Category-Based YouTube Request Pattern Characterization

Shaiful Alam Chowdhury and Dwight Makaroff

Department of Computer Science,
University of Saskatchewan, Saskatoon, SK, Canada S7N 5C9
{sbc882,makaroff}@cs.usask.ca
http://www.cs.usask.ca

Abstract. Media content distribution systems make extensive use of computational resources, such as disk and network bandwidth. The use of these resources is proportional to the relative popularity of the objects and their level of replication over time. Therefore, understanding request popularity over time can inform system design decisions. As well, advertisers can target popular objects to maximize their impact.
Workload characterization is especially challenging with user-generated content, such as in YouTube, where popularity is hard to predict *a priori* and content is uploaded at a very fast rate. In this paper, we consider category as a distinguishing feature of a video and perform an extensive analysis of a snapshot of videos uploaded over two 24-hour periods. Our results show significant differences between categories in the first 149 days of the videos' lifetimes. The lifespan of videos, relative popularity and time to reach peak popularity clearly differentiate between news/sports and music/film. Predicting popularity is a challenging task that requires sophisticated techniques (e.g. time-series clustering). From our analysis, we develop a workload generator that can be used to evaluate caching, distribution and advertising policies. This workload generator matches the empirical data on a number of statistical measurements.

Keywords: Workload characterization: multimedia applications: content distribution; time-series clustering

1 Introduction

YouTube and other user generated content (UGC) sites have altered the way people watch Internet video. YouTube was the 4^{th} most accessed Internet site in 2007 [6], and its use was increasing over time in a power-law manner. Recent studies support two central observations: 1) increasing number of videos/users [8, 16] and 2) dissatisfying experiences of users in watching YouTube videos [13]. Other studies [10, 14, 15] suggest that YouTube is the most bandwidth intensive service of today's Internet, accounting for 20-35% of Internet traffic.

Much research has been done investigating request characteristics from both client [11, 19] and server perspectives [2, 5, 8, 9] in order to enable improved service. However, none of this earlier work considered the categories of video objects. This aggregate data may not tell the whole story.

A proper understanding of YouTube's workload will aid in the design of new systems, as well as capacity planning, and network management for similar types of systems. The methodology we have developed is useful for UGC sites that have a single cache for the region of requests captured. YouTube itself operates on such a global scale that a single cache would not be sufficient. Rather, multiple regional caches satisfy regional demand patterns that have been shown to differ between different regions in the world [3]. If regional request data was available through the standard API, we could account for multiple caches in our analysis.

In this paper, the time-varying global viewing patterns of a sample of YouTube videos from their upload time are analyzed, considering video category.[1] We present the results of one data collection period (5 months of views of videos uploaded in 2 consecutive days). We show that different categories exhibit different viewing patterns in terms of overall popularity and detailed popularity over time. In fact, it is possible to predict the future popularity of some categories of videos at very early ages, because of correlations over time. We confirmed that the number of views of the popular videos follows a Zipf distribution for most categories, whereas views of the unpopular videos follow a heavy tail distribution. We also find that the uploading trends in YouTube have changed over time. People are now uploading more user generated content (UGC) compared to earlier observations. We also show that time-series clustering can be successfully used to understand the growth patterns for the categories where early popularity cannot be used to predict popularity in the rest of the measurement period.

These observations contribute to a better understanding of the popularity dynamics of YouTube videos, enabling realistic testing scenarios for developing and evaluating various design parameters for UGC sites. Request patterns for different categories may vary around the world; our dataset and analysis provide a case study that shows that global category differences persist, and therefore, will exist in each region. Our analysis enables the development of category-specific workload generators which can be combined to form the input for simulators and prototype systems. While developing and evaluating a comprehensive workload generator remains as future work, we have a strategy for generating synthetic requests on a category basis and present preliminary results which match reasonably well for two categories: News and Music.

The remainder of the paper is organized as follows. Related work is described in Section 2. Section 3 explains the data collection methods. Request patterns are discussed in Section 4, and we use views over time to develop a workload generator in Section 5. Section 6 provides conclusions and future work.

2 RELATED WORK

Previous request characterization and video popularity analysis has been used to investigate the feasibility of different content delivery streaming techniques, and to design and evaluate caching policies/systems for UGC sites. Our work leverages this research to investigate *category* popularity over time.

[1] as defined by the uploader

YouTube video request traffic was captured at the packet level at the University of Calgary over a 4 month period [11]. They investigated video popularity properties, usage patterns, and transfer behaviours as measured from the client edge of the distribution network. The traces examined contained data from both completed and incomplete requests. Their analysis suggests that appropriate caching decisions not only can improve end user experience, but also reduce network bandwidth usage. Another study [19] observed the traffic of YouTube videos between a university campus and the YouTube server. Approximately 25% of the videos in the trace were requested more than once, leaving a long tail in the distribution. Three different content delivery techniques were analyzed: P2P based distribution, proxy caching and local caching. Proxy-caching outperformed the other techniques, and P2P based distribution sometimes exhibited worse performance than local caching. These two results can be biased by the measurement locations that restrict the context of the studies and the proposed solutions. For instance, it is claimed that video requests in YouTube follow a Zipf distribution [11], which is different from other works that consider global request patterns [1, 5]. For our purposes, global access patterns are essential.

