Characterizing Videos and Users in YouTube: A Survey

Shaiful Alam Chowdhury and Dwight Makaroff
Department of Computer Science
University of Saskatchewan
Saskatoon, SK, CANADA
(sbc882,makaroff)@cs.usask.ca

***Abstract*—Web 2.0 has reshaped the way people interact with Web sites. People are now able to view content created by other users as well as publish their own content on Web 2.0 sites, instead of downloading content created by a single author. Understanding the characteristics of the Web 2.0 sites has become a subject of immense interest to the Internet service providers, content makers and on-line advertisers. This understanding is also important for the sustainable development of the content distribution systems. As an approach to comprehend the characteristics of Web 2.0, significant amount of research has been done in investigating the characteristics of YouTube, the most popular web 2.0 site. In this paper, the characteristics of YouTube, based on earlier works, are studied from both video and user perspectives along with some open research issues. This kind of study is instrumental to understand the driving aspects of YouTube and other similar user generated content (UGC) sites.**

***Keywords*-YouTube videos; YouTube users; Web 2.0; User generated content**

I. Introduction

YouTube, the most popular user generated content (UGC) site, was the 4^{th} most accessed site in 2007 Internet, with more than 40 million videos and 20 million users. The estimated cost for bandwidth was $2 million per month, as approximately 10% of all Internet traffic were coming from YouTube [1]. The numbers of videos and users in YouTube were increasing by following a power law curve [1]; subsequent measurements by Alexa[1] claimed that YouTube is the 3^{rd} most accessed site in 2012 Internet—after Google and Facebook. Recent studies ([2], [3], [4]) suggest that YouTube now accounts for 20-35% of the entire Internet traffic with approximately 448 million videos and 47.3 million uploaders [5]. As stated by Cheng *et al.* [1], a survey that was conducted in 2007 shows that YouTube video delivery speed was slower than most of the surveyed sites that are similar to YouTube. This disappointing performance of YouTube video distribution technique is still a subject of concern [6], which also poses challenges to other increasingly popular UGC sites like Dailymotion, Metacafe etc.

The rapidly increasing number of videos and users in YouTube led researchers to characterize the patterns of traffic and user interactions in YouTube so that appropriate content delivery techniques can be designed. In this paper, video characteristics of YouTube along with their growth patterns are studied. Users behaviours, in particular those of YouTube uploaders, are also investigated, which depicts the different styles of interaction with YouTube by different kinds of users. In addition, the way to detect content duplication and predict comment ratings are explained in this paper. These kinds of analyses can be helpful for designing similar new systems, capacity planning, network management and appointing appropriate advertisement policies. Potential drawbacks of the earlier works along with future research direction in this area are also presented in this paper.

The rest of this paper is organized as follows. Section II describes the work done on analyzing YouTube videos in terms of video characteristics, traffic measurements, and video distribution challenges. Section III analyzes the research that differentiates behavior from the individual user perspective. Section IV discusses the primary characteristics that distinguish YouTube from other video distribution sites. Open research issues in this area are addressed in Section V. Finally, Section VI concludes this paper.

II. Characterizing YouTube Videos and Request Traffic

In this section, YouTube is characterized from different aspects including videos and users of this site. This Section is organized as follows. Subsection II-A describes research investigating features of YouTube videos. Growth pattern of YouTube videos is described in Subsection II-B. Subsection II-C shows the way to detect duplicate content in YouTube. Concerns over playback quality of YouTube videos along with a solution to improve the quality are presented in Subsection II-D. Characteristics of YouTube traffic under campus networks and regional popularity of YouTube are presented in Subsection II-E and II-F respectively. Finally, section II-G discusses an approach to evaluate YouTube traffic and develop a workload generator, based on that evaluation.

A. YouTube Video Characteristics

A detailed investigation on the characteristics of YouTube videos is done by Cheng *et al.* [1]. Information of approximately 2.6 million videos was collected by following the related video links of some popular videos whereas

[1] http:/www.alexa.com

the estimated number of uploaded videos in YouTube was around 42.5 million by that time. The main characteristics considered were video length, category, active life span and relationship to other videos.

