

Music Video Redundancy and Half-Life in YouTube

Matthias Prellwitz¹ and Michael L. Nelson²

¹ HTW Berlin, 10318 Berlin, Germany

² Old Dominion University, Norfolk VA 23508 USA

Abstract. YouTube is the largest, most popular video digital library in existence, and is quite possibly the most popular digital library regardless of format type. Furthermore, music videos are one of the primary applications of YouTube. Based on our experiences of linking to music videos in YouTube, we observed that while any single URI had a short half-life, music videos were always available at another URI. For this study we collected 1291 music videos and found that very few had zero or one copies in YouTube at any given time, and some had several thousand copies at any given time. Furthermore, individual URIs had a half-life of anywhere from 9 to 18 months, depending on the publication date and remaining commercial potential.

1 Introduction

YouTube is the leading video hosting service on the Web, with an *Alexa Traffic Rank* of three in November 2010 [1]. Similar services of have much lower ranks, e.g., dailymotion.com (Alexa Traffic Rank: 105), vimeo.com (166), myvideo.de (79). Due to its high popularity, YouTube is also a pioneer in struggling with copyright infringement as users upload (music) videos where they do not have the appropriate publishing rights, and copyright owners – mostly music record companies – identify these violations and petition YouTube to remove the offending videos. This ongoing publishing and removal of user-contributed music videos means that any particular music video likely has several functioning versions on YouTube at any given time, any specific URI is subject to be removed for a variety of reasons. For example, a copy of the music video “The Rolling Stones - Satisfaction” at <http://www.youtube.com/watch?v=214szPQBUYc> was removed on April 9, 2010 and the error message “This video is no longer available due to a copyright claim by ABKCO” is shown instead. Using YouTube’s search functionality and querying the original video title in quotation marks returns 304 results (as of 12/10/2010) that could be understood as alternative available copies of the song.

We selected 1291 YouTube music videos from three different sources (U.S. Top 40, music blogs, and Rolling Stone’s “500 Greatest Songs”) and track the number of available copies of each music video as well as the half-life of individual URIs for a 10 week period. We quantify what we had observed anecdotally:

individual copies of music videos are regularly uploaded and removed for a variety of reasons, but most music videos have multiple copies extant on YouTube at any given time.

Given its success, it is no surprise that YouTube has been studied and discussed innumerable times. Cheng et al. [4] characterizes the collection of videos at large, and also shows that music is the primary category (at 22.9%). Sharing on YouTube has been studied (e.g., [3]), as well as how people discover videos on YouTube [5]. YouTube’s impact on popular culture and politics has been studied [8] [2], as well as ways for extracting the content and context for preservation purposes [9] [7]. But to the best of our knowledge, the half-life and redundancy of music videos in YouTube has not been studied.

2 YouTube HTTP Mechanics

A YouTube video URI is typically `http://www.youtube.com/watch?v=VIDEOID` where VIDEOID is a eleven character alpha-numeric identifier of the video. Dereferencing a YouTube URI returns a 200 OK HTTP response code and the necessary HTML to play the video. Dereferencing a video’s Atom feed at `http://gdata.youtube.com/feeds/api/videos/VIDEOID` returns a 200 OK status code and the XML feed entry with metadata: video title, associated user, duration, tags, and optionally allowed or denied countries for seeing the video.

Once a video becomes unavailable, a 303 See Other HTTP response code is returned³ along with a Location: response header that gives a URI for an HTML page explaining why the video is unavailable: removed or blocked; classified as controversial; contains mature content; authentication required (i.e., private); captcha redirect⁴. We consider all but the last reason to indicate an unavailable video. Dereferencing these video’s Atom feed URI results in an 403 Forbidden HTTP response code and the metadata describing the video is now unavailable.

3 A Dataset of YouTube Music Videos

Since we were constructing a dataset of music videos that would be drawn from multiple sources (U.S. Top 40, music blogs, and Rolling Stone’s “500 Greatest Songs”), we needed to be able to reliably gather metadata such as publication year and genre, for the music videos. For this purpose, we used Discogs⁵, a free service providing detailed information about music releases with releases year, genre, track list, format, etc.

The first up to 20 search results identifying a musical “release” on Discogs were recorded and each HTML page was parsed out for publication year and

³ After this data reported in this experiment was taken, YouTube switched to conventional 404 Not Found HTTP response codes for unavailable videos.

⁴ See: <http://en.wikipedia.org/wiki/CAPTCHA>

⁵ <http://www.discogs.com/>

genre. The release having the lowest publication date was chosen, and its publication year and genre(s) selected for the monitor set item.