2.5 million YouTube videos were obtained using related video links [6] in a study at Simon Fraser University. Access patterns of the popular videos did follow a Zipf-like distribution, in spite of having a heavy. This indicated that the YouTube network is similar to small world networks, and P2P techniques could be successfully applied, contradicting earlier findings [19]. Their dataset is likely to be biased to popular videos because of the crawling approach, and popularity over time is not investigated.

A recent approach to investigate growth patterns in YouTube video requests was to use Google charts to collect views over time [9]. They analyzed the time-varying viewing patterns of popular videos, deleted videos and randomly selected videos. Popular videos usually experience a huge number of views on a single peak day or week. Unfortunately, using the Google charts API is not sufficient to have a proper, fine-grained understanding of the dynamics of video popularity as Google charts API always returns 100 data points, regardless of video age.

Recent work was done on nearly 30,000 videos, collected by using the recently uploaded standard feed provided by the YouTube API [2]. Their collection procedure claims to have an unbiased dataset; the *Most Recent* standard feed returns video information randomly that are uploaded recently. Most of these videos experienced their peak popularity within fewer than six weeks of their uploading time. Video collection based on keyword search is shown to be biased to popular videos, suggesting that the method of data collection is important.

3 DATA COLLECTION

No prior work measures the daily views of different categories of YouTube videos from the first day of their uploading time. We modified previous unbiased data collection methods [2], since we speculate that the first week since uploading deserves more investigation, even though this may expose day-of-week effects.

Moreover, similar numbers of videos from all the categories are needed for appropriate comparison between different categories.

Multiple crawlers were deployed to obtain data used in our analysis:

(1) Most Recent crawlers. 15 crawlers were deployed on March 3^{rd}, 2012 (a Saturday) to collect video IDs for 15 different categories,[2] by restricting each crawler's queries to a single category from those available for upload on that date. Though a video may be assigned to more than one category, we use the categories selected by the YouTube API. All crawlers collected video information for 24 hours. The *Most Recent* standard feed provides video information randomly, reducing bias. A similar procedure was followed on March 4^{th}, 2012. After two days, 71,208 videos' information was obtained. The dataset size is limited by the YouTube API, returning information for at most 100 different videos to each crawler every 1 or 2 hours.

(2) Video view collection crawlers. Video view collection using two separate crawlers was started from March 4^{th}, 2012 and March 5^{th}, 2012. This continued for 149 consecutive days (approximately 5 months). The crawlers ensured a 24-hour difference between view collections. Normalization was performed on the first day's views. Due to network connection failures, some video views on days 20 and 58 of the measurement period were not captured. Fortunately, those days are not that important for most of the videos; most significant events occur very early. Thus, 147 day's views are analyzed. After 149 days, the number of videos in the dataset fell from 71,208 to 47,711 (an average deletion rate of 33%). Table 1 shows the summary of our dataset. Howto, Film, Entertainment and Tech videos experience the highest deletion rates.

(3) Uploading rate crawlers. Another crawler was developed that collected category names of videos provided by YouTube's *Most Recent* standard feed. The crawler ran for 5 months, starting from February 2^{nd}, 2012 and collected approximately 365,000 unique videos' information. This allows us to estimate the short-term current category-specific uploading rates. While not an accurate representation of the entirety of YouTube, it does give some insight.

4 VIDEO REQUEST ANALYSIS

4.1 Time-Varying Category Popularity

Figure 1 shows the cumulative distribution functions (CDF) of time-to-peak for the videos from different categories with at least 100 views; a video with a very small number of views has no actual growth pattern. One consequence of this restriction is that the number of videos in each category is significantly reduced,

[2] https://developers.google.com/youtube/2.0/reference#YouTube_Category_List. Last accessed: 09-05-13.

Table 1. Categories and Number of videos

Category	Number of videos (Day 1)	Number of videos (Day 149)	Deleted videos Pct
Howto	4773	1772	62.87
Film	4654	2346	49.59
Ent.	4991	2528	49.34
Tech	4942	2682	45.73
Games	4711	2966	37.04
People	4310	2730	36.65
Autos	4714	3245	31.16
Comedy	4744	3467	26.91
News	4623	3432	25.76
Travel	4918	3698	24.80
Sports	4812	3733	22.42
Music	4774	3477	21.93
Nonprofit	4624	3691	20.17
Education	4710	3801	19.29
Animals	4908	4143	15.58
Total	**71208**	**47711**	**33.00**

down to 42% for News and Sports and 18% for Animals and Travel. We define *time-to-peak* as the day in which a video experienced the most views [2].

