CSE 422: Artificial Intelligence

Final Lab Project Report

Section – 03, Group - 10

“YouTube Video Recommendation System Using K-Nearest Neighbor (KNN) Algorithm”

Himel Dey (Author)
CSE, BRAC University
ID: 14101101
Dhaka, Bangladesh
himelhimu6@gmail.com

Rifat Islam (Author)
CSE, BRAC University
ID: 17341012
Dhaka, Bangladesh
rifat.islam2012@gmail.com

Ahmed Jawad Khan (Author)
CS, BRAC University
Dhaka, Bangladesh
a.j.ay0nz@gmail.com

Omar Bin Parvez (Author)
CSE, BRAC University
ID: 11101071
Dhaka, Bangladesh
omarbinparvez@gmail.com

Abstract

A recommendation system is a sophisticated information filtering system that attempts to predict an item based on the preferences of the user. A music recommender will predict the preferable names of the videos of the users liking. In this paper we describe how our model works to recommend the videos for the particular user. We made this system to make a user find his/her likeable videos with ease. We collected datasets of YouTube videos from www.kaggle.com. We found out the average ratings on a scale of 1-10 and converted it into percentage. The video IDs having percentage above 80% are selected to be worthy of recommendation and used to train our dataset for the KNN algorithm. The algorithm then predicts the preferred videos. We learned that the algorithm predicts videos with neighbor distance of 1, which is a major result. An implication of this system is for a user to find videos with very less complication while also saving their precious time.

Keywords—recommendation; ratings; neighbor; distance; KNN (key words)

1. Introduction

YouTube is the largest platform for uploading videos. Anything can be found in it ranging from live news videos to latest music videos. Currently there are many recommendation systems which use the Collaborative Filtering, Content Based Filtering and Hybrid Filtering. Current knowledge we have about that, many algorithms have been used in measuring user similarity or item similarity in recommender systems. For example, the K-Nearest Neighbor (K-NN) [1]a approach and the Pearson Correlation as first implemented by Allen. [2] There are algorithms like Bayesian classifiers, genetic algorithms and relevance feedbacks. The Bayesian classifiers require a huge training set of data to give accuracy. Genetic algorithms are very slow but the KNN algorithm that falls under the category of relevance feedback requires a small training dataset to work and predict with more accuracy. The gap in the previous work is that, they did not classify to recommend videos in the category of age. We filled in the gap by predicting videos according to age difference.

2. Method

We used the threshold of rating 81 as the adjustable parameter to get our data. The videos which had ratings above 81 were considered to be useful for our predicting model. The materials used were python 3.4, anaconda 64 bit and data from www.kaggle.com. In the Fig1 the working procedure and model have been suggested:

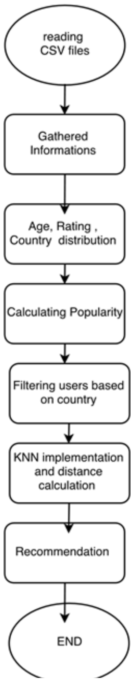

reading CSV files
Gathered Informations
Age, Rating , Country distribution
Calculating Popularity
Filtering users based on country
KNN implementation and distance calculation
Recommendation
END

Figure 1: Flowchart

First of all we take input from the CSV files which contains our dataset. The important columns in the dataset are kept like age, country and ratings. Other columns are discarded. The overall popularity is then calculated. Users are filtered on the basis of countries; the processed data is then given as input to get output by running the KNN algorithm. The outputs are the recommendations. The calculations in the algorithm are carried out using the formula shown in fig2.

$$D(a,b)=\sqrt{\sum_{i=1}^{n}(b_i-a_i)^2}$$

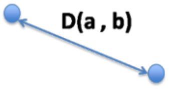


Figure 2: KNN Formulation

This formula calculates the Euclidean distance between two values (difference in ratings in our case) having neighbor distance of 1. Another formula shown in figure 3 is also used to find the cosine distance.

$$sim(A,B)=\cos(\theta)=\frac{A\cdot B}{\|A\|\|B\|}$$

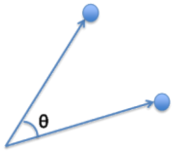


Figure 3: Sine Distance Formula

The technique used is the KNN algorithm. The calibration was done by making the neighbor distance stay at 1. It is because, for values greater than 1, the accuracy decreases and produces inappropriate results which are also our limitation. We assumed the neighbor distance to be 1 to be accurate.

3. Results

In figure 4 the statistics of number of users versus the age is shown. Here count is the number of people that have the ages given in x axis.

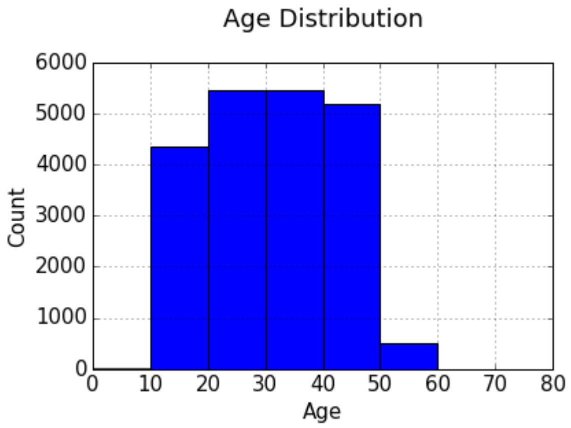


Figure 4: Users versus Age Graph

In figure 4 the number of users is 5200 between the ages of 20 to 30. By using the information in the figure an analysis is carried out of users and their taste according to their age differences. Another figure 5 has the rating distribution.

