Youtube Trending Analysis Visualisation

Michael Giovannoni and Omar Iltaf

Abstract—This paper seeks to present a design for visualization of YouTube data that will allow users to see the impact of the time of day of posting on their video's success

Index Terms—YouTube, video impact, viewcount

1 Introduction

Our project involves determining the optimal factors a YouTube video must possess in order to be successful. To quantify success, we will use data from YouTube's trending video list and analyze how these videos have become successful. The Youtube trending list is a front page selection of videos that Youtube has generated using their proprietary algorithm. Youtube has not explained exactly what factors can get a video on the trending list but it is likely some combination of likes, dislikes, comments, views, and category. Videos that end up on the Youtube trending list receive a high level of additional attention and views. Such factors we will look at include: the time of day the video was published, like to dislike ratio, and the most effective video category. We will be looking at YouTube data from five different countries: Canada, Germany, France, the UK and the USA). With the growing influence of the YouTube platform, it is increasingly difficult for videos to garner more attention. This problem is best addressed with a visualization as it makes use of a large amount of data, pulling out the necessary details to display to the user. In this way, anybody would be able to make sense of what the data is communicating with ease. Some potential users of this visualization could simply be any one person with a YouTube account looking to increase the number of views they receive on a video they upload. Users could also include digital marketers and companies trying to increase the success of their promotional campaigns. The general approach we intend to take with this visualization is to present our selected data in a form easily understandable by the user: bar charts, donut charts, and a word cloud. We also intend to help the user to navigate through the data by allowing them to select different factors to display within the visualization.

1.1 Visualization Tasks

With the increased use of Youtube as a media platform, more and more content creators are looking to make their mark on the online video landscape. Youtube is not only a platform for creative expression but a marketplace where people make their careers. The goal of our visualization is to provide Youtube content creators with the information they need to make informed decisions that will allow them to maximize the impact of their video. We want to visually illustrate the impact that time of day has on the success of a video and what time of day a content creator should post their content in order to maximize their exposure. To do this we will provide several graphs that show the impact of a video based on the time of day it was posted. Impact can be measured in several different ways. The first is the view count of the video. After view count, statistics such as video count can be used to measure impact. By illustrating these statistics Youtube creators will be able to see the impact their choice of publish time has on video performance. To give creators as much information as possible we will enable them to choose the country or region which they can use to target their specific audience.

In addition to publish time, we want to give creators the opportunity to explore what categories make up the Youtube trending video list as well as what video tags are the most popular. A simple bar chart can measure category popularity while a word cloud made up of the most popular video tags will illustrate what tags are common.

The third task we wish to illustrate is the makeup of how users interact with videos on the trending list. To do this we will compare the average number of likes a video receives with averages dislikes and comments.

- *Michael Giovannoni is a computer science student at Oregon State University. E-mail: giovanmi@oregonstate.edu.*
- *Omar Iltaf is is a computer science student at Sheffield University. E-mail: ltafo@oregonstate.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

2 Related Work

Many papers have been written about the problem of visualizing data from social networks. In Slingsby's *Interactive Tag Maps and Tag Clouds for the Multiscale Exploration of Large Spatio-temporal Datasets* [1], social media tag maps are discussed. This approach is another method for visualizing social media data. By creating a tag map of popular YouTube tags, the effects of tags on video's popularity can be visualized. These tag maps would show the most impactful video tags on the YouTube trending video list and allow users to target their video tags more precisely. Similarly, in Vigas' *Many Eyes: A Site for Visualization at Internet Scale* [2], approaches for visualization of internet data are discussed. The ManyEyes tool would be another method for visualizing YouTube data.

3 Methods

3.1 Data Source

The data we will use is the Trending Youtube Video Statistics published by kaggle.com. This data consists of several months of statistics on the top 200 most popular Youtube videos as defined by Youtube in their trending video section. Data sets for several different regions are available including the United States, France, Great Britain, and Canada.

The data set consists of a flat record. For each video entry, data such as views, likes, dislikes, comment count, and description are provided. The full structure of the data is illustrated in figure 1.

