[[1]](#footnote-1)

**Movies Rating Prediction Based upon Crew Experience**

**Chaitra, Essa, Majeed, Saajan, Sai Sree**

*Abstract*— In this project, the main idea is to find a way of analyzing the rating of a movie based upon the experience of the crew working on the movie it. Movie rating refers to the ranking of a movie in regards to its quality. There are numerous ways of ranking the movies. One of the just happens to be by using the crewmembers working on the movie. Experience of the crewmembers goes well in hand in with the quality because the more experience the crewmembers are the better quality of the movie. Thus, theoretically, the best movies usually have the best crewmembers who in turn have a lot of experience. Looking at the experience and the quality of the crewmembers should assist in finding a metric for approximating the rating as close to the overall viewership rating. This metric approach will be advantageous because the producers of the movie will know what to expect in terms of returns, income, viewership rating and the ability of the movie to become an instant classic.

# **INTRODUCTION**

Choosing a movie to watch is tough nowadays because of the bewildering number of movies that exist in today’s society. Movies are quite common with most people having to look at the quality and theme of the movie in order to make a decision of whether to watch it or not. Therefore, the people making the movies have to take in to consideration all that is necessary in order to ensure the best quality possible. In this project, the interest is with the movies that are yet to receive a rating and are rather new. New movies should receive a rating based on the similarity with other movies with special consideration on the content of the movie characteristics.

The importance of movie rating is that the producers of the movie get to understand exactly what was the best parts of the movie and what were the worst parts of the movie. This way, the producers of the movies are in a position to improve on the quality of the movies and in future, entertainment based on the movies becomes better compared to when there is no rating system. In addition, the rating information will help the producers make the best kinds of movies while considering all the aspects that the viewership demands from the movies. It is also worth noting that the movie viewership will grow because of the better ratings, improved quality, and the desire for viewers to get more movies will increase thereby initiating a growth in the movie entertainment industry.

# **Literature review and proposed plan**

This project deals with the definition of a metric system that will be used to measure the success of the particular movie. The two considering factors will be the income rate of the movie and the ratings from the viewers of the movie. Income rate refers to the amount of money that viewership of the movie manages to raise and comparing it to the budget of the movie production does give a good indication of the success of the movie. This usually puts the movie in the sense of a good or service produced and the profit it generates given the income.

The quality success of the movie takes in to account the overall outreach of the company based on the number of viewers and their geographical distribution. Such is usually very hard to account for unless there is a way of tracking the viewership at cinemas, the online viewership, and the sales of the original copy and the geographical distribution of the viewership. Theoretically, the most successful movies will get the best ratings.

To quantify the ratings of the movies using the crewmembers who worked on it, the consideration will be on the ratings and the categories of the movies. That metric will be addressed in a vector that demonstrates the experience or accomplishment of people in specific sorts of films: performance, movement, spoof, et cetera. In the vector of experience, we need to consider institutionalizing by the number of films that each on-screen character/boss or diverse has performed. Without further ado, given the headings of the all inclusive community that are behind a movie, we require figure a rating from a particular customer. Here. The consideration takes a gander at the vector of slants of the customer – by characterization – with a gathered vector of experience by class of the all inclusive community behind that film. The film vector should consider the part of each person - a head of a film and his then again her experience should be more vital than the experience of a cinematographer.

It is normal to want to know the impact of the crewmembers on the movie. This is because the crewmembers are actually the persons who will work on the movies and make it what it will be in terms of final product. Thus understanding the impact of the crewmembers is advantageous to the producers who will conduct a recruitment of the crewmembers beforehand. Theoretically, the experience of the crewmembers should enhance the overall quality of the movie because of the fact that it is the crewmembers who form the bulk of the production team and they will be primarily responsible for the execution of the plan to produce the movie.

# **Methodology**

**Vector analysis (Cosine distance vs rating):-**

Primary intention of this analysis is to decompose user’s preferences and movies into a set of latent factors (genres), and verify if it is possible to correlate the rating that the user gave to a movie with the movie itself. We are expecting something really simple: A shorter distance between the vector of the user and the vector of the movie should imply a better rating. Vector analysis required pre­processing of the ratings matrix before it could be used in a matrix function. The preprocessing functions collected data summary statistics across rows and columns of the matrix.

In this subject, we are trying to create vectors of experience per film personnel per movie and used the average rating for the weights of each movie in a specific genre. We are planning to tally up the weights per genre and then normalize the weighted experience vector by dividing it by the count of the most popular genre keeping the rating of the that genre constant.

For an example, in the figures below let us consider 2 directors and their weighted experience vectors, we use the same method to calculate the vector for each movie. We will only consider the experience of the crew previous to the year of release of the movie.



Figure 1



Figure 2: User Vector

We are interested in the relation between the Cosine distance and the ratings. To get the cosine distance we measured the distance between two vectors, that is the user vector and the movie vector of an inner product space that measures the cosine of the angle between them. We imported the scipy. spatial. distance library that defines the cosine distance as where u.v is the dot product between the two vectors

u = user vector, and v = movie vector

Initial attempts with vectors gave us confusing results but after the review of the data we got positive insights. We realized that:

a. Introduction of actors without any filter would introduce irrelevant information into the model.

b. Intuitively, the work of directors, writers and producers is 100% involved in the movie unlike actors and actresses, who have a different participation based on the importance of the character.

