Social Network Mining YouTube Videos

Austin Voecks

***Abstract*—YouTube videos are categorized into a number of classes such as `music`, `entertainment`, `sports`. This paper will explore how these categories are related, how strongly they are related, and the techniques used to gain this information. Data mining, graph analysis, and visualization techniques will be leveraged to accomplish this goal.**

***Keywords*—social networks, data mining, YouTube**

I. Introduction

YouTube is the preeminent public video sharing platform on the internet today. It's vast array of examples of human behavior can provide insight into how we as a society categorize activities, and how those categorizations are related to each other. Each video has a number of attributes we can mine, including variable number of related videos. We can use the related video information to construct a social network of videos that show their relationships.

The related videos are determined by YouTube algorithmically, not by hand, so this approach is not as organic as human labeled data. However, the machine labeling allows access to a much greater data set.

This work seeks to explore the relationships between videos through first order attributes and infer second order relationships such as how similar or dissimilar video categories are.

II. Data Set

Researchers at Simon Fraser University, BC conducted a number of YouTube crawls [1]. The attributes provided in this data set are:

`video_id`	`uploader`	`age`
`category`	`length`	`views`
`rate`	`ratings`	`num_comments`
	`related_ids`	

In particular, the `related_ids` attribute can have a variable number of values. Each video has a unique video id, and the related ids are denoted using these same ids. `video_id` and `related_ids` compose the nodes and edges of the network respectively.

The "data set" referred to here is composed of many smaller data sets spanning the initial years of YouTube's presence on the internet. This implicit time series information allows further analysis into how findings change over time.

The data set has many rich attributes but poses the following complications to analysis:

1) **Mixed Value Types**
   The types of values between attributes range between integers, floats, strings, and lists. This does not pose a major problem to general graph analysis tools but would make most machine learning techniques difficult to apply without extensive data transformations.
2) **Sparsity**
   The data sets represent only a small fraction of YouTube's complete video library. This means that the likelihood of a given related video appearing elsewhere in the same data set is low. Having graph nodes without accompanying attribute information would complicate analysis and were pruned.
3) **Variable Dimension**
   Approached directly, the number of attributes in each element ranges from 10 to 29 due to the variable number of related videos. This is less of a problem after the network has been constructed, but it does mean that not all nodes will have the same degree. Despite the variability, the vast majority of the elements in each data set contain the maximum of 20 related videos.
4) **Time Series**
   Much of the analysis done here relies on working with multiple data sets at a time. Because of the randomized nature of the crawls, it's possible for an element to appear multiple times with different attribute values as those values change over time.

Research completed under the supervision of James Hearne

III. Goals

The main goal of this work is to understand the relationships between categories of videos. The data exploration phase of research will hopefully illuminate other interesting relationships using the other attributes. Given high confidence relationship measures, this work will explore how those relationships change over time.

IV. Methods

A. Preprocessing

The data sets provided ranged in size from 10,000 to 750,000 elements. A balance had to be found between network size and network sparsity. Smaller networks were more manageable to visualize and run analysis on but were more sparse. Conversely, larger networks have a higher chance of containing nodes with more relationships but take more time to generate and process.

Related videos referenced in existing elements that do not exist elsewhere in the data sets were pruned. This was accomplished by reading every element of the data set into a hash map then iterating over every related video for each element. If a given related video's ID did not exist as it's own entry in the hash map, it was removed from the current element's related video list. This not only removed null entries from the graph, but greatly reduced the size of the output graphs.

V. Exploration

The data in its raw form was very difficult to interpret and gain any insight into. Text processing tools were used first to under stand the categories and relationship sparsity. Table 1 shows the distribution of categories for a canonical data set, that was found to be representative of the other data sets used.

Category	Count	Percentage
Music	16456	28.35%
Comedy	6040	10.41%
Sports	5503	9.48%
Film & Animation	4421	7.62%
People & Blogs	4272	7.36%
Gadgets & Games	3038	5.23%
News & Politics	2197	3.78%
Travel & Places	1368	2.36%
Autos & Vehicles	954	1.64%
Howto & DIY	903	1.56%
Pets & Animals	632	1.09%
Unlabeled	584	1.01%
Total	58047	100.00%

Figure 1. Category Distribution

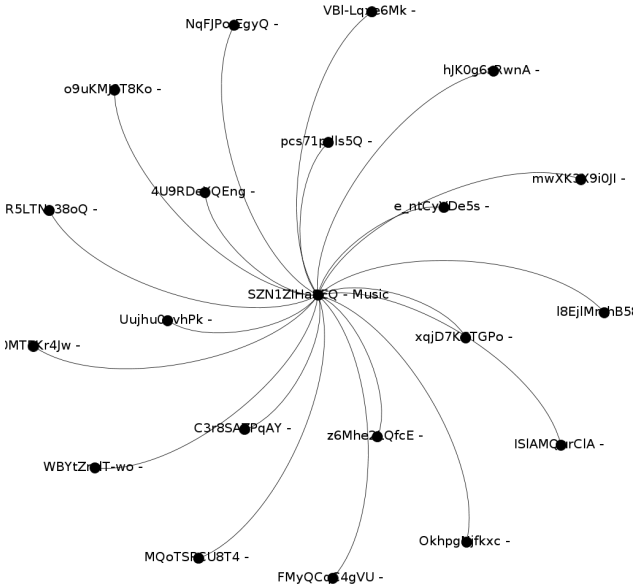

Figure 2. Example Network Construction

Gephi is visualization and exploration framework tailored to network and graph data. After initial data preprocessing, Gephi was able to show the general relationships between categories for several data sets. It was unable to show the entirety of the larger data sets at once, but contains filtering options to reduce the number of nodes considered.