The Top 40 songs of the US Singles Charts of September 25, 2010 [11] were chosen. Due to the relative homogeneity of the publication year and genre distribution of the 49 set items (there are more than 40 items to account for variations in artist and song metadata), the figure is omitted. Most of the songs were re-released in the current and previous years.

Three blogs from Blogger.com were selected that provide music reviews: `f-measure.blogspot.com` `youtube-music-videos.blogspot.com` `silaswillrock.blogspot.com`. The YouTube URIs were extracted from their Atom feeds, and the artist and title information was extracted from the YouTube HTML pages. For URIs not having this metadata, its video title was queried in quotation marks against Yahoo! search engine with the parameter “site:Last.fm”, and the title and artist can be extracted from the structured URI returned by Last.fm In total 742 items were created. As expected, figure 1 shows a greater range of genres and publication years than the Top 40 dataset.

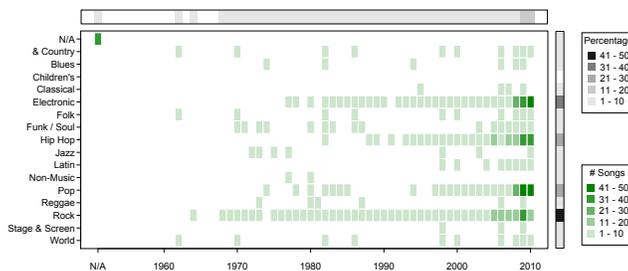


Fig. 1. Set distribution: Genre/Publication Year, Dataset: Music Blogs at `blogspot.com`, against Discogs. Number of items: 742

Rolling Stone Magazine published a list of “The 500 Greatest Songs of All Time” [10], chosen by musicians, critics and industry figures. The song titles were mapped to YouTube URIs in the same manner as the Top 40 list. As expected from a list of this nature, Figure 2 shows a shift to older (i.e., “classic”) songs as well as the genre roots of popular music (e.g., country, blues, funk/soul).

4 Experiment Methodology

We searched YouTube’s GData API [6] with the query “ARTIST TITLE”. The search is not a fielded search; it searches the entire page for the terms. The GData API allows retrieving of only the first 1,000 items, even if the total result size is larger (as indicated in a `OpenSearch`⁶ element). We queried with the default chunk size of 25, up to the total result size or 1000 items.

⁶ <http://www.opensearch.org/>

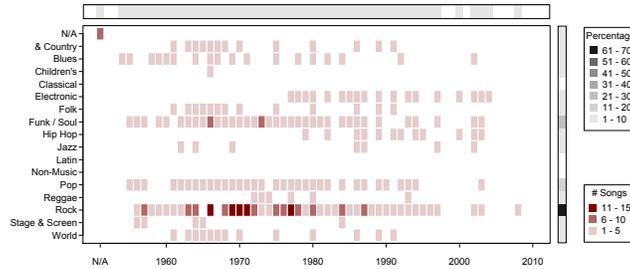


Fig. 2. Set distribution: Genre/Publication Year, Dataset: The 500 Greatest Songs of All Time by Rolling Stone, against Discogs. Number of items: 500

After retrieving the whole list, the total result size value was stored according to the item with the current date. Afterwards, each feed item containing information about a video id was taken and processed: if a ‘uri’ element – identified by the video watch id – did not exist so far, a new record was inserted into the database with the static video properties: uploader (user), duration and publication date. If a ‘uri’ was not updated on the same day (as a ‘uri’ can be result of several monitor set items), its variable data – video title, rating, statistic, and comment counts, allowed/denied countries, and update date – were checked against a last dated database record in the corresponding table, if any, and a new tuple with the actual date was inserted according to the ‘uri’. On change, the ‘uri’ is updated to the current date to prevent redundant processing if it appears in other search results.

For each feed item and transformed ‘uri’ object, its rank position is stored in relation to the search term, i.e. the ‘item’ object, with the current date. That allows keeping track of each URI over time with its (dis)appearance within the first up to 1,000 crawlable results, and its rank change in the feed.

A video URI can also fail to appear in any of the feed results of all items on a day. The reason might be that the rank of a video URI was lowered and its current position is greater than 1,000 that is not possible to observe due to the GData limitation. Another reason can be that a video is no longer available. To check the state of these videos, a cron job, starting after the previous one terminated, takes all ‘uri’ objects not updated on the current date and dereferences each URI. Receiving a 200 OK HTTP status, the ‘uri’ record is just updated with the current date in its ‘last crawled’ property. With a 303 See Other redirect, the ‘uri’ object is set to inactive with the current date, and the reason (parsed from the HTML page) is stored.