Results suggest that the uploading rate of YouTube videos could be fitted with a power law curve, and out of 15 categories Music and Entertainment videos were found to be uploaded most frequently. In case of video lengths, almost 98% of the videos were found within 600 seconds. This may be due to the limit imposed by YouTube on video length in that time, which is why the longer videos were found in several episodes. Perhaps not surprisingly, no correlation is found between video length and video popularity. In spite of having a heavy tail portion in the popularity distribution curve of YouTube videos, distribution only for the popular videos in YouTube follows Zipf distribution. This implies that popular videos of YouTube are as popular as Zipf's law predicts. With respect to *active life spans* of videos, investigation suggests that most of the videos have been watched frequently only in a short span of time. These characteristics can be fitted well by a Pareto distribution, which indicates the low probability of watching a video after its active life span. Finally, the YouTube video network is found to be similar to the small-world network as the graph of related videos in YouTube exhibits similar characteristic path length and clustering coefficient to small-world networks.

Considering the small-world properties of YouTube network, this paper concludes that the peer-to-peer technique, with proper modifications, can be employed to save YouTube as well as other similar sites. Even in case of proxy caching, approximately 80% hit-ratio can be achieved with only 8GByte of disk space, using prefix caching of related videos. That is, if a group of videos are significantly related to each other, then a user is likely to select another video from the same group after finishing the current one. However, the first set of data collection was based on some standard feeds provided by YouTube API that only return popular videos. As a consequence, collection of information of videos by using related links of those videos has a high probability that the dataset contains information for popular YouTube videos only. Although this kind of dataset can be used to evaluate the caching policies, it is unlikely to have appropriate understanding of YouTube-like sites by ignoring the characteristics of dominating number of unpopular videos. For example, while fitting the viewing pattern of YouTube videos with Weibull and Gamma distributions, it is likely that the shape and scale parameters for both of the distributions might be changed if an unbiased data set is used. Moreover, for unbiased data set, the tail section of the distribution would be found longer than that is found in this paper.

B. Popularity Growth Pattern of YouTube Videos

In order to observe the time-varying popularity of YouTube videos that is crucial for efficient object caching, Borghol *et al.* [7] collected information of 29,791 YouTube videos by using the Most Recent standard feed provided by the YouTube API. Their collection procedure was good enough to have an unbiased dataset; the Most Recent standard feed returns video information randomly that are uploaded very recently, regardless of their number of views. Their investigation shows that most of the videos achieve their peak popularity within less than six weeks from their uploading time. Moreover, as an approach to investigate whether or not the current popularity of a video is an indicator of future popularity, Pearson's correlation coefficient was calculated between added views at consecutive snapshots. The correlation coefficient between snapshots two and three is found to be very weak (0.09). Interestingly, this coefficient becomes approximately 0.7 between the snapshots eight and nine, which further approach to 1 between snapshots sixteen and seventeen. This observation suggests that current popularity of an older video can reflect its immediate future popularity, which is not the case for a very young video.

YouTube videos were grouped according to whether they were after, before or at the age at which they experienced their peak popularity. Increasing viewing pattern is observed for most of the videos before the peak, which gradually falls and approaches to a constant pattern after the peak. Unlike most of the earlier works, in this work video information was collected randomly, which helps to characterize YouTube videos with the least possible known bias. However, it would have been better if the prediction of future popularity were conducted only for the popular videos. It is expected that viewing patterns of the unpopular videos follow a different distribution than the popular ones.

Figueiredo *et al.* [8] applied a novel technique, Google charts, to collect the number of views over time for YouTube videos. Then the time varying viewing patterns of popular videos, deleted videos and randomly selected videos were analyzed. Their analyses show that, very interestingly, the videos that were deleted because of the copyright violation, tend to get most of their views much earlier in their life times. Results also suggest that the popular videos usually experience huge number of views on a single peak day or week. For instance, for half of the videos in the popular, deleted and random datasets, it takes at most 65%, 21% and 87%, respectively, of their lifetimes until they experience at least 90% of their total views. For 50% of the total views it takes 26%, 5% and 43% respectively for the previously mentioned three datasets. In addition to popularity over time, they also investigated the impact of different types of referrers, both internal and external, that can positively influence the views of a video. Results show that out of all different referrers, Featured and Social referrers have significant impacts on the number of views of Youtube videos.

However, the dataset that was collected using Google charts API is not appropriate to have a proper understanding of the dynamics of video popularity since the Google charts API shows the views of a particular video at most at 100 different points, regardless of the age of the video. This procedure limits the details of the viewing pattern of a video presentable. Moreover, it is another research issue to identify whether one referrer might influence the number of views from other referrers. For instance, a popular video may experience further popularity growth from Social referrer after being featured by YouTube. Similarly, it may first receive a large number of views from Social referrer; thus leading it to be featured by YouTube. Although this paper shows that the videos that violate copyright laws experience most of the views in the very early of their lifetimes, whether or not these videos are deleted immediately after their peak popularity, is not mentioned.

