nodes, we performed the following experiment. We first selected a set of PlanetLab nodes from four different regions (US nodes - 197, Europe (EU) nodes - 243, Asia (AS) nodes - 62, South America (SA) nodes -17, 519 in total). Each of these 519 PlanetLab nodes requests 100 YouTube videos randomly chosen from a trace collected in the US (see Section 5), and for each video request we obtain the related list recommended by YouTube on its video *watch page*. To analyze the difference in related lists obtained from YouTube, we make use of the following two measures:

- Content Change (CC) count: For this measure, we count the number of different videos between two related lists for a specific video irrespective of the position of the video in the list.

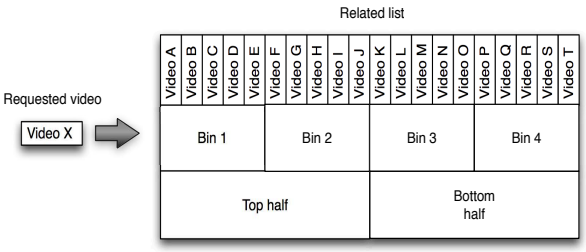


Figure 3: Related Video List abbrevations used in the paper.

- Order Change (OC) count: This measure focuses on the change in the order videos appear in the related list. This measure counts the differences in the related list position wise.

For each video we perform a complete comparison between the related lists retrieved by all PlanetLab nodes. In this specific case, if n PlanetLab nodes retrieve the related list of 100 YouTube videos, then $100 * n(n-1)/2$ comparisons between related lists are performed and the average for both measures (CC and OC) is calculated. Figure 2 shows how we determine the differences between related lists in terms of Content Change and Order Change counts. In this case, the same video (Video X) is requested from two different PlanetLab nodes (node A and node B). Figure 2(a) shows an example where $CC = 2$ and $OC = 4$, while for 2(b) $CC = 0$ and $OC = 2$.

For our analysis we obtained each of the two measures mentioned above for bins of 5 videos from the related list, i.e., 1-5, 6-10, 11-15, and 16-20 related videos (see Figure 3). The goal of determining these measures for subsets of the related list is to identify if the differences are higher or lower in the top half of the related list or bottom half. Also, binning in subsets of 5 videos has an additional effect. The number of video related to the current video must be expected to be rather big in many cases, while the list is always only a few videos long (usually 20). By performing a loop count analysis on the video requested by a user from related lists [3], we have learned that not all potentially related videos are offered in the related lists by YouTube. By binning, we can understand whether the recommended list is derived from a single small pool of related candidates, ordered by some mechanism, or whether videos higher on the recommended list are chosen from a smaller pool of more closely related videos.

In addition to analyzing related list differences between nodes in a region, we also analyze the number of related list differences in the same node for the same video requests for 5 consecutive days. We perform this analysis to better understand two characteristics; How often does the related list change even if the videos are requested from the same node? If it differs, how much of the related video list changes for the video requested from the same node? This investigation allows us to analyze how client based proxy caching and prefetching of related videos gets affected by the related list changes.

We have found that the video size of the related list offered by YouTube has a skewed distribution as shown in Figure 4

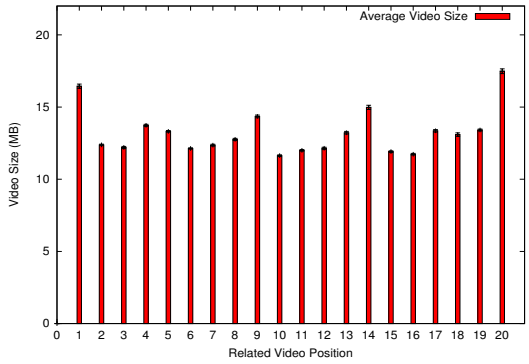


Figure 4: Average Related Video Size Distribution

with an average videos size of 13.23 MB over 2000 related videos from 100 requested videos. Figure 4 shows the average video size of all the related videos at corresponding related video positions and the error bars indicate the 95th percentile of the average video size indicating the skewness in the video size. Considering the scenarios of how many related videos to cache based on the videos size of the related videos and the cache limit is out of scope of this paper. Also, related video list offered by YouTube might differ based on the popularity, view count and region of the video requested. In this paper, we have not considered the related list differences based on these factors and is subject for our future work.