Time to reach peak popularity is not the same for all categories. News and Sports categories follow a similar distribution with the shortest time to reach their peak. Approximately 85% of News and Sports videos peak within the first 4-5 days of their lifetimes. As well, between 50% and 60% of the videos in almost every category experience their peak viewing on the first day.

Other categories such as Music, Film, Howto, Tech and Education follow similar patterns and many videos in these categories reach peak popularity much later. The remaining categories follow similar distributions, and peak distributions of these categories lie within the previous two groups.

The significance of time-to-peak can be enhanced the CDF of total views over time for all videos in a subset of categories (Figure 2). Music and Film videos experience relatively fewer views early in their lifetime. Film videos follow an almost constant viewing rate for the entire measurement period. News and Sports videos, however, experience a significant portion of the total views early.

It is important to understand if the peak day differs significantly from other days of a video's lifetime in order to determine if our previous statistic is helpful. Figure 3 shows the complementary cumulative distribution function (CCDF) of the most distant day x after the peak such that the views on day x is at least 50% of the peak views, defined as follows:

$$x \ = \ max(i) : view(i) \geq 50\% \times view(peak) \wedge (i \ > \ peak) \tag{1}$$

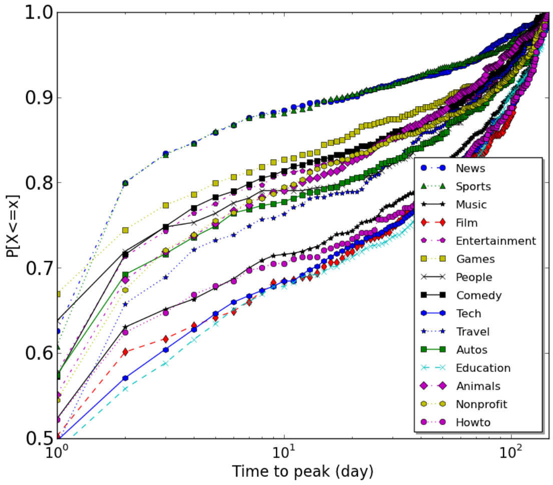


Fig. 1. CDF of time-to-peak

where $view(i)$ is the views on day i and $view(peak)$ is the number of views on the peak day. Only videos with more than 100 views are considered. Figure 3 shows the peak day as a unique point in the lifetime of videos for faster-growing categories (e.g., News and Sports). These categories experience a popularity burst, and quickly decline to a lower viewing rate.

Many Music, Film, Howto, Education and Tech videos that reach peak popularity comparatively lately do not have that drop in their popularity (Figures 1 and 3), so time to reach peak popularity is proportional to the active lifespan of a video. For example, over 75% of the News and Sports videos *never* experience half of their peak days' views after the peak day (Figure 3), but fewer than 50% for Film and Tech videos have this characteristic. The stability of Film and Tech videos suggests that a longer measurement period would increase the difference between these categories and News/Sports.

We are also interested to know if the categories that reach peak popularity faster than others also experience differing numbers of views. Figure 4(a) depicts the 95^{th} percentile of views of selected categories over time. We show the 95^{th} percentile to remove the potential effect of outliers. This shows the minimum percentage of popular videos (5%) during the first 100 days of data collection and the relative popularity of the categories for those popular videos.

These graphs illustrate how viewing patterns of different categories change throughout the early part of their lifetimes. Although the most similar dataset collected [2] indicates that the views of Music category exceeds all other categories within their 8-month measurement period,[3] our dataset shows that pop-

[3] We used another crawler to collect categories for the videos which remained.

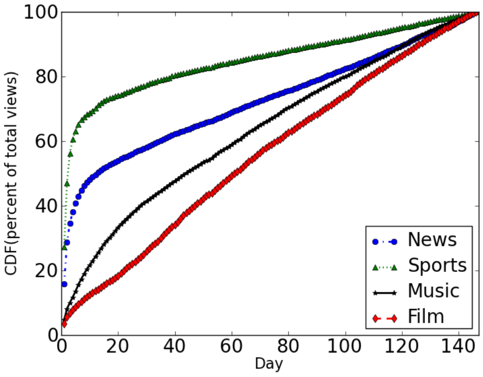


Fig. 2. Percent of total views over time

ular News, and Sports videos enjoy higher viewing rates than any other types of videos for the first couple of days since publication. Almost all categories have at least 5% of their videos experience a high initial viewing rate, but after these few peak days, views for most of the categories become very low, except Music, Film and Tech, showing the variations in active life spans of different categories.