C. *Content Aliasing in YouTube*

Pedro *et al.* [8] investigated content duplication and overlap in YouTube. In order to detect duplicate scenes among different videos, content-based copy detection tools (CBCR) has been used. Sets of graphs were formed such that the edges in a graph represent highly related videos in YouTube. The components of the fingerprint-based CBCR can be described as three steps: fingerprint generation module, reference content database and search module. In fingerprint generation module, all the videos are transformed into a sequence of points in the fingerprint feature space. Reference content database is a database of known fingerprints that can be developed using supervised trainings. Finally, in the search module step, fingerprints for all incoming video streams are compared with the reference content database. In order to evaluate the effectiveness of the CBCR technique, a pilot experiment was conducted for a known database, which confirms 90% accuracy of CBCR.

As the final step to investigate the content redundancy in YouTube, 703 queries for YouTube keyword-based search were collected by using top 10 gaining weekly queries provided by Google Zeitgeist. After filtering, 579 queries were used to collect 28,216 video's information. Result suggests that almost 16% of the YouTube videos suffer from content duplication along with significant amount of overlapping among videos. Not surprisingly, it was found that popular videos suffer more from content duplication than comparatively unpopular videos. It is claimed that video duplication happens mainly for two reasons. Firstly, many users re-upload popular content so-called "user copied content or UCC", in order to increase their popularity as a uploader. Secondly, many users upload different versions of a video with the subtitle in their own language, which is referred as multilingualism.

However, some cases of duplication—videos with common a descendant for example—were not considered, although common ancestor of different videos was considered during the detection process. Most importantly, impact of content aliasing on the original videos are not presented, which might be very crucial for the on-line marketers. For instance, Cha *et al.* [9] shows that total view counts from different copies of a single video can be more than two orders of magnitude that of the original video. Likewise, the dataset is not rich enough to estimate the actual amount of content duplication in YouTube.

D. *Playback Quality Concerns/Potential Solutions*

Dissatisfying experience of YouTube users in watching videos along with a promising solution are illustrated by Khemmarat *et al.* [6]. At first, an experiment was conducted to evaluate user experience in watching YouTube videos—how often a user experiences pauses during video playback and how long the pauses are. The information of pause frequency was collected automatically by examining video download traces. Twelve volunteers from twelve different environments representing different network access technologies were asked to use the Wireshark network protocol analyzer to capture YouTube traffic. A model was developed to estimate the number of pauses in playback assuming that a fragment of a video should arrive at the client before playing that fragment. From the sample dataset, it was found that 10 out of 12 environments contained playbacks with pauses, and 41 of 117 playbacks contained pauses, which represents approximately 35% of the total playbacks. This observation portrays that YouTube users experience noisy playbacks, possibly more significantly for higher quality videos. This problem can be intolerable as high definition videos become increasingly popular in YouTube.

The authors suggest that prefetching can be applied to solve this problem. Unlike caching where content is only stored locally after being requested by a client, prefetching retrieves a content from the source before it is requested by a client. Two different kinds of prefetching agents (PA) were considered: PF-Client and PF-Proxy. PF-Client is dedicated only for one client is located at the client whereas PF-Proxy is located at proxy server and serves for all the client under the same proxy. All the YouTube requests from a client are directed to the PA. The PA serves the client with the prefix of the video if the video is available in the local server, and starts retrieving the remaining part of the video from the YouTube server. If the prefix is not found locally, the PA retrieves the whole video from YouTube and sends it to the client. Two different referrers were used to select videos for prefetching: YouTube search results list and related video lists, as these two lists were found as the two most frequently used referrers. Although these two referrers return up to 25 videos' titles in the list, it was found quite challenging to estimate the actual number of videos that need to be prefetched for optimal performance. The top N videos were selected for prefetching where the value of N was varied in different parts of the experiment.

