Link Prediction to recommend YouTube videos

1st Appala Avinash
Computer Science and Engineering
Shiv Nadar University
Greater Noida , India
aa760@snu.edu.in

2nd Seeram Narasimha
Computer Science and Engineering
Shiv Nadar University
Greater Noida , India
sn701@snu.edu.in

3rd Kolluru Jeshwanth
Computer Science and Engineering
Shiv Nadar University
Greater Noida , India
kk345@snu.edu.in

I. Abstract

Youtube is a prominent video sharing community and sharing site in internet. Later the prominence has transformed you tube from a video site to a big social network , connecting subscribers, commenters, and content creators together to form a huge network of intermingling social circles We'd like to learn more about how videos in these various social circles influence one another and what roles they play in their communities. In this paper, we use this data to create and compare several link prediction algorithms that recommend new related videos to users. To address this, we examined a YouTube dataset containing graphs of related videos and examined the emergent roles and communities within this network to see how their interactions influenced the rest of the graph's connections. We discovered that using K nearest neighbours on a combination of RolX roles and genres performed the best at predicting accuracy.

II. Key words

Link prediction, Supervised Learning, RolX, Logistic regression Random forest, Naive bayes, Standard vector machine, K nearest neighbour, Training set, Validation set, Testing set, accuracy, precision, F1 score

III. Introduction

YouTube is an amazing application where it helps you to watch, share and create videos online. On average every month six billion hours of videos are streamed on YouTube. It was created in 2005 and it is one of the leading websites in the internet now .There are millions of videos in you tube it is hard to predict link in the you tube, so there are some ways to predict the links in you tube .Link prediction is one of the most important research topics in the field of graphs and networks. The objective of link prediction is addresses the predicting the existence of missing relations or new ones

One of the main applications of link prediction is the detection of hidden social relationships, which is a friend-ship suggestion mechanism used by some online social networks. In such cases, hidden relationships could be existing social ties that have not yet been established in a social network or social ties that have been missed during the evolution of the social network [4]

These link prediction techniques divided into two ways like supervised [1] and unsupervised [5].Additionally the strategy used for a particular method influence how its performance is evaluated

Usually,unsupervised approaches assigns a score for each node with base on neighbourhood nodes and path information methods . According to these experimental results path information methods have higher accuracy than the neighbourhood nodes.But these global methods is time consuming and not feasible for large scale networks [6]

On the other side the methods based on supervised learning they consider the problem as classification problem [3] As a result, network information such as structural information and node attributes are used to construct a feature vector for each pair of nodes. These vectors are then used to train various classifiers to determine whether or not a link exists between two nodes.

In this paper, we aim to propose a model for predicting YouTube links in order to assist the user in finding additional videos.The lack of genre diversity in YouTube's suggested videos is a widespread complaint among viewers. A YouTube user might benefit from the platform by receiving recommendations for videos with comparable roles and attributes but from different genres, so improving their viewing experience and expanding their worldview.

We are trying to implement this using various supervised machine learning approaches for link prediction using features extracted from communities, including RolX role counts.A link prediction model will point to the relative relationships between communities

IV. Literature Review

A. Data set

The data set we will use is Statistics and social network of YouTube videos, which can be found at https://netsg.cs.sfu.ca/youtubedata/. Each dataset file is a directed graph of crawled YouTube videos, with each video corresponding to a node.

Each node contains information about the uploader, category, length, view count, and other relevant information that we may find useful when analysing the various video roles. A directed edge in the graph exists from node a to b whenever a video b is among the first twenty videos in video a related video list. We didn't have to collect any data for our dataset, but we did have to aggregate all of the graph files.

B. Charecterization of the youTube video community

The topology of YouTube is mainly focused in this work to learn more about its structural qualities as well as the nature of social links between users and between users and videos. It also examines network features such as user profiles and video popularity to emphasize the impact of social interactions on a content-sharing network such as YouTube.

This is also important in our work because it deals with the relatedness of videos, with the genre being the most important attribute. Although the tags are selected algorithmically rather than directly by human decision, this article indicates that these videos are highly influenced by social ties. Previous YouTube research has largely looked at the YouTube network through the eyes of users and subscribers, rather than through the eyes of videos and how they interact, as we do.