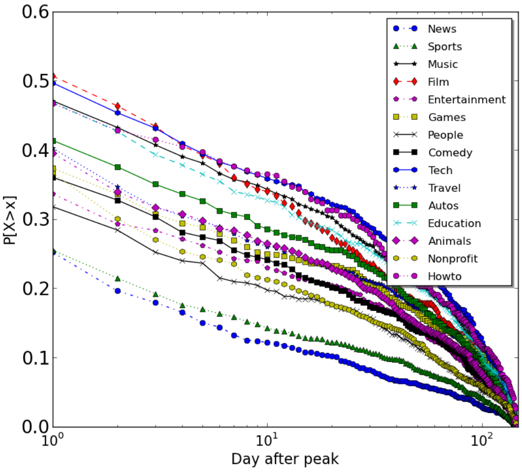
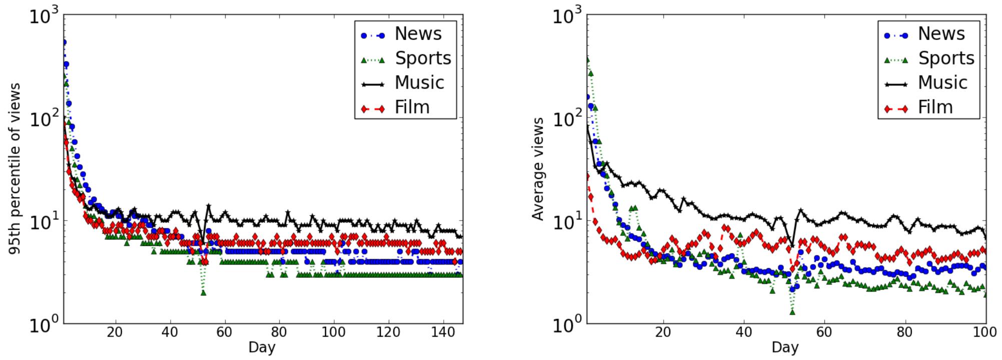


Fig. 3. CCDF of time-after-peak

Although similar results can be observed from the average views per day (Figure 4(b)), the high variance of views may distort the statistic. The most popular video in the dataset is a Sports video, (24 times the 2^{nd} most popular Sports video), increasing the early average views of Sports videos substantially.



(a) 95th percentile (b) Average

Fig. 4. Views per day

4.2 Fractions of Popular Videos

The percent of videos with different views of the YouTube categories are shown in Table 2. Only approximately 10% of the Music videos enjoy *fewer* than 10 views; this value is over 30% for Howto, People, Autos, Comedy, and Travel. Music, News, Sports, and Film contain most of the popular videos in our dataset ($>$ 1.11% with over 10,000 views). The most unpopular videos are in the Travel category, followed by Comedy and Animals. Only 0.44% of the People videos had more than 10,000 views, in spite of the highest uploading rate (shown later). Although uploaders currently upload more UGC videos, users are still not attracted to UGC videos compared to UCC (user copied content) videos.

4.3 Current Uploading Rate

In order to design a request generator for YouTube, the category uploading rate must be known. In 2007, Music was in the top position in number of uploaded videos followed by Entertainment, Comedy, Sports and Film [6]. Manual sampling revealed that these categories are now dominated by UCC rather than UGC content; most of the videos in YouTube were likely UCC then as well.

Figure 5 shows the current uploading trend of YouTube videos obtained by crawler 3. The uploading trend in YouTube has changed over time. The People category is at the top position with approximately 24% of all the new videos, which was at the 6^{th} position in 2007, only 8% of all the videos. Samples from the People category contain comparatively more UGC objects than other categories.

Table 2. Relative Video Popularity

Category	**≤10 views**		**11-100**		**101-1000**		**1001-10000**		**10001-100000**		**> 100000**	
	Pct	Num	Pct	Num	Pct	Num	Pct	Num	Pct	Num	Pct	Num
Music	10.4	363	48.7	1694	32.9	1143	6.4	222	1.29	45	0.29	10
News	18.9	647	39.6	1358	31.6	1085	8.4	289	1.4	48	0.15	5
Sports	20.8	776	46.0	1717	26.1	975	6.0	223	1.04	39	0.08	3
Tech	22.6	605	47.3	1268	24.6	660	4.9	130	0.63	17	0.07	2
Film	23.1	541	49.5	1162	20.8	489	5.5	128	1.07	25	0.04	1
Entertainment	27.8	702	46.9	1185	20.6	521	3.9	98	0.75	19	0.12	3
Howto	43.8	776	34.6	613	17.0	302	4.0	71	0.45	8	0.11	2
Nonprofit	24.1	890	48.0	1773	23.5	867	3.9	142	0.46	17	0.05	2
Education	24.7	940	48.8	1856	21.7	825	4.3	165	0.37	14	0.03	1
Animals	25.6	1060	56.5	2340	15.5	643	2.1	85	0.34	14	0.02	1
Games	27.5	816	49.4	1464	19.1	566	3.4	102	0.51	15	0.1	3
People	29.5	806	49.9	1363	17.7	483	2.4	66	0.4	11	0.04	1
Autos	30.6	992	41.5	1345	23.2	752	4.1	132	0.68	22	0.06	2
Comedy	32.3	1121	51.1	1771	14.1	488	2.1	72	0.35	12	0.09	3
Travel	33.8	1248	48.9	1808	15.4	571	1.8	65	0.14	5	0.03	1

