# FOM Hochschule für Oekonomie & Management

University Centre Essen

## Group Project

in the study course Big Data Consulting Project

on the Topic

# Prediction Analysis on Movie Recommendation

by

Agnishwar Das
Abhishek pandve
Sagnik Patra
Tamal Chakraborty

| | |
|---|---|
| First Assessor Prof. | Prof. Dr. Benjamin Niestroj |
| Matriculation number | 547965 |
| | 547963 |
| | 550496 |
| | 557644 |

| | |
|---|---|
| Delivery date | 2021-08-13 |

**Table of Contents**

## List of Figures:

# 1. Introduction

## 1.1 Importance of the topic:

Recommender systems was officially introduced in the mid 1990's to help people choose the product that suits them best from the bucket of options available with them. The main idea which led to their development was that we people often depend upon the opinions of our peers before trying out something new for example let's say buying a phone or a laptop or getting the reviews before going to a movie or a doctor as well. Till date we have vast recommender system developed for various areas using different recommendation approaches but yet there are still a few limitations of recommender system that need's to have worked upon so in this paper we focused on how to minimize the limitations and how to make this movie recommender system a open source with the user's interaction involved. Movie recommendation system provide the user with movie suggestions that are more likely to be watched by user using some means like the user's past behavior or user's profile ,or user's demographic data etc. Recommendation systems are special types of expert systems in the sense that they filter by combining the knowledge of the expert in a given domain with the user's preferences to filter the available information and provide the user with the most relevant information. The user's past history to recommend new items where as collaborative approach uses the preferences of other people with similar tastes for recommending item to the user. Using this model we can predict the genres and the period of the movies that the user prefers based on the user's personal information, so these system collect the information about a user and provide suggestion based on these preferences. Collaborative filtering is a technique used by recommendation systems, predicts and recommends items such as information ,products or services that the user might like. Collaborative filtering based movie recommendations predict a likeness score or a list of top N recommend movies for a given user based on ratings from many users. This paper proposes an algorithm for movie recommendation systems that utilizes the genres of the movies as well as the ratings of the movies to increase rating prediction accuracies. Similarly trust or reputation systems give support to the users by providing information on reliability. So the recommendation systems suggest items to the users by providing ratings to that user by estimating the choice of movies. Basically we have taken the inspiration from Netflix a well known OTT platform where user's get recommended using their watch behavior upon genre's, artist, squeals. So in order to get along we needed a dataset which we took from movie lens. Movie Lens has a stable benchmark dataset of over 1700 movies with 100000 ratings from 1000 users which helps the users to find movies we like and rate the movies to build a custom taste profile. We explored the dataset with expressive search tools.

## 1.2 Problem Statement :

Existing recommendation system are proprietary. Due to lack of an open source recommendation system, cost and time both are being worn.

## 1.3 Research Question :

How can we recommend movies of interest to a user that offers personalized input?

## 1.4 Planned output of the assignment :

To build an open source Movie recommendation system using Machine Learning algorithm.

## 1.5 Background Research :

Movie recommendation system doesn't have any proper open source model to recommend user suitable movie's and the one which is available still now has some limitations which we have came across in our paper. Some limitations which we came across are mentioned under :

**A. Sparsity Problem**

Sparsity issues are one of the major issues facing recommender systems, and data rarity has a significant impact on recommendation quality. In general, system data such as MovieLens is displayed in the form of a user entry matrix ( ) filled with ratings given to the movie and is displayed as no. The number of users and items increase the dimension of the matrix and the rarity evolves. The main reason for data scarcity is that it is generally rare for most users to be able to use most items without a rating. Collaborative filtering goes through this problem because it depends on the evaluation matrix in most cases. Many researchers have tried to alleviate this problem. Still This area needs more research.

**B. Cold Start Problem**

A cold start problem refers to a situation immediately after a new user or entry enters the system. There are 3 kinds of problems with Cold Start which are problems with new users, problems with new items, and problems with the new system. In this case, it is very difficult to give recommendations for new users because very little information is available about user's, and collaborative filtering is not possible because ratings are not generally available for new items. We make Recommendation helpful in case of new items and new users. However, since the content -based method does not rely on information about previous ratings of other users who recommend the item, it can provide recommendations for new items.

**C. Scalability**

Scalability is a characteristic of a system and demonstrates its ability to properly handle increasing amounts of information. It is clear that the data in the guideline holder system is exploding with the massive increase in information over the Internet. Therefore, handling is a big challenge, as demand is constantly increasing. Part of the guideline holder system algorithm handles computation , which increases the number of users and items also increases as well. CF calculation is exponential and expensive which may lead to inaccurate results. The proposed method to address this scalability issue and accelerate the establishment of recommendations is roughly based on the mechanism. Even with the performance improvement, times in most cases will be less accurate.

**D. Over Specialization Problem**

Users are sometimes limited to getting recommendations similar to those already known or defined in the profile, which is a matter of over-specialization. It prevents users from discovering new items and other available options. However, the versatility of the recommendations is a desirable feature of any Recommendation System . After solving the problem using a genetic algorithm, users are presented with a wide set of choices.

Despite of all the limitations we have came across with the vector's which solved these limitations to build a cosine model, then we overcame with the issue of open sourcing this model.