4. ANALYSIS RESULTS

In this section, we present the results from the experiment described in Section 3. We present the related list differences in terms of bins of 5 videos (1-5, 6-10, 11-15, 16-20), and for each of the bins, we provide the percentage of videos with related lists difference (N) from 0 to 5.

In the Content Change count analysis, N=0 indicates that the related video list of the bins are the same in both nodes, whereas N=5 indicates that the related list does not match at all between the nodes for that bin. Similarly, in the Order Change count analysis, N=0 indicates that the related video list is the same in all positions of the related list between nodes, while N=5 indicates that the related list content differs in all the positions between the nodes for that bin.

4.1 Regional Related List Differences

Figure 5 shows the results for the Content Change related list differences between nodes within a region (US, EU, AS, and SA), while Figure 6 shows the results for the Order Change related list differences. It can be seen from Figure 5 that, in each region, related list differences occur between nodes of the same region and the differences grow larger for the 11-15 and 16-20 bins. Figure 5 also shows that for bin 1-5, N=0 between the nodes of a region occurs for 25% to 35% of the comparisons. This percentage of similarity reduces as we move towards higher order bins. For example, in the US region (see Figure 5(a)), the percentage of video comparisons yielding N=5 is 0.84% for bin 1-5 whereas it is 19.34% for bin 16-20. This trend remains the same across all regions and nodes from around the globe.

The Order Change for each of the bins for each region shown in Figure 6 provides result that are different from the Content Change results presented in Figure 5. Also, the percentage of videos with N=5 in Order Change for each related video position bin is higher than the Content Change results presented in Figure 5. For example, the percentage of videos with N=5 in Content Change for US region (Figure 5(a)) for bin 1-5 is about 0.84%, whereas it is about 19.76% for the N=5 in Order Change (Figure 6(a)). This shows that, for the same nodes in each region, though the related video list provided might be same, the position of the videos in the list vary considerably from node to node. Another interesting observation is that the percentage of videos with related list differences are increasing as we move from lower bins to higher bins in the same fashion for Content Change (Figure 5) and Order Change (Figure 6).

4.2 Same Node Differences

As mentioned in Section 3, we are also interested in the related list differences in the case of daily video requests from the same node. We analyze the related list differences for this case to understand how much these differences would affect the efficiency of client-based caching and prefetching. Client based caching and prefetching have been shown to reduce the latency of video requests [2] and improve the viewer's experience.

Figures 7 and 8 show the related list differences for the same node on five consecutive days for Content Change and Order Change counts, respectively. The results show that the related lists change on a daily basis and the trend is similar to the one shown in Figures 5 and 6. I.e., the related video differences are low for the upper related list bins and increases for lower bins. The interesting observation from Figures 7 and 8 is that the number of related list differences remains the same for all related list bins across all regions. The number of differences of related lists for the same set of videos in the US region is the same as that in the EU, AS, and SA region.

This is interesting because it shows that on one hand YouTube uses its regional caches to prepare the related list and does not generate these list centrally. But, on the other hand, the number of list changes for a video that is requested twice from the same node in a 24-hour interval is always the same. In combination with the user behavior mentioned above, this mechanism will prevent the low-popularity videos (the ones from the long tail of the popularity distribution) from being stored on regional caches. But it is counterproductive for the performance of caches since changes in the related list decrease the potential for cache hits.

5. IMPACT OF RELATED VIDEO DIFFERENCES

In this section, we analyze the implications of the related list difference results presented in Section 4 on the efficiency of caching and prefetching in terms of reducing bandwidth consumption and latency.

We first demonstrate why caching and prefetching videos from the related lists can be beneficial. To demonstrate this, we analyze a network trace collected for 3 days at a campus gateway. The details of the trace are provided in Table 1, which shows that the number of videos with "related video" tag are about 45.5% of the total requests received in the trace. This shows that almost half of the requests from YouTube viewers are made by selecting a video from the related list.