4.4 Category Popularity Distributions

Figure 6 shows the Rank-frequency distribution for the 6 categories that showed the most interesting patterns. Other categories followed one of these patterns. Previous studies [1, 6] showed that although requests for popular YouTube videos follow a Zipf-like distribution, a Weibull distribution fits better because of the heavy tail section, indicating a large number of very unpopular YouTube videos. After considering video categories, only News videos follow a Weibull distribution (and first 80% with better accuracy), because of the comparatively flatter head section of News access pattern. This is consistent with *fetch-at-most-once* behaviour [12], as expected in watching News videos. For all the other categories, request distributions of only the popular videos follow a Zipf-like distributions and the tail sections of these categories can be fitted to a Weibull distribution with a high goodness of fit (R^2). Our dataset indicates a very light tail. The number of videos exhibiting Zipf behaviour differs between the categories.

Another measure that we calculated was the CCDF of total views over the measurement period. There were a substantial number of videos in certain categories that had at most 1 view, potentially skewing the popularity measures. The HowTo and Autos category had 17% and 12.6% of videos with at most 1 view, respectively, while 9% of HowTo videos had 0 views. There is a section of completely unpopular videos that get published, but never viewed. Figure 7 shows the CCDF of the total views for a selected number of categories. We truncate the x-axis to see the behaviour of views for unpopular videos more clearly. Entertainment is used as an example of a group of categories that had very similar CCDFs. The shape of the distribution of total views is very similar

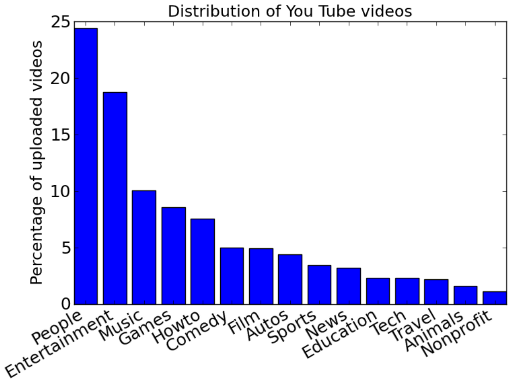


Fig. 5. Category Uploading Rate (365,000 videos)

in these categories, but that of views over time is not. Music has very few videos below 20 views, but HowTo has almost 50% of the videos below 20 views.

5 TOWARDS A WORKLOAD GENERATOR

5.1 Predicting Popularity

As an approach to predict future popularity of videos, Pearson's correlation coefficient (Equation 2) is calculated between the added views[4] at different snapshots of the measurement period.

$$r_{xy} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2} \qquad (2)$$

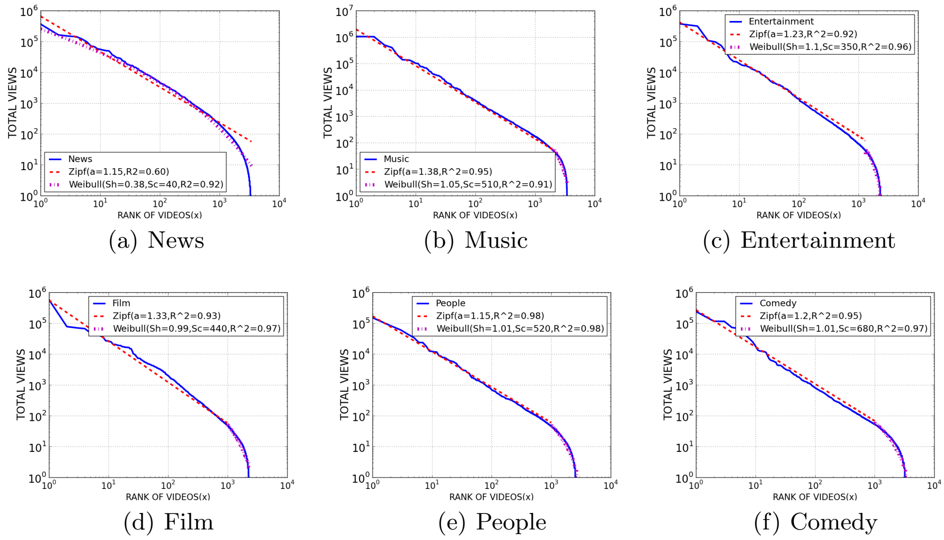
A high correlation coefficient between early views and and the rest of the period implies that prediction of future views of individual videos is achievable [17]. We got very encouraging results for some of the categories including Sports, Travel, Howto, Tech and Games.[5] However, for other categories like Film, News, Entertainment the coefficients are very poor, indicating significant changes in the set of popular videos. Music shows a bit different characteristics than others (good correlation with the rest of the measurement period if we take first 10 days as our first snapshot). Figure 8 explains why early views of Sports (Film) videos can (cannot) be used as a good predictor of future views.

