IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

Movie Recommendation System: Content-Based and Collaborative Filtering

S. K. Raghuwanshi[1*], R. K. Pateriya[2]

[1]Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal, India
[2]Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal, India

**Corresponding Author: sraghuwnashi@gmail.com*

Available online at: www.ijcseonline.org

Received: 20/Mar/2018, Revised: 28/Mar/2018, Accepted: 19/Apr/2018, Published: 30/Apr/2018

***Abstract*—** Since last decade a huge amount of information is transferred over the internet on day to day basis. However, all the information is not relevant to each user and is also difficult to find the right content for the user as per his/her need. Recommender system works as a guide to find or suggest right items for users. A movie recommendation system is predicting or suggest a movie which user might like using his/her previous watch list or history. After Netflix prize competition many academician and researchers have shown interest to develop new and better filtering techniques for the movie recommendation. This paper studies the two most fundamental techniques: content-based and collaborative filtering methods of information retrieval and shows their application for movie recommendation with pros and cons. An experiment was carried out over MovieLens 100K dataset to show the implementation of discussed methods. The obtained results have shown that Item-Item based neighbourhood collaborative filtering method is better among implemented three techniques with 0.786 MAE and 0.985 RMSE values.

***Keywords*—** Content-Based Filtering, Collaborative Filtering, Movie Recommendation.

I. INTRODUCTION

The Internet has become a place to share and upload content which leads to having a huge amount of meaningful information available online. However, this explosive growth leads other problems like high processing cost and system overload. Information filtering systems or Recommendation system(RS) answer the stated question and have become an essential part of the current web-based system of e-commerce, entrainment, business or almost every field. There are a wide range of applications for recommendation system and become famous in last decade. The primary use of RS is predicting preference of user over an item (book, movie, news article, music, DVDs or webpage) based on their history. These systems make use of user's feedback and information to improve their suggestions in future. For example, Amazon analyses that if a large number of users buy a product, A also buy product B then it starts to recommend product B for a new user who bought product A.
Movie recommendation has become most widely used and popular information filtering system. It helps the user to get his/her right multimedia content. Movie recommendation systems leverage user history and feedbacks to predict and recommend new movies from a huge movie library. Figure 1 depicts a typical movie recommendation system.

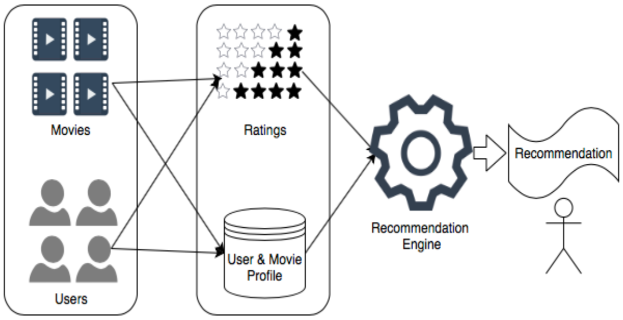


Fig 1: Movie Recommendation System

Extensive work has been done in the field of movie recommendation in recent time especially after the Netflix prize competition. Researchers have proposed new and enhanced filtering algorithms to get better movie recommendations. These techniques are primarily categorized as content-based, collaborative and hybrid techniques. The majority of existing systems followed collaborative filtering as their state of the art[1][2] and classified as neighbourhood-based, model-based and hybrid collaborative filtering. The primary advantage of collaborative filtering is that it does not need any domain-specific knowledge to generate a prediction. However, these

systems suffer from some limitations too like cold start, new user problem, sparsity and scalability. Some techniques have proposed to develop model-based Collaborative Filtering techniques to overcome above-stated limitations.[3][4][5]
This paper is an attempt to present a review of content-based and collaborative filtering techniques for movie recommendation system. We have shown the implementation of two fundamental techniques (Content based and Collaborative Filtering) of information filtering with their pros and cons in the field of movie recommendation. Rest of this paper organized as section 2 discusses the literature of recommendation system using content-based and collaborative filtering. Section 3 presents an experimental model and result part, and section 4 discusses conclusion and future scope of the work.

II. BACKGROUND AND RELATED WORK

Recommendation systems are to predict preference of user over an item using user history, item profile and user item interaction details. Recommendation system techniques mainly categorised in three classes as content-based filtering(CBF): Recommendations are based on the similarity between the contents of target item and the items liked by user in the past, collaborative filtering(CF): leverages the ratings or feedbacks of other similar users to target user for recommendation and hybrid filtering: combines the best practices of content-based and collaborative filtering to improve accuracy of prediction.

A. Content-Based Filtering

Content-based filtering recommendation systems match users to items which are similar to those liked or preferred by the user in the past. The similarity is computed based on the attributes of the item, and other users play little role in the recommendation process. Content-based systems use a different source of information (item description and user profile) and process them to provide recommendations. The content-based system includes pre-processing and user profile learning(offline) and online prediction components. The offline component is generally a classification or regression model which then used to generate online predictions.

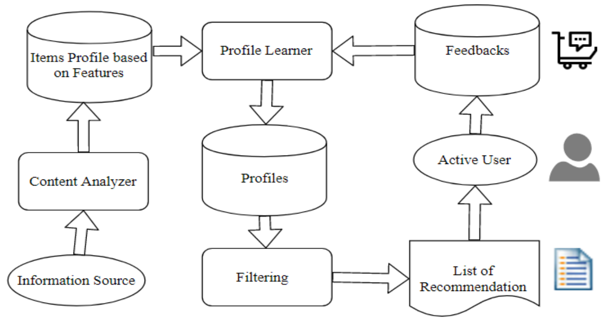


Fig 2: Framework for Content-Based Filtering

The steps of content-based recommendation system are as follows:

1) *Feature extraction and preprocessing:* Item features are extracted from different sources and converted into keyword based vector representation. This step is highly domain specific and needs to be handled carefully because a proper extraction of feature information is essential for effective functioning of a recommender system.[6] For example, a movie feature vector may comprise a different set of genre, actor list, director etc. A MovieGenre matrix can be constituted using binary value where 1 shows the presence of genre and 0 otherwise. Refer figure 4 to get MovieGenre matrix.

2) *User Profile Generation:* Content-based systems are user specific, and therefore a user-specific profile has to be created to identify user interests over items. This model can be created using his/her browsing history (implicit feedback) or using an explicit feedback system(rating). These feedbacks are then used with item attributes to learn user inclination toward item attributes or to generate a user profile. This step is very much similar to classification or regression modelling.[6] For example, in Movie recommendation system a user to genre inclination matrix is formed using MovieGenre and rating matrix to generate a user profile. (Refer figure 4 UserGenre Matrix)

3) *Recommendation:* This is the final step of content-based filtering and predicts the recommendations from the model learned in last stages. This step is online, and predictions are generated in real time.[6] To predict recommendations of user preference toward a movie, the user profile is mixed with Movie genre list and a weight vector. Prediction formula can be written as:

$$P_{u,i} = \sum_{i=1}^{\#\,of\,Genre} w * UserGenre_u * MovieGenre_i$$

Where w, represents the genre weight vector in documents.

B. Collaborative Filtering (CF):

Collaborative filtering has become state of the art recommendation system in recent time. Many commercial companies like Amazon, YouTube, IMDB, MovieLens and Netflix are making great use of these techniques to provide better and effective recommendations. The working principle of collaborative filtering is based that if two users share a common interest in past are more likely to have same preferences in future too.[7][8] For example, let two users A and B share common preferences then they tend to have similar ratings for items (similarity can be calculated using any similarity metric), and it is very likely that the rating in which the only one of them has specified is also expected to be same for another user.

The input of collaborative filtering methods is the user-item rating matrix and can be classified as memory based and model-based collaborative filtering. These techniques differ in how they process the input rating matrix. Memory-based methods are also known as neighbourhood-based the technique makes use of entire rating matrix and predicts preference of user over item by use of similarity measures whereas model-based methods make use of data mining and machine learning approaches to generate prediction[9][10].