Figure 1 shows the hit ratio of different prefetching

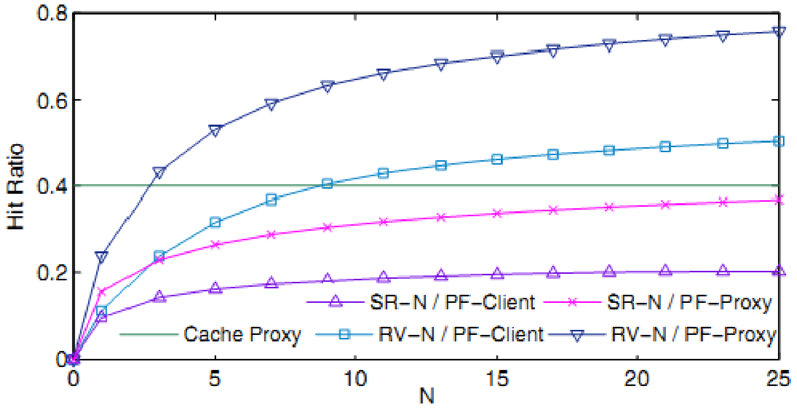


Figure 1. Performance of different prefetching techniques [6]

techniques against different values of N. For example, SR-N/PF-Client represents the hit ratio of the PF-client agent that prefetches top N videos when the search referrer is used. From Figure 1, it can be seen that PF-Proxy—one agent for all the local clients—outperforms all other techniques when videos are selected for prefetching from the related video list provided by YouTube. Moreover, the top 15 videos are enough to store so that 75% hit ratio can be obtained. Although this part of examination was conducted for infinite cache size, similar results were found for reasonable amount of cache size. Interestingly, this paper found that combination of caching and prefetching can increase the hit ratio by 5-20% as compared to the prefetch only mode.

However, the amount of data required to resume from pausing was estimated as the actual amount used by YouTube was unknown. Moreover, it would have been better if the performances of prefetching were examined by using other referrers like Most Viewed, and Top Rated. Besides, the performance of prefetching was not compared to other potential techniques like batching, although batching improves playback quality as well as reduces the requirement of network bandwidth [10].

E. YouTube video Traffic Under Campus Networks

Zink *et al.* [11] examined You Tube traffic between YouTube server and University of Massachusetts. Three different periods were used for this measurement in 2007 and 2008. The result shows that only approximately 25% of all requested videos were requested more than once. Based on this observation, three different content delivery techniques were examined: proxy caching, client-based local caching and P2P-based distribution. Results suggest that local caching can improve the overall system performance. Surprisingly, P2P-based caching shows worse performance than the client-based caching architecture. In this paper, proxy caching is found to exhibit an effective low-cost solution. The results of their overall simulation illustrate that, compared to the other types of content delivery method for YouTube, caching is more effective to decrease network traffic as well as video access time.

As a cache replacement policy in the proxy server, the oldest video clip was replaced with newer requested video clip. The size of the proxy cache was varied between 100 MB and 150 GB, and it was observed that when the cache size changes from 100 MB to 1 GB the performance increases 10%. Finally, maximum performance was found when the cache size was 100 GB.

A similar experiment was conducted by Gill *et al.* [12] by collecting the traffic information of YouTube videos in University of Calgary campus network. They investigated file properties, usage patterns, and transfer behaviours of YouTube videos along with the social networking aspects. Their analysis suggests that appropriate caching decisions not only can improve the end user experience, but also reduce network bandwidth requirement to access YouTube.

F. Regional Popularity of YouTube

Brodersen *et al.* [13] investigated relationship between locality and popularity of YouTube videos. The number of daily views for more than 20 million videos were collected. Including official states and minor territories, this paper considered 250 different regions for the analyses. Surprisingly, results suggest that there are about 40% of YouTube videos that enjoy at least 80% of their views in a single region. This evidence indicates that YouTube videos tend to become popular in a locally confined area, rather than in a globally wide region. *The difference of YouTube popularity among different regions is obvious, which is found to be followed Zipf distribution.* Not surprisingly, different categories were found to exhibit different patterns of global and local popularity. This observation portrays that the topic of a video is very important in order to attract the viewers from all over the world. Likewise, strong correlation is found between the location of a video's uploader and its regional popularity. For instance, because of similar interests, videos uploaded from USA exhibit similar popularity in UK, Mexico, and Canada. On the contrary, videos uploaded in Japan and Brazil enjoy on average 90% of their views in their uploading region.

The impact of social sharing on YouTube videos popularity is investigated as well. Although the amount of social sharing experienced by YouTube videos is different for videos with different number of lifetime views, very surprisingly, the impact of social sharing is found significant for unpopular videos, while for popular videos the social sharing becomes less prominent. This paper also shows that, on average, a video tends to become popular and to peak in its own focus location (where a video has most number of views in its lifetime), and only then this video becomes popular in other regions.