5 Results

The Top 40 dataset was initiated on October 1, 2010, Top 500 dataset on November 7, 2010, and the music blogs dataset on November 13, 2010. For the total

of 1,291 monitor set items, 902,869 YouTube video URIs were discovered by December 12, 2010.

Figure 3 shows the Top 40 dataset with its items in descending median result size. The lines to each item indicate the item's maximum and minimum total result size returned over the observation period. For example, the first item with the highest median of results (83,298) has a variation over time from a minimum of 43,945 to a maximum of 123,239 total results given. The last item with the lowest median of 35 results developed a minimum of 26 and a maximum of 66 over time. There was never a time where no videos were available for any song in this collection.

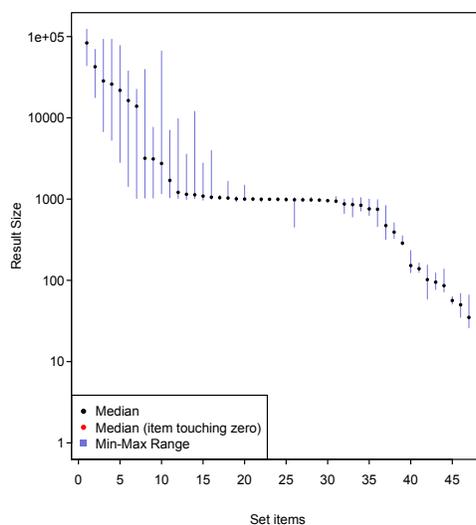


Fig. 3. Total Result Size Dataset: Top 40 US Single Charts of 9/25/10

A similar high variation of total result sizes of 1,000 or more also applies to the music blog dataset (Figure 4). For example, the first item with a median size of 255,581 varies over the given time up to 31,817 between maximum and minimum total result size. The range of items touching zero with their median or minimum result size are shown in detail in the sub plot: from 50 items that retrieved at least once zero results, 44 items never had a result returned. This could be because the video title does not accurately describe the song.

Finally, the Rolling Stone dataset also starts with a high total result size of median 145,076 for the first item and nearly 80% having a median of at least 100 copies. (Figure 5). Only 1%, five items retrieved zero results at least once, and four never received a result, which might be due to the accurate complete artist and song title information, as well as the popularity of songs on the list.

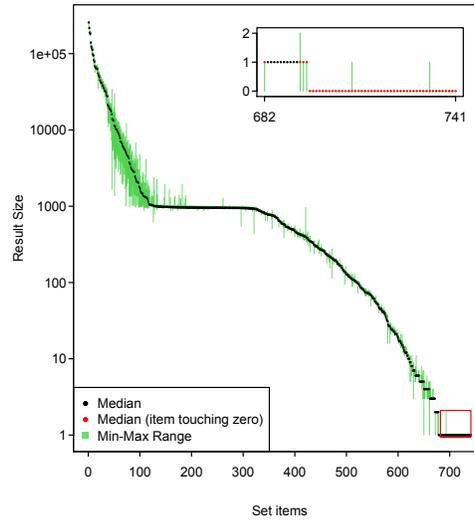


Fig. 4. Total Result Size Dataset: Music Blogs at blogspot.com

We kept a list of all unique URIs discovered from the daily result sets described above so we could trace how long individual URIs persisted. Rooted from that set, the daily availability of those URIs was measured over time. Normalized and aggregated monitor set-wise, figure 6 shows the removal rate over the observation period. Due to the different start dates of the three monitor sets, different durations are present. A continuous unavailability rate can be concluded for each set, visualized by its median values. Furthermore, the plot shows a higher rate of removal for the Top 40 set, presumably because of their current economic potential they are more actively policed by their copyright holders.

Applying linear regression to each monitor set and predicting its progress is an interesting aspect of the half-life of each collection. Figure 7 shows the monitor set-wise aggregated regression.

Nearly half of the video removals (48.8%) were a result of third-party claims by copy right holders. For 23.3% of the removed videos, YouTube removed a video or discontinued the user account, e.g. due to violations against one of its policies or its terms of service. For only 13.2% did users voluntarily remove the video or close their account. The remaining group summarizes observed crawling errors or status changes of video, e.g. the user set a video to private.