# 2. Theoretical Foundation:

## 2.1. What is recommender system?

Since the past few years, the recommender systems are the most intelligent technologies in our daily live. We use this unavoidable technologies from e-commerce to online advertisement. This system is designed to recommended things to the users based on many factors. Recommender systems are the algorithms that suggest the relevant items to the users such as articles to read, products to buy and movies to watch. Suggesting movies on Netflix, products on Amazon, are the real-world example of the recommender systems. The history of the recommender system is traced back to 60's, that describes the historical cycle of the recommendation system. The recommender system is used in different field that contains structured and unstructured data, text data, commerce data etc. Data scarcity, Scalability, Vulnerability and Diversity are the main challenging factors in the recommendation system.[1]

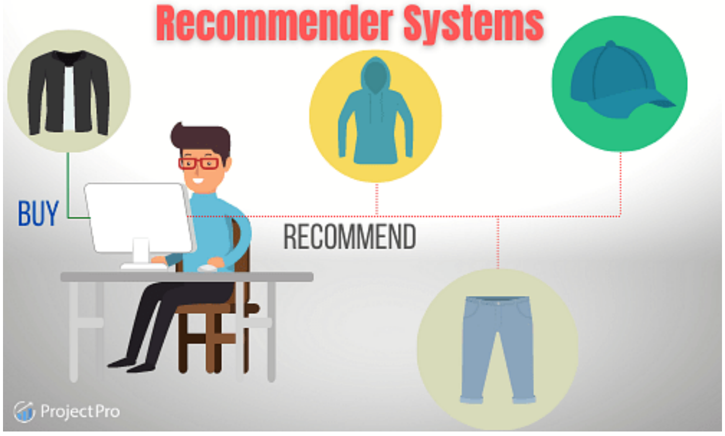

**FIG-1: Recommender System**

[1] Kumar, Manoj & Yadav, D.K. & Singh, Ankur & Kr, Vijay, "A Movie Recommender System: M OVREC ", 2015 International Journal of Computer Applications. 124. 7-11. 10.5120 /ijca2015904111

A buyer bought a Nike jacket from any e-commerce company, and he satisfied with this product. He rated a good score. Other Nike items like pants, caps and hoodies are recommended to the same user to buy these products.

Benefits of the recommender systems:

- Benefits users to find their interesting items
- Personalized content
- Revenue and sales increase
- User satisfaction growth
- Turnover increase
- Help websites to improve user engagement.[2]

## 2.2. Types of recommender systems:

There are mainly four types of recommender systems that is used in media, Entertainment, ecommerce and other industry: Collaborative Filtering, Content based Filtering, Knowledge based Filtering and Hybrid recommender system.

### 2.2.1 Collaborative filtering:

Collaborative filtering uses similarity to suggest which items to recommend to the users. It is based on the assumption that people will like similar kind of products as they liked in the past. First, similar users in taste is classified, then the products rated by these similar users are recommended. Some examples are movie recommendation by Movie lens, Products recommendation by eBay and so on. The most useful collaborative filtering approach is K-nearest Neighbours (KNN) method which identifies the k most similar users and recommended items rated by these neighbors. Other methods are Bayesian networks, Latent Dirichlet allocation that are useful for collaborative filtering. In fig-2, there are two users (one is male and another is female) who watched the two similar movies. This defines that they are

[2] R. Subramaniam, R. Lee and T. Matsuo, "Movie Master: Hybrid Movie Recommendation," 2017 International Conference on Computational Science and Computational Intelligence (CSCI), 2017, pp. 334-339, doi: 10.1109/CSCI.2017.56.

similar kind of users to watch. If another movie is watched by her, then that movie is recommended to him.[3]

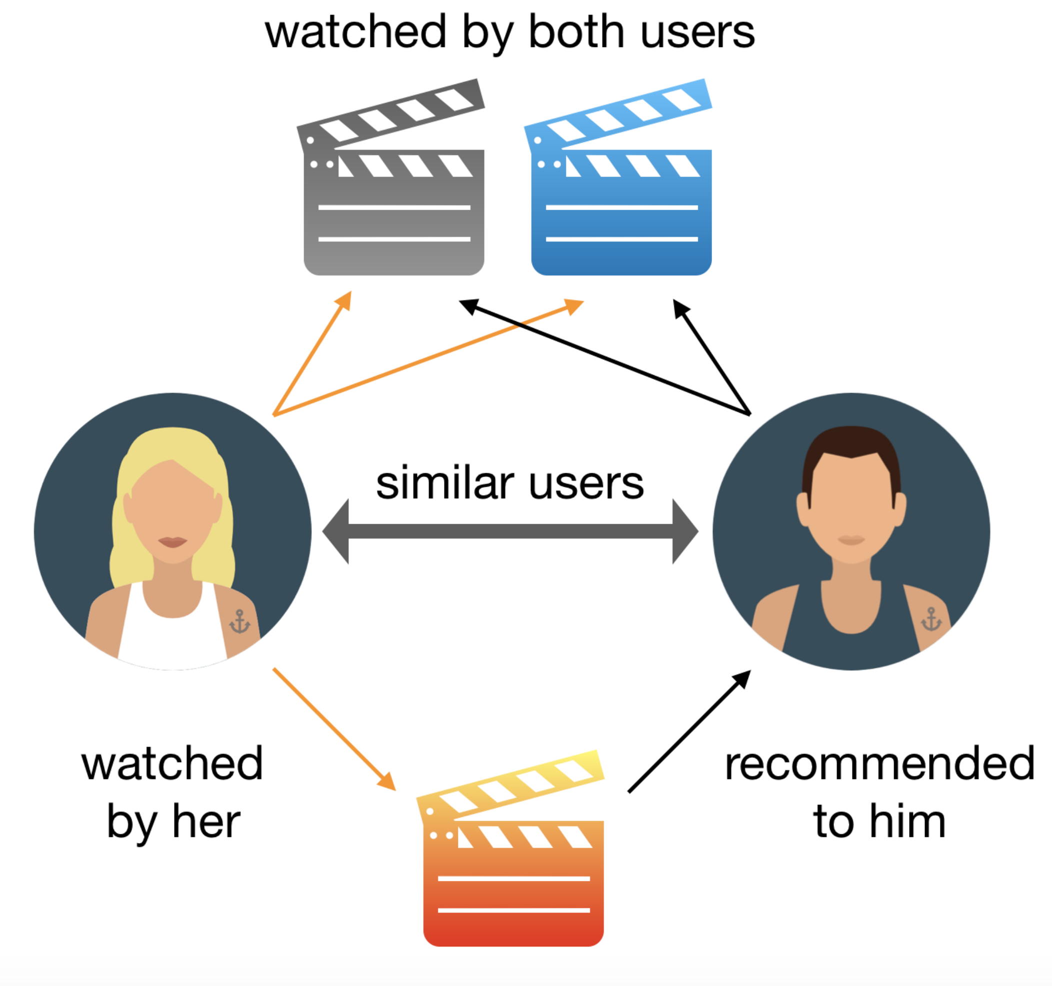

**FIG-2: Collaborative filtering recommendation system**

**2.2.2. Content based filtering:** Content based recommenders will use data predominantly about the items. We need to have an information about the users' preference and choice, so that we can recommend the new items with similar keywords that describes the items by the user. In content-based recommender system, the algorithms used are such that it recommends users preferred items that the user has liked in the past or now. In fig 3, there is a female user.

