International Journal of Computational Science and Engineering.
ISSN 2249-4251 Volume 10, Number 1 (2020), pp. 1-10

http://www.ripublication.com

Netflix Ranking by Combination of K-Nearest Neighbour and Singular Value Decomposition

Ms.Monica. M
(PG Student)
Master of Computer Applications
Kongu Engineering College
Perundurai, Erode, Tamil Nadu, India.
gowrimoni3655@gmail.com

Mrs. Chitra. K
Assistant Professor
Master of Computer Applications
Kongu Engineering College
Perundurai, Erode, Tamil Nadu, India.
k_chitra@kongu.ac.in

Mr. Dinesh
(PG Student)
Master of Computer Applications
Kongu Engineering College
Perundurai, Erode, Tamil Nadu, India.
rdineshpp@gmail.com

Mr. Dinesh Kumar. R
(PG Student)
Master of Computer Applications
Kongu Engineering College
Perundurai, Erode, Tamil Nadu, India.
dineshkumarsoftinfo@gmail.com

Abstract

There are a lot of social media applications are surviving in the world, the reality shows in the television are extending day to day, whereas user's reviews and ratings on television shows are viewed and reviewed by users on public media. So as users are very much interested to make their time on these social media. Peoples are used to making their free time to spend on applications like YouTube, Twitter, Whatsapp, Facebook, Instagram and so on. Like these media, there is a new application that gets trending on the internet is ***Netflix streaming service.*** Netflix is considered to be an Ad-Free viewing television shows. There is so many numbers of reviews, commands and ratings for many channels on Netflix. To find the ***Television Rating Point (TRP)*** for each channel manually is difficult. By using ***K-Nearest Neighbors (KNN)*** for classifying the channel according to view count and ***Singular Value Decomposition (SVD)*** for finding the TRP rating.

Keywords: Component; Netflix streaming service, Television Rating Point (TRP), K-Nearest Neighbors (KNN), Singular Value Decomposition (SVD).

I. INTRODUCTION

Netflix is a global online video streaming service that offers updated movies, TV shows and reality shows where users can watch over the internet at where they are. It has millions of members and subscribers across many countries. Users can subscribe to Netflix in order to watch movies and TV series online. It has become one of the favorite channels for so many people all over the world. Netflix has a package system where the user can select the package according to their needs. The titles available to stream mainly depend on your Netflix package. Each Netflix package has a different library. Netflix is one of the ads-free TV shows. This gives you fresh content always. Where Netflix will keep on update the movies and shows regularly, once you subscribe to Netflix, you are free to stream all movies and TV shows available at no additional cost.

Users can watch Netflix videos either on Smart TVs or on Netflix video streaming apps on their android phones, apple phones and tablets. The benefit of using smartphones or tablets to watch some programs is that you don't need to leave your Netflix movies shows in the middle when you move out of home and you can watch the whole thing anywhere, while on the mobile with the Netflix mobile app. For this, you just need to be subscribed to Netflix and you need the right phone or tablet. all smartphone device owners all can access the Netflix mobile app.

Although experts say that only a handful of Android systems support the Netflix streaming so far. Netflix says it is working fast to bring video streaming to more mobile systems but it is hampered by the lack of standardized streaming playback features across Android phones. The price for Netflix steaming starts at just $7.99 a month and you won't find a better movie-viewing application.

To download Netflix Mobile app, just search for the app in the Android Play Store, Apple App Store or Windows Phone 7 Marketplace. Register and Log in with your Netflix username and password, and instantly stream movies already listed in your Instant Queue. You can also download the app without being a subscriber and try the company's free month-long trial to see if you like the service.

The Netflix subscribers base is continuously growing both in the USA and abroad. The reports say that there is a total of 131.7 million subscribers both inside and outside the US as of September 2018. To be precise, there were millions of Netflix users in the United States and millions of subscribers internationally. Most of them use the regional Netflix service but some people also use VPN service in Australia to get access to more content. Although now there are many competitors in the market Netflix is still the leader in most of the regions. Let's now discuss the top reasons behind the success of Netflix. Netflix company's service achieved an availability rate of 99.97% in 2017, according to reports. Netflix has given users comfort to watch something from the comfort of their home while traveling or while at work. As this company is growing, its services are getting even better.