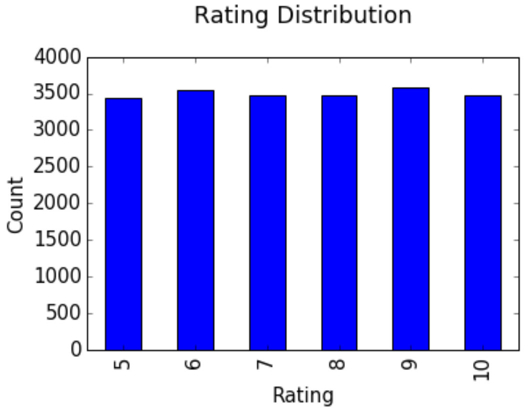
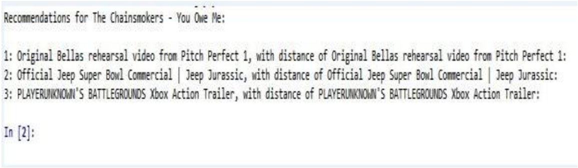


Figure 5: Rating Distribution

The rating distribution is shown in figure 5. It shows number of users that have given a particular rating. For example the number of users that have given rating 5 is 3400. Using this information the overall rating is calculated.

Output

The results are shown in the figure 6. In figure 6 recommendations are given for a user who has seen the video The Chainsmokers. The recommendations are Original Bellas, Official Jeep Super and Playrunknown’s.



```
Recommendations for The Chainsmokers - You Owe Me:

1: Original Bellas rehearsal video from Pitch Perfect 1, with distance of Original Bellas rehearsal video from Pitch Perfect 1:
2: Official Jeep Super Bowl Commercial | Jeep Jurassic, with distance of Official Jeep Super Bowl Commercial | Jeep Jurassic:
3: PLAYERUNKNOWN'S BATTLEGROUNDS Xbox Action Trailer, with distance of PLAYERUNKNOWN'S BATTLEGROUNDS Xbox Action Trailer:

In [2]:
```

Figure 6: Sample Output

4. Discussion

In our paper we have used Pearson’s R correlation coefficient and cosine distance calculation method of K nearest neighbor algorithm, the knowledge of cosine distance has been gathered from [3]. To complete this analysis we have followed a few papers. Susan [4] describes in her paper that she used KNN algorithm to find out the distance which will help his system to recommend Books. According to Susan, virtually everyone has an online experience where a website makes personalized recommendations in hopes of future sales or ongoing traffic. They have use Pearson’s R correlation coefficient to measure the linear correlation between two variables, in their case, the ratings for two books. The pattern we have used from Susan [4], we let our system read the CSV file of user, videos and rating to analysis data. Then we drop the unnecessary columns of this CSV files. We have created relations among user data, rating data and video data. Finally after applying KNN algorithm, our system provides recommendation for a randomly selected video or music which is based on the rating of that music or video which has been given by users. Our final results mean the recommendation for a particular video or music.

Our paper is unique in a sense that we have used country, age and rating classifications which are missing as a whole on other papers we have followed. The agreement between Susan [4] and our paper is basically both of us have used KNN, Pearson’s R correlation coefficient and cosine similarities to calculate distance. After having this experiment we can ensure that one can have a clear recommendation of videos and music which are based on age, country and rating simultaneously. This paper will really help individuals to find relevant videos and music.

5. Conclusion

To conclude, it’s safe to say that the strongest observations we have seen are the recommendations given with the neighbor distance of 1. Most of the time, the recommendations from this system matched other recommendation systems like that of YouTube.

[5] The purpose of this paper was to show that KNN algorithm using supervised learning can be used to build a recommendation system with fair accuracy. However, accuracy drastically decreases with increasing value of neighbor distance. Also the accuracy falls with a larger dataset.

Reference

[1] **B. Sarwar, G. Karypis, J. Konstan and J. Riedl**, “Application of Dimensionality Reduction in Recommender System A Case Study”, *ACM WebKDD 2000 Workshop*, Minneapolis, 2000. Available: http://www.dtic.mil/docs/citations/ADA439541

[2] **R. B. Allen**, “User models: theory, method, and practice”, *International Journal of Man-Machine Studies*, vol. 32, no. 5, pp. 511-543, 1990, ISSN 0020-7373, https://doi.org/10.1016/S0020-7373(05)80032-X. Available: (http://www.sciencedirect.com/science/article/pii/S002073730580032X)

[3] **Understanding KNN Output and functions**: http://cs.carleton.edu/cs_comps/0910/netflixprize/final_results/knn/index.html

[4] **Building A Book Recommender System** – The Basics, kNN and Matrix Factorization https://datascienceplus.com/category/introduction/?tdo_tag=Python

[5] **The Chainsmokers Song: You Owe Me**: https://www.youtube.com/watch?v=_hMQe2U4c6w