C. RolX

We use RolX to figure out which roles are present in our data, as it has already been demonstrated to be a reliable and scalable method of defining node roles via unsupervised learning. This study created an algorithm for extracting roles from a graph, which could subsequently be understood and used to categorize and find comparable nodes. we are interested in it since this offered a way to extract additional node attributes that could be utilized to discover graph section relationships. This can easily be reused to predict if two nodes are likely to form a link, which is equal to assessing whether two videos are likely to be related, utilizing these node attributes on YouTube videos.

D. Link prediction using supervised learning

The link prediction using supervised learning were mainly divided into three types of features: proximity, aggregated, and topological. These variables, which are integrated to help forecast link development,, specific qualities of nodes, define the similarity between nodes and relational structures between nodes.

Based on the study above we are trying to analyze a new link prediction model to provide good performance for predicting links

V. Mathematical Background and algorithms

A. Role of Rolx

To determine the role of a specific video, we can use the Rolx algorithm to automatically discover the structural role of the video in the YouTube network.

In the initial step Role describe each node as a feature vector.It can use any set of features that can be deemed important.by using Structural feature algorithm the feature extraction is done.For a particular node it extracts the the local and ego net features based on counts of links adjacent to b and within adjacent to ego net of v. Again, RolX is flexible in terms of a feature discovery algorithm, so RolX 's main results would hold for other structural feature extraction techniques as well.

After extracting features we have n vectors of f numerical entries each.So we can use soft clustering in the structural feature space and specifically , an automatic version of matrix factorisation after that generate a rank r approximation .There are many methods to do such approximation Rolx is not stick to any one of it. However , Here we should take care of computational complexity [2]

B. Feature Selection

In Link prediction using supervised learning approach this paper constructs a model for predicting links by selecting which features are approximate to use.and these include proximity, topological and aggregate features.

Proximity features includes key names and key words that can be found in the description of the video, like key word in a name that brings similarity between nodes.

Aggregate features are those which sum the value for particular metrics with in a set of nodes

Topological feature that contain the features of underlying connections in the graph. Let G = (V,E) be a directed unweighted graph , we denote set of node v's as direct neighbours of N(v)

1. Shortest distance

$$S(A,B) theshortestpathconnectingAandB$$

2. Common neighbours:

$$C(A,B) = |N(A) \cap N(B)|$$

3. Jaccard Distance:

$$J(A,B) = \frac{|N(A) \cap N(B)|}{|N(A) \cup N(B)|}$$

C. K nearest neighbour algorithm

KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. KNN captures the idea of similarity with some mathematics like calculating distance between points in a graph

The KNN algorithm first loads the data and initialises a value K to the chosen number of neighbours and calculate the distance between the query example and current example from the data and after that add that distance and the index of the example to an ordered collection and sort the collected elements and pick the first k entries from the sorted collection and get the labels of selected k entries and our case is classification so return the mod of k labels.

Here we should reminder that the we decrease the value of k from K to 1, the predictions become less stable and similarly if k value increases the predictions become more stable due to majority voting.In cases where we are taking a majority vote among labels, we usually make K an odd number to have a tiebreaker.

D. Random Forest algorithm

Random forest is a versatile, user-friendly machine learning algorithm that produces excellent results most of the time even without hyper-parameter tuning. Because of its simplicity and diversity, it is one of the most widely used algorithms.

The random forest is a supervised learning algorithm . The ”forest” it creates is an ensemble of decision trees, which are typically trained using the ”bagging” method. The bagging method is based on the idea that combining learning models improves the overall result.

Random forest constructs multiple decision trees and merges them to produce a more accurate and stable prediction. Random forest has a significant advantage in that it can be used for both classification and regression problems, which comprise the majority of current machine learning systems. You can also use random forest to handle regression tasks by using the algorithm’s regressor. While growing the trees, Random Forest adds more randomness to the model. When splitting a node, it looks for the best feature among a random subset of features rather than the most important feature. As a result, there is a wide range of diversity, which leads to a better model in general.

E. Navie bayes

Navie bayes is a supervised learning algorithm.It is based on Bayes theorem.It is a simple machine learning classification algorithm.It is depended on the baves theorem for calculating probability and conditional probability.