Filtering was done by requiring nodes to have a higher degree. There is a positive linear distribution for degree over the nodes; by filtering the nodes that had degree less than half the average degree we could restrict the graph to more highly connected nodes and work with roughly half of the original nodes.

Initial exploration showed that videos generally form cliques with other videos in the same category. Though not as common as intra-category relationships, a few obvious directed inter-category relationships became visible:

1) $Music \rightarrow Entertainment$
2) $Music \rightarrow Comedy$
3) $Entertainment \rightarrow Comedy$
4) $People \rightarrow News$
5) $Sports \rightarrow Entertainment$
6) $Sports \rightarrow News$

Next, we went back to the data preprocessing stage to make new graphs directly targeted at inter-category relationship mining.

A. Building a Network

Each video in the data set has a unique identifier, a category attribute, and 0-20 related videos. Figure 2 presents how we use this information to construct a network by treating the related videos as edges in a directed graph.

For this construction to be useful, further processing was required. Firstly, the videos referenced in a data point's related videos may not appear else where in the data set. Videos referenced without corresponding information cannot be used since they lack the category information required for analysis. So, we prune the network of all edges where this property holds. This pruning may produce nodes in the network that have corresponding category information, but no longer have any edges. These nodes are also not of any use and are pruned as well.

Tables 3 and 4 present the result of this pruning process on the number of edges and nodes remaining in the network. We can see that very large proportions of the nodes and edges are being lost to the pruning process. The Future Work Section describes possible approaches to reduce this cost.

Original Nodes	167	1,440	8,689	58,047
Post Pruning	39	869	5,239	37,735
Percent Pruned	76.64	39.65	39.15	34.99

Figure 3. Pruning on Nodes by Number of Nodes

Original Edges	3,185	27,334	166,725	1,112,785
Post Pruning	61	7,049	37,898	247,760
Percent Pruned	98.08	74.21	77.27	77.35

Figure 4. Pruning on Edges by Number of Edges

B. Filtering on Network Attributes

This processing greatly reduces network size, but not by enough. Gephi [2] cannot produce meaningful visualizations for networks greater than a few thousand nodes. Even if it was possible, the size of the resulting networks would likely be difficult to interpret through visual inspection.

Further filtering may be done based on a number of network or node attributes, including: degree, edge weight, diameter, density, average degree, and max degree. These attributes may also be used to compare filtered networks to full networks to ensure that the filtered networks are still representative of the information originally available.

Network attributes proved to be poor predictors of representativeness between network sizes. Most of the attributes are depending on the size of network and vary greatly between different graph sizes. This can be seen in Table 5, there is large variability in the values of each network attribute between network sizes.

Network Size	167	1440	8689	58047
Diameter	2	6	24	36
Density	0.002	0.003	0.001	0
Average Degree	0.731	9.811	8.736	8.736
Max Degree	5	43	45	95

Figure 5. Network Attributes by Number of Nodes

Another consideration when sampling networks is how category representation changes in smaller samples. Ideally, the smallest sample that still gives an accurate representation would be chosen. Figure 6 shows the distribution of categories for data sets of varying sizes. These sizes were taken from the data sets provided. Treating the 58 thousand video data set as the baseline, we can see that samples as small as 1440 generally have the same category representations.

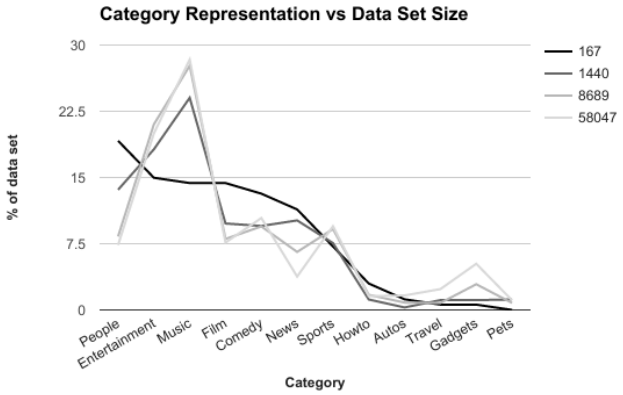

Figure 6. Category Representation

VI. Results

A. Further Analysis

Video-to-video relationships provide some interesting information but have some inherent limitations. Videos in the same category are broken into many disparate clusters, making analysis on a per-category level difficult. Additionally, the networks produced for video-to-video relationships are very large, making visualization difficult and network attribute calculation expensive.