[3] Soares, M., Viana, P. Tuning metadata for better movie content-based recommendation systems. Multimed Tools Appl **74,** 7015–7036 (2015). https://doi.org/10.1007/s11042-014-1950-1

A movie is watched by this user. Another movie is similar to this movie by Genre, Actor, Directors, Production. Then, this similar movie is recommended to this user.

When the user sign up on any website for the first time, there is no information about the users' taste and choice to build a profile for the users. The recommend system will ask the user about a few movies they like and show them the results which are the most similar to those movies. We are going to build wo types of Content-based recommendation system. That are

- **Plot description-based recommender system-** This recommender model compares the descriptions and watchwords of different movies, then provides recommendations that have the most similar plot description.
- **Metadata-based recommender system-** This recommender system reads a lot of features such as genres, actor, directors, cast, production and crew and provides recommendation that are most similar to these aforesaid features.[4]

[4] Pazzani, Michael J., and Daniel Billsus. "Content-based recommendation systems." In *The adaptive web*, pp. 325-341. Springer, Berlin, Heidelberg, 2007.

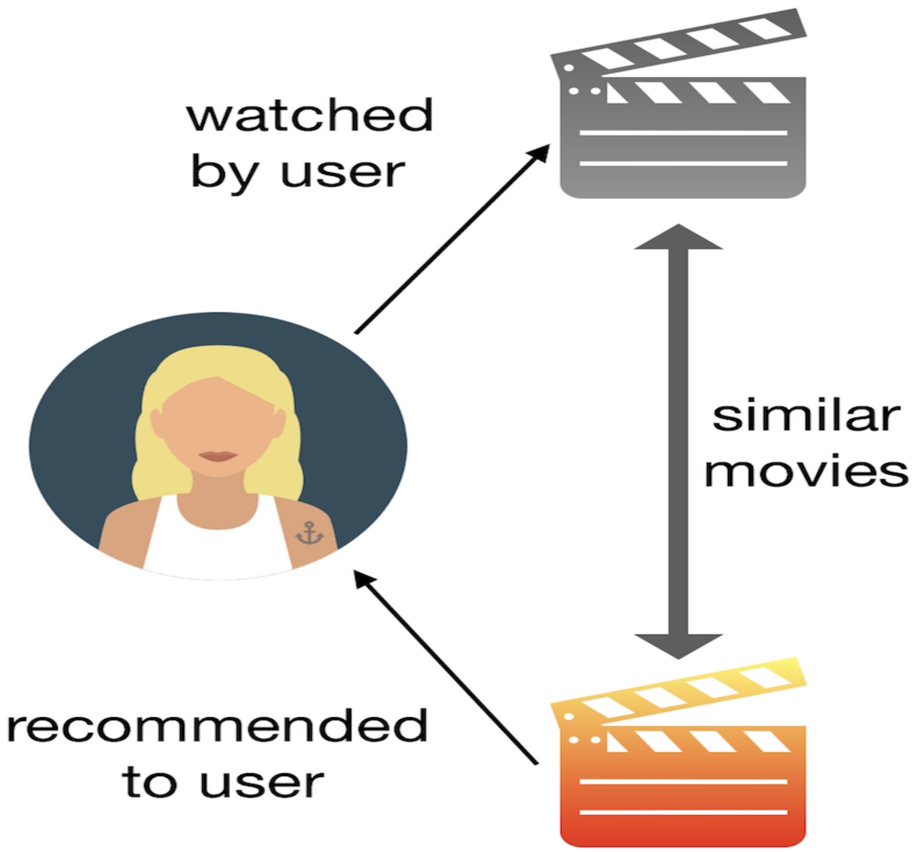

**FIG-3: Content-based filtering**

### 2.2.3. Knowledge-based recommenders:

A knowledge-based recommender system recommends based not on a user's rating history but on past purchasing activity or a specific query made by the user. This recommenders are used for items that are very rarely bought like in real estate business. Real estate is generally a onetime purchase in a lifetime for a family. For this we have to build a system that asks for a specifics and preferences, then recommend the items to the user that fulfill their choices. For example, we could ask a user about their requirements for a property like number of rooms, locality and Budgets and so on. Based on that we could recommend the good property that satisfy them.[5]

[5] Burke, Robin. "Knowledge-based recommender systems." *Encyclopedia of library and information systems* 69, no. Supplement 32 (2000): 175-186.

### 2.2.4. Hybrid recommenders:

Hybrid recommenders are robust systems that combines types of recommender system. Every recommender system has its own advantages and disadvantages. Hybrid recommender system try to ignore the disadvantages of a model against an advantage of another. For example, In Movie Lens when the user sign in for the first time, Movie Lens overcomes the problem of collaborative filters by using a content-based recommender. When the user moderately start watching and rating the movies, series and other items, Netflix brings its collaborative filtering algorithm into play. This is a successful recommender system in nature.[6]

## 2.3. Tf-idf vectorizer:

The basic idea of Content based recommender is to recommend items that are most similar with what user liked in the past. The main mission of content-based recommender system is to calculate the similarity between items. The most famous method to model the item is vector Space Model. The model extracts bag of words from the items and calculate the weight by TF-IDF. The full form of TF-IDF is term Frequency-Inverse Document Frequency. It takes the above-mentioned point into consideration and assign weights ti each words according to the following formula. In document j, for every word I, the following formula applies:

$$W_{i,j} = tf_{i,j} * \log(N/df_i)$$

- $W_{i,j}$ is the weight of word I in document j

[6] Tarus, John K., Zhendong Niu, and Abdallah Yousif. "A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining." *Future Generation Computer Systems* 72 (2017): 37-48.