II. OBJECTIVES OF STUDY

The objective of the study:

1. To find the Netflix channels TRP rank, by calculating the number of users count and rating of the channel.
2. We use a classifier algorithm KNN for classifying the NETFLIX dataset.
3. To calculate and predict the TRP Rank we use SVD (Singular Value Decomposition) algorithm

k-Nearest Neighbor (kNN)

The *k*-nearest neighbors' algorithm is one of the simplest machine learning algorithms. It is simply based on the idea that objects are 'near' each other will also have similar characteristics. Thus if you know the features of one object, We can predict its nearest neighbor object." k-NN is an improvisation over the nearest neighbor technique. It is based on the idea that any new instance can be classified by the majority vote of its 'k' neighbors, - where *k* is a positive integer, usually a small number.

kNN is one of the most simple and supervised machine learning algorithms. It is called Memory-Based Classification as the training examples need to be in the memory at run-time [1]. We can make use with continuous attributes the difference between the attributes is calculated using the Euclidean distance. A major problem when dealing with the Euclidean distance formula is that the frequency of the large value swamps the smaller ones. For example, let us consider a patient who is seeking with heart disease records the cholesterol measure ranges between 100 and 190 while the age measure ranges between 40 and 80. So the cholesterol measure will be higher than the age. To overcome this problem the continuous attributes are normalized so that they have the same influence on the distance measure between instances [2].

Singular Value Decomposition (SVD):

Let A be an m × n matrix. It will be a decomposing a matrix into another matrix[3]:

$$\mathbf{M=U\Sigma V^t}$$

Property[1]: A factorization $M=U\Sigma V^t$ is said Singular Value Decomposition for M, We consider this matrix of the value multiplying matrix **M**, which the **M** gets decomposed into other three different matrices:

$$\mathbf{M}_{m\times n}=\mathbf{U}_{m\times m}\Sigma_{m\times n}\mathbf{V}^t_{n\times n}$$

Where U is an m × m orthogonal matrix, Σ is an m × n pseudo diagonal matrix whose elements nonnegative, and V is an n × n orthogonal matrix. The diagonal elements of

the matrix Σ are called the singular value of M.

where

$$M=U\Sigma V^{t} \longrightarrow 1$$

it will become

$$M^{*}=V\Sigma^{*}U^{*} \longrightarrow 2$$

Because of U and V are real or complex unitary matrix their transpose will be their inverse, whereas **Σ** is a diagonal matrix so their transpose of the diagonal matrix will be the same when we multiple the 2nd equation with 1st equation we get:

$$R=M^{*}M=V\Sigma^{*}\Sigma V^{*} \longrightarrow 3$$

Where in this equation 3 which implies Σ^{2} =Eigen value of R, arrange the Eigenvalue in decreasing order because singular values are in decreasing order and find the Eigenvectors. To calculate SVD, We consider finding eigenvalues and eigenvectors, the eigenvalues and eigenvectors are:

$$AA^{T} \text{ and } A^{T}A.$$

The eigenvectors of $A^{t}A$ to be considered as columns of V, the eigenvectors of AA^{t} to be considered as columns of U. Also, the singular values in Σ are square roots of eigenvalues from AA^{t} or $A^{t}A$.[6] The singular values are the diagonal entries of the Σ matrix and are arranged in descending order. The singular values are always real numbers. If the matrix M is a real matrix, then U and V is also considered as real.