The existence of a given feature in a class is assumed to be independent to the presence of any other feature by the Naive Bayes classifier. It is majorly used in high-dimentional training dataset.one of the most simple and successful classification techniques for developing fast machine learning models capable of making quick predictions.Probability classifier means predicts based on probability of an objects

F. Standard vector machine

The support vector machine algorithm’s goal is to find a hyperplane in an N-dimensional space (N — the number of features) that clearly classifies the data points. There are numerous hyperplanes that could be used to separate the two classes of data points. The goal is to find the plane with the most margin. Maximizing the margin distance provides some reinforcement, allowing future data points to be classified with greater certainty. They will use support planes and vectors for this in this case.

VI. Classification and Approach

We will assume Link prediction as a binary classification problem and we check there is a connection between vertex and giving a score of 1 or 0 bases on the features between those nodes .In our data set it uses n(n-1)/2 points here n is the number of nodes in our graph G

In paper Link prediction using supervised learning it includes the multiple approaches we used including standard vector machine with a linear radial basis function kernel and random forest decision trees ,K nearest neighbours and naive Bayes and in this standard vector machine performs the best in overall. [3]

For evaluating each model we are testing with links between popular you tube videos from our data set from 2007. We then measured the recall , precision and F1-score of our models to see how accurate we can predict link prediction

A. Graph Formation

Our data set contains the information , in the first column contains the video id which is a 11 digit string and unique, Second column contains the user name of uploader ,third column contains the age i.e an integer number of days between the date when the video was uploaded and 15th february 2007, fourth column contains the category which is a string of video category choosen by the uploader ,fifth column contains the length which is an integer number of the video length , sixth column contains an integer number of views ,seventh column contains the rate i.e a float number of a video rate ,eight column contains the rating an integer number of the ratings ,ninth column contains comments an integer number of comments tenth column contains related id’s upto twenty strings of the related video ids

To develop a graph, we must first hash the video id string to provide a unique integer id. We can then retrieve the graph’s id from the dictionary id. After that we must add all of the nodes to the graph and create an undirected graph based on the video id dictionary.Finally we will load the data set file and constructs a directory

Using networkx we can plot the graph and with the help of the networkx library also we can calculate the clustering coefficient display the graph information

We analysed our data by visualizing various graph attributes.The data set includes several depths of the crawls through the related video through different time frames

Some statistics extracted from the graph are represented below . we done the lowest two depth crawl

- Graph with 3356 nodes and 3595 edges
- Graph info Graph with 3356 nodes and 3595 edges
- Graph clustering coefficient 0.042347428040800064

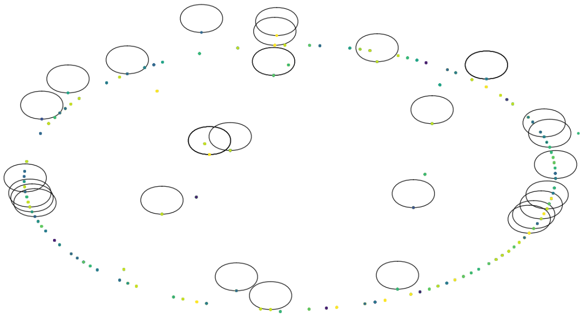

Fig. 1. Graph obtained from data set 0222

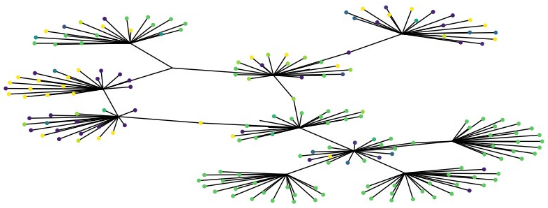

Fig. 2. Cluster

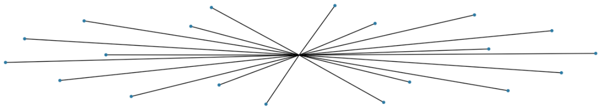

Fig. 3. A node with 20 related id’s

We relied on the given genres to identify graph communities when clustering nodes. For the sake of simplicity, we assumed that videos from the same genre would naturally form related clusters.