3.2 Data Manipulation

To create this visualization we made use of several different tools. To analyze and organize our data we wrote scripts in Python using the Numpy and Pandas libraries. Pandas allowed us to quickly parse the large csv file with pandas.read_csv() and then select the columns we needed while Numpy was essential for calculating the data's statistics. The pandas.to_datetime() function allowed us to quickly parse the timestamp column provided in the data set and with Python's datetime library we were able to convert the column into easily manipulate datetime objects. Using the numpy.where() and numpy.mean() functions, we could select the columns we needed and calculate the relevant statistics. For example, to select the indices of all videos published between two different times we used np.where(np.logical_and((timeList $>$ lastHour), (timeList $<$ nextHour))). To generate a word cloud, we used the tags column of the data and generated a dict that contained all words paired

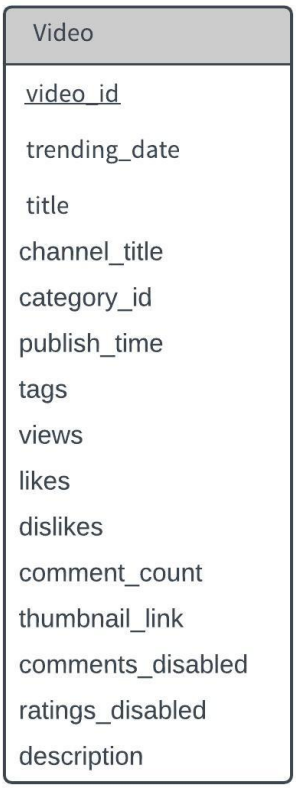

Fig. 1. A visualization of the Trending YouTube Video Statistics data set

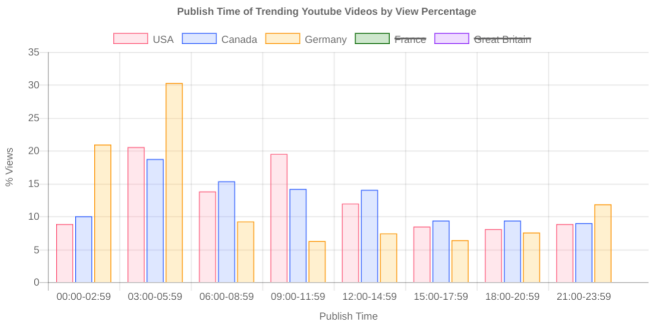

Fig. 2. A visualization of publish time compared to average view count for the Trending YouTube Video Statistics data set

Fig. 3. A word cloud of popular tags

with the count of how many times they appeared in the tags. We took the top 150 of these words for our wordcloud. After analyzing the data plain Python was used to format the output in a form that was readable by our chosen chart creation tool

3.3 Visualization Construction

To create the graphs we used the Chart.JS charting library. This library allowed us to easily produce different types of graphs all of which were visually pleasing and interactive. These were integrated into our visualisation webpage using Javascript. The D3 Cloud library was also used to create a word cloud as Chart.JS does not support this feature. The webpage itself with the visualisation settings panel was created with a mixture of HTML, Javascript, and CSS. The jQuery library was used to handle the the user interactions with the site and to respond accordingly. This includes displaying the appropriate visualisation based on the users? selection. The Materialize framework was utilised so as to provide the site with a well-designed, responsive and modern interface. These were all imported in the HTML head section, allowing them to be used throughout the code.

Bar charts were used to illustrate publish time and categories, a donut chart was used to show the makeup of the video's likes, dislikes, and comment counts, and a word cloud was used to visualize video tags.

4 Results

The results show a higher view count for videos published between midnight and noon, illustrated in figure 2. The highest view counts are on videos published between 3:00 AM and 6:00 AM with the lowest being between 6:00 PM and 9:00 PM. Clearly videos published early in the day are more successful. As we can see in figure 4, the opposite pattern is observed when looking at the most popular time of day for publishing videos. A strong majority of videos are published within the afternoon and evening hours.

One interesting feature of the data is that the most successful publish time is common to all countries. We expected for different countries to have different top publish times due to differing timezones so this was unexpected. Several possibilities can explain this discrepancy. It is possible that the United States is so impactful in the mediasphere that they choose the most successful publish time even on videos from other countries. The most successful slot, 3:00 - 6:00 AM PST, is when most of the east coast of the United States is waking up, a common time to access videos and a time accepted by many to be the best time of day to post internet content. We must also wonder if the data we have is labeled accurately. The times are all supposed to be in PST, Youtube's timezone, but looking at the distribution it is possible that the timezones are that of the country where the video was published.