Property[2]: Matrix M is symmetric if and only if there exist a diagonal matrix D and an orthogonal matrix P with[5]

$$M=PDP^{t}$$

Suppose that matrix M is a $m \times n$ matrix. Then, matrix A^*A is a symmetric matrix according to property[1], and by the property[2] it can be obtained a factorization

$$A^t A=PDP^t$$

Where *D* is a diagonal matrix whose entries are the Eigenvalues of A^TA, and P is an orthogonal matrix such that the matrix P is the eigenvector corresponding to the eigenvalues on the diagonal matrix D. According to property [1] if given a matrix *A* then $M=U\Sigma V^t$ is the Singular Value Decomposition for M, where U and V is an orthogonal matrix, and Σ is a pseudo diagonal matrix. Where $\sigma1 > \sigma2 > \sigma3 > ...$ By arranging and normalizing it in decreasing order such that the value was in equal to 1[4]

$$\mathbf{A=U\,\Sigma\,V^T} \text{ and } \mathbf{A^T=V\,\Sigma\,U^T}$$

$$\mathbf{A^TA = V\,\Sigma\,U^TU\,\Sigma\,V^T}$$

$$\mathbf{A^TA = V\,\Sigma^2V^T}$$

$$\mathbf{A^TAV = V\,\Sigma^2}$$

Singular value decomposition brought a matrix on Σ are sorted from the largest to the smallest i.e of decreasing order then the best possible to the matrix A can be taken by the first p rows and columns of matrix Σ. Taking *p* rows and *p* columns of the matrix Σ not only eliminates the zero vector but also delete some singular values that are relatively small when comparing to other values. [8].

III. PROPOSED WORK

A. DATA COLLECTIONS:

The proposed system is tested with the data's which are collected from the kaggle datasets. The kaggle datasets contain plenty of data and information about the various different applications, wherein these kaggle datasets we focus on a dataset based on **NETFLIX STREAMING SERVICES**. Where these records are collected and made use of it.

	title	rating	ratingLevel	ratingDes
2	White Chicks	PG-13	crude and sexua	80
3	Lucky Number Slevin	R	strong violence	100
4	Grey's Anatomy	TV-14	Parents strongly	90
5	Prison Break	TV-14	Parents strongly	90
6	How I Met Your Mother	TV-15	Parental guidan	70
7	Supernatural	TV-16	Parents strongly	90
8	Breaking Bad	TV-17	For mature audi	110
9	The Vampire Diaries	TV-18	Parents strongly	90
10	The Walking Dead	TV-19	For mature audi	110
11	Pretty Little Liars	TV-20	Parents strongly	90
12	Once Upon a Time	TV-21	Parental guidan	70
13	Sherlock	TV-22	Parents strongly	90
14	Death Note	TV-23	Parents strongly	90
15	Naruto	TV-24	Parental guidan	70
16	The Hunter	TV-25	language and br	100

Figure 1: NETFLIX Dataset

To trust the information given by the customer about channels that are available on Netflix, the data consideration is around 1000 data, from this collection 70% is for training and 30% is for testing the data is done. The Netflix serves categories like newly realized movies, web series, Hollywood and boll wood serials, reality shows, games, etc… The rating about channels on Netflix is getting extracted from the dataset and the testing of the proposed system is performed in it.

B. DATA PREPROCESSING:

As the dataset from the kaggle.com, where the data is in the form of numeric values where there are empty values and NA is there So the data which is collected for our work is cleaned and relevant data is provided, from the trained data we can make a process.

C. ALGORITHM FOR KNN

- A positive value k is assigned, along with a new object
- We select the k value in our dataset which is closest to the new object

- We find the most common classification of these entity value
- The classification which we give to the new object

