Analysing Youtube Datasets Structure and Relation

Ruyuan Zuo, Haihan Jiang

Table of Contents

Abstract:

Most video websites today like Youtube.com, used a recommendation algorithm to tailor experience for its users. While there are benefits with personalized website recommendation, there also has been an argument on whether a recommender system narrows a personal view with online experience. In this project, we studied the youtube data structure using network analysis. We developed a database to store the dataset and then used a page rank algorithm to assign each video with a page rank score. By combining videos with its relative rank score, we aim to propose a solution for the problem by adding an indicating factor. With such a factor, similar to scholarly articles having an impact factor, users can use the information provided by the vidoes recommendation at their own discretion and allows the users to avoid or select certain types of video.

1. Introduction

1.1. Background

Most video viewing websites use recommendation systems. A recommendation system gives its user suggestions for the videos to view next. However, a personalized recommendation is also said to have an aggravating factor to narrow the view of the users. A personalized recommendation will only reinforce the opinions or views of the users, but not to provide alternative opinions. While the development of personalized recommendations was aimed to help the user experience, these defaults should also be considered when designing systems for recommendation. In this project, we propose to let users have a choice to have information on the weight of the videos that they are viewing, similar to an impact factor, to inform the user with the choice of the video recommendation.

1.2. Problem statement

The three main questions we aim to answer in this project is how the youtube dataset looks like.we want to understand the structure of the youtube video network by visualization techniques. A network visualization by Gephi can give information on the clustering and direction of the recommendation videos. A SQL query can see what are the videos that are related. The second question we consider is how the recommendation was done. We implemented a naive recommendation algorithm called PageRank and we analyze how the rank of the videos is. With the ranking information, we then consider how the rank is related to underlying assumptions the rank entail. Is a personalized recommendation system narrow the view of its users.

1.3. Inputs and Outputs

For SQL, the input is converted to a structured dataset by a Java program written to populate the database according to the ER design. The output for the database is any query the user wants to search with information provided.

For PageRank, the raw data was changed to a graph data structure. This graph data was then used by the PageRank class to calculate the page rank for each node. The output will be a matrix containing nodes and its output.

Our final output will consider to concat the SQL queries with the page rank to provide information on the selected entries.

1.4. Goal and Objectives

First, we want to prove that our assumption with the recommendation system is correct. We aim to provide a page ranking with the entries of videos we present in the database. The goal is to provide users with the information on the calculated weight by the webpage for the user's discretion if needed. We also want to justify if this recommendation system will reinforce the opinion for the users.

1.5. Challenges

One challenge we faced is data cleaning. The dataset we downloaded was in an unstructured format. The related videos need to be structured to key value format in order for performing further analysis. The second challenge is the data size is very large, and it takes a very long time to run the analysis on the dataset, especially with the Gephi tool.

1.6. Related Works

In 1998, Larry Page, the founder of Google proposed a PageRank algorithm as an effective way to rank web search results[1]. In the paper, Page described in detail about the functionality of PageRank with search engines. The linked webpage is ranked by its connections with other webpages. This algorithm can also be considered a personalized keyword search. With its wide utilities, this algorithm is chosen as our main method when implementing a recommendation system.

In researching on this paper, we also looked into papers like on youtube recommendations. There have been numerous papers [Nie et al,2,3] that studied the social effect of using youtube. Alghout youtube data has been found useful in studying the social iterations and also impact with human perception. There has not been study on providing user information with how the video is ranked.

2. Data

2.1. Data Description

The total entries is 32636 nodes. We chose the first layer of the crawling data from a public available datasource from from Simon Fraser University British Columbia, Canada http://netsg.cs.sfu.ca/youtubedata/. This set of data has eight categories. video ID is an 11-digit string, which is unique and contains no repeated ids for each entries.Uploader is a string for the video upload username. Age is an integer number of days between the date when the video was uploaded. Category is a string containing video category information. Length is an integer number for the video length. Views is an integer for the number of views. Rate is a float number containing information about the video rating. Ratings is an integer number of the ratings. Comments is an integer number of the comments. Related IDs contain all related videos ids for the current video.