5.2 Three-phase Characterization

The three-phase characterization of Borghol *et al.* [2] considers average viewing rate over time to be constant when the videos are grouped into at-peak, before-peak or after-peak on a particular day, because similar view distributions exist for

[4] Added views is the number of views on a particular day

[5] Sports is 0.99 for the first day's views and the rest of the measurement period



(a) News (b) Music (c) Entertainment

(d) Film (e) People (f) Comedy

Fig. 6. Number of views against rank for categories

the entire measurement period. This fairly simple approach requires only three fixed distributions plus the fixed peak distribution for the entire modelling.

Figure 9 shows the average viewing rate for News videos grouped at their peak phases. Showing results for only one of the three phases is enough, as the three-phase characterization method can only be applied when a constant rate is found for all three phases. We observed similar results for all other categories, which suggests that the viewing rates over time are not constant for any YouTube categories. The high and highly variable average views for News videos at the end of the measurement period is because very few videos reach peak popularity around that time. Otherwise, a decay in viewing rate is observed for the first two months, contradicting the time-invariant nature observed previously [2]. For some of the days, no videos were at their peak popularity.

5.3 Time-Series Clustering

This category variation led us to model the growth patterns differently. We decided to investigate whether the popularity growth patterns of videos in a specific category follow similar shapes. This can be considered as a time-series clustering problem and becomes challenging as different videos reach peak popularity at different times. Inspired by a study on viral videos [4], we translate all the time-series so that the x-axis is centred on the peak day, since most of the significant events happen around the peak periods.

Another challenging issue is to select the appropriate time-series clustering algorithm. We are particularly interested to identify similar shapes of the views

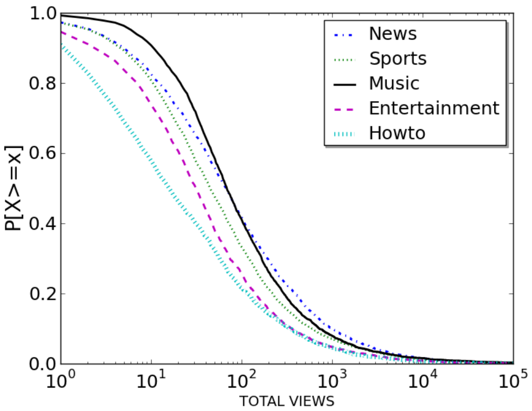
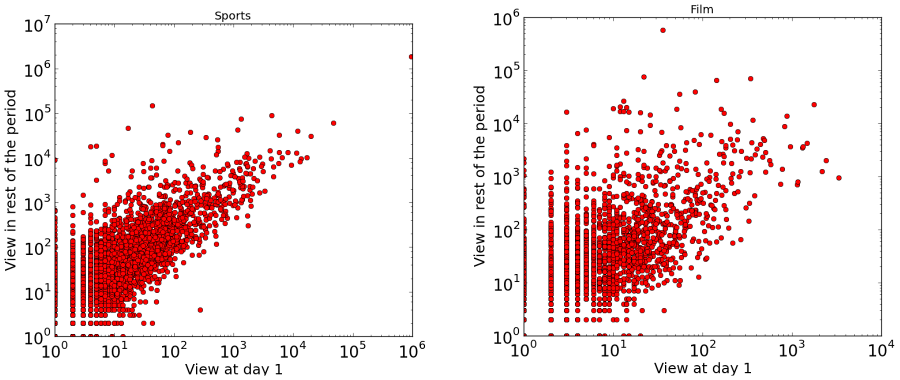


Fig. 7. Selected CCDF of total views

per day, regardless of the time to peak. Moreover, the algorithm should not be affected much by outliers. We selected K-SC clustering [18], which has been found to be accurate in identifying the growth patterns of other Web content. Unlike K-means clustering, K-SC cluster centroids are not distorted by outliers. Instead of considering Euclidean distance between the curves, K-SC applies a scale and shift invariant distance metric [7]. We evaluated the performance of K-SC algorithm for multiple categories. Only Music is shown. The clustering was performed for the top 2000 videos in order to present more accurate results.



(a) Sports (b) Film

Fig. 8. View changes of videos between two different snapshots

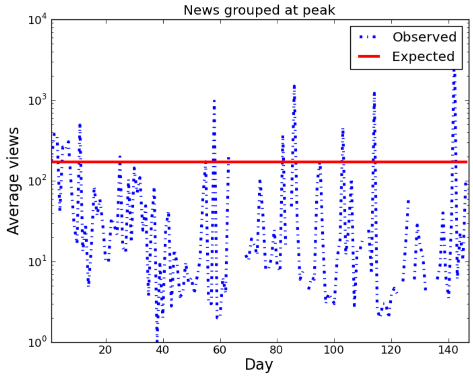
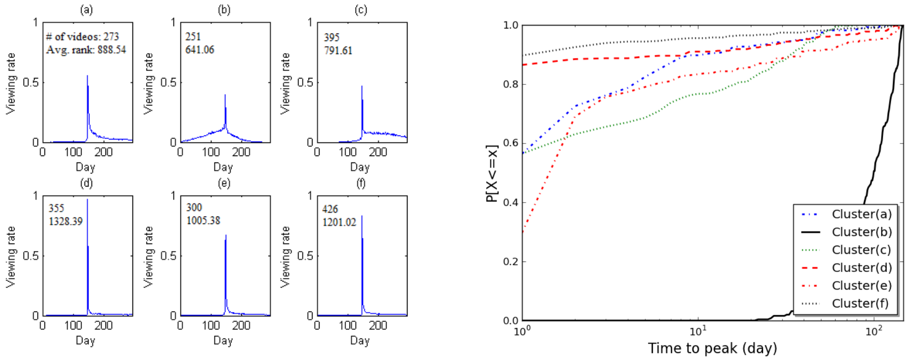