The findings can be instrumental for local caching mechanisms and advertisement policies of YouTube and similar content distribution sites. However, although it is claimed that News, Sports, and Politics videos are expected to exhibit regional popularity, unfortunately the actual names of the categories that were found to exhibit such phenomenon were not mentioned in this paper.

G. YouTube Workload Analysis and Generation

Abhari *et al.* [14] design a workload generator for YouTube, and then evaluate the performance of proxy caching with two different datasets. The first dataset ('popular dataset') is collected by using the standard feed Most-Viewed in a day and Most-viewed in a week provided by the YouTube API. On the other hand for the second dataset ('regular dataset'), Most-Discussed, Most-Viewed, Recently-featured, and Top-Rated standard feeds were used first. Data collection was continued by following the related links of the first two datasets and thus ensuring a significant amount of video information. This paper then characterizes the properties of YouTube videos. Because of the similar crawling approach, distribution of video lengths and correlation between length and popularity are found similar to Cheng *et al.* [1]. Likewise, popularity of YouTube videos was found to be fitted with heavy tail Weibull distribution. The amount of time that a video file remains in the most popular video list is also examined. The short active life span of the popular videos is confirmed by observing the daily Most viewed list provided by the YouTube API.

Based on these observations, two different workload generators were developed: server workload generator and client session generator. The server workload generator simulates the files available on the YouTube server, whereas the client session generator simulates user accessing the server by selecting a video from available videos. The client session generator was designed in a way that videos with the larger value of view counts are more likely to be selected by the client. Poisson distribution was used to generate subsequent requests from a client. The performance of proxy caching was measured according to the request patterns generated by the workload generator. Figure 2(a) and 2(b) depict the performance of proxy caching considering infinite cache size and finite cache size respectively. When the cache size was considered finite, Least Recently Used (LRU) technique was applied for video replacement. Figure 2(a) shows that with a higher percentage of requests, both daily and weekly traces have a higher hit ratio, and better hit ratio is found for longer traces (weekly) than shorter traces (daily). On the other hand, Figure 2(a) shows that hit ratio achieved by proxy caching and LRU policy are in the range of 12% to 90% for different cache sizes.

Similar to the dataset collected by Cheng *et al.* [1], datasets of this paper are also biased to the popular videos and suffer from similar problems. Moreover, while generating subsequent requests from a client by the workload generator, it was considered that a client does not send a new request without watching the earlier requested video completely, which is not a general case for YouTube videos. Likewise, the type of video object was not considered at all in this paper, which might be crucial to understand the actual growth pattern of YouTube videos.

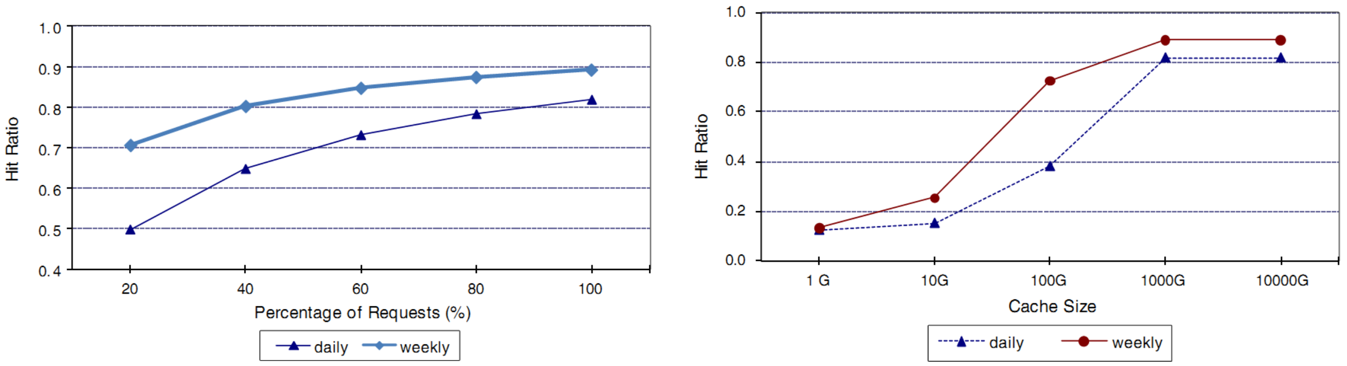
III. Characterizing YouTube User Behaviour

Actions other than viewing videos have an impact on the traffic generated for the YouTube network. This section explores research correlating this user activity into observations about popularity of YouTube videos. The impact of comments and the way to predict comments ratings are illustrated in Subsection III-A. Subsection III-B shows the characteristics of YouTube uploaders whereas Subsection III-C analyzes behaviour of different categories of registered YouTube users.