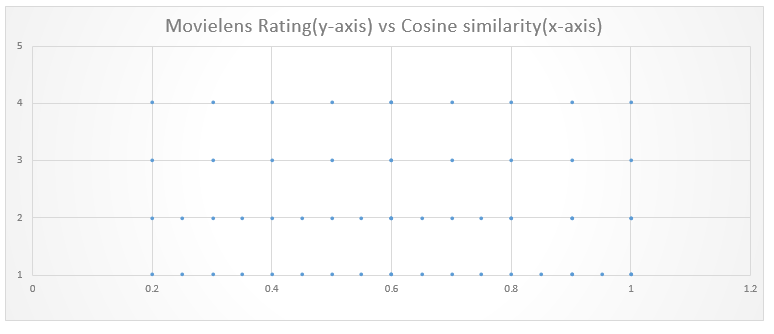
**Results**:

Initial attempts with vectors gave us confusing results but after the review of the data

we got positive insights. We realized that:

* Introduction of actors without any filter would introduce irrelevant information into the model.
* Intuitively, the work of directors, writers and producers is 100% involved in the movie unlike actors and actresses, who have a

different participation based on the importance of the character.



the average distance vs. the rating shows:

Vertical axis:- number of ratings

Horizontal axis: cosine similarity

Vertical axis:- number of ratings

Horizontal axis: cosine similarity

**Naïve Bayes Classifier (Prediction): -**

Statistical classifier that predicts class membership probabilities such as the probability that a given tuple belongs to a particular class. It is a robust classifier with respect to missing feature values, which make it well suited to the task of rating prediction. To apply the naive Bayes classifier to rating prediction, we independently teach one classifier for each user u. We train the classifier for u using all movies rated by u, excepting one target movie m. Then we can predict the rating of m using the rests of the movies. To learn the naive Bayes rating predictor, we estimate P (Director is in Movie) and P (rating=1 | Director is in Movie). We smooth the probabilities by adding prior counts to avoid zero probabilities.

P(rating=1|DinMovie) =

**Results:**

20M ratings.

100850 ratings not considered for not matching (IMDb & Movielens)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Bayes OK 32828

Bayes NoOK 56580

Total 89408

Number of Users 800

Average prediction success per user: 21%

**Methodology**: - With crew experience we can get the parameters for the classifiers. Classifier will return a probability for each rating. Rating with the maximum probability will be answer for classifier.

# **Dataset**

Dataset’s willing to use would be: - Movielens and IMDB.

**MovieLens (ratings): -**

The GroupLens Research provides a number of collections of movie ratings data collected from users of MovieLens. The data provide movie ratings, movie metadata like genre, year and demographic data about the users such as the occupation, gender, zip code, and age.

The MovieLens data sets are divided into three smaller datasets: a. MovieLens latest consists of latest ratings by 1000 users on 1700 movies.

b. MovieLens 20M consists of 10 million ratings applied to 10,000 movies by 72,000 users.

c. 1M MovieLens data set.

We are not using the third one which is 1M MovieLens. We are trying to spread the data sets across 3 tables: ratings, user information, and movie information.

**IMDb**: -

The IMDb (Internet Movie Database) is an online database of information related to movies, television shows, actors, production crew and video games. Information is stored in a lists format that is very descriptive of the film personnel and the individual provides a number of collections of movie ratings data collected from users of MovieLens. filmographies. The lists contained included but not limited to the directorslist, producers list, actors list and so on.

**Data Challenges:**

Trimming of unwanted data from dataset was bit difficult initially but while processing the project we got clear ideas of what all sub data are required. Inconvenience in parsing and stacking the data like IMDb data dealt with into little knots. For example, boss rundown, on-screen characters list, and creators rundown. Various naming benchmarks like movies and sorts. For example, The Magnificent Seven Vs Splendid Seven, The. Seven versus Se7ven. Data Size is large at 4 gigabytes, IMDb records generally are at least one gigabyte, and the set up should be to portray businesses and deal with them in python shows.

The datasets for the ratings also do not consider the theme of the movie and the execution of the story telling. Movie consumers are attracted by more than just the actors, the crewmembers and the production company. Movie consumers demand a good story and the best kinds of stories are usually the ones that the users end up being emotionally invested in the movie. All these factors coalesce to ensure that a given movie gets its overall quality rating from the viewers rather than just crewmembers’ experience.

# **Conclusion & future-work**

## With the Cosine isolate as opposed to rating vector examination and classifiers strategy, we can fulfill the target in sensible measure of time. This means that it is possible to get a metric that can approximate the movie rating but essentially, the process is going to be fairly hard given the huge amount of data and the challenges of datasets. In addition, movie production can be very unique depending on the producer, the budget, the story telling ability of the actors, the experience of the crewmembers and the duration of the production phase and the theme of the movie.

**Future work:**

* Improving the vectors: Considering the actors/actresses dataset, but evaluating the amount of "work" for each move based on the role. i.e. (leading role v. supporting role)
* New parameters for the classifier using the cosine similarity.
* Baseline rating for the vector analysis.

**References**

* [*http://grouplens.org/datasets/movielens/*](http://grouplens.org/datasets/movielens/)
* [*http://www.imdb.com/interfaces*](http://www.imdb.com/interfaces)
* *Kim, M., & Park, S. O. (2013). Group affinity based social trust model for an intelligent movie recommender system. Multimedia tools and applications,64(2), 505-516.*
* *Levandoski, J. J., Sarwat, M., Eldawy, A., & Mokbel, M. F. (2012, April). Lars: A location-aware recommender system. In 2012 IEEE 28th International Conference on Data Engineering (pp. 450-461). IEEE.*
* *Ramos, M., Calvão, A. M., & Anteneodo, C. (2015). Statistical patterns in movie rating behavior. PloS one, 10(8), e0136083.*

1. [↑](#footnote-ref-1)