Movie Popularity and Target Audience Prediction Using the Content-Based Recommender System

Mr.Mabubasha[1], Thorlikonda Siva Prasad[2], Vishnumolakala Sarvan Manikanta[3], Sabavath Mohan Naik[4]

[1]Assistant Professor, Dept. of Computer Science and Engineering, R.V.R. & J.C College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India

[2,3,4]B. Tech Student, Dept. of Computer Science and Engineering, R.V.R. & J.C College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India

Abstract: The movie is one of the integral components of our everyday entertainment. The worldwide movie industry is one of the most growing and significant industries and seizing the attention of people of all ages. It has been observed in a recent study that only a few movies achieve success. Uncertainty in the sector has created immense pressure on the film production stakeholder. Moviemakers and researchers continuously feel it necessary to have some expert systems predicting the movie success probability preceding its production with reasonable accuracy. A maximum of the research work has been conducted to predict the movie's popularity in the post-production stage. To help the movie maker estimate the upcoming film and make necessary changes, we need to conduct the prediction at the early stage of movie production and provide specific observations about the upcoming movie.

Keywords: Movie, audience, prediction, moviemakers, film production.

I. INTRODUCTION

The worldwide movie industry is a fast-moving revenue generating industry, and the multi-billion dollar has been involved in this industry. A large number of people are associated with this industry, and massive investment is required as qualitative and quantitative. In 2019 the total box office revenue of the United States and Canada was $11.32 billion [59]. However, in-ground reality, few numbers movies has been achieved success. Film producers and researchers constantly feel it essential to have some expert systems that predict the movie's success chance leading its production with appropriate accuracy. The movie industry is massive and diverse. A sign cant number of parameters from different dimensions are involved in creating a movie. Representing an upcoming movie's success or degree of success is a highly complex task. Research works have been conducted to predict movie popularity. Earlier, several works have been conducted on post-production or post release forecast. However, it is not benecial as the investor has already contributed their funds to the lm production. The early production stage and pre-production prediction with satisfying accuracy have been benecial to secure investment. A forecast made soon after the cast, director, and storyline have been nalized would assist the investor in making a facial decision. After a rigorous study, we have seen signicant research on movie hit prediction before the official release. Predictions performed shortly before or following, the official release (the last stage in lm production) may have additional data to use and produce a more precise prediction [66]. Still, they are considerably delayed for investors to estimate any critical decision. Early-stage (production) forecast of movie success

is the most biennial. Very little work has been performed to forecast movie success at an early stage of movie production. The early-stage forecast of previous works' accuracy is not significantly good. Maximum of the works are performed only to focus success probability of the upcoming movie. Some of them classify the problem into a binary problem (hit/ op), and in some work, they classify the problem into a multiclass problem. Movie Makers start creating a new movie while targeting a specie audience group or groups most of the time. Audience age is one of the essential criteria for the target audience. Some movies are created by targeting the junior audience group. Some movies target teenage audiences, sometimes target the mid-age and senior audience group, and some movies are for all. Suppose we could predict whether the upcoming movie would be famous among the target audience group or not. Movie makers would be beneted if we could measure the inuence of the upcoming movie among all the age groups at the early stage of the movie production. Then, movie Maker could make necessary changes if needed. The movie hit forecasting and target audience prediction of the upcoming movie at the early stage of the movie production are interrelated and meaningful. The outcome of this work could reduce the risk involved in the movie industry.

II. LITERATURE REVIEW

[1] L. Sharma and A. Gera, ``A survey of recommendation system: Research challenges," *Int. J. Eng. Trends Technol.*, vol. 4, no. 5, pp. 1989_1992, 2013.