A. Predicting Comment Rating in YouTube

The predictability of comment ratings was investigated by Siersdorfer *et al.* [15]. More than 6 million comments on 67,000 YouTube videos were collected to analyze the dependency between comment ratings and sentiment expressed in a comment. To calculate both positive and negative sensitivity of a comment, the publicly available SentiWordNet thesaurus was used. In SentiWordNet, a word is represented by three sentivalues called positive, negative, and neutral. The sentivalues are in the range of [0, 1] and sum up to 1 for each triple. For example, a triple (0.875, 0.0, 0.125) represent a good word in SentiWordNet whereas (0.25, 0.375, 0.375) usually represents a bad word. Sentivalue for a comment is calculated by computing the averages of positive, negative and neutral values that are found for each word in that comment.

Observations suggest that the distribution of ratings is asymmetric for positive and negative ratings in YouTube, which suggests that the YouTube users tend to cast more positive votes than negative. Interestingly, 50% of the comments



(a) Infinite cache size

(b) Finite cache size

Figure 2. Performance of proxy caching [14]

are found neutrally evaluated by the users. As expected, this paper shows that negatively rated comments tend to contain more negative sentiment terms than positively rated comments and vice versa. Category dependencies of ratings are also investigated in this paper, and it is found that because of the impartial nature of the Science videos, they present a majority of neutral comments. Very interestingly, Politics videos are found to have significantly more negatively rated comments compared to other categories. On the contrary, Music videos enjoy more positively rated comments than all other categories.

Different categories of YouTube tend to attract different kinds of users and produce more or less discussion as a function of the controversy of the topics. It is also depicted that the rating of a comment can be predicted to some extent. The findings are useful for promoting interesting comments even in the absence of community feedback. In other words, automatically predicted comment ratings can be helpful as a supplementary ranking criterion for search results. Predicting comment ratings can also be useful to predict a video's future popularity as Chatzopoulou *et al.* [16] found strong correlation among YouTube videos' total views, number of comments, number of ratings and number of favourites. However, this work would have been much better if the results were verified by using another tool besides SentiWordNet. 67,000 videos are not enough to draw a conclusion. Many more videos' information could have been collected using the YouTube API.

B. YouTube Uploaders

Ding *et al.* [5] examined the uploaders' behaviours in YouTube extensively. Data analysis shows that the number of videos uploaded by the users follows Zipf like distribution and, very surprisingly, it shows that this uploading rate follows 80-20 rule, which means 80% of the videos are uploaded by only 20% of the uploaders. Not surprisingly, numbers of subscribers for the uploaders also follow Zipf like distribution. While analyzing the geographical locations of the uploaders, this paper shows that approximately 31% of the videos are uploaded only by the USA users. After analyzing the social network in YouTube, results suggest that the social users in YouTube not only upload more videos but also their videos are watched more than the non-social users. This paper also demonstrates that male users usually upload more videos than female users. Finally, comparison between user copied content(UCC) and UGC videos in YouTube is done. It shows that most of the popular uploaders usually uploads more UCC videos, although their UGC videos are more popular, in terms of number of views, than their UGC videos. Some of the findings of this paper are very important. For instance, it shows that the top 20% of the most popular uploaders attract approximately 97% of the total views. Moreover, it also suggests that 20% of the uploaders only upload to a single category, and more than 85% of the uploaders upload more than 50% of their videos only to their three top categories. These findings can be very useful in order to predict the future popularity of videos at their very early age.

This paper has some mentionable drawbacks. Most of the results presented in this paper are based on estimation. For instance, identification of UGC and UCC was done by examining sample of the videos only, and then conclusion was drawn for the whole dataset. It would be a worthwhile research to identify a video's category, UGC or UCC, by a methodological approach, which would be able to give more accurate results. Moreover, the crawling approach is not presented clearly. For example, a seed user was selected first to collect its uploaded videos, and then all the related videos were crawled to capture their uploaders. This process was repeated many times with a new seed. Unfortunately, how the seed was selected is not mentioned in this paper. Likewise, BFS approach to crawl YouTube social network can be biased to capture only the information of high degree users. However, findings of this paper address some issues that need further investigation and thus illustrating some of the open research issues in this area.