2.2. Data Cleaning

We clean the data by removing any incomplete entries with NAN as the focal node, and remove any connected node that has no name. Using Java and python, we are able to structure the data to the specific key value format for our use.

2.3. Data Visualization

We used Gephi as our tool for visualization. The below figure one shows the network at degree filter at 25.

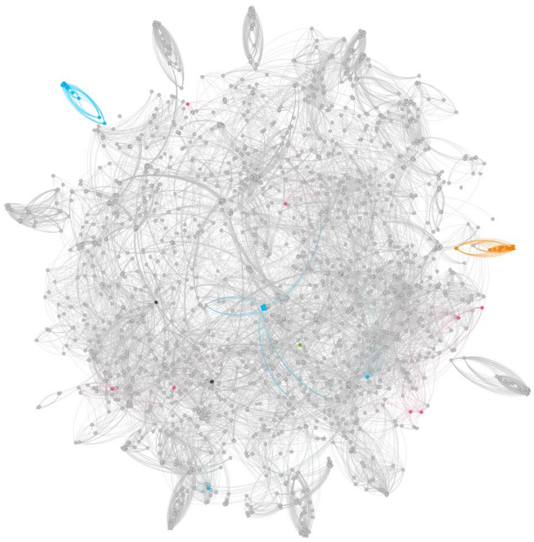


Figure 1: Data Network Visualization by Gelphi

The darker nodes are nodes with higher levels of connectivity, while the lighter nodes are ones with less connectivity. We can visualize that there are a few clusters of nodes that have very high connectivity, and those connections are cluster based.

We also visualized some of the categorical information provided below. We first looked into the categorical data and analyzed the distribution shown below. We can see that the top few most popular categories for this video set is entertainment, sports and people & Blogs.

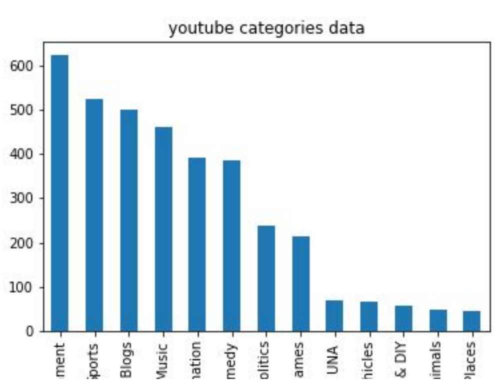


Fig 2. Categories data

We then looked into both the length, comments and rate distribution and plotted these features below. We can see that most video is less than 250 seconds, most video has less than 10 comments and the videos in this dataset seem to have a higher rating.

