Network Analysis on Youtube Recommendation System

Project Proposal

Sarah ABDELBAR, Caterina CONZ, Stefano MONETA, Chiara PALMA

ABSTRACT

The aim of this project is to analyze the YouTube network, performing link prediction to better understand how the platform recommendation systems works, and clustering, to verify if the assigned video category is a good approximation of how video cluster together.

1 MOTIVATION AND PROBLEM DEFINITION

According to viewing figures recently revealed by YouTube, people around the world spend a billion hours watching content on the platform. The American video sharing website is equipped with one of the most sophisticated and largest scale recommendation systems in existence. The algorithm that lies behind it is essentially a combination of directives tailored by Google engineers along with learned behaviors that have been training on the blur process of machine learning which makes the recommendation system itself hard to disentangle. This begs key questions about the functioning of the algorithm, for marketers, content creators and in general for users. That is indeed, what brought our group to dedicate this project to the analysis of the whole YouTube network in order to have a clear idea of it as a whole, identify genre clusters of each video and perform link prediction to fully understand the aforementioned algorithm.

RankBrain, the component of Google's algorithm at the base of YouTube recommendation system, has truly revolutionized the way search results are determined. The idea of links as a ranking signal, originated in 1996 with Google's PageRank algorithm [8], which then paved the way for several different structures and implementations for ranking algorithms, and RankBrain is between these. This system, introduced in 2015, allows YouTube to better understand the user's intent behind a search query.

Significant efforts have been made in the literature to disentangle RankBrain. Having reached consensus regarding the particular problem we wanted to address, we reviewed a series of scientific papers to have a better understanding of the state-of-the-art methods currently being used in the field. J. Paolillo and Ro et Al. provide in their papers [7] [8] a general overview of the YouTube network as a social network, describing the degree distribution and the shape of the network. Moreover, they try to identify influence sets of the videos, a factor that could have key strategic applications, such as measuring effectiveness of advertisements.

Baluja et Al., present in their paper [2] an analysis of the user-video graph with the aim of providing personalized video suggestions for users. By doing so, they challenge RankBrain, producing a new algorithm that they call Adsorption, which is able to efficiently propagate preference information through a variety of graphs, then testing their resulting recommendations on a three month snapshot of live data from YouTube.

We also investigated the previous approaches researchers have implemented to analyse music video recommendations specifically: Matsumoto et Al. present in their paper [6] a novel method which constructs a network, that not only represents relationships between music videos and users but also captures multi-modal features of these, such as audio, visual and textual features. In addition to this, they derive a scheme for link prediction, considering local and global structures of the network. In doing so, they use multiple link prediction scores based on both local and global structures. Also, they use LP-LGSN to predict the degree to which users desire music videos, and then also here they perform testing on a YouTube dataset.

2 METHODOLOGY

Our goal is to analyze the YouTube network to understand the existing relationships and clusters of videos. This is a crucial step to generate video recommendations for users as well predicting the number of views of the videos. For that, our project proposed some link prediction algorithms that lead to video recommendations as well as number of views predictions for the videos in the YouTube network. This project is complementary to prior work that dives deep on user profiles from a social network perspective rather than focusing on video features. In order to reach our final goal, the below point summarize the proposed flow for our project.

2.1 Data Preprocessing

For a better understanding of the network, the first steps would include Graph Visualization and computation of number of nodes, edges and cluster coefficients. This will allow us to visualize the structure of the graph and the density of its different clusters. Data Prepossessing techniques including node embeddings like Node2vec will follow as a step to prepare the data for feature engineering. To dive deeper into the structure of the network, Centrality algorithms such as Betweeness Centrality, Degree Centrality, and Eigenvector Centrality will also be computed.

2.2 Feature Extraction and Selection

Features to be included in our Link Prediction model include node features that are already in the dataset like views or comments aggregation. Other features also include similarity metrics between descriptions of the videos using the keywords, or similarities in video authors. As for the topological features generated by the study of the graph, we can extract node role features using RolX [5], as it has been proven to be very effective in prior work on the same topic. The RolX (Role eXtraction) is an unsupervised learning approach used to automatically determine the roles of every node in the network. These roles capture the structural behavior of the graph. The way the approach works is that it extracts connectivity features such as degree for example, and then uses those features to generate additional ones by aggregating the features of the neighboring nodes. The logic behind this algorithm is that nodes with similar roles will most likely have similar neighbors. The roles generated by

RolX are very important for graph visualization, network transfer learning, as well as node similarity tasks like the approach proposed by our project.Previous work showed experimentation results that prove the ability of roles to perform network learning without the availability of class labels in the target graph. In our project, Roles generated by RolX will be used as node similarity features.

2.3 Generating Video Recommendations for users using Link Prediction

Supervised machine learning algorithms to be implemented include Logistic Regression, a tree-based algorithm like Random Forest, K-Nearest Neighbors, as well as Naive Bayes algorithms. The results generated by the different models will be compared to complement the work done by Han Launch, James Li and Justin Xu in their paper published by Stanford University. [1] This Link Prediction problem is treated as binary classification as we try to predict whether a link connecting two nodes exists, hence producing recommendations of similar videos for the user to see.