Of the US video categories, music is by far the most popular. Music has the highest views at 5.9 million average views. Of the 16 categories that are featured on the YouTube Trending videos list (not all categories appear), the News and Politics category is by far the least popular with an average view count of 581,000. We see a similar distribution of category views between the different countries with a few exceptions. Science and Technology is far more popular in Great Britain than in the other countries we observed as is News and Politics.

The word cloud presents an interesting view of what tags are popular. Most are unexpected, words such as "the" and "to", but several stand out. "Makeup" is featured quite prominently reflecting the increased popularity of makeup videos on Youtube and we also see words like "iphone", representing the technical review side of the site.

5 Contributions

Michael Giovannoni was responsible for analyzing the data and constructing the necessary data tables for creating the graphs. In addition,

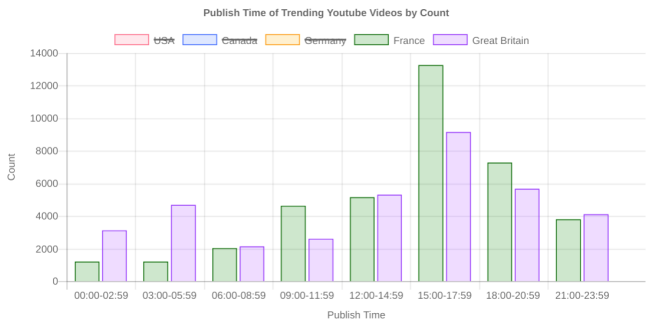

Fig. 4. A visualization of publish time compared to number of videos for the Trending YouTube Video Statistics data set

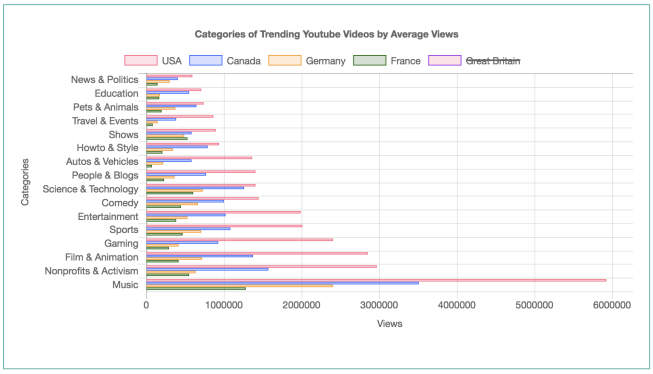

Fig. 5. A visualization of mean views per video category for the Trending YouTube Video Statistics data set

he write most of the report and compiled the final submission.

Omar Iltaf was responsible for the visualization applet. He designed the site and wrote all of the html, CSS and most of the javascript.

6 Conclusion

Our visualization allows YouTube content creators to easily see the impact that time of day has on video success and what categories are most popular. Users are able to select the factors that they are interested in and the data is displayed in simple charts. This visualization enables YouTube channel creators to tailor their video publishing to their audience and increase their exposure.

The most difficult part of this project was dealing with the data. The dataset we were provided was incomplete. In some places values were missing and in others extraneous values were provided, causing the read_csv function to give us errors.

If the project were to be continued, it would be helpful to find another dataset, that of non-trending Youtube videos. If we were to compare the two sets we could better visualize exactly what factors get a video to the top of Youtube. As it is, the project shows the makeup of the trending video list but it does not explain exactly which factors are the most influential in the Youtube trending video algorithm. The scope of the algorithm is outside the realm of information visualization but it may be an interesting project for the future.

This visualization technique may be suitable for other kinds of social network datasets. For example, the same graphs could be used to visualize the most effective times to make a reddit post and the resulting upvotes and downvotes received on those reddit posts. It may also be effective for data from Facebook and a variety of other social networks.

References

[1] A. Slingsby, J. Dykes, J. Wood, and K. Clarke. Interactive tag maps and tag clouds for the multiscale exploration of large spatio-temporal datasets. In *Information Visualization, 2007. IV '07. 11th International Conference*, pp. 497–504, July 2007. doi: 10.1109/IV.2007.71

[2] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, Nov 2007. doi: 10.1109/TVCG.2007.70577