Fig. 9. Average views over time for News videos at peak

Figure 10 (a) shows the six clusters for Music videos found by K-SC. Forcing K-SC to select fewer than six clusters drops the accuracy significantly, as we lose some of the interesting patterns. However, more than six clusters does not significantly improve the accuracy as similar clusters repeat.

The cluster shapes for News videos (not shown) are very similar to Music, except very little difference between cluster *(a)*. However, the number of videos in each cluster differ between these two categories, complementing our earlier findings. 46% of Music videos are contained within the slower-decaying clusters; this drops to 15% for News videos.



(a) Music clusters

(b) Peak distribution of music clusters

Fig. 10. Cluster information for music videos

An important question that must be answered is whether a particular cluster is more biased to popular videos than others. This can be answered by taking the average rank value of all the videos in a cluster. The central limit theorem suggests that the average rank of each cluster should be 1000 if it is not biased. For News videos, the average rank values are similar for each cluster (near 1000). For Music videos, the clusters with slower decay contain more popular videos, with average rank values of approximately 750. Popular Music videos observed a sharp decay with less frequency than popular News videos.

5.4 Performance of K-SC

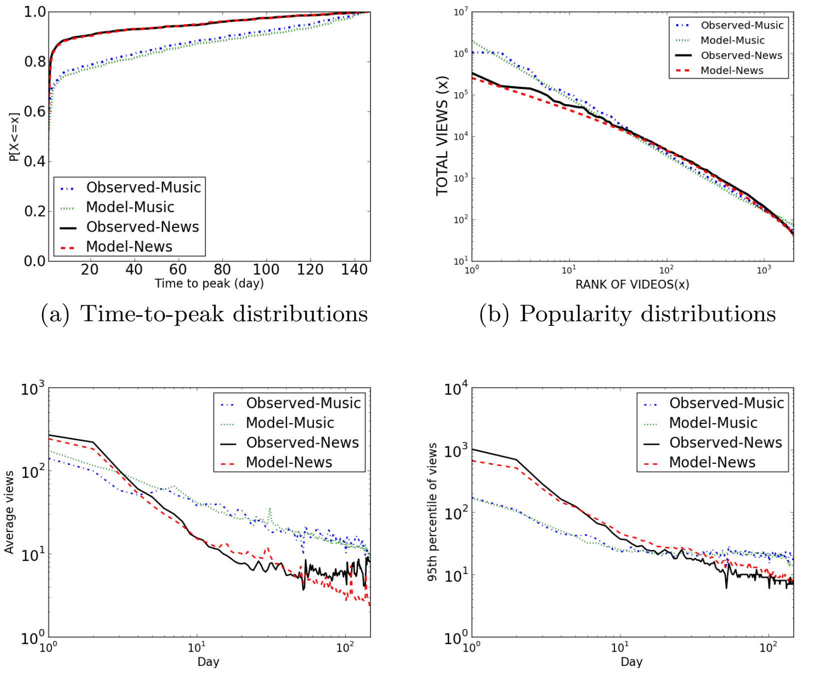
In order to evaluate the performance of K-SC, we designed a synthetic workload generator for News and Music videos. The synthetic data should show similar characteristics to the empirical YouTube data if the clustering of K-SC is accurate. The workload generator can be described as follows. A rank value is assigned to each of the videos as suggested by the chosen distributions for Music and News respectively. Then centroid/cluster is assigned to the videos based on the distribution we observed earlier. We also imposed a little bias for the popular videos before selecting the appropriate cluster in order to match our observed average rank value. As the peak distributions are conspicuously different among the clusters in a category (Figure 10(b)), each of them are considered separately in the request generator, so that the accuracy of K-SC can be verified.

We test similarity between the synthetic and empirical data from four different perspectives: 1) The total view distribution, 2) time-to-peak distribution, 3) Average daily views over time, and 4) 95^{th} percentile of views over time. Figures 11(a) and 11(b) indicate very good matches between synthetic and empirical data for metrics 1 and 2, which does not in itself indicate high accuracy of K-SC. We imposed the distributions for these two cases from our observations,. Metrics 3 and 4 show, however, that the clusters found by the K-SC algorithm for both categories represent most of the videos growth patterns (Figure 11(c) and 11(d), respectively).