- $Df_i$ is the number of documents that contain the term i
- N is the total number of documents[7]

## 2.4. Cosine Similarity:

We all know about vectors. They can be 2D,3D OR any other D. In 2-D, the dot product of vectors is equal to the product of the two vectors' magnitudes and the cosine of the angle between the two vectors. The resultant of the dot product of two vectors lie in the same plane of the two vectors. The dot product may be a positive real number, Zero or the negative real number.

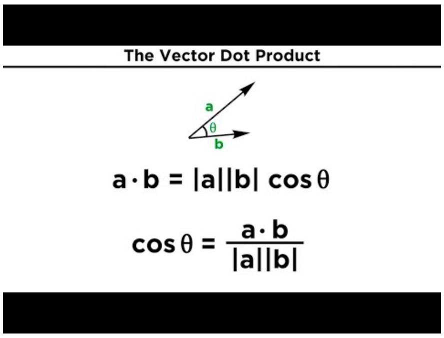

**Fig-4: The Vector Dot Product**

If the two vectors are perpendicular to each other, the dot product will be zero. The dot product is very important when defining the similarity because it is directly related to it. The definition of similarity between two vectors u and v is the ratio between their dot product and the product of their magnitudes.

[7] Sunandana, G., M. Reshma, Y. Pratyusha, Madhuri Kommineni, and Subbarao Gogulamudi. "Movie recommendation system using enhanced content-based filtering algorithm based on user demographic data." In 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1-5. IEEE, 2021.

$$similarity = cos(\theta) = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \sqrt{\sum_{i=1}^{n} v_i^2}}$$

**Fig-5: Similarity**

According to the definition of similarity, if two vectors are identical, the similarity will be equal to 1, and  if the two vectors are orthogonal, it will be 0. The similarity is a number ranged between 0 and 1. It tell us how much the two vectors are similar to each other.

# 3. Research Design:

## 3.1 Literature research

Recommender frameworks have become a significant examination field since the rise of the primary paper on community-oriented separating in the mid-1990s. As a general rule, recommender frameworks are characterized as the supporting frameworks which assist clients with discovering data, items, or administrations (like books, motion pictures, music, advanced items, sites, and TV programs) by amassing and investigating ideas from different clients, which mean audits from different specialists, and client credits.

Notwithstanding, as scholastic explores on recommender frameworks has expanded essentially throughout the most recent ten years, more investigates are needed to be pertinent in reality circumstances. Since the examination field on recommender frameworks is still wide and less developed than other exploration fields. In like manner, the current articles on recommender frameworks should be looked into toward the up-and-coming age of recommender frameworks. Notwithstanding, it would be difficult to restrict the recommender framework's investigations to explicit disciplines, considering the idea of the recommender framework explores. Along these lines, we inspected all articles on recommender frameworks from 37 diaries that were distributed from 2001 to 2010. The 37 diaries are chosen from the main 125 diaries of the MIS Journal Rankings. Additionally, the writing search depended on the descriptors "Recommender framework","Re-tribute framework", "Personalization framework", "Communitarian sifting", and "Substance separating".

## 3.2 Uses of content analysis

Content examination has also been greatly assisted by the development of computer software that identify and match terms and assist in coding. Many of these programmes changed manual content analysis, for the most part as it has become uncomplicated to convert documents to searchable files. Numerous understudies additionally use a portion of the components of content investigation, in spite of the fact that they frequently not understand, then, at that point looking through bibliometric information bases, for example, google researcher and web of science then seeking references and information on subjects. systematic reveal of literature is driven by content analysis and can form the bases of meta/analyses of data from different publication therefore content can rudimentary research tool that can be applied in a range of different settings and since the first stage in research project development as a result of literature identification.

while the lead of content analyses is countless, the main downside is the potential influence of the researcher. Historically, content analysis has regarded as an impartial and unremarkable means of generating a quantitative illustration of the content of different media that is by counting the number of times a word materialize or the amount of space assigned to a particular theory or story in the media However, it has set off increasingly acknowledge that researcher prejudice and inferences with respect to messages, analysis and interpretation are two things which is in acceptance of the research goals or hypotheses. let alone their interpretation, means that the perspective has necessary matter of judgement bias has the data collection with potential.

In tourism content analysis has been used in different conditions. One of the earliest studies as conducted by Richter (1980) who conduct article of CB media research from few daily newspapers in the Philippines. It stated that tourism might have been used as a publicity tool in the Philippines to promote the countries beauty and ignore the fact that martial law had been proclaimed.

In systematic literature review Content analysis can be a functional to identify trends as part of the conduct. For example, to find how various theme will come out and how some themes will work in the content analysis we have to do systematic analysis on content analysis with the help of system analysis and meta-data analysis.