A recommender system is a set of tools for information retrieval. It improves access and proactively recommends items and services that match users' tastes by considering their explicit and implicit preferences and behaviors. Recommender systems have become very popular in the e-commerce field. Today, the internet is flooded with diverse information that makes it very difficult for the end-users to reach out for what they need. Recommender systems provide tailored views to users who are constantly adapted to the users' changing tastes. Although many recommendation techniques have been developed in multiple domains, recommender systems still face problems and challenges that hinder their precision. This paper provides a comprehensive summary of the key challenges and problems when developing recommender systems and summarizes the latest research achievements and directions to resolve them. In addition to that, we go beyond this by presenting the evaluation techniques used to judge the performance of recommender systems.

[2] N. Das, S. Borra, N. Dey, and S. Borah, ``Social networking in web based movie recommendation system," in *Social Networks Science: Design, Implementation, Security, and Challenges*. Cham, Switzerland: Springer, 2018, pp. 25_45.

Movie Recommendations Systems are a common practice by most of the online stores today. The web based movie recommendation systems makes predictions about the responses of the users based on their search history or known preferences. Recommendation of items is usually done based on the properties or content of the item or collaboration of the user’s ratings, and by using intelligent algorithms that include classification or clustering techniques. Accurate prediction of what the customer may likely to busy or the user my visit is of utmost important, as it benefits both the service providers and customers. This chapter provides the evolution, fundamental concepts, classification, traditional and novel models, requirements, similarity measures, evaluation approaches, issues, challenges, impacts due to social networking, and future of movie recommendation systems.

[3] P. Nagarnaik and A. Thomas, ``Survey on recommendation system methods," in *Proc. 2nd Int. Conf. Electron. Commun. Syst. (ICECS)*, Feb. 2015, pp. 1603_1608.

In recent years recommendation systems have changed the way of communication between both websites and users. Recommendation system sorts through massive amounts of data to identify interest of users and makes the information search easier. For that purpose many methods have been used. Collaborative Filtering (CF) is a method of making automatic predictions about the interests of customers by collecting information from number of other customers, for that purpose many collaborative base algorithms are used. CHARM algorithm is one of the frequent patterns finding algorithm which is capable to handle huge dataset, unlike all previous association mining algorithms which do not support huge dataset. This paper covers different techniques which are used in recommendation system and also proposes a new system for efficient web page recommendation based on hybrid collaborative filtering i.e. using collaborative technique and CHARM algorithm which are coupled with the pattern discovery algorithms such as clustering and association rule mining.

[4] M. A. Hameed, O. Al Jadaan, and S. Ramachandram, ``Collaborative _ltering based recommendation system: A survey," *Int. J. Comput. Sci. Eng.*, vol. 4, no. 5, p. 859, 2012.