Here difficult to see the connected nodes in the graph centre the peripherals show that the related videos of each nodes are likely to be of the same genre.So we assume that the videos with same genres will be likely to form cluster so we this information as a feature representing communities into our link prediction algorithm instead of extracting them using other methods

We ran ROlX on our graph to categorise our nodes into specific roles to be used as feature for our link prediction algorithms .The recursive features of Role incorporated both local and global network structures helpful in encapsulating the structure of the network

VII. Results

Displaying our results by running the data sets with various supervised machine learning algorithms Random forest, logistic regression, k nearest neighbour and naive bayes

The training data set is the sample of data used to fit the model the actual data set that we use to train the model the model sees and learns from the data

The validation Data set the sample of data used to provide an unbiased evaluation of a model fit on the training data set while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation data set is incorporated into the model configuration

Test set is the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Now we are currently calculating the accuracy, precision, recall, and f1 score for training, validating, and testing data sets

Accuracy - Accuracy is the most intuitive performance measure which is simply a ratio of correctly predicted observations to total observations. One might believe that if we have high accuracy, our model is the best.

Precision - Precision is defined as the proportion of correctly predicted positive observations to total predicted positive observations.

Recall - The ratio of accurately predicted positive observations to all observations in the actual class is referred to as recall.

F1 score - F1 score is the weighted average of Precision and Recall. As a result, this score considers both false positives and false negatives. It is not as intuitive as accuracy, but F1 is usually more useful than accuracy, especially if the class distribution is uneven. For getting our validation set we first extracted 1 percent of node pairs from our original graph using them as our validation set to test performance on unseen node pairs .

Due to the fact that our validation set had such a few number of 1 labels the precision was very low but we can observe that when we are running recall values the models performed better but the k nearest neighbour gave the high recall values and also by comparing the F1 scores the k nearest neighbour performing the better

TABLE I
Results for Random forest

Rolx	Random Forest
Training accuracy	0.6036134453781513
Training precision	0.7154840964697659
Training recall	0.34403361344537814
Training F1 score	0.4646464646464646
Validation accuracy	0.8615885925347161
Validation precision	0.010281945162959131
Validation recall	0.35891647855530473
Validation F1 score	0.019991198843276548
Test accuracy	0.4411992170207319
Test precision	0.0017828370944522305
Test recall	0.7362204724409449
Test F1 score	0.0035570603844288255

TABLE II
Results for K nearest neighbour

Rolx	KNN
Training accuracy	0.9587755102040816
Training precision	0.9246788460683647
Training recall	0.9989195678271309
Training F1 score	0.9603665574073219
Validation accuracy	0.855044926661221
Validation precision	0.024830969903667805
Validation recall	0.9367945823927766
Validation F1 score	0.048379575658661696
Test accuracy	0.8346943020657211
Test precision	0.001815176169330005
Test recall	0.2204724409448819
Test F1 score	0.0036007072817874942

TABLE III
RESULTS FOR LOGISTIC REGRESSION

Rolx	logistic regression
Training accuracy	0.5875030012004802
Training precision	0.6222699365920756
Training recall	0.4453301320528211
Training F1 score	0.519137383321438
Validation accuracy	0.7310260326029051
Validation precision	0.0068069920364801
Validation recall	0.4650112866817156
Validation F1 score	0.0134175731127466
Test accuracy	0.6824453701283809
Test precision	0.0024339476953033
Test recall	0.5708661417322834
Test F1 score	0.0048472287223373

TABLE IV
RESULTS FOR NAIVE BAYES

Rolx	Naive Bayes
Training accuracy	0.5272388955582233
Training precision	0.5202629087856543
Training recall	0.69937575030012
Training F1 score	0.5966673153145772
Validation accuracy	0.3606829562808538
Validation precision	0.0040748440748440751
Validation recall	0.6636568848758465
Validation F1 score	0.008099954541071453
Test accuracy	0.2321042834512958
Test precision	0.0017403966162806823
Test recall	0.9881889763779528
Test F1 score	0.0034746736437005965

TABLE V
RESULTS FOR STANDARD VECTOR MACHINE KERNAL

Rolx	SVM Kernal
Training accuracy	0.5843337334933973
Training precision	0.6176204667983792
Training recall	0.4428331332533013
Training F1 score	0.5158222980437123
Validation accuracy	0.7323667294100934
Validation precision	0.006841126461211477
Validation recall	0.4650112866817156
Validation F1 score	0.0.013483881525118637
Test accuracy	0.6828293926577026
Test precision	0.0024368928775503344
Test recall	0.5708661417322834
Test F1 score	0.004853069147867996

But comparing the performance on the test set the k nearest neighbour did not perform well to new graph .The precision and recall values are dropped very low here.

VIII. SYSTEM SPECIFICATIONS

We run our code in a system with specifications

- Processor type: Intel i5 10th gen.
- Processor base speed: 1.0GHz.
- RAM: 8GB.