Duration	3 days
Start date	Feb 6th 2012
Requests	105339
Videos with "related_video" tag	47986

Table 1: Trace characteristics

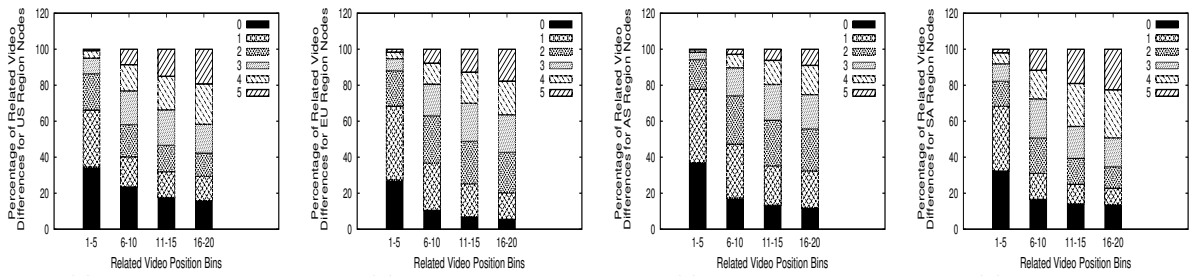
From the trace described in Table 1, we can deduce the percentage of videos selected from the related list based on their positions. Figure 9 shows the percentage of video requests for all related list bins. The result from Figure 9 shows that users usually request videos from the top half of the related list which make up to 80% of the video requests. The remaining 20% of the requests are from the bottom half of the related list, which shows that users usually find their required video or select one of the videos from the top half of related list without scrolling down the whole list.

The question we want to answer is, how do the related list differences between the nodes in a region and the related list differences on the same node for consecutive days affect the efficiency of caching and prefetching related videos? How much of the related list should be cached or prefetched to improve the efficiency of caching and prefetching?

5.1 Impact of Regional Differences

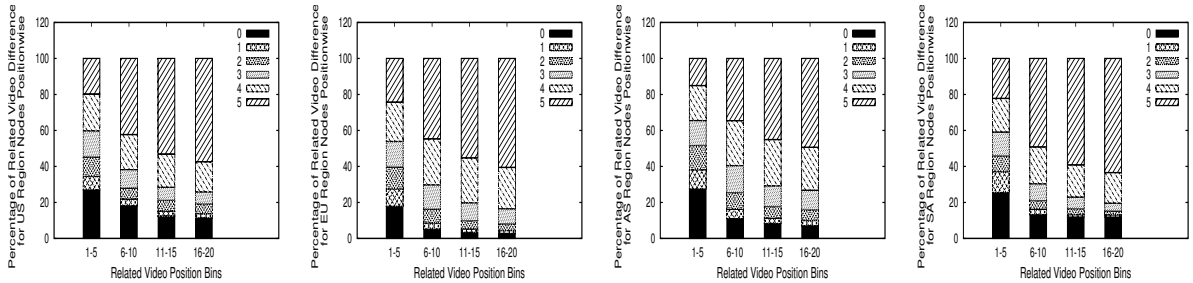
To answer these questions, we analyze the results from Figure 5 and Figure 9. In Figure 5, we can see that about 35% of the time there are related list differences of at least 2 for the top related list bin, which includes the most requested position of the related list as shown in Figure 9 (60%). Assuming an approach in which the Top 5 related videos of every video requested are cached or prefetched, $\sim$ 21% of the time three of the five cached videos will be of no use for subsequent viewers who are served from the same cache. This would require the download of an additional 21% of related videos from a higher level cache or an origin server, increasing backbone network resource consumption. This number increases for lower related list bins. For example, in the bottom half bins there is a related list difference of at least 3 about 65% of the time. About 20% of all requests account for these bins, which leads to the fact that approximately 13% of the cached videos are either replaced or never requested. Though the percentage of additional load due to caching or prefetching is less for the bottom half of the related list, the percentage of related requests suggests that caching or prefetching the top half of the related list increases the hit rate. Hence, less load is imposed on the backbone network compared to caching or prefetching the bottom half of the related list.