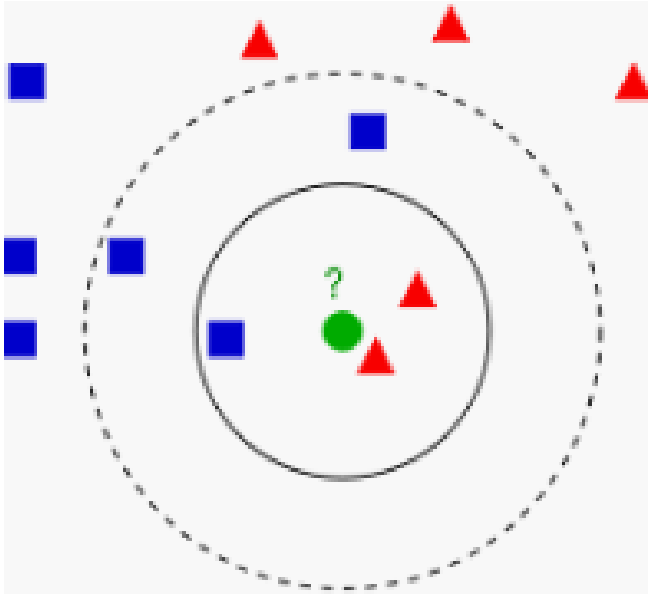

Figure 2: KNN algorithm

D. **FEATURES OF KNN[12]**

- KNN stores the training dataset which it uses as its representation.
- A model which is not necessary for KNN
- KNN makes predictions by calculating the similarity between an input sample and each training data.

IV. EXPERIMENTAL WORK:

In the process of Netflix streaming service with the collected data's we use to perform (KNN) it classify the data according to the algorithm by assigning K as a center point and get classified

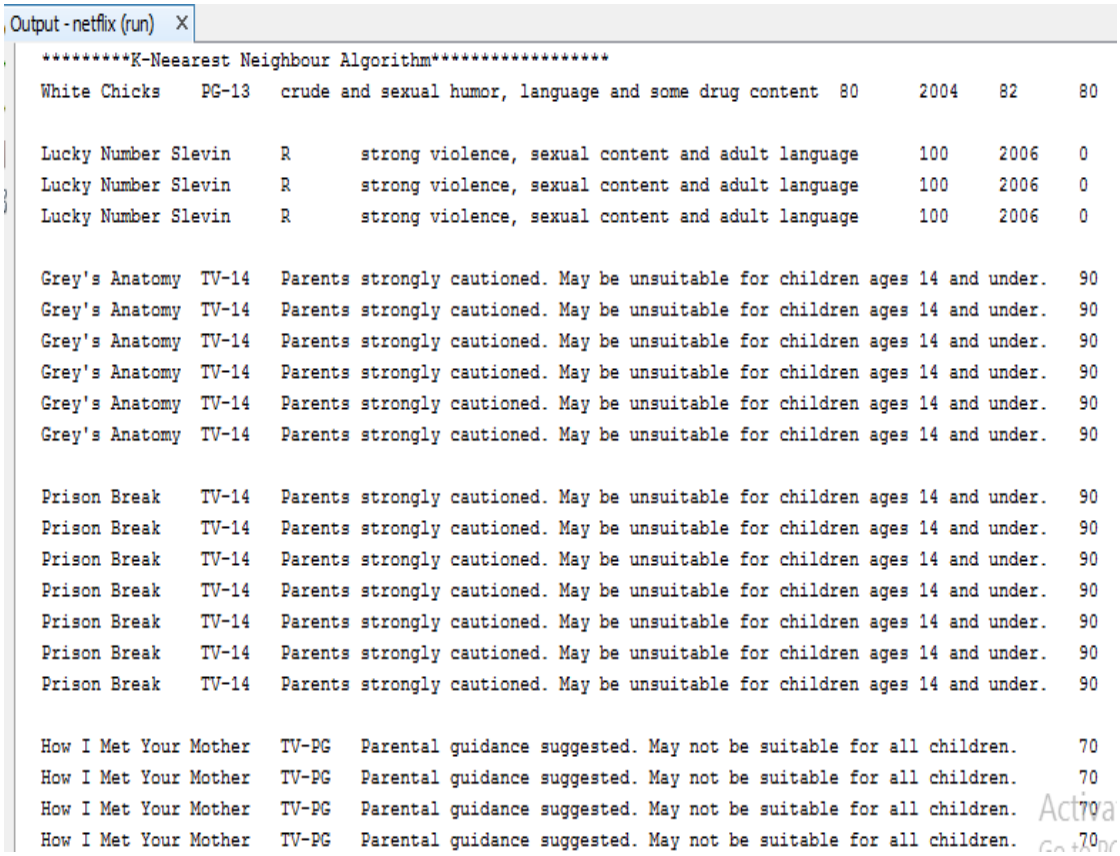

Figure 2: Using KNN ALGORITHM

Singular Value Decomposition (SVD) is a method to reduce the dimension of the data than its real data, whereas because of singular value which are arranged in the order of descending order which is possible to reduce its dimension Latent semantic analysis (LSA)[3][4] is a technique to process the pent-dimensional space of the data relating to latent semantic space [10][11]. LSA is an extension of the vector space model that uses SVD to reduce the dimensioning method. within the LSA there is a truncated SVD algorithm