And the time took by the each alogorithm is

- plot graph: 17.32 minutes
- Standard Vector Machines : 14.636 minutes
- Naive Bayes : 16.372 minutes
- Logistic Regression : 16.985 minutes
- K Nearest Neighbours : 15.382 minutes
- Random forest : 15.945 minutes

IX. ANALYSIS

Based on our calculations, we discovered that K -Nearest Neighbors performs best when using Rolx roles. The reason for this is that videos with similar structural roles and genres in the graph are more likely to be related to each other, and a directed edge between the two corresponding nodes is more likely to form.

So we believe that K nearest neighbours performed so well overall because it focuses on finding videos that shared these similarities, and thus the features included were well suited for algorithm results. Because KNN focuses so heavily on similarities and the edges in our YouTube network are based on presence in the related video list, it stands to reason that an algorithm for determining what other nodes are similar to the current node would perform so well on a link prediction task based on related videos.

It is not surprising that the Random Forest algorithm performed well in predicting related videos, and thus links, in the graph, for very similar reasons, as the algorithm also considers nodes based on similarity, except between decision trees rather than the nearest neighbours.

With the notable exception of Random Forest, the performance of almost all of the machine learning algorithms we used showed little change after we incorporated aggregate features into our algorithms, decreasing in some metrics.Because the aggregate features include sums and differences in comment and view counts, we suspect that the performance difference is due to the fact that videos are not always related to one another based on comment and view counts, so incorporating this information is mostly irrelevant.This also explains why KNN performance has decreased, because comment and view count similarities do not fully contribute meaning to video relatedness and thus detract from link prediction accuracy. It also explains why Random Forest performance did not suffer, as the multiple decision trees could simply ignore these additional meaning less features, accounting for only those f that dealt directly with accurate link prediction, such as roles and genres.

Finally according to our results K nearest neighbour algorithm works fine when comparing to remaining algorithms we performed Random forest, logistic regression Navie bayes and standard vector machine

X. CONCLUSION

In conclusion, because our link prediction problem is essentially a problem of determining whether two videos are related, the algorithms that were better suited to determining similarity, such as K Nearest Neighbour and Random Forest were performed better overall.Furthermore, only the features that directly dealt with similarity were useful for our predictions, as videos with similar genres in the YouTube network were more likely to be related to each other, as not only the videos, but the videos to which they were related, were similar in many ways. Overall the accuracy precision recall and f1 scores of our models are not too low in any case we hope that our algorithms were working fine.But we assumed that videos with similar genres were more likely to be related to each other but it can't be true in every case.

XI. Future work

Due to the time and memory constraints we run and check our algorithms in only two time frame. In future we try to run with several time frames with high computational resources and time and also we would try different classification techniques and also try to use community extraction instead of genre.

References

[1] Santos, R.L., Rocha, B.P., Rezende, C.G. and Loureiro, A.A., 2007. Characterizing the YouTube video-sharing community. Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil, Tech. Rep.

[2] Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H., Basu, S., Akoglu, L., Koutra, D., Faloutsos, C. and Li, L., 2012, August. Rolx: structural role extraction and mining in large graphs. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1231-1239).

[3] Al Hasan, M., Chaoji, V., Salem, S. and Zaki, M., 2006, April. Link prediction using supervised learning. In SDM06: workshop on link analysis, counter-terrorism and security (Vol. 30, pp. 798-805).

[4] Valverde-Rebaza, J.C. and Andrade Lopes, A.D., 2014, June. Link prediction in online social networks using group information. In International Conference on Computational Science and Its Applications (pp. 31-45). Springer, Cham.

[5] Lichtenwalter, R.N., Lussier, J.T. and Chawla, N.V., 2010, July. New perspectives and methods in link prediction. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 243-252).

[6] Liben-Nowell, D. and Kleinberg, J., 2007. The link-prediction problem for social networks. Journal of the American society for information science and technology, 58(7), pp.1019-1031.

[7] Li, L., Wang, H., Fang, S., Shan, N. and Chen, X., 2021. A supervised similarity measure for link prediction based on KNN. International Journal of Modern Physics C, 32(09), p.2150112.

[8] Lu, Z., Savas, B., Tang, W. and Dhillon, I.S., 2010, December. Supervised link prediction using multiple sources. In 2010 IEEE international conference on data mining (pp. 923-928). IEEE.