2.4 Clustering of YouTube Video

Different clustering algorithms could be also used to partition the graph into natural groups so that the nodes in the same cluster are closer to each other than to those in other clusters, identifying in this "social groups". Some of the approaches we could test are K-means, Greedy Agglomeration, Markov Clustering Algorithm using random walks, and clustering with Minimum-Cut Tree.

2.5 Predicting number of views

Having number of views as one of the node features in our dataset, another potential application in our project is to come up with a model that predicts the number of views for videos based on the view flow of the user and the related videos in our graph. In other words, suppose that we reach a point where we know a lot about the relationships between the videos, we would be able to predict what video the user might watch next, assuming he starts at a random video, generating likelihood predictions that can be translated into number of views. With this, we will be predicting the popularity of the different videos in the YouTube network. In simpler terms, given some basic features of a video, we can predict whether it might go viral or not.

3 MODEL EVALUATION

3.1 Database Used

Our analysis of the Youtube network will be conducted on the dataset "Statistics and Social Network of YouTube Videos" retrieved on the Multimedia and Wireless Networking Group at Simon Fraser University website [1]. The database collects data about YouTube videos, crawled across different videos. In particular, information from the YouTube API together with publication data, category and related videos are extracted and available for academic use.

We plan to use this dataset to build a directed graph where each node corresponds to a video, and node a is connected to node b if b is in the related video list (considering just the first 20 entries) of video a. The database contains multiple folders, one for each day in which the data ere crawled, and the total amount of unique videos crawled across 2 years exceeds one million. Therefore, we decided to analyze videos collected on one day, till the third depth. Our approach is similar to what has been previously used in [1], but considering just one day instead of one, and going till further than depth two.

In order to test link prediction algorithms, we plan to create a train and a test dataset, randomly removing some edges from the graph. Differently, to analyze the different clusters present in the network, we are going to use the entire dataset.

3.2 Evaluation of the algorithms

3.2.1 Link Prediction Algorithms. Link prediction algorithms will be evaluated in the following way. We are going to consider every pair of not connected nodes in the train set, predict the likelihood that they are going to be connected and compare it with the actual edges in the test dataset. The evaluation metrics typically used in link prediction are much the same as those used in any binary classification task. Moreover, they can be divided into two broad categories: fixed threshold metrics and threshold curves [4]. In particular, we are going to use the F1 score, which takes into account both precision, i.e. the ratio of true positives to all predicted positives, and recall, i.e. the ratio of true positives to all actual positive values. Indeed, this metric is particularly well-suited for unbalanced classification problems, such as link-prediction.

$$F1 = 2 \cdot \frac{p \cdot r}{p + r} \; where \; p = \frac{tp}{tp + fp}, \; r = \frac{tp}{tp + fn}$$

Finally, we are going to evaluate the performance using threshold curves such as the ROC curve and the precision-recall curve.

3.2.2 Clustering Algorithms. We could use several graph clustering evaluation metrics, such as the ones proposed in [3], i.e. Average Isolability (AVI), Average Unifiability (AVU) and Average Normalized Unifiability and Isolability (ANUI), which takes into account both of them. In addition, it could be interesting to compare the clusters identified with the category of each video as found in the YouTube metadata.

3.2.3 Number of views prediction. In order to evaluate the accuracy of our views prediction we could compare the value predicted with the updated number of views, available on a separate dataset at the same source. In particular, metrics used in regression problems could be used, such as the Mean Squared Error (MSE).

[1] https://netsg.cs.sfu.ca/youtubedata/

REFERENCES

[1] Aung, H. L., Li, J., and Xu, J. (2017). Links prediction between youtube videos using role features and node attributes. *Stanford University*.
[2] Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., Ravichandran, D., and Aly, M. A. E. (2008). Video suggestion and discovery for youtube: taking random walks through the view graph.
[3] Biswas, A. and Biswas, B. (2017). Defining quality metrics for graph clustering evaluation. *Expert Systems with Applications*, 71:1–17.
[4] Davidson, J., Livingston, B., Sampath, D., Liebald, B., Liu, J., Nandy, P., Vleet, T. V., Gargi, U., Gupta, S., and He, Y. (2012). Link prediction: fair and effective evaluation. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.
[5] Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H., Basu, S., Akoglu, L., Koutra, D., Faloutsos, C., and Li, L. (2012). Rolx: Structural role extraction mining in large graphs.

[6] Matsumoto, Y., Harakawa, R., Ogawa, T., and Haseyama, M. (2019). Music video recommendation based on link prediction considering local and global structures of a network. *IEEE Access*, 7:104155–104167.

[7] Paolillo, J. C. (2008). Structure and network in the youtube core. page 156.

[8] Ro, Y., Lee, H., and Won, D. (2020). Youtube graph network model and analysis.