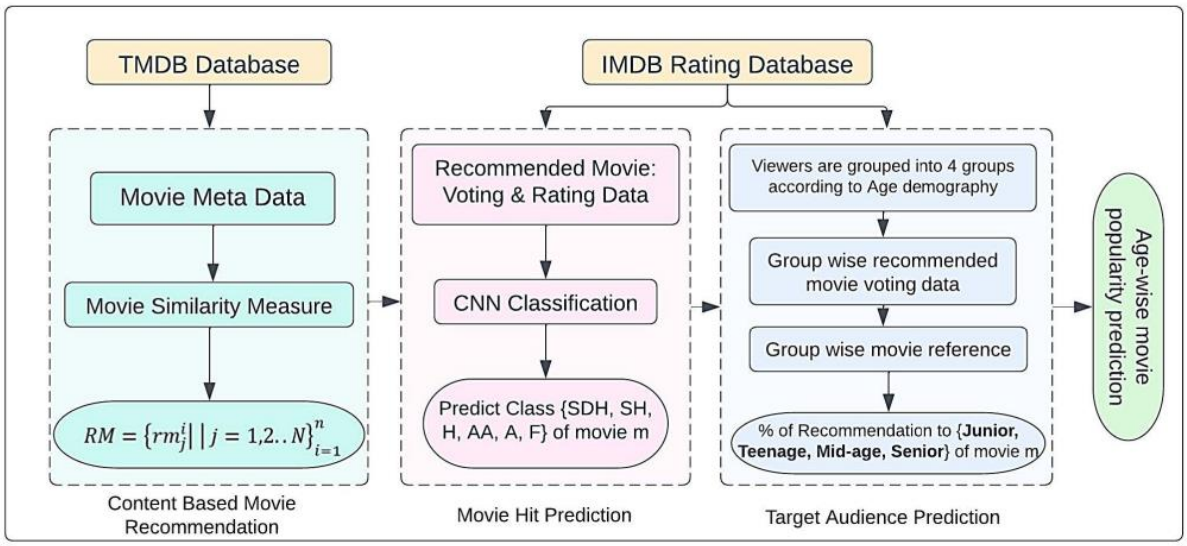
In the era of big data, recommender system (RS) has become an effective information filtering tool that alleviates information overload for Web users. Collaborative filtering (CF), as one of the most successful recommendation techniques, has been widely studied by various research institutions and industries and has been applied in practice. CF makes recommendations for the current active user using lots of users' historical rating information without analyzing the content of the information resource. However, in recent years, data sparsity and high dimensionality brought by big data have negatively affected the efficiency of the traditional CF-based recommendation approaches. In CF, the context information, such as time information and trust relationships among the friends, is introduced into RS to construct a training model to further improve the recommendation accuracy and user's satisfaction, and therefore, a variety of hybrid CF-based recommendation algorithms have emerged. In this paper, we mainly review and summarize the traditional CF-based approaches and techniques used in RS and study some recent hybrid CF-based recommendation approaches and techniques, including the latest hybrid memory-based and model-based CF recommendation algorithms. Finally, we discuss the potential impact that may improve the RS and future direction. In this paper, we aim at introducing the recent hybrid CF-based recommendation techniques fusing social networks to solve data sparsity and high dimensionality and provide a novel point of view to improve the performance of RS, thereby presenting a useful resource in the state-of-the-art research result for future researchers.

[5] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, ``Item-based collaborative _ltering recommendation algorithms," in *Proc. 10th Int. Conf.WorldWide Web (WWW)*, 2001, pp. 285_295.

In order to overcome the limitations of data sparsity and inaccurate similarity in personalized recommendation systems, a new collaborative filtering recommendation algorithm by using items categories similarity and interestingness measure is proposed. In this algorithm, first the items categories similarity matrix is constructed by calculating the item-item category distance, and then analyzes the correlation degree of different items by using interestingness measure, last an improved collaborative filtering algorithm is proposed by combining the information of

items categories with item-item interestingness and utilizing improved conditional probability method as the standard item-item similarity measure. Experimental results show this algorithm can effectively alleviate the dataset sparsity problem and achieve better prediction accuracy compared to other well-performing collaborative filtering algorithms.

III. SYSTEM ARCHITECTURE



Algorithms:

- TF-IDF
- KNN
- RF
- LR

TfidfVectorizer

Generally Machine Learning algorithm only deals with numeric format data, in text classification when we are using text data we need to convert the data in to numeric vector format. Generally this process may can be done with two ways, one is CountVectorizer and another one is TfidfVectorizer, in CountVectorizer it will take the word count of the text, it is very basic one. But in the TfidfVectorizer it will take the Tf*IDf score as its numerical data for vector model.

K- NEAREST NEIGHBOUR ALGORITHM (KNN):

KNN is slow supervised learning algorithm, it take more time to get trained classification like other algorithm is divided into two step training from data and testing it on new instance . The K Nearest Neighbour working principle is based on assignment of weight to the each data point which is called as neighbour. In K Nearest Neighbour distance is calculate for training dataset for each of the K Nearest data points now classification is done on basis of majority of votes there are three types of distances need to be measured in KNN Euclidian, Manhattan, Minkowski distance in which Euclidian will be consider most one the following formula is used to calculate their distance.

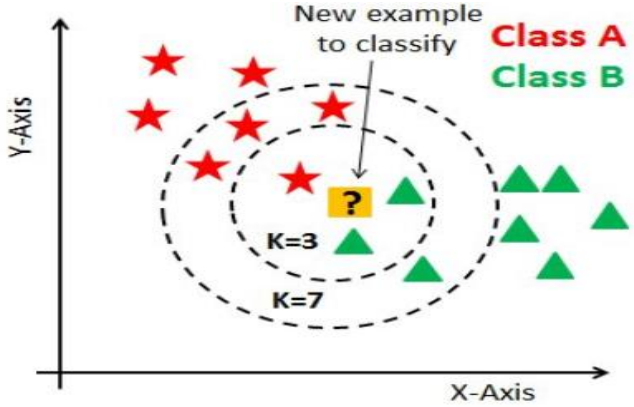
$$Eucledian\ Distance = D(x, y) \qquad (1)$$
$$= (x_i - y_i)_{2k_i} = 1$$

K=number of cluster

x , y=co-ordinate sample spaces

The algorithm for KNN is defined in the steps given below:

1. D represents the samples used in the training and k denotes the number of nearest neighbour.
2. Create super class for each sample class.
3. Compute Euclidian distance for every training sample
4. Based on majority of class in neighbour, classify the sample



K- Nearest Neighbour

Logistic regression

Logistic regression is a statistical method for analyzing data set in which there are one or more independent variables that determine an outcome. The outcome is measured with dichotomous variable (in which there are only two possible outcomes). For the cases when there are more than two labels, the strategy, which is called “One versus all”, is used. In this strategy every category is binary classified against its inverse(a fictional category that states that the example does not belong to the current category). The category with the highest score is picked as a result of a classification. Logistic regression is one of the simplest machine learning techniques. It is easy to implement and easy to interpret. It is usually a good idea to implement logistic regression classifier before proceeding with a more complex approach because it gives you an estimate of how well machine learning algorithms will perform on this specific task. It also helps to eliminate some basic implementation bugs regarding data set treatment.

Random Forest

In our experiment, we use random forest as a classifier. The popularity of decision tree models in data mining is owed to their simplification in algorithm and flexibility in handling different data attribute types. However, single-tree model is possibly sensitive to specific training data and easy to overt . Ensemble methods can solve these problems by combine a group of individual decisions in some way and are more accurate than single classifiers. Random forest, one of ensemble methods, is a combination of multiple tree predictors such that each tree depends on a random independent dataset and all trees in the forest are of the same distribution. The capacity of random forest not only depends on the strength of individual tree but also the correlation between different trees. The stronger the strength of single tree and the less the

correlation of different tress, the better the performance of random forest. The variation of trees comes from their randomness which involves bootstrapped samples and randomly selects a subset of data attributes.

Below is the step by step Python implementation. ...

Step 2 : Import and print the dataset.

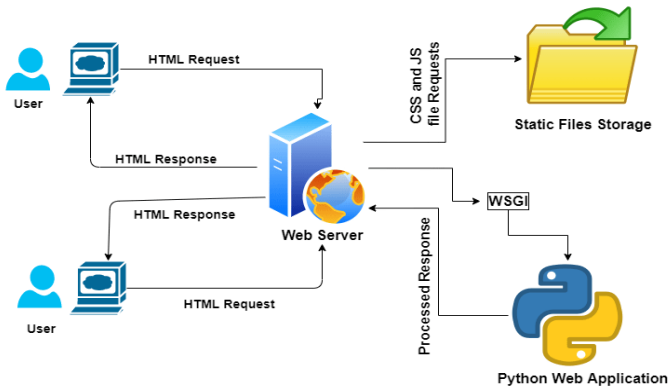
Step 3 : Select all rows and column 1 from dataset to x and all rows and column 2 as y.

Step 4 : Fit Random Forest regressor to the dataset.

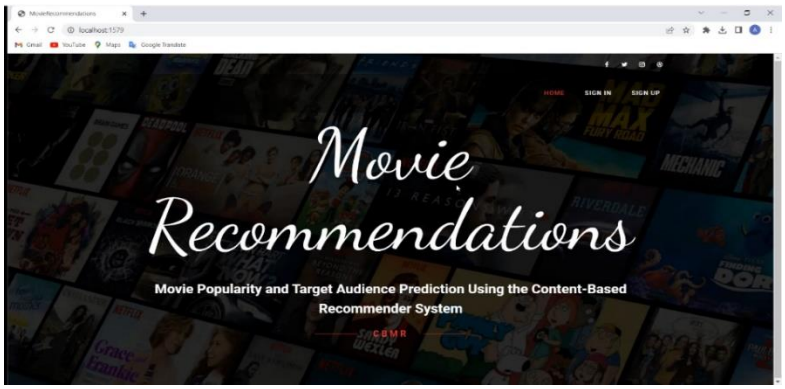
Step 5 : Predicting a new result.

Step 6 : Visualising the result.

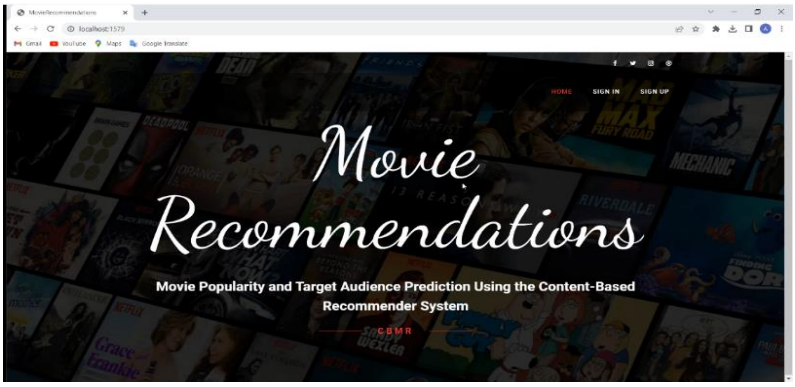
Deployment Model



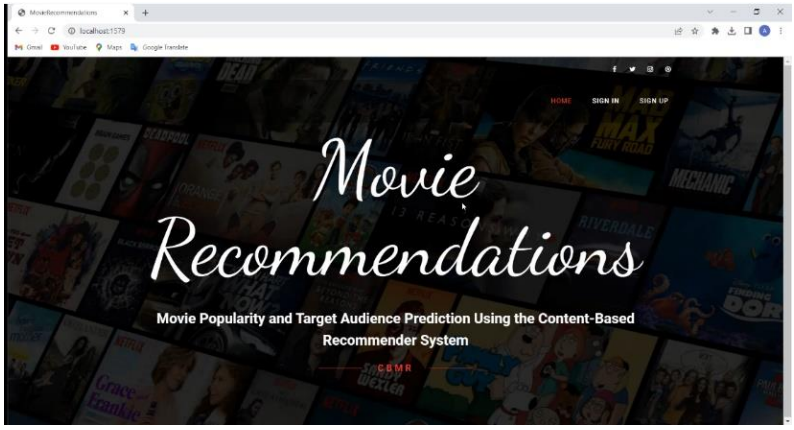
IV. RESULTS



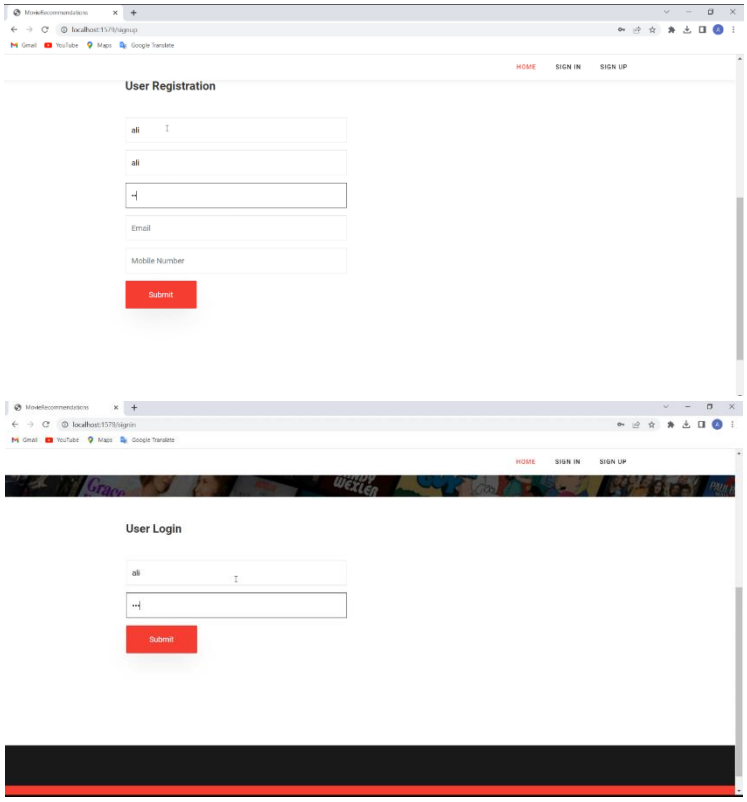
Result of my Project in chrome

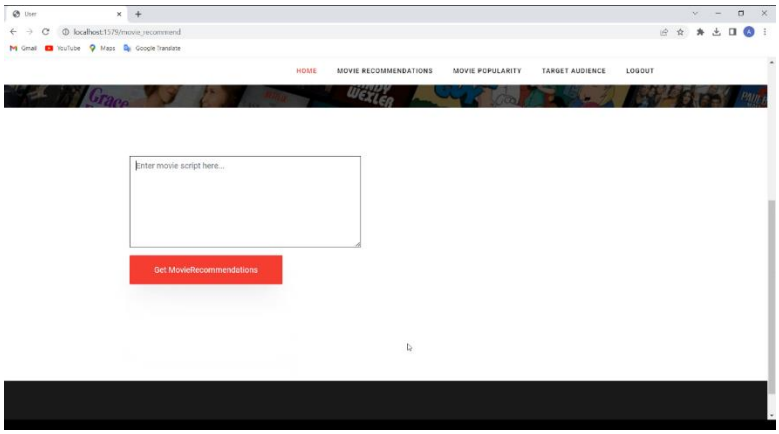


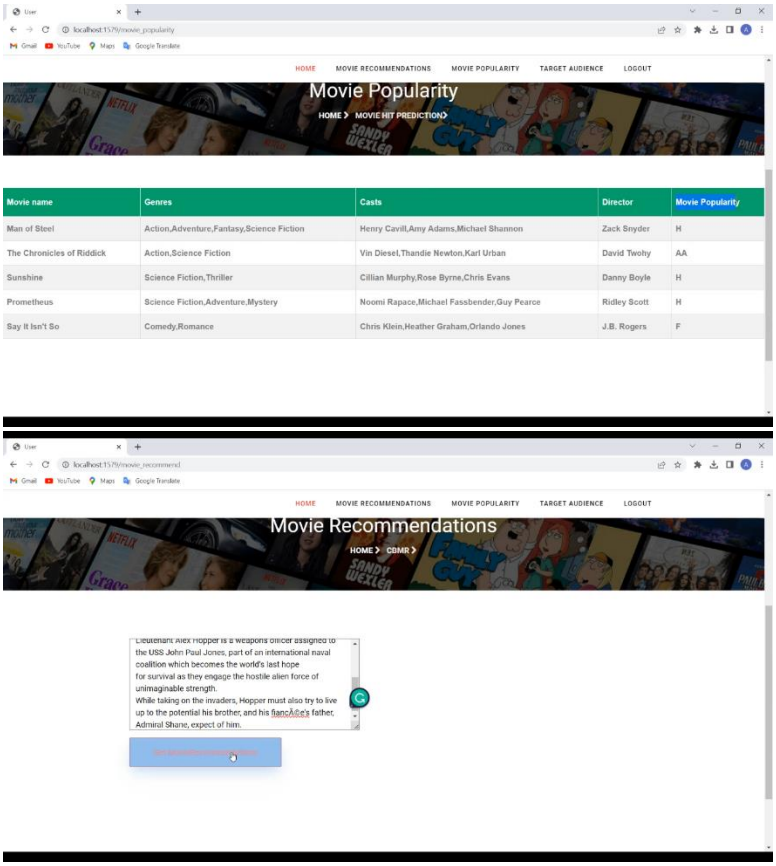
Result of my Project in Opera



SCREENSHOTS







V. CONCLUSION

A substantial amount of financing is consumed in every box-office movie. However, most movies fail to achieve success. Earlier, the most significant number of works have been done on post-production or post-release forecast. The estimate does not influence as the investor has already consumed their funds on the _lm production. The pre-production or early production stage forecast needs high accuracy and the best time to ensure investment. The objective of our study is to propose an expert system that could help the movie maker execute necessary changes if needed at the appropriate time. Our system can food cost the level of popularity of the upcoming movie before the production has started for the earliest stage of the production and with significant accuracy.

REFERENCES

[1] L. Sharma and A. Gera, ``A survey of recommendation system: Research challenges," *Int. J. Eng. Trends Technol.*, vol. 4, no. 5, pp. 1989_1992, 2013.

[2] N. Das, S. Borra, N. Dey, and S. Borah, ``Social networking in web based movie recommendation system," in *Social Networks Science: Design, Implementation, Security, and Challenges*. Cham, Switzerland: Springer, 2018, pp. 25_45.

[3] P. Nagarnaik and A. Thomas, ``Survey on recommendation system methods," in *Proc. 2nd Int. Conf. Electron. Commun. Syst. (ICECS)*, Feb. 2015, pp. 1603_1608.

[4] M. A. Hameed, O. Al Jadaan, and S. Ramachandram, ``Collaborative _ltering based recommendation system: A survey," *Int. J. Comput. Sci. Eng.*, vol. 4, no. 5, p. 859, 2012.

[5] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, ``Item-based collaborative _ltering recommendation algorithms," in *Proc. 10th Int. Conf.WorldWide Web (WWW)*, 2001, pp. 285_295.

[6] Y. Koren and R. Bell, ``Advances in collaborative _ltering," in *Rec-ommender Systems Handbook*. Boston, MA, USA: Springer, 2015, pp. 77_118.
[7] J. L. Herlocker, J. A. Konstan, and J. Riedl, ``Explaining collaborative _ltering recommendations," in *Proc. ACM Conf. Comput. supported Coop-erat. Work (CSCW)*, 2000, pp. 241_250.
[8] M. J. Pazzani, ``A framework for collaborative, content-based and demographic _ltering," *Artif. Intell. Rev.*, vol. 13, no. 5, pp. 393_408, 1999.
[9] R. Van Meteren and M. Van Someren, ``Using content-based _ltering for recommendation," in *Proc. Mach. Learn. Inf. Age, MLnet/ECML2000 Workshop*, vol. 30, 2000, pp. 47_56.
[10] P. B. Thorat, R. M. Goudar, and S. Barve, ``Survey on collaborative _ltering, content-based _ltering and hybrid recommendation system," *Int. J. Comput. Appl.*, vol. 110, no. 4, pp. 31_36, Jan. 2015.