Categorized different publications on some themes like crisis in tourism to detect different themes emerged over time and some themes, such as biodiversity loss, while noted in other literatures part of tourism literature on crisis had not been part of same literature while noted in other literature which Is in different way of objectives in different conditions. With some other procedures like meta-analysis we can adapt and control those literatures and categorized them in part of content analysis of data

But similar study managed on medical tourism and different subcategories of data which is carried by the systematic analysis that found that over half of all literature on the subject had been conducted in the two years prior to the review and that there was remarkable lack of overlap between the tourism and medical. Narrative literature review do not compulsory use an overt structured methodology to work through the applicable deliberate survey is hence a thorough a reproducible exploration technique which is as of late coming to be used to an extraordinary content in the travel industry research. In specific cases the aftereffects of an efficient survey can likewise give the premise of a meta/examination in which different factual techniques are utilized to join the consequences of various investigations that arrangement with the very factors to create a more grounded end than that can be gotten from any of the investigations all alone. Content investigation has been frequently utilized in the travel industry to distinguish various understandings of ideas. led a substance examination of ecotourism definitions. Through their examination they recognized segments that there by and large settled upon, for example, non/destructive or instructive and contrasted their discoveries and the consequences of a quantitative report wherein ecotourism administrators were interrogated concerning different issues encompassing the theme. exceptionally powerful found that according to a worldly point of view, the information showed that the years going from where the most useful as far as the improvement of eco the travel industry definitions, and all the more explicitly some of the most exhaustive of these were authored. Preservation, training, morals, maintainability, effects and neighborhood benefits were the factors that were better addressed in the later definitions showing a changing accentuation in how the term has been conceptualized additional time. attempted a substance investigation of articles on visiting companions and family members the travel industry to likewise distinguish subjects and patterns in the writing on the idea over the long run.

Ethnicity utilized substance investigation to look at the substance of blog sections to recognize positive and negative view of Austria as a travel industry objective, and contrasted these and other examination into Austria objective picture. Site content is additionally a critical core interest. to recognize the picture portrayals of Macau on the Web by dissecting the substance of an assortment of web data sources. attempted a comparative investigation of Dubai/based sites, however including pictures just as text. more explicit topical investigation of sites has embraced by picture utilized a substance examination to take a gander at the food part of sites in as, while the sites of Canadian clinical the travel industry organizations that has promoted global medical care at the end of the day picked the commercial hub. Business and government data, for example, yearly reports and file additionally give a rich wellspring of data that can be contemplated by means of content examination. Chen and Peng the C (2 data from China Ks top G= lodging the board organizations. Coles consolidated a substance investigation of 99 aircrafts documentation with key/data with 66 carriers including three of the four market/driving low passages carriers. more the methodology has used by who taken a gander at online data and yearly reports of voyage/line organizations to recognize the e1tent to which biosecurity measures were used in polar cruising. utilized a substance examination of the

Intergovernmental Board on Environmental Change 3IPCC4 appraisals to produce a period series investigation of the e1tent to which the travel industry had been perceived in the evaluations and corresponding to that themes.

Content examination can be acted in three unique techniques: ordinary, coordinated, and summative. There are three distinct methodologies, they expect to comprehend and investigate the importance of content. They do have explicit contrasts, which is overwhelmingly in the coding framework.

### 3.2.1. Customary Content Analysis

It is called inductive class advancement; this methodology is utilized when the current hypothesis or examination on some random subject is restricted. Here information is utilized as a source to show up at classifications instead of utilizing any of the previous classifications. In this methodology, the explores depend totally on the information to show up at new bits of knowledge. A large portion of the subjective examination strategies utilize this way to deal with contemplate and dissect.

### 3.2.2. Coordinated Content Analysis

In this methodology, research depends on a current hypothesis. This methodology of content examination is utilized to approve or additionally investigate the all existing hypothesis. This strategy should be possible. One way is to begin coding the information dependent on the foreordained codes from the prior approach. Another way is to survey the current codes and allot new codes for the content that couldn't be ordered in the past technique. The guided substance examination plans to center and stretch out the prior hypothesis to decide the key ideas.

### 3.2.3. Summative Content Analysis

In this methodology, the expressions of text will be at first tallied and looked at, trailed by additional translation of the substance. The summative substance investigation targets tracking down the hidden implications of the content or words. In this methodology, the examination begins via looking for a specific book and checking the occasions it shows up and further attempts to comprehend the major setting for the utilization of the words either unequivocal or in its roundabout terms. Summative substance investigation is a nonreactive technique for contemplating the wonder of revenue.

The methodologies of content investigation rely upon the examination purposes that may require diverse exploration plans and different procedures of examination. The examination should take the decision of utilizing a regular or summative or coordinated methodology in the wake of thinking about the reason and the strategies.[8]

## 3.3. Quantitative content analysis:

Quantitative substance investigation is an examination strategy wherein highlights of printed, visual, or aural material are methodologically ordered and recorded with the goal that they can be broke down. Broadly utilized in the field of correspondence, it likewise has utility

[8] Creswell JW. Research design: qualitative, quantitative, and mixed methods approaches. Thousand Oaks: Sage; 2014.

in a scope of different fields. Integral to content investigation is the way toward coding, which includes adhering to a bunch of guidelines about what highlights to search for in a content and afterward making the assigned documentation when that element shows up. Leading an effective substance investigation requires cautious regard for unitizing (sectioning the writings for examination), testing (choosing a fitting assortment of units to dissect), dependability (various specialists making codes reliably), and authenticity (using a coding plan that enough tends to the foreordained miracles). Quantitative substance examination is the conscious and replicable appraisal of pictures of correspondence, which have been existed numeric characteristics as demonstrated by genuine assessment rules, and the assessment of associations including those characteristics using quantifiable procedures, to portray the correspondence, draw enlistments about its significance, or construe from the correspondence Quantitative examination is portrayed as an exact evaluation of wonders by get-together quantifiable information and performing obvious, numerical, or computational strategies. Quantitative evaluation aggregates data from existing and potential clients utilizing examining strategies and passing on the web reviews, online outlines, studies the aftereffects of which can be portrayed as mathematical. After cautious view of these numbers to expect the conceivable predetermination of a thing or association and make changes as necessities be. Quantitative result research is the part organized in the human sciences utilizing the quantifiable frameworks utilized above to amass quantitative information from the evaluation study. In this examination strategy, agents and analysts pass on numerical structures and speculations that relate to the entirety under question. Quantitative appraisal designs are pragmatic, elaborate, and usually, even investigational. The outcomes accomplished from this examination philosophy are legitimate, unquestionable, and reasonable. Information assortment happened utilizing a planned system and drove on more noteworthy models that address the whole people.