6 Summary

We collected 1291 music videos from three collections: a Top 40 U.S. Singles chart (49 videos), a series of blogs that link to music videos (742 videos), and the Rolling Stone list of the “Top 500 Greatest Songs of All Time” (500 videos). We have shown that for music videos in YouTube, one can expect the URI for

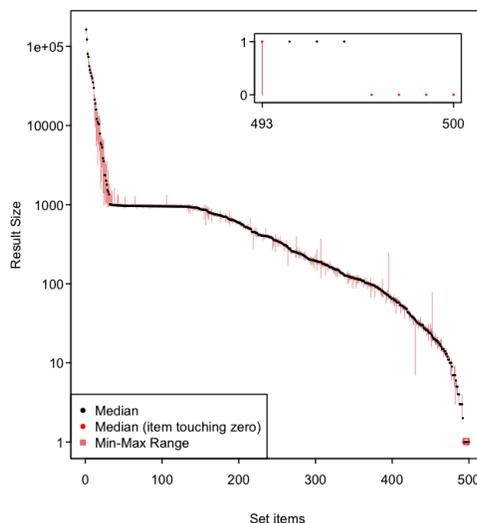


Fig. 5. Total Result Size Dataset: The Top 500 Songs of All Time

any given video to be short lived, with half-lives of 9 (Top 40) to 18 months (Greatest 500) calculated from our datasets. This suggests that the more recent and popular the song (and thus, the more economic potential it represents), the more likely there will be thousands of copies at any given time, as well as copyright holders aggressively requesting their takedown. YouTube provides their Content ID⁷ software suite as a method to help copyright owners to identify when their intellectual property is being used (and optionally, to monetize its use). Despite its use, although individual URIs come and go quite frequently, the music video persists, in aggregate, quite well in YouTube.

References

1. Alexa. Youtube.com Site Info, 2010. <http://www.alexa.com/siteinfo/youtube.com>, last checked: 12/11/2010.
2. R.G. Capra, C.A. Lee, G. Marchionini, T. Russell, C. Shah, and F. Stutzman. Selection and context scoping for digital video collections: an investigation of youtube and blogs. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 211–220, 2008.
3. X. Cheng, C. Dale, and J. Liu. Understanding the characteristics of internet short video sharing: YouTube as a case study. Technical Report Arxiv preprint arXiv:0707.3670, 2007.
4. X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pages 229–238, 2008.

⁷ <http://www.youtube.com/t/contentid>

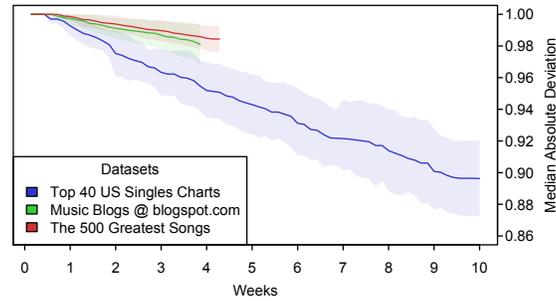


Fig. 6. Unavailable URIs

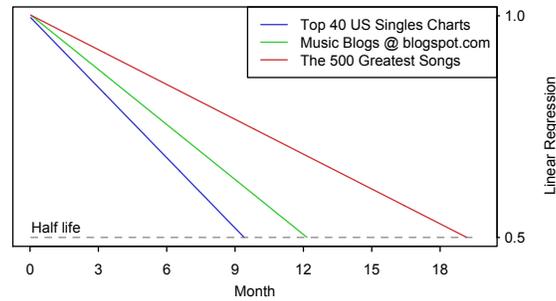


Fig. 7. Predicted Half-Life of Collection

5. S.J. Cunningham and D.M. Nichols. How people find videos. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 201–210, 2008.
6. Google. Developer’s Guide: Data API Protocol API Query Parameters - YouTube APIs and Tools - Google Code, 2010. https://code.google.com/apis/youtube/2.0/developers_guide_protocol_api_query_parameters.html#qsp, last checked: 12/12/2010.
7. G. Marchionini, C. Shah, C.A. Lee, and R. Capra. Query parameters for harvesting digital video and associated contextual information. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 77–86. ACM, 2009.
8. B. Sayre, L. Bode, D. Shah, D. Wilcox, and C. Shah. Agenda Setting in a Digital Age: Tracking Attention to California Proposition 8 in Social Media, Online News and Conventional News. *Policy & Internet*, 2(2):2, 2010.
9. C. Shah. Tubekit: a query-based youtube crawling toolkit. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 433–433, 2008.
10. The Rolling Stones Magazine. The RS 500 Greatest Songs of All Time : Rolling Stone, 2004. <http://web.archive.org/web/20080622145429/www.rollingstone.com/news/coverstory/500songs>, last checked: 11/13/2010.
11. top40-charts.com. USA Singles Top 40 @ Top40-Charts.com, 2010. <http://top40-charts.com/chart.php?cid=27&date=2010-09-25>, last checked: 12/13/2010.