C. User Categories in YouTube

In April 2006, YouTube announced the Director program in response to a video length limitation that was imposed to prevent copyright violations. A user could apply for a Director account after proving himself/herself as a legitimate creator of his/her uploaded content, which allowed to upload videos longer than 10 minutes. Then Musicians and Comedians accounts were introduced for publishing performer information and schedule of show dates. In 2008, Guru and Reporter accounts were added. Guru is for those who likes to post videos that teach skills and how to do something whereas Reporter was for the people who likes to share news and events occurring around them. Finally, Non-profit and Politician accounts were introduced.

Biel *et al.* [17] analyzes these user categories of YouTube along with their uploading rate, viewing rate and social aspects. In YouTube, a user is labeled as standard user when he/she first registers for this site. Then a user can change his label as Director, Comedian, Musician, Guru, or Reporter after applying for any of these profiles. Statistics suggest that most of the users (almost 90%) in YouTube do not belong to any of these special categories and continue their watching or sharing as standard users. Results suggest that in spite of their lower numbers in YouTube, the special users contribute more than the standard users, considering uploading, watching and subscribing. This indicates that only the active users are interested about these categories, and many of the users are still not aware of these categories because of the poor advertisement of YouTube. Similar to previous studies, this paper also shows that, in YouTube, male users dominate female users in terms of uploading videos, watching videos. On the contrary, it is found that female users are more social in YouTube than male users as they have more subscribers, subscriptions, and they also favourite more videos than male users. Interestingly, this paper shows that although Politicians and Reporters uploads many videos, their number of watched videos is very low, which indicates that these people are more interested in releasing their work or spreading their messages, rather than exploring people's interests. This paper is a good example to show how YouTube can be used to reveal the attitudes and behaviours of different kinds of people.

IV. Comparison Between YouTube and other Non-UGC sites

Cha *et al.* [9] shows that characteristics of YouTube are significantly different than non-UGC sites. In case of content uploading rates, as of June 9^{th}, 2008, the largest on-line movie data-base IMDb carried only 1,039,447 movies and TV episodes, whereas approximately 65,000 videos were being uploaded daily in YouTube. This statistics implies that to produce the same number of videos as listed in IMDb, it takes only 15 days for YouTube. While comparing the video publishers between YouTube and non-UGC sites, based on Lovefilm, this paper shows that in YouTube there are some publishers who post more than 1000 new videos over a few years whereas it usually takes more than 50 years for a single producer to produce 100 movies in the film industry.

Not surprisingly, YouTube videos are found shorter than non-UGC by two orders of magnitude, although the length of YouTube videos varies according to the category. In order to compare the viewing rates, information was collected from Netflix and Yahoo! Movies whereas views of Science & Technology videos were collected from YouTube. For Netflix movies, customer ratings were used to estimate the number of views since information about views are not provided by Netflix. Results suggest that lot of YouTube videos have no views, while all the movies in Netflix and Yahoo! Movies have been watched at least once. However, the scale of consumers per video is very different for YouTube and non-UGC. The views distribution of YouTube spans more than 6 orders of magnitude, while the number of ratings per movie in Netflix and Yahoo! Movies span about 4 orders of magnitude. This observation illustrates the natural diversity of in YouTube uploader and consumer population.

Unlike the non-UGC sites, it is found that most of the popular Science & Technology videos in YouTube have incoming links from external sites. Surprisingly, in spite of that enormous number of incoming links, the authors observed that only 3% of the total views comes from these external sites. This paper also suggests that video popularity in YouTube follows a power-law distribution with an exponential cutoff. Although findings of this paper illustrate the differences between YouTube and non-UGC sites very clearly, it is not clear why the authors considered Science and Technology videos while comparing with Netflix and Yahoo! Movies. YouTube itself has a category named Film & Animation, which should be the perfect selection for this part of the analysis.

V. Open research issues

Scalability is considered as one of the most important issues in YouTube like sites because of the freedom in video uploading. Peer-to-peer techniques, in spite of their promising solution to the scalability problem, can not be deployed without appropriate modifications, especially with their incentive mechanisms. Imposing restrictions in download speed for example, which is employed by the incentive mechanisms, can contribute oppositely to the current popularity of YouTube and other UGC sites. This area of video distribution needs further extensive research. Along with P2P techniques, other well known video distribution approaches like batching and patching can be investigated. For example, it would be worthwhile to investigate the performance of batching for YouTube live streaming; batching has been proved as potential candidate to improve the playback quality for this kinds of video distributions. For caching mechanism, multilayer caching policies can improve the buffering delay. Given the regional popularity of videos [13], different local caching approaches can be developed. In case of local cache miss, the request will be forwarded to the central server such that videos in central server are cached in a way that reflects the global popularity of videos.