6 CONCLUSIONS AND FUTURE WORK

In this paper, we analyzed global daily viewing patterns of a representative subset of YouTube videos from upload time until they were 5 months old. We discovered significant time-varying popularity differences between categories. Most videos exhibit their peak viewing day very soon after publication and then there is a decay; relatively few videos ever approach peak popularity again. Video categories that reached their peaks later were more stable. This is expected and matches our intuitions. We developed an analysis method that permits quantification of these differences on a particular dataset. The confirmation of Zipf distributions for the total views of popular videos in nearly every category indicates that caching would be effective. One limitation is the accuracy of category identification, especially for those videos that belong to multiple categories.

We determined the relative trends of category-specific viewing patterns in the first few months since upload. Some categories contain a non-trivial number of videos which are still popular 5 months after upload date, whereas other categories dwindle to nothing. Some categories have videos which exhibit stationary behaviour that allows prediction of future popularity. Popularity changes around peak time can be captured by appropriate time-series clustering. Unfortunately, scale and deployment issues make direct applicability to YouTube impractical. Our methodology and analysis could be used to help design, configure, and deploy any category-specific UGC site. We developed a workload generator that matches with the empirical data for several categories; similar clusters exist in each category, but different numbers of videos belong to each cluster.



(a) Time-to-peak distributions

(b) Popularity distributions

(c) Daily Views over Time

(d) 95th percentile of views over time

Fig. 11. Modelling Distributions

As future work, we are in the process of building a complete workload generator that encompasses more aspects of user-generated content video requests. In particular, we will incorporate category-specific introduction of new content over time to drive simulations and/or prototype content distribution networks to evaluate different design policies for storing and delivering videos.

Acknowledgments. The authors would like to acknowledge the support of the University of Saskatchewan's Dean's Scholarship Program.

References

1. A. Abhari and M. Soraya. Workload Generation for YouTube. *Multimedia Tools and Applications*, 46(1):91–118, January 2010.
2. Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. Characterizing and Modelling Popularity of User-Generated Videos. *Performance Evaluation*, 68:1037–1055, November 2011.
3. A. Brodersen, S. Scellato, and M. Wattenhofer. YouTube Around the World: Geographic Popularity of Videos. In *WWW*, pages 241–250, Lyon, France, April 2012.
4. T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer. Catching a viral video. In *IEEE Data Mining Workshops*, pages 296–304, Sydney, Australia, December 2010.
5. M. Cha, H. Kwok, P. Rodriguez, Y. Ahn, and S. Moon. Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems. *IEEE/ACM Trans. Netw.*, 17(5):1357–1370, October 2009.
6. X. Cheng, C. Dale, and J. Liu. Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study. Technical report, Cornell University, arXiv e-prints, July 2007.
7. K. K. W. Chu and M. H. Wong. Fast time-series searching with scaling and shifting. In *ACM PODS*, pages 237–248, Philadelphia, PA, May 1999.
8. Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, and A. Ghose. Broadcast Yourself: Understanding YouTube Uploaders. In *ACM IMC*, pages 361–370, Berlin, Germany, November 2011.
9. F. Figueiredo, F. Benevenuto, and J. Almeida. The Tube over Time: Characterizing Popularity Growth of Youtube Videos. In *ACM WSDM*, pages 745–754, Hong Kong, China, February 2011.
10. A. Gember, A. Anand, and A. Akella. A Comparative Study of Handheld and Non-handheld Traffic in Campus Wi-Fi Networks. In *PAM*, pages 173–183, Atlanta, GA, March 2011.
11. P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube Traffic Characterization: A View From the Edge. In *ACM IMC*, pages 15–28, San Diego, CA, October 2007.
12. K.P. Gummadi, R.J. Dunn, S. Saroiu, S.D. Gribble, H.M. Levy, and J. Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *ACM SOSP*, pages 314–329, Bolton Landing, NY, October 2003.
13. S. Khemmarat, R. Zhou, L. Gao, and M. Zink. Watching User Generated Videos with Prefetching. In *ACM MMSYS*, pages 187–198, San Jose, CA, February 2011.
14. C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian. Internet Inter-Domain Traffic. In *ACM SIGCOMM*, pages 75–86, New Delhi, India, August 2010.
15. G. Maier, F. Schneider, and A. Feldmann. A First Look at Mobile Hand-held Device Traffic. In *PAM*, pages 161–170, Zurich, Switzerland, April 2010.
16. S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro. How Useful are Your Comments?: Analyzing and Predicting YouTube Comments and Comment Ratings. In *WWW*, pages 891–900, Raleigh, NC, April 2010.
17. G. Szabo and B. Huberman. Predicting the popularity of online content. *CACM*, 53(8):80–88, August 2010.
18. J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *ACM WSDM*, pages 177–186, Hong Kong, China, February 2011.
19. 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.