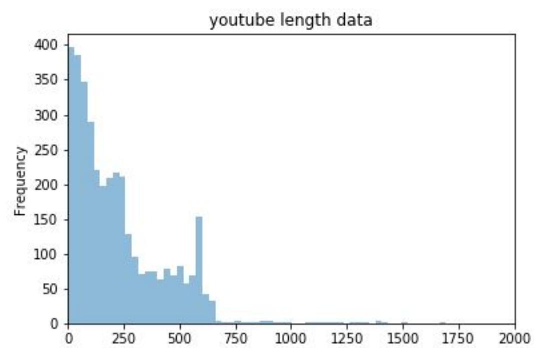


Fig 3. Categories data

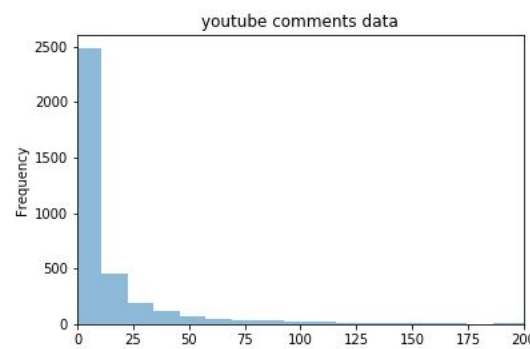


Fig 4. Comments data on youtube dataset

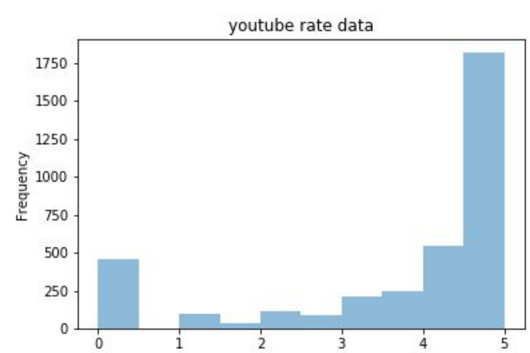


Fig 5. Rate data on youtube dataset

3. Python Evaluation

3.1. Page Rank algorithm and Run Time Analysis

As discussed in the related work section, PageRank Algorithms has a wide range of usages with graph structured data and ranking nodes according to their connections.

The following is a pseudocode for the algorithm:

Pseudocode:

- Assign weight matrix for directed graph as 1/(num of node)
- Calculate the individual node according to the sum of the probability of each current income node over the total out node of the income node.

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)},$$

Equation 1, see reference [1].

- Iterative until convergence
- Added in damping factors to force convergence

(Also see attached code file for implementation details, code_display.py)

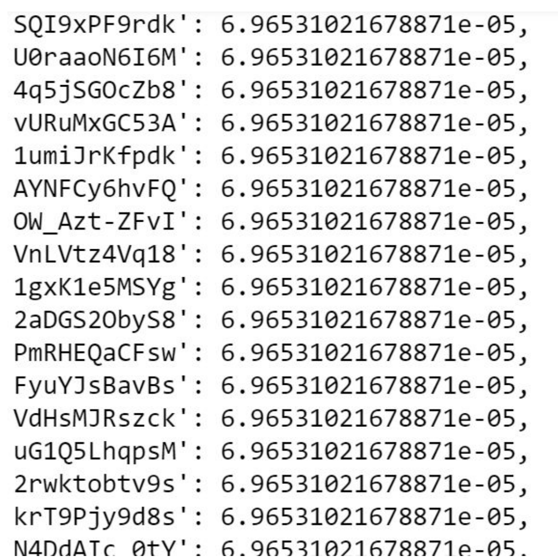
Each iteration requires a whole matrix multiplication, so the complexity is the size of the table, which is O(n^2) time complexity, where n is the size of nodes in the network.

3.2. Result with Page Rank

We selected the top five ranks and studied their connecting nodes and the categories, as shown by the Figure 6 and Table 1 below. We find that all the top ranked videos are in the category sports.

Video	Page Rank (e^-0.5)	Categories
1ALjw5wrhzU	9.248	Sports
y0_XLRcKH_Y	9.207	Sports
Er3K59aVJmM	9.050	Sports
LcIVYYHxoEA	8.693	Sports
8240cBUuP-c	8. 447	Sports

Table 1: Top five node by PageRank



```
SQI9xPF9rdk': 6.96531021678871e-05,
U0raaoN6I6M': 6.96531021678871e-05,
4q5jSGOcZb8': 6.96531021678871e-05,
vURuMxGC53A': 6.96531021678871e-05,
1umiJrKfpdk': 6.96531021678871e-05,
AYNFCy6hvFQ': 6.96531021678871e-05,
OW_Azt-ZFvI': 6.96531021678871e-05,
VnLVtz4Vq18': 6.96531021678871e-05,
1gxK1e5MSYg': 6.96531021678871e-05,
2aDGS2ObyS8': 6.96531021678871e-05,
PmRHEQaCFsw': 6.96531021678871e-05,
FyuYJsBavBs': 6.96531021678871e-05,
VdHsMJRszck': 6.96531021678871e-05,
uG1Q5LhqpsM': 6.96531021678871e-05,
2rwktobtv9s': 6.96531021678871e-05,
krT9Pjy9d8s': 6.96531021678871e-05,
N4DdAIc_0tY': 6.96531021678871e-05,
```

Figure 6: Page rank matrices for all the nodes

We also find the related videos for the top ranked videos and we also find that the majority of their related videos are also in the categories Sport. This confers with our assumption that a highly ranked video will skew the recommendation to certain categories. The result is that once a high weighted node is identified, the related recommendations, as shown by the page rank, will also be of that category. What this means is that if we weren't told with page rank, we will see once we clicked on some videos, suddenly all the videos on our thread become to a certain degree.