Category-to-category networks address both of these short comings. Category networks only have a number of nodes equal to the number of categories, in this case twelve. Likewise, the number of edges is reduced from $O(20 * |videos|)$ to $O(2^{|categories|})$ or approximately from 1.2 million to 4096 for the largest data set considered, containing $58,047$ videos.

Additionally, these networks allow higher level analysis compared to video-to-video networks, which in turn allows deeper insight into how categories are related.

B. Intra-Category Clustering

All categories are strongly biased towards self-relationships, as opposed to an even distribution of edges to all categories. Figure 7 shows the percentages of edges for each category that are self-loops. This information gives us a measure of the exclusivity of each category. As an example, we can infer that categories like Travel & Places are more likely to overlap with other areas of interest than, say Sports videos.

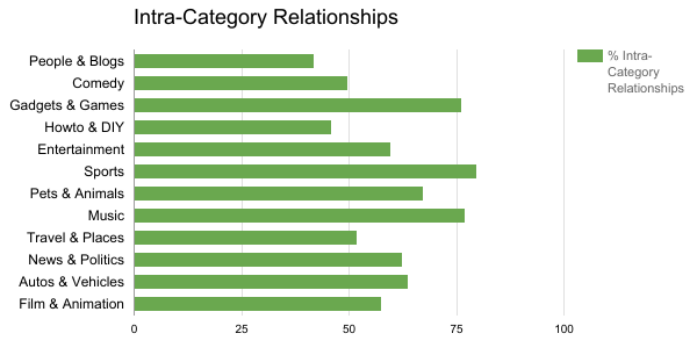

Figure 7. Intra-Category Relationships

C. Categories

Video-to-video networks were useful for initial exploration but were difficult to use for measuring inter-category relationships. To this end, new graphs were constructed from the data with the following characteristics:

1) Each node represents a category from the original data set
2) Edges represent the strength of the relationship between nodes, measured by the following function:

$$weight(E) = \frac{|\texttt{edges A to B}|}{|\texttt{edges from A}|} + \frac{|\texttt{edges B to A}|}{|\texttt{edges from B}|}$$

where E is the edge between nodes A and B.

If we treat categories as nodes, and each video as an edge, we can construct a new network that shows the strength of relationship between each category. In Figure 8, edge weight is assigned according to the preceding function.

We can see from the line weight that some categories have strong relationships, and almost all categories have at least

weak relationships with all other categories. As a note, the weights shown in Figure 8 are heavily skewed towards self-loops, which is why it's difficult to tell the difference between some of the weaker weights. Likewise, any weight above 0 results in a line, meaning that even if only 0.001% of Sports videos have a related video in the Pets & Animals category, there will still be an edge between them.

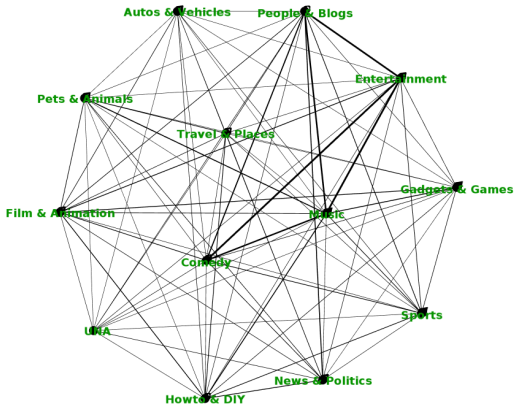

Figure 8. Full Category Relationship Network

The strongest relationships are not entirely surprising: we can see that Entertainment is related to Music and Comedy, and Music is also related to People & Blogs. The least strongly related categories were Pets & Animals to Autos & Vehicles and Music to Autos & Vehicles.

It's important to remember that these are directed edges and the strength of the relationships are partially determined by the starting node. As an example, 11.74% of Music videos were related to Entertainment videos, however 16.37% of Entertainment videos were related to Music videos.

VII. Future Work

With more time and potentially different processing techniques, it would be valuable to explore how the relationships between categories change over time. This would be possible with the current data sets, since they were created through crawls at different times. The timing spans 3 years in total, so there should be enough information to detect a trend if one exists.

Further investigation should be conducted into constructing representative sample networks. Sampling networks by node attributes resulted in a high percentage of the edges, and therefore relationships, to be lost. Some kind of agglomeration of data sets could be done to increase the chances of finding the related videos referenced by each node. Smarter pruning might target low degree nodes, or nodes whose edges do not lead to other nodes in our data set.

VIII. Conclusion

We have seen that social network mining approaches can work well for understanding large graphs. Additionally, these tools can be applied to data sets that are not traditionally seen as social networks. Pruning networks without affecting their properties and representativeness is difficult, but can be achieved by characterizing networks by node attributes.

References

[1] X. Cheng, C. Dale, and J. Liu. (). Dataset for statistics and social network of youtube videos, [Online]. Available: http://netsg.cs.sfu.ca/youtubedata/ (visited on 02/16/2017).

[2] (). Gephi, the open source graphviz platform, [Online]. Available: https://gephi.org/ (visited on 04/16/2017).