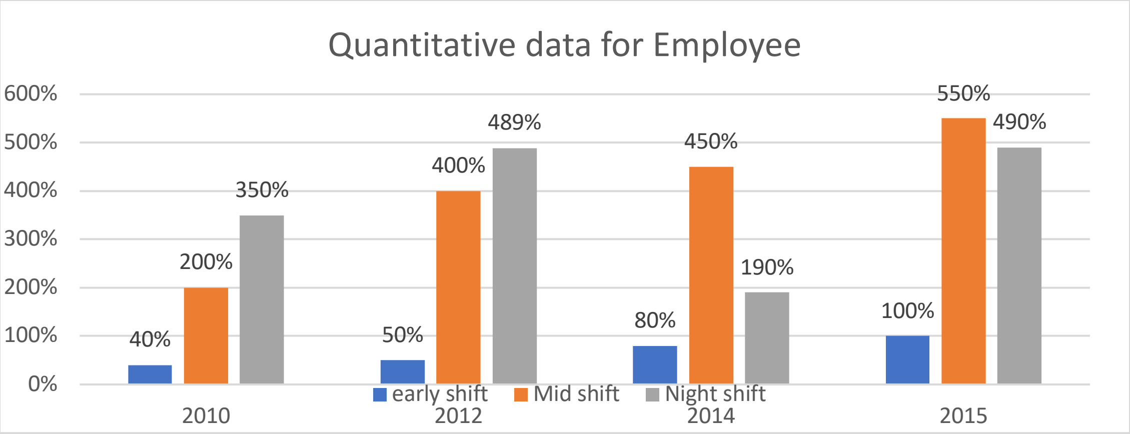

**Fig-6: Qualitative data for employee**

### 3.4. Primary quantitative research methods:

Fundamental quantitative assessment is the most by and large used system for coordinating measurable reviewing. The specific component of fundamental assessment is that the researcher work around social occasion data directly as opposed to depending upon data accumulated from as of late done research. Fundamental quantitative assessment setup can be isolated into three further indisputable tracks, similarly as the cooperation stream. An advantage of utilizing quantitative information is that they can without much of a stretch be summed up by utilizing insights or diagrams, and along these lines can measure up to other exploration discoveries and are not difficult to dissect. In view of this they are additionally simple to recreate. Be that as it may, the subjective strategy utilized permitted the examination to be taken past the limits of a research facility and gives rich knowledge into what distinctive social qualities can mean for mentalities. Quantitative information will in general be less significant than subjective information, as subjective information can give us more insight regarding research. In any case as I would see it, quantitative information is the best strategy for information assortment, as it is more logical and can be utilized in insights which is vital in research.Expressive Design This kind of quantitative exploration depicts the current status of a variable or a marvel. The expert doesn't begin with an idea notwithstanding he would like to cultivate one exclusively after the data is assembled. Data grouping is for the most part observational in nature. Correlational arrangement researches the association between factors using authentic assessments. In any case, it doesn't simply look for a conditions and consistent outcomes in the circumstance, which is the explanation it can moreover be considered as observational to the extent data arrangement. Quantitative substance examination is an exploration strategy wherein highlights of text based, visual, or aural material are methodologically arranged and recorded so they can be dissected. Generally utilized in the field of correspondence, it likewise has utility in a scope of different fields. Integral to content examination is the way toward coding, which includes adhering to a bunch of guidelines about what highlights to search for in a content and afterward making the assigned documentation when that element shows up. Directing a fruitful substance investigation requires cautious regard for unitizing (dividing the writings for examination), testing (choosing a suitable assortment of units to break down), dependability (various specialists making codes reliably), and legitimacy (utilizing a coding plan that sufficiently addresses the predefined marvels).[9] To develop the frontend for our model we have deployed Node.js which is primarily used for non blocking or event driven servers due to its single threaded nature. It is mainly used for traditional web sites and back end API services as well as we have also used PostgreSQL and React.JS, so firstly the json file was created where the hosted site name is displayed using icons, size and image type so for creation of the image a logo was created which then was engraved into .png format with src command. The json file is contributed with a standalone display mentioning the start url. Theme color with the color number according to binary has been taken into account with the background color as well. The header file was also inputted as a logo in order to display. The coding is encoded with UTF-8 encoder and deployed with an web app on the priority basis. Event Listener is also added into function event to create a event log. A search app bar is created in order to take the user input which is routed in the localhost. The route path is then diverted to user query where Snackbar command comes into play with the open message, save login and elevation described as 6 with the variant filled with success, then the path is diverted into search query followed by movie query with class name as classes which will again redirect if not found then switch to route. Review is then taken into account with post review or typography with rating button value with classes rating which is disabled upon login accompanied by button on click. Header classes is inserted into toolbar where the input base is placed upon placeholder as “Search Movie”, root classes is describes as input root and then input classes and on change set value to e.target.value, onekeydown search it simply redirects to profile link into the avatar with classes name defined as user, button variant as contained to login page. Home page is redirected using class name container with circular progress style displayed movie and block margin, cardnedia is classified as movie inherit movie_id with text decoration, rating given as read only value as 5nor greater than that. Box command is played next next to profile avatar to create the box in order for a search bar input. Movie name or title displayed to be inputted by the user for the recommendation system. It is displayed as text with the movie cast.map where login is mapped using user id.

[9] Babbie E. The practice of social research. 14th ed. Belmont: Cengage Learning; 2016.Google Scholar

# 4. Source Correction:

**Dataset:**

The data is collected from Movie Lens and contents the following variations:

There are two datasets, the first data set contents:

- Movie ID
- Cast
- Crew

The second dataset contents the following features:

- Budget
- Genre
- Homepage
- ID
- Keywords
- Original language
- Original Title
- Overview
- Popularity
- Production Companies
- Production Countries
- Release Date
- Revenue
- Runtime
- Status
- Tagline
- Title
- Average ratings of the movie received
- The count of the votes received

We have built two models and have trained and tested them on the same data set.
However, one is traditional (The demographic Filtering Model) and the other is a machine learning model. The objectives are same for both; however, the results are surprising.