Content aliasing in YouTube is another issue of concern for the on-line marketers, which is responsible to distort the popularity of the original videos. Although the way to detect duplicate content is suggested by Pedro *et al.* [8], no work has been done that investigate the way to eliminate duplicate content from YouTube at their very early ages, which can help to minimize the impact of duplicate videos on the original video's popularity. Considering the types of video objects while analyzing the time-varying popularity can contribute to properly understand the growth pattern of videos. It is likely that some of the categories, News and Sports for examples, experience most of the views at their very early ages. This kind of analysis might not only improve the caching mechanism but also can be beneficial for appointing appropriate advertisement policies.

VI. Conclusion

In this paper, we investigated several aspects of YouTube including video characteristics, content redundancy, playback quality, and users behaviours. Although some of the results are found very similar among different studies, all of them are presented in this paper in order to identify the results that do not need further verification. On the contrary, some of the results are found contradictory among different studies that need further examination. For example, Cheng *et al.* [1] suggest that P2P technique with appropriate modification is the best candidate to solve the scalability issue of YouTube, where exactly opposite result is found by Zink *et al.* [11]. This study can be helpful for future research on YouTube, as potential drawbacks of the earlier works are also analyzed.

References

[1] X. Cheng, C. Dale, and J. Liu, "Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study," Cornell University, arXiv e-prints, Tech. Rep., July 2007.

[2] A. Gember, A. Anand, and A. Akella, "A Comparative Study of Handheld and Non-handheld Traffic in Campus Wi-Fi Networks," in *PAM 2011*, Atlanta, GA, March 2011, pp. 173–183.

[3] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian, "Internet Inter-Domain Traffic," in *ACM SIGCOMM 2010*, New Delhi, India, August 2010.

[4] G. Maier, F. Schneider, and A. Feldmann, "A First Look at Mobile Hand-held Device Traffic," in *PAM 2010*, Zurich, Switzerland, April 2010, pp. 161–170.

[5] Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, and G. A, "Broadcast Yourself: Understanding YouTube Uploaders," in *IMC 2011*, Berlin, Germany, November 2011, pp. 361–370.

[6] S. Khemmarat, R. Zhou, L. Gao, and M. Zink, "Watching User Generated Videos with Prefetching," in *MMSYS 2011*, San Jose, CA, February 2011, pp. 187–198.

[7] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing and Modelling Popularity of User-Generated Videos," *Performance Evaluation*, vol. 68, pp. 1037–1055, November 2011.

[8] J. Pedro, S. Siersdorfer, and M. Sanderson, "Content Redundancy in YouTube and its Application to Video Tagging," *ACM Transactions on Information Systems*, vol. 29, no. 3, pp. 13:1–13:31, July 2011.

[9] 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 Transactions on Networking*, vol. 17, no. 5, pp. 1357 –1370, October 2009.

[10] H. Shachnai and P. S. Yu, "Exploring wait tolerance in effective batching for video-on-demand scheduling," *Multimedia Systems*, vol. 6, no. 6, pp. 382–394, November 1998.

[11] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of YouTube Network Traffic at a Campus Network - Measurements, Models, and Implications," *Computer Networks*, vol. 53, no. 4, pp. 501–514, March 2009.

[12] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube Traffic Characterization: A View From the Edge," in *IMC 2007*, San Diego, California, October 2007, pp. 15–28.

[13] A. Brodersen, S. Scellato, and M. Wattenhofer, "Youtube Around the World: Geographic Popularity of Videos," in *WWW 2012*, Lyon, France, April 2012.

[14] A. Abhari and M. Soraya, "Workload Generation for YouTube," *Multimedia Tools and Applications*, vol. 46, no. 1, pp. 91–118, January 2010.

[15] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. S. Pedro, "How Useful are Your Comments?: Analyzing and Predicting Youtube Comments and Comment Ratings," in *WWW 2010*, Raleigh, NC, April 2010, pp. 891–900.

[16] G. Chatzopoulou, C. Sheng, and M. Faloutsos, "A First Step Towards Understanding Popularity in YouTube," in *2010 INFOCOM Workshops*, San Diego, CA, March 2010, pp. 1–6.

[17] J. Biel and D. Gatica-Perez, "Wearing a YouTube Hat: Directors, Comedians, Gurus, and User Aggregated Behavior," in *ACM Multimedia 2009*, Beijing, China, October 2009, pp. 833–836.