We also look into the impact of related list differences based on their order by looking into the results of Figure 6. We have shown the advantages of the related list reordering approach based on the content in the cache [3]. The results from the position centric reordering approach shows a significant increase in hit rate, where the position of the video selection from the related list is held constant. The results of Figure 6 shows that there is about 50% to 60% related list differences of at least 3 in order for the Top 5 related list across all regions, which negates the advantages of the



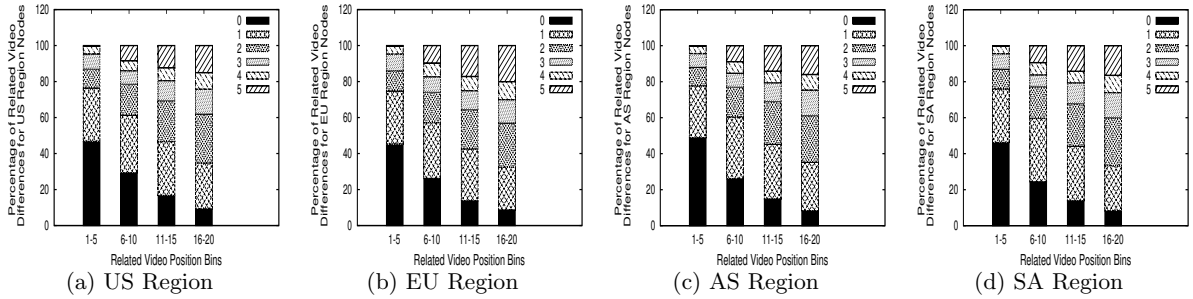
(a) US Region (b) EU Region (c) AS Region (d) SA Region

Figure 5: Average Content Change Related Video Differences



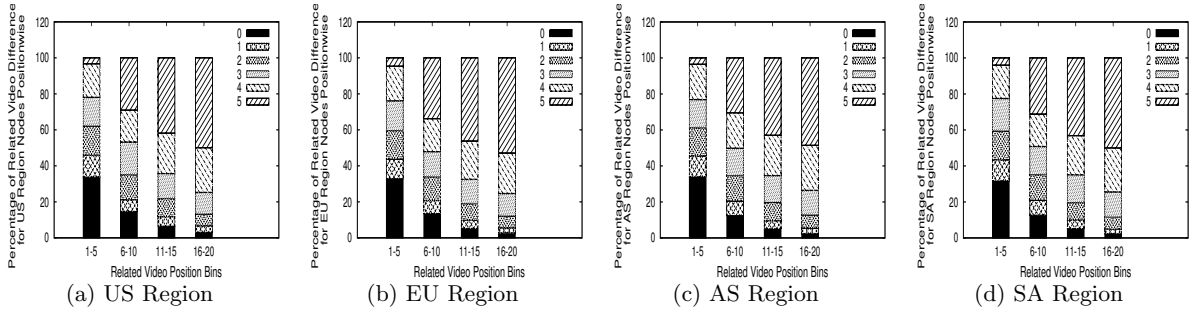
(a) US Region (b) EU Region (c) AS Region (d) SA Region

Figure 6: Average Order Change Related Video Differences



(a) US Region (b) EU Region (c) AS Region (d) SA Region

Figure 7: Daily Average Content Change Related Video Differences



(a) US Region (b) EU Region (c) AS Region (d) SA Region

Figure 8: Daily Average Order Change Related Video Differences

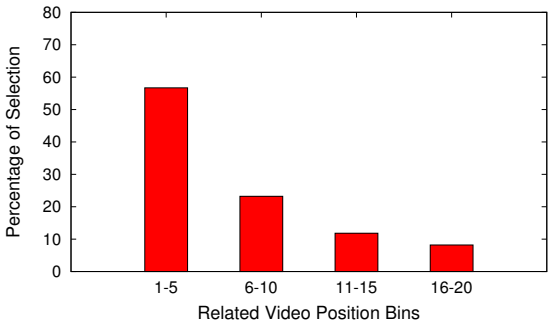


Figure 9: Related Video Position Analysis

proposed position centric reordering approach. This related list differences increases to about 90% when we move to the bottom half of the related list recommended by YouTube.

5.2 Impact of Client Differences

So far, we have only looked into situations where the cache is located at the edge or gateway of a network, but it is sometimes advantageous to perform caching at the client to take advantage of the prefetching of related videos as shown in [2]. The results from the caching and prefetching of related videos for a cache located on the client shows a hit rate of 42% for related videos prefetched up to the top 10 and 48% hit rate when top 20 related videos are prefetched. These results show significant reduction in video latency for each request as we know that clients tend to request videos from the related videos list recommended by YouTube. But, as shown in Figure 7, the related video list changes daily for the same set of video requests from the same clients. This result leads to an additional set of videos being prefetched on a daily basis, which increases the load on the network. As shown Figure 7, the percentage of related list differences of at least 3 is about 20% for the top half of the related video list and it is about 40% for the bottom half of the related video list. From the results provided by [2] for the client based prefetching, the hit rate improves by $\sim 8\%$ by prefetching the top half of related video list compared to prefetching the bottom half. But, this also results in $\sim 20\%$ additional prefetching of videos when we consider the related video differences that occur daily for each client. Combining these two results, we suggest that prefetching only the top half of the related videos leads to a higher hit rate and also reduces the amount of videos that have to be additionally prefetched.

6. CONCLUSION

In this paper, we present an in-depth analysis of one significant part of YouTube's recommendation system, the related video list. Earlier work has shown that this list is often used by viewers to request subsequent videos on YouTube. Thus, caching and prefetching videos from that list has the potential to improve the efficiency of YouTube's video distribution system. This assumption is only true under the condition that the related list does not change between different requests (either in time or location) for a video. Our analysis shows that there are significant changes in the related lists delivered to clients, which has a diminishing effect on efficiency of caching and prefetching. The good news is that, changes are smaller for higher parts of the list (positions 1-5), which are also the most popular ones in terms of viewer selection.

The related list offered by YouTube steers user behavior. The related lists can steer people who are mainly searching and browsing off the long tail, and within a set of active pools. When operating an edge cache, knowing the top videos and the related list pools, the system admin do not have to bother with cache misses because the related list directs the user to those remaining misses on the tail. The paper re-engineers YouTube's current strategy for related video list generation. The list may change at any time, but previous statements from YouTube indicate that their challenge is to reduce the number of cache misses in the long tail of videos. The related lists are a part of a strategy to achieve this, since they provide a way of keeping users watching videos from a common pool. It does not hurt caching performance when that pool changes daily, but it has to be of limited size. Understanding this gives ISPs a better chance of dimensioning their caches and networks.

In future work, our intent is to analyze what parts of a related list transmitted to the clients are already stored on a YouTube cache. To perform this analysis we will make use of measurement technique presented in [3] that allows us to identify if a video is delivered from a cache or not. Also, we would like to investigate if the related videos offered differs based on different factors such as popularity, view count, region etc., of the video requested.

7. REFERENCES

[1] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *IMC*, October 2007.

[2] S. Khemmarat, R. Zhou, L. Gao, and M. Zink. Watching User Generated Videos with Prefetching. In *MMSys*, February 2011.

[3] D. K. Krishnappa, M. Zink, C. Gridwoz, and P. Halvorsen. Cache-centric Video Recommendation: An Approach to Improve the Efficiency of YouTube Caches. In *MMSys*, February 2013.

[4] R. Zhou, S. Khemmarat, and L. Gao. The Impact of YouTube Recommendation System on Video Views. In *IMC*, November 2010.

[5] R. Zhou, S. Khemmarat, L. Gao, and H. Wang. Boosting Video Popularity through Recommendation Systems. In *DBSocial*, June 2011.

[6] M. Zink, K. Suh, Yu, and J. Kurose. Characteristics of YouTube Network Traffic at a Campus Network - Measurements, Models, and Implications. *Elsevier Computer Networks*, 2009.