This page rank result showed that our assumption about skewed recommendation is valid, and Figure 7-8, with the related videos, showed that the assumption with narrow the viewers video choice is also corrected.

	videoID	uploader	age	category	length	views	rate	ratings	comments	relatedID
31016	l-0twwKwDUQ	bmiu4evr	740.0	Sports	85.0	4405.0	2.50	2.0	3.0	1ALjw5wrhzU
31157	tiKPf5d3fH0	qwermish	740.0	Sports	146.0	4251.0	4.33	3.0	12.0	1ALjw5wrhzU
31221	p4pcJOAm17Y	MrFootballFanatic	740.0	People & Blogs	320.0	4096.0	2.44	9.0	2.0	1ALjw5wrhzU
35974	6eg_DmSyx3I	arsenal4ever2006	741.0	Sports	32.0	2371.0	5.00	2.0	3.0	1ALjw5wrhzU
35998	iAywBdeF9j4	massilia74	740.0	Sports	27.0	9586.0	5.00	6.0	11.0	1ALjw5wrhzU
41357	DF7Be9Pvfkg	Hazboom100	648.0	Sports	50.0	6557.0	4.20	5.0	1.0	1ALjw5wrhzU
41377	J1W2QKqXO0Y	Myfootballvideos3	740.0	Sports	63.0	20780.0	4.47	15.0	4.0	1ALjw5wrhzU
41397	Z9GguOdO1Ps	robbenpt	740.0	Sports	27.0	8361.0	4.00	4.0	6.0	1ALjw5wrhzU
41417	rQnALf0CxFw	trabzonumbenim	742.0	Sports	28.0	288.0	0.00	0.0	0.0	1ALjw5wrhzU

Figure 7: (Partial) Related video for 1st ranked video

	videoID	uploader	age	category	length	views	rate	ratings	comments	relatedID
388	bk5WiqF0AVM	SensNetworkDotCom	494.0	Sports	104.0	8276.0	5.00	2.0	9.0	Er3K59aVJmM
548	Bp0tGnQgeZw	SensNetworkDotCom	494.0	Sports	9.0	296.0	0.00	0.0	0.0	Er3K59aVJmM
594	tD9Jr7GP67c	SensNetworkDotCom	739.0	Sports	57.0	5300.0	4.44	9.0	38.0	Er3K59aVJmM
30599	DXXtJ_Xi79c	NYR21135	741.0	Sports	88.0	46.0	0.00	0.0	0.0	Er3K59aVJmM
30619	vN6_8Bj3Mw8	SyntaxZ	739.0	Sports	27.0	1910.0	1.50	2.0	2.0	Er3K59aVJmM
30639	95EjlJdbxlw	eassona	740.0	Sports	55.0	441.0	5.00	2.0	0.0	Er3K59aVJmM
30676	NaN	jedi24	646.0	Sports	60.0	722.0	0.00	0.0	0.0	Er3K59aVJmM

Figure 8: (Partial) Related video for 2rd ranked video

3.3. Node analysis

In the following section, we introduced some network analysis on the properties of the network for the youtube data. As seen by the average clustering and average degree, the network is comparibaley sparse with a larger input node (N = 32753) and edges (E = 67911).

The following explains the calculation of these parameters. The average degree is calculated by the average number of edges for one node. The average clustering is calculated by the number of edges connecting a node neighbor divided by the total number of edges between the nodes [4].