$$M_{m \times n} = U_{m \times p} \Sigma_{p \times p} V^{t}_{p \times n}$$

which in this process where Σ a singular value is sorted in descending order with the factorization matrix of A, where let we consider P its first row and column to matrix Σ, by selecting the first row and column of Pit not only removes the zero vector also it removes the smaller values.[13][14].

In the process of Netflix streaming service let us consider the matrix M to be an M × N users review which in this process whereas LSA algorithm which decomposes the matrix M into M × p orthogonal matrix U, where p × p will be a pseudo diagonal matrix $\tilde{\Sigma}$, where p × n orthogonal matrix $\tilde{V}\ t$, therefore, P × N reduce the form matrix M, because P can be set much smaller than m

V. RESULT AND CONCLUSION

Whereas with this Netflix dataset by using the SVD algorithm (TRP) Television Rating Point are calculated to know the channels TRP Rate

TRP Rank

Title	TRP	TRP Rank
13 Reasons Why	770.6666666666667	1
Girlboss	662.6666666666666	2
Prison Break	625.3333333333334	3
Shameless (U.S.)	574.0	4
Grace and Frankie	560.0	5
Grey's Anatomy	536.0	6
Pretty Little Liars	532.0	7
The Vampire Diaries	522.0	8
New Girl	522.0	9
Criminal Minds	446.66666666666663	10
Friends	446.66666666666663	11
The Iron Giant	445.99999999999994	12
The Office (U.S.)	445.0	13
Gossip Girl	443.33333333333337	14

Figure 3: Netflix channel rank

The system can withstand huge volumes of customer ratings about channels on Netflix and can provide a (TRP) Television Rating Point. This study and work are based on the data extracted from the kaggle Netflix streaming service dataset.

FUTURE ENHANCEMENT:

This paper could be further studied for improvements. The opinions matter a lot while mining the sentiments from social media, any forums or websites and so on. The proposed system helps to give a result of Television Rating Point(TRP) and can find the users favorite channel. In the future, extend feature-based opinion mining focus on the Users Text review. Also, like to extend the work to find out the strength of various features which help to increase the sentiment text scores

REFERENCES

[1] Alpaydin, E. (1997), Voting over Multiple Condensed Nearest Neighbors. Artificial Intelligence Review, p. 115–132.

[2] Bramer, M., (2007) Principles of data mining: Springer.

[3] Khumaisa Nur'aini1, Ibtisami Najahaty2, Lina Hidayati3, Universitas Indonesia(2015) COMBINATION OF KMEANS AND SVD.

[4] Alter O, Brown PO, Botstein D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, **97**, 10101-6.

[5] Golub, G.H., and Van Loan, C.F. (1989) Matrix Computations,2nded.(Baltimore: Johns Hopkins University Press).

[6] Greenberg, M. (2001) Differential equations & Linear algebra (Upper Saddle River, N.J. : Prentice-Hall).

[7] Strang, G. (1998) Introduction to linear algebra (Wellesley, MA: Wellesley-Cambridge Press).

[8] R. L. Burden, J. D. Faires. *Numerical Analysis*. Brooks/Cole Cengage Learning, 2011.

[9] G. H. Golub, C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, 2013.

[10] S. C. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. "Indexing by latent semantic analysis". *Journal of the American Society of Information Science*, 41(6), pp. 391-407, 1990.

[11] S. T. Dumais. "Latent semantic analysis". *Annual Review of InformationScience and Technology*, 38 (1), pp. 188-230, 2005.

[12] Dasarathy, B. V., "Nearest Neighbor (NN) Norms, NN Pattern Classification Techniques". IEEE Computer Society Press, 1990.