RESEARCH ARTICLE OPEN ACCESS

Movie Recommendation System Using Sentiment Analysis

[1]Ms.Shamli, [2]P.Yogesh, [3]D.Dedeepya Sumanjali Bonigala

[1]Assistant professor, School of Computing, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India- 600073.
[2,3,4,5] Student, School of Computing, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India- 600073.
[1]shamli.cse@bharathuniv.ac.in[2]yokee1852003@gmail.com[3]dedeepya.bonigala@gmail.com

Abstract:
In that case, certain actions are calming for people in the world and watching films is one thing. Nevertheless, with a lot of movies out there globally choosing one becomes difficult for users. They have to spend time on looking and selecting the movie. It involves much effort and time. In this way, recommendation systems create stuffs. Within this study, we are developing a movie recommendation system by combining KNN and collaborative filtering methodologies.Recommendation systems are usually built using collaborative filtering content-based or hybrid approaches.The various techniques employed in this system include user based recommendation and matrix factorization for collaborative filtering.

Keywords: recommendation systems, matrix factorization, collaborative filtering, content based filtering, memory based, model based, architecture design,cosine similarity

1. Introduction:

It is through recommendations that we are able to make decisions in life. These suggestions are often sought from experienced individuals due to the valuable insights that they provide. The need for recommendations in modern technology is also a matter of efficient results requiring experience[And in the world of digital technology, there exists a sort of preference which goes by name recommendations but now it should be noted that unlike earlier times these recommendations does not only apply on humans but also on machines]. Humans get knowledge by learning while the machine learning derives insights from data and outcomes similar to human learning.The bottom line is recommendation systems which exist across many domains [3]. Through extensive research there have been numerous recommender systems and models.One famous model includes collaborative filtering as consideration of recent activities becomes crucial [4]. Many algorithms utilize different data set to forecast specific user preferences. Some examples include Over-The-Top (OTT) platforms, music streaming services, e-commerce sites recommending related items based on prior purchases made by users or even using AI for image recognition. Specific recommendation techniques like collaborative filtering are generally divided into memory-based and model-based methods. Memory-based approaches require inputs from the user rating matrix whose dynamic updates reflect user preferences.Nowadays, given the huge amount of information available that can only be processed with the help of computers, recommendation systems are essential in a data rich society. These systems are crucial in assisting consumers go through huge amounts of stored data on websites, cloud platforms and servers or even in digital forms [2]. One good example is movie recommendations which are under high demand and expected to increase over time.The film industry has been around for over one hundred years and movies can be found from almost every corner of the world. Nevertheless, most users want to find specific requirements like language, theme, genre among others from this pool. Here is where a movie recommendation system comes in by using machine learning algorithms to analyze datasets. These systems provide recommendations based on various and more. A familiar example in daily life is observed on Over-The-Top (OTT) platforms, where users receive pop-ups of recommended movies based on their viewing or search history, tailored to their interests.The underlying technique of recommendation systems serves the dual purpose of meeting customer needs while benefiting content providers by delivering personalized content to users.

2. Related Work:

Even if one narrows it down to genres, thousands of films are still left. In recent years, the rise in popularity of Over-The-Top (OTT) platforms like Netflix and Amazon Prime has changed the way users can access and consume content. Previously, they had to rely on others' comments about movies in order to find an appropriate movie for their needs. However, finding a suitable movie becomes harder when you have unique interests. So there is now a need for a perfect recommendation system.

Recommendation systems that address this issue adopt different approaches: collaborative ones, hybrid ones and content-based ones. Collaborative approach uses data about similar users to make recommendations. Content-based approach involves using information about just one user to create output so as to match individual tastes/preferences.

$$\begin{bmatrix} 5 & 1 & 4 & 5 & 1 \\ & 5 & 2 & 1 & 4 \\ 1 & 4 & 1 & 1 & 2 \\ 4 & 1 & 5 & 5 & 4 \\ 5 & 3 & 3 & & 4 \\ 1 & 5 & 1 & 1 & 1 \\ 5 & 1 & 5 & 5 & 4 \end{bmatrix} \approx \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1K} \\ u_{21} & u_{22} & \cdots & u_{2K} \\ u_{31} & u_{32} & \cdots & u_{3K} \\ u_{41} & u_{42} & \cdots & u_{4K} \\ u_{51} & u_{52} & \cdots & u_{5K} \\ u_{61} & u_{62} & \cdots & u_{6K} \\ u_{71} & u_{72} & \cdots & u_{7K} \end{bmatrix} \times \begin{bmatrix} v_{11} & v_{21} & v_{31} & v_{41} & v_{51} \\ v_{12} & v_{22} & v_{32} & v_{42} & v_{52} \\ \vdots & & & & \\ v_{1K} & v_{2K} & v_{3K} & v_{4K} & v_{5K} \end{bmatrix} \approx \begin{bmatrix} 0.2 & 3.4 \\ 3.6 & 1.0 \\ 2.6 & 0.6 \\ 0.9 & 3.7 \\ 2.0 & 3.4 \\ 2.9 & 0.5 \\ 0.8 & 3.9 \end{bmatrix} \times \begin{bmatrix} 0.0 & 1.5 & 0.1 & 0.0 & 0. \\ 1.3 & 0.0 & 1.2 & 1.4 & 0. \end{bmatrix}$$

2.1 Existing Modal

2.1.1 Matrix decomposition

It is an effective approach for small scale project. This algorithm use matrix decomposition for recommending movie. We will make vectors with the given rating of the user and use this for produce result.

factors, including the lead actor, genre, reviews, search history, language, region,

2.1.2 Clustering

Matrix factorization becomes less efficient as the size of the system increases. However, clustering allows for an unsupervised method that is more useful in building large systems. This makes clustering a more scalable solution compared to supervised techniques which need constant vigilance because they cannot handle larger volumes of data as their datasets expand. It then assigns every user into a cluster or group based on their taste similarity.

The users are scattered throughout the system before clustering is done but after clustering, the users fall into distinct groups, usually three. Users who fall within this category will then be recommended similar kinds of information hence making it easier to give out such recommendations and improve the overall system performance.

2.1.3Deep learning approach

Nowadays, neural networks have become very popular and applied in different machine learning models especially on big platforms like YouTube. Youtube video recommendations is quite difficult due to its enormous scale and external influences that come with it.

3. Available dataset:

The information for the movie recommendation system was gotten from the official IMDB website and supplemented by some additional data from Kaggle. However, managing big data is problematic because it is characterized by repetition and null values. Python as a powerful tool in machine learning provides several libraries that come with it to solve such issues. Numpy and Pandas are vital in dealing with datasets by helping in importing them, examining and selecting or removing specific details. This enhances the reliability of the model when its used to represent relevant information.

Name	Domain	Users	Columns	Ratings	Null Values
IMDB 5000 Movie Dataset	Movies	5028	14	5 Star	15
The Movies Dataset	Movies	45548	26	5 Star	64
List of movies in 2018	Movies	259	5	5 Star	0
List of movies in 2019	Movies	224	5	5 Star	0
List of movies in 2020	Movies	241	5	5 Star	0

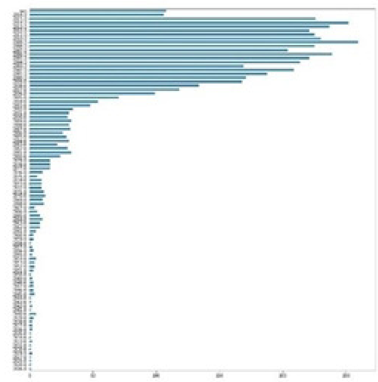
4. Data Cleaning and Exploratory Data Analysis

When data cleaning is mentioned, it means either deleting or modifying inaccurate, incomplete, outdated, duplicated or unformatted data among others so that it can be analyzed. Such like kind of information can hinder the process or bring about wrong outcomes in most cases. Depending on the storage format and a desired result different ways of cleaning may be applied.

Data cleaning does not only involve wiping off some previous figures to let new ones fit inside; it also implies making more accurate a dataset without necessarily losing value.Exploratory Data Analysis (EDA) refers to a technique used for assessing datasets

In our project, we used different sources of data such as movie_metadata.csv and credits.csv. First, we uploaded the movie_metadata.csv which contained the head function to display the first ten elements in a dataset. Moreover, using the shape and columns functions let us access its dimensions and column details respectively. A Figure 1 graphically shows that relationship between years and number of movies released from1916 to 2016.The immense sizes of data set made it necessary to simplify it by creating another dataframe. This only held onto essential elements from movie_metadata.csv and replaced any NaN values with "unknown." Similarly, valuable information was extracted from movies_metadata.csv as well as credits.csv focusing on Hollywood movies up to 2020.

Fig 1: Plot between Year vs Number of movies released



5. Collaborating filtering

Collaborative filtering is an item-based recommendation system where items are published or recommended based on similarity between the item and users[3]. It works by collecting ratings on a feature from each user in which then those similarities are found out for finalized recommendations. It is a user-oriented approach[2].It figure outs the users with the same opinion and then after catching the similarity in reviews it recommends the particular movie.

Advantages:

1. Dependent on the ratings, thus making it content independent.
2. 2. It can suggests fortuitous recommendations based on the similarity of users.
3. 3. It also considers the experience to create real life assessment.

Disadvantages:

1. If the initial ratings are contradictory to the later ones then ambiguity arises.

2. Variations in review cases are difficult to group in agree or disagree nature.

3. Difficult in tackling sparcity situations.

Divided into two major categories:

Memory based method ,Model based method

3.1 Memory based method

Each data item is examined continuously by the system to propose and group it into specific primary classes, working on the basis of similarity. For example, in movie recommendation context the approach is based on cosine similarity. The cosine coefficient, also known as vector similarity, treats user ratings as points in a vector model. Then it calculates the cosine of angle (θ) between these points [2].

Model Based method

A theoretical model for user rating behavior was proposed. It has been using raw data for last many years to make recommendations.

6. Content-based filtering program

Content-based recommends systems are designed to predict user attributes or characteristics using features from an item that a user has positively respondedto.

Movies	User 1	User 2	User 3	User 4	Action	Comedy
Item 1	1		4	5	Yes	No
Item 2	5	4	1	2	No	Yes
Item 3	4	4		3	Yes	Yes
Item 4	2	2	4	4	No	Yes

The last two columns, Action and Comedy, delineate the types of movies. However, if we are provided with these genres alone, discerning the specific preferences of users for each genre becomes challenging. To address this, it is crucial to identify user preferences based on their reactions to movies within a particular genre. Once we have insight into a user's preferences, we can embed this information into a feature-generated vector within the embedded platform. Recommendations are then tailored according to the user's preferences.

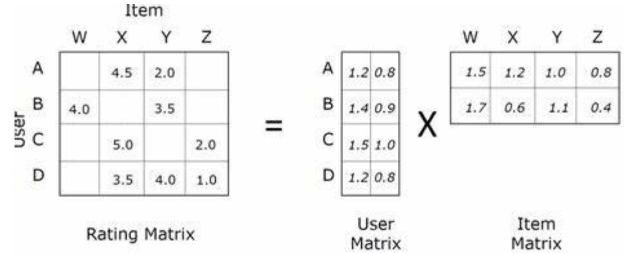
During the recommendation process, matching metrics are computed by comparing the item's feature vectors with the user's preference vectors derived from their past records. Subsequently, the top recommendations are made. Notably, content-based filtering operates independently, not relying on other users' data for the recommendations of a specific user.

Naive Bayes: algorithms are widely used in sentiment analysis, spam filtering, recommendation systems etc. They are quick and easy to use but their biggest disadvantage is that the need for forecasts is independent. In many real-life situations, forecasts rely, this precludes the operation of the distinction.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)},$$

7. Matrix Factorization

In this project we have used Similarity score and cosine similarity to recommend items or movies to the users based on what he or she has searched for. We can also use the concept of Matrix Factorization. In recommender systems, matrix factorization is a type of collaborative filtering algorithm. The user-item interaction matrix is decomposed into the product of two lower dimensionality rectangular matrices by matrix factorization algorithms. According to Simon Funk, this method family became well known during the Netflix prize challenge due to its effectiveness. Matrix factorization is a technique for representing users and objects in a lower-dimensional latent space, Refer Fig 3. Matrix factorization is used in collaborative filtering to determine the relationship between item and user entities. We'd like to predict how users will rate items based on the feedback of customer reviews so that users can get recommendations based on the forecast.



7.1 Mathematical Concept of Matrix Factorization

Create a fixed of users (U), objects encompasses all consumer scores. The primary goal is to pick out K latent features, resulting within the product matrix S obtained through the enter*k*okay).

Matrix P signifies the affiliation among users and features, while matrix Q represents the affiliation among gadgets and capabilities. The prediction of an item's score is finished with the aid of calculating the dot made of the vectors corresponding to U_i and I_j.

To acquire matrices P and Q, initialization precedes the calculation of the distinction, forming matrix M. Iterative steps are then employed to reduce this difference, utilising gradient descent to are searching for a local minimum.

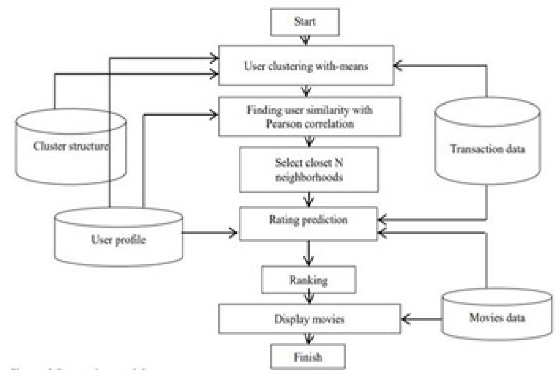
The loss function, represented by using lij^2, is calculated as the squared distinction between the actual and predicted rankings. The gradient of this characteristic is used to update both p_ik and q_kj for my part, aiming to decrease the overall loss. Iterative updates keep till the overall loss converges to a minimal.

8. Architecture Design

The device is built at the Python anThe the front-end is developed using HTML, CSS, and JavaScript. HTML gives the fundamental structure, whilst CSS is hired for styling factors along with color, layout, and font, improving the overall user interface. JavaScript adds interactivity to the website, permitting customers to engage seamlessly.Python serves as the backend language, controlled thru the Django framework. MySQL storage is employed for the database. The machine operates thru layers, with the primary layer being the User Interface (UI) layer evolved with HTML, CSS, and JS. This layer helps person interaction with the backend.User movements on the the front end cause requests to the backend via REST API and HTTP requests. Various HTTP methods, inclusive of

GET, POST, PATCH, and DELETE, are applied. The GET technique fetches facts from the server with out a request body, whilst POST is used when the server accepts records enclosed inside the request frame. DELETE removes records from the database, and PATCH updates precise attributes of current records. The PUT approach is likewise used for updates but involves replacing the prevailing facts.

The backend strategies these requests based totally on the endpoints and HTTP methods, executing operations at the database. The backend then sends a reaction to the frontend, which displays the statistics as a consequence. This establishes a seamless waft of statistics between the user interface and the backend, enabling effective communique and gadget functionality.



8.1 Database Design

Database is the foundation of any website. The database should follow ACID properties to work properly. The database used in the Project is MYSQL. Tables names are User Table, Movie Table, User Similarity Table, Movie Type Table.

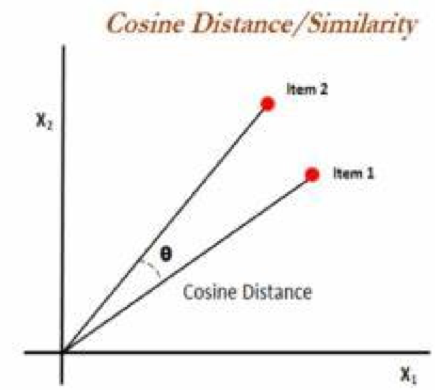
9. Implementation, Results and Deployment

The dataset become acquired from Kaggle, explaining why it best includes data up to 2016. For the years 2018, 2019, and 2020, film data changed into extracted from Wikipedia. However, Wikipedia lacked a style column for films, so the TMDB API was utilized to fetch style records through GET requests. The resulting facts, acquired in JSON format, became processed the usage of the important thing "style" to retrieve the style of the respective film. The lambda feature changed into employed throughout preprocessing, and the finalized statistics become saved in a CSV record.

Subsequently, a sentimental model was trained the usage of the evaluations.Txt report, permitting the machine to decide whether or not a given review is effective or poor. Various libraries, along with TFidfVectorizer and NLTK, had been used to convert overview textual content into vector layout. A Multinomial Naive Bayes version was trained and saved as a pkl document for destiny use.

In the final internet site, car-recommendations for movie names had been implemented using JavaScript and the available information. Upon coming into a film call and pressing enter, the device provides records including identify, evaluation, genre, rating, release date, and greater. Additionally, details about the solid, pinnacle consumer reviews, and the emotions (superb or poor) are displayed. The internet site also gives top encouraged movies based on the searched film. This functionality is implemented the usage of Python, JavaScript, and Ajax.

To decide item similarity, the device utilizes similarity ratings, numerical values starting from zero to 1. These ratings investigate how similar two items are by means of comparing their textual content statistics. Cosine similarity is especially employed, permitting the system to degree similarity irrespective of file size. It calculates the cosine of the attitude among two vectors projected in a multi-dimensional space. The better the cosine similarity, the smaller the attitude, indicating more similarity among the gadgets. This metric is important for recommending items with similar textual records, ensuring correct and applicable tips



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

10. Working of recommend systems

There are 4 primary types of recommender structures.

A. Context Recommender System:

The idea of "context" is multifaceted and studied in diverse fields. In the context of Recommender Systems, parameters range primarily based at the type of advice device, which includes for movies or tourism. Contextual records consists of elements like date of viewing, area, watch time, season information, social connections, and massive occasions. Traditional advice systems frequently neglect contextual information, impacting decision-making accuracy. Research has explored the selection of suitable contextual features, emphasizing their impact on accuracy and computational load.

B. Network-Based Recommender System:

To overcome the limitations of Collaborative Filtering, latest research has centered on community-based advice systems. Trust-primarily based recommendation strategies involve constructing neighborhoods based totally on trust relationships between customers, improving suggestions. Multidisciplinary processes include contextual facts into recommender structures, considering factors like time, location, and context. However, predicting the same context within the future is challenging, leading to improved information sparsity. Different methods, including content material-primarily based processes, intention to seize contextual statistics for advanced predictive accuracy.

C. Trust-Based Recommender System:

Trust is a complicated component inspired by way of diverse elements consisting of human relationships, mental elements, and the have an impact on of others' evaluations. Building fashions for agree with-based recommendation structures entails thinking about elements like transitivity, asymmetry, and personalization. Trust can be transferred between people, however it isn't always absolutely compatible with mathematical experience because of versions in individuals' reviews and backgrounds. Modeling the prevalence of trust entails formal frameworks, thinking about factors like reliability values, distrust, and uneven trust.

D. Content-Based Recommender System:

Content-based recommender structures attention on item capabilities and user preferences. Recommendations are made by way of considering the content material details, and similarity rankings are used to degree how near two gadgets' textual content statistics is. Cosine similarity is a typically used metric to determine similarity between documents, regardless of their size. The better the cosine similarity, the smaller the perspective, indicating greater similarity.

In precis, recommender systems encompass various strategies, consisting of context, community-primarily based strategies, trust-based totally fashions, and content-based totally methods, every addressing particular demanding situations and enhancing recommendation accuracy in unique contexts.LITRATUE SURVEY:

Recommender systems cope with the undertaking of records overload by presenting personalized tips and services. Various advice systems had been evolved, utilising collaborative filtering, content material-based totally filtering, or hybrid approaches. Collaborative filtering, the most mature and widely used method, recommends gadgets based totally on figuring out users with similar hobbies and the usage of their alternatives to make hints to an lively person. This method has been carried out in diverse areas, together with Group Lens for information retrieval, Ringo for social media filtering, and Amazon for improving hints via collaborative filtering.

Content-based techniques attention on content material sources and user capabilities, dismissing contributions from different users. Interactive content-based totally and filtration-primarily based filters are commonly used independently or combined over the years. In every other recommendation gadget, hints are generated based on users with comparable choices to the modern consumer. This approach extends to group recommendation procedures. The metaphor is employed to estimate the weight of each person-targeted metric, combining them right into a unmarried price for a extra dependable judgment.

To ease user comments, a way is brought, permitting customers to explicit critiques about guidelines in place of answering a prolonged listing of questions. This technique may be prolonged to various parameters without users being aware that the algorithm isn't totally customized, leading to specific effects based on their selections.

Datasets from Kaggle.Com are used for benchmarking new recommendation strategies. Five different datasets were hired on this undertaking:

IMDB 5000 Movie Dataset: Encompassing all Hollywood movies until 2017, which includes the ones from OTT platforms.

The Movies Dataset: Comprising credit.Csv and movies_metadata.Csv documents for large film information.

List of Movies in 2018: Featuring films launched in 2018 at the side of their ratings.

List of Movies in 2019: Including films released in 2019.

List of Movies in 2020: Encompassing information of movies released in 2020.

10. Conclusions

We meticulously wiped clean and analyzed the statistics acquired from Kaggle and IMDB to prepare it for schooling our version. Conducting exploratory facts evaluation allowed us to extract valuable insights, get rid of missing values, and carry out vital facts preprocessing steps.

Upon travelling our internet site, users can input a movie call in the seek bar, triggering vehicle-suggestions associated with their query. Upon pressing input, customers are redirected to a new web page showing complete facts approximately the searched movie. This facts consists of info consisting of style, rating, superstar forged, user critiques, sentiments of the critiques, and recommended films.

It's essential to note that our task currently specializes in Hollywood movies, however its capability can be prolonged to cover any kind of film with the necessary facts.

We have compiled a detailed record explaining how the film recommender gadget operates, which includes an exploration of various recommender gadget types and a survey to beautify its improvement. The report delves into distinct sorts of recommender structures, highlighting the distinctions among movie recommender systems and vacationer recommender systems. Filtering kinds employed in recommender structures are also discussed, at the side of insights into the datasets used in this challenge.

11. References

[1] Raval, Nirav, and Vijayshri Khedkar. 2019. "A Review Paper On Collaborative Filtering Based MoiveRecommedation System." INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH 8. www.ijstr.org.

[2] Bhatt, Bhumika, Premal J Patel, Hetal Gaudani, and Associate Professor. 2014. "A Review Paper on Machine Learning Based Recommendation System." International Journal of Engineering Development and Research. Vol. 2. www.ijedr.org.

[3] Hande, Rupali, Ajinkya Gutti, Kevin Shah, Jeet Gandhi, and Vrushal Kamtikar. n.d. "IJESRT INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY MOVIEMENDER-A MOVIE RECOMMENDER SYSTEM." International Journal of Engineering Sciences & Research Technology. https://doi.org/10.5281/zenodo.167478.

[4] Adomavicius, G., Tuzhilin, A.,"Context-Aware Recommender Systems", In Recommender Systems Handbook, ed. F. Ricci,L. Rokach, B. Shapira, P. Kantor, 217–256. Berlin: Springer Verlag, 2011.

[5] Ante Odic, Marko Tkalcic, Jurij F, An drej Kosir. "Relevant Context in a Movie Recommender System: Users' Opinion vs. Statistical detection", CARS -2012.

[6] Rahul Gupta, Arpit JainJ, Satakshi Rana, Sanjay Singh "Contextual Information based Recommender System using Singular Value Decomposition", ICACCI, 2013, pp. 2084-2089.

[7] Raval, Nirav, and Vijayshri Khedkar. 2019. "A Comprehensive Review on Collaborative Filtering-Based Movie Recommendation Systems." INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH 8. www.ijstr.org.

[9] Bhatt, Bhumika, Premal J Patel, Hetal Gaudani, and Associate Professor. 2014. "An In-Depth Review Paper on Machine Learning-Based Recommendation Systems." International Journal of Engineering Development and Research. Vol. 2. www.ijedr.org.

[9] Hande, Rupali, Ajinkya Gutti, Kevin Shah, Jeet Gandhi, and Vrushal Kamtikar. n.d. "IJESRT INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY MOVIEMENDER - A MOVIE RECOMMENDER SYSTEM." International Journal of Engineering Sciences & Research Technology.
https://doi.org/10.5281/zenodo.167478.

[10] Adomavicius, G., Tuzhilin, A. "Context-Aware Recommender Systems." In Recommender Systems Handbook, edited by F. Ricci, L. Rokach, B. Shapira, P. Kantor, 217–256. Berlin: Springer Verlag, 2011.

[11] Ante Odic, Marko Tkalcic, Jurij F, Andrej Kosir. "Relevant Context in a Movie Recommender System: Users' Opinion vs. Statistical Detection," CARS - 2012.

[12] Rahul Gupta, Arpit Jain, Satakshi Rana, Sanjay Singh. "Contextual Information-based Recommender System using Singular Value Decomposition," ICACCI, 2013, pp. 2084-2089.