Total Node	32752
Edges	67911
Strongly Connected Components	30903
Weakly Connected Components	172
Average Clustering:	0.16543
Average Degree:	2.074

Table 2: Network Node Analysis Results

We also vaulted the in- and out-degree analysis for the network and plotted below. These results again showed that compared to the number of nodes, the connections are not as strong as we would have expected, but they do show some clusters and connections. For the median in degree is at one and the average out degree is around 0. This shows that more data is needed for the next step, if provided enough computation power.

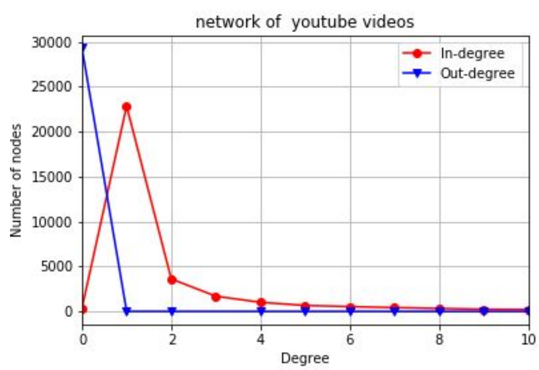


Figure 9: In and out- degree analysis for the network

After seeing the cluster visualization, we also performed a centrality analysis. The result shows confirmation with the node analysis on the sparse organization of the matrice. This would suggest that the skewed influence for page rank, and other recommendation algorithms will be stronger if a more dense network is obtained, thus increasing the reinforcement of similar contents with recommendations.

Top closeness centrality	1umiJrKfpdk: 1.0 vURuMxGC53A: 1.0
Top eigenvector centrality	1umiJrKfpdk: 0.2241 vURuMxGC53A: 0.2241

Table 3: Centrality Analysis

	videoID	uploader	age	category	length	views	rate	ratings	comments	relatedID
68654	krT9Pjy9d8s	EA	741.0	Gadgets & Games	775.0	117.0	5.0	1.0	1.0	IqlxYO7YCI8
68854	1umiJrKfpdk	EA	742.0	Gadgets & Games	77.0	1043.0	5.0	9.0	7.0	IqlxYO7YCI8

Figure 10. Top central node categories

4. Data Storage and Sql Query

4.1. Data Scanning

The input raw data file's type is txt file. We used the find and replace function in the Notepad to eliminate white space and unnecessary symbols for clean data storage and data reading.

In the reading process, we implemented Java scanner class for reading through txt file line by line. To make easy data storage management, we used Linkedlist for storing attribute values. With complete data cleaning storage, it is able to connect to mySql database and load data into it.

4.2. MySql Database

Based on 2.1 data description and 2.3 data visualization, we developed a Entity Relationship diagram to show how entities are related to each other.

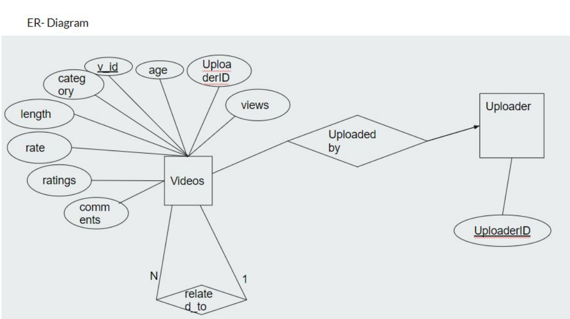


Figure 11. Entity Relation Diagram

There are two entities in our database: Videos and Upload.The Primary key for Videos and Uploader are v_id and Uploader respectively. For each video, there is a set of videos related to it, and it is uploaded by the uploader. We developed our create table statement based on the ER diagram. Then we applied a JDBC driver to connect the database and storage of data using Java.