# 5. Comparison between the Machine learning and the Traditional Model:

**Demographic Filtering model:**

Demographic filtering is typical to media and product recommendation system. The legacy Amazon website had a content-based filtering system. In fact, many start-ups still use the same method.

The Demographic Filtering model uses the following pre-requisites to get started:

- A metric score or rate of the movie
- To Calculate the score of every movie
- Recommend the best rated movie after sorting the scores

The Demographic Filtering Model uses the following equation:

**Weighted Rating = ((v/v+m)*R) + ((m/v+m)*C))**

Where:

- V as in the number of votes for the movie
- M as in the minimum votes required in the chart to be listed
- R stands for the average rating of the movie
- C as in the mean of the votes across the report

From the pre- processed data, we already have the vote count and the vote average.
We calculate C as:

C = df2[‘vote_average’].mean()

Then we display C’s output.

Cascading the same process, we can generate the output of m:

M – df2[‘vote_count’].quatile(0.9)

We can also filter out the movies which qualifies for the chart:

q_movies = df2.copy().loc[df2[‘vote_count’] >= m]

We will apply weighted_rating() and define the new feature score, we will apply this function to our data frame after calculating the value.

The score feature is derived from sorting the data frame.

**Machine Learning Model:**

The Demographic Recommendation system works, although a very basic approach. There is nothing wrong with a basic recommendation system, however, it does not use the resources to its full extend.

To tackle this issue, we came up with our own machine learning model built from scratch. The advantages of using this machine learning model over a traditional recommendation system is:

- Uses more meta data
- Provides comparatively better recommendation

- Can be integrated with any tabulated data
- Industry agnostic
- Uses top of the line technology (all open-sourced)
- Upgradeable to fit more data
- Ground up built for open sourcing
- Customizable

Things change when we use our own machine Learning model. It uses a cosign model based on “Plot Description Based Recommender” which is data agnostic. It can transform any tabular data after it is preprocessed into Numpy.

We compare pairwise similarity score of all movies based on plot descriptions and recommend movies based on similarity scores. The cosign model will feed in more data than any other model. Thus, the algorithm will have more metadata to work with. This was the same approach used by Apple Music’s station feature.

For the model to work we have to compute Term Frequency of a word in the document. Which is given as (term instance/total instance). Inverse Document Frequency contents the term log (number of documents/documents with term). The Importance of each variable which appears on the process is equal to TF * IDF.

But we do not have to go through all this since Scikit learn gives us a built in tfidVectorizer that can produce TF-IDF matrix with a very little lines of code.

Let us take a detailed look:

```
#Import TfIdfVectorizer from scikit-learn
from sklearn.feature_extraction.text import TfidfVectorizer

#Define a TF-IDF Vectorizer Object. Remove all english stop words such as 'the', 'a'
tfidf = TfidfVectorizer(stop_words='english')

#Replace NaN with an empty string
df2['overview'] = df2['overview'].fillna('')

#Construct the required TF-IDF matrix by fitting and transforming the data
tfidf_matrix = tfidf.fit_transform(df2['overview'])

#Output the shape of tfidf_matrix
tfidf_matrix.shape
```

The output seems to be: **(4803, 20978)**

Which means 20978 words were used to describe 4803 movies. This is not even the full dataset. This test was done to have an idea of the data we are using.

We are calculating and comparing three similarity scores:

- Euclidean
- Pearson
- Cosine Similarity Score

We are focusing on the Cosine Similarity Score for this report; however, we should keep in mind, there is no study which states the best score out of these three.

Mathematically the cosine Similarity is defined as follows:

$$\textbf{Similarity} = \cos(\Theta) = (A * B)/ (||A|| \, ||B||) = (\Sigma A \, . \, B)/ (\sqrt{\Sigma A^2} \sqrt{\Sigma B^2})$$

We can take a look on the background process of measuring the data:

```
/opt/conda/lib/python3.6/site-packages/surprise/evaluate.py:66: UserWarning: The evaluate() method
is deprecated. Please use model_selection.cross_validate() instead.
'model_selection.cross_validate() instead.', UserWarning)
/opt/conda/lib/python3.6/site-packages/surprise/dataset.py:193: UserWarning: Using data.split() or
using load_from_folds() without using a CV iterator is now deprecated.
UserWarning)
Evaluating RMSE, MAE of algorithm SVD.

------------
Fold 1
RMSE: 0.9053
MAE:  0.6951
------------
Fold 2
RMSE: 0.8933
MAE:  0.6896
------------
Fold 3
RMSE: 0.8991
MAE:  0.6906
------------
Fold 4
RMSE: 0.8924
MAE:  0.6882
------------
Fold 5
RMSE: 0.8921
MAE:  0.6893
------------
------------
Mean RMSE: 0.8964
Mean MAE : 0.6906
------------
------------
Out[33]:
CaseInsensitiveDefaultDict(list,
{'rmse': [0.9053249354442916,
0.8933118114597819,
0.8991051561187587,
0.8924047744398809,
0.8920738097595919],
'mae': [0.6951043276181835,
0.6896295679940669,
0.6905941154058887,
0.6881763472727346,
0.6893100283898707]})
```

We choose to use sklearn’s linear_kernel() instead of cosign similarities since it is faster.

These are the following steps to define our cosine similarity recommendation system functions:

- Get the movie index
- Generate the index of the movie when provided with the title
- Get the cosine similarity score
- Sort the list of tuples based on the cosine similarity score
- Return to the indices corresponding to the title

In Jupyter notebook, it looks like the following:

```
# Function that takes in movie title as input and outputs most similar movies
def get_recommendations(title, cosine_sim=cosine_sim):
# Get the index of the movie that matches the title
idx = indices[title]

# Get the pairwsie similarity scores of all movies with that movie
sim_scores = list(enumerate(cosine_sim[idx]))

# Sort the movies based on the similarity scores
sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)

# Get the scores of the 10 most similar movies
sim_scores = sim_scores[1:11]

# Get the movie indices
movie_indices = [i[0] for i in sim_scores]

# Return the top 10 most similar movies
return df2['title'].iloc[movie_indices]
```

## 6. Projection of Results:

***<u>For the Demographic(traditional) Filtering Model:</u>***

get_recommendation('Argo')

The output:

```
4629                 Some Guy Who Kills People
    3409                      Criminal Activities
     4438                            Circumstance
      2001                                The Crew
941     13 Hours: The Secret Soldiers of Benghazi
     2241                            Passchendaele
      3679                            Take Shelter
   2685                    Exorcist II: The Heretic
   4720                      The Birth of a Nation
      3328                              Persepolis
               Name: title, dtype: object
```

This will return with the ten best possible movies. A traditional method, still works well.

***<u>For our Machine Learning Model, we need to call the cosine model:</u>***

get_recommendations('Argo', cosine_sim2)

The output:

```
  1328              The Town
611     The Sum of All Fears
   693             Gone Girl
 1660          Runner Runner
    871                Gigli
 1346         Reindeer Games
2975       Good Will Hunting
   991             Fair Game
  1187       Bridge of Spies
    1308              Enough
  Name: title, dtype: object
```

If we look closely, the movie input is same, which in this case is the movie 'Argo', but the results generated are different. This is because, the cosine model is working with more metadata and has more parameters.

In fact, the model learns from the users searches and looks for patterns to generate more results, all sandboxed into the system. The model does collect other user information.

The model accuracy is over 91% surpassing and satisfying the industry standard. Although falling short in compare to Netflix's algorithm.

## 7. Conclusion:

We can safely conclude that our model delivers the product with exceptional results. With 91% accuracy, it surpasses the industry standard.

The report also fills the gap in the market of open-sourced recommendation system which can be adapted to any environment and is tangible. In addition to that, the website created is unique in nature and does not collect any user's data, thus maintaining user's privacy. The website is familiar to use and as tested, users have found it to be very straight forward without losing most of the important features like sorting and maintained personal history.

The combination of our powerful machine learning algorithm and esthetically designed website makes this venture a perfect pair.

## 8. Future Scope of the project:

Even though this paper focuses on Movie recommendation system, we have also kept in mind the reusability of the model. It is a hectic task to create a Machine Learning model for each mutually exclusive task, hence we need more open-sourced industry agnostic machine learning models.

The primary objective of making the cosine similarity model was keeping the upgradability in mind. The model is created in such a way, that it can be fit into any tabular data.

As we analyze the industry, most of the data needs preprocessing, hence our model can accommodate the data and run the cosine similarity without or a little difficulty.

Another main objective of creating the model is, open sourcing it. By open sourcing the model we create unlimited possibilities.

In fact, we have tested this model with:

- Traffic simulation
- Color pallet recognition

To our surprise, both outcomes have impressive results with little to no code change or multiple trained iterations. This just shows how powerful our Machine Learning model can get. We have hence open-sourced it on git-hub and will also publish the code on Kaggle.

## 9. Bibliography:

- Wang, Zan, Xue Yu, Nan Feng, and Zhenhua Wang. "An improved collaborative movie recommendation system using computational intelligence." Journal of Visual Languages & Computing 25, no. 6 (2014): 667-675.
- Reddy, S. R. S., Sravani Nalluri, Subramanyam Kunisetti, S. Ashok, and B. Venkatesh. "Content-based movie recommendation system using genre correlation." In Smart Intelligent Computing and Applications, pp. 391-397. Springer, Singapore, 2019.
- Basu, Chumki, Haym Hirsh, and William Cohen. "Recommendation as classification: Using social and content-based information in recommendation." In Aaai/iaai, pp. 714-720. 1998.
- Khadse, Vivek P., Syed Muzamil Basha, N. Iyengar, and R. Caytiles. "Recommendation Engine for Predicting Best Rated Movies." International Journal of Advanced Science and Technology 110 (2018): 65-76.
- M. M. Reddy, R. S. Kanmani and B. Surendiran, "Analysis of Movie Recommendation Systems; with and without considering the low rated movies," *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1-4, doi: 10.1109/ic-ETITE47903.2020.453.
- Lee, S. J., T. R. Jeon, G. D. Back, and S. S. Kim. "A Movie Rating Prediction System Based on Personal Propensity Analysis." In *Proc. of kIIS Fall Conference 2008*, vol. 18, no. 2, pp. 203-206. 2008.

- Lee, S. J., T. R. Jeon, G. D. Back, and S. S. Kim. "A Movie Rating Prediction System Based on Personal Propensity Analysis." In *Proc. of kIIS Fall Conference 2008*, vol. 18, no. 2, pp. 203-206. 2008.

- Lekakos, George, and Petros Caravelas. "A hybrid approach for movie recommendation." *Multimedia tools and applications* 36, no. 1 (2008): 55-70.
- Armstrong JS. Significance tests harm progress in forecasting. Int J Forecast. 2007;23(2):321–7.
- Babbie E. The practice of social research. 14th ed. Belmont: Cengage Learning; 2016.Google Scholar
- Creswell JW. Research design: qualitative, quantitative, and mixed methods approaches. Thousand Oaks: Sage; 2014.

Declaration in lieu of oath:

I hereby declare that I produced the submitted paper with no assistance from any other party and without the use of any unauthorized aids and, in particular, that I have marked as quotations all passages which are reproduced verbatim or near-verbatim from publications. Also, I declare that the submitted print version of this thesis is identical with its digital version. Further, I declare that this thesis has never been sub-mitted be-fore to any examination board in either its present form or in any other similar version. I herewith agree/disagree that this thesis may be published. I herewith consent that this thesis may be uploaded to the server of external contractors for the purpose of submit-ting it to the contractors' plagiarism detection systems. Uploading this thesis for the purpose of submitting it to plagiarism detection systems is not a form of publication.

Essen, July 23, 2021

___________

(Location, Date)

Agnishwar Das

Sagnik Patra

(Signature)