4.3. SQL Query

MySql ensures basic search functionality in the database. It ensures users to define their duration and entity search.

vid
1HxER8CPamY
1lYXY_eTNY8
1umiJrKfpdk
20Er3QOq8ks
26h06irOVIE
26ZRpwosc2s
2aDGS2ObyS8
2AYRJ_uRW5U
2B9TW4NG1vw
2cidEv9vchE
2EfuhjnvIsA

ratings
0
17
57
0
0
18
5
4
15

Figure 12 Examples of search output

select vid from videos where ratings>1 & ratings <5 ;

select ratings from videos where category = "Sports" & length<5;

Besides original attributes' values, we also hope to provide more valuable information to users, researchers or some people who want to know the weight of each video played in the recommendation system. So, we added pagerank values associated with videos into our database.

vid	uploaderID	age	category	length	views	rate	ratings	comments	pageRank
1umiJrKfpdk	EA	742	Gadgets & Games	77	1043	7	9	7	6.96531
4q5jSGOcZb8	EA	742	Gadgets & Games	92	1059	5	6	13	6.96531
AYNFCy6hvFQ	EA	742	Gadgets & Games	74	454	5	3	13	6.96531
OW_Azt-ZFvI	EA	742	Gadgets & Games	291	227	3.67	3	5	6.96531
SQI9xPF9rdk	EA	742	Gadgets & Games	68	1518	4.79	14	17	6.96531
U0raaoN6I6M	EA	742	Gadgets & Games	61	1128	4.67	9	6	6.96531
vURuMxGC53A	EA	742	Gadgets & Games	105	884	5	5	9	6.96531

Figure 12. Videos table with pagerank values

5. Conclusion

5.1. Findings

Overall, we consider our method successful in the sense that the page rank result for the most connected nodes did show a tendency for connecting to the same categories. What this entails is when someone uses these heavy weighted vidoes and the recommendations for the following videos they will get is also in the similar categories. We also see from the centrality analysis, that the clustering of the most centraled node also has a tendency to connect to the videos in the same categories.

Using pageRank algorithm to study the relationships between videos than using SQL is a much more effective method in terms of learning about clustering and connections. Videos with higher rank will induce a strong preference for recommending the same categories, and it might be useful to consider informing its user with an indicator factor like ranks by PageRank.

5.2. Future Work

Results on the centrality analysis showed that this network has relative low centrality, and most of the nodes were unclustered within the dataset. The result will be more conclusive if we considered all three layers of the crawling data.

PageRank was developed in 1988. Since there are numerous changes and improvements to the modern search and recommendation systems. Real websites also use much more complicated algorithms than PageRank. For further analysis, we might want to consider a more complex and up-to-date algorithm for analysis.

For future recommending programs, we suggest that they will not just look at one result from one particular algorithm, but to consider diversifying their result with a portion that is randomly selected. In other words, we suggest big companies like Google show their user results that came from multiple algorithms. For example, if we rank by PageRank, then all the categories will be in Sports. However, if you add in a factor to consider centrality, then, we will have a more diversified result containing both categories Sport and Game.

For larger companies, this would be easier to do because of the large database and available data. By studying the network properties at different angles, the recommendations will return more well rounded results.

6. Roles and Contribution

6.1. Group Work

- Discuss project selection
- Design ER diagram
- Discuss design choices and measurements
- Find dataset

6.2. Individual Work

- Ruyuan Zuo
  - Analyses using Python
  - Visualization using Gephi
  - SQL schema

- Haihan Jiang
  - Data Cleaning
  - Data Reading and Parsing with Java
  - Connection data with MySql

7. Reference

[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999, November 11). The PageRank Citation Ranking: Bringing Order to the Web. Retrieved December 10, 2020, from http://ilpubs.stanford.edu:8090/422/

[2] Nei, B., & Zhang, H. (2014). Social interaction based video recommendation: Recommending YouTube videos to facebook users. Retrieved December 10, 2020, from https://ieeexplore.ieee.org/abstract/document/6849175

[3] Kura, S. (2020, August 20). Applied Network Analysis using Python. Retrieved December 10, 2020, from https://towardsdatascience.com/applied-network-analysis-using-python-25021633a702

[4] D. L. Hansen, B. Shneiderman, M. A. Smith, and I. Himelboim, “Calculating and visualizing network metrics,” Analyzing Social Media Networks with NodeXL (Second Edition), 17-May-2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128177563000066. [Accessed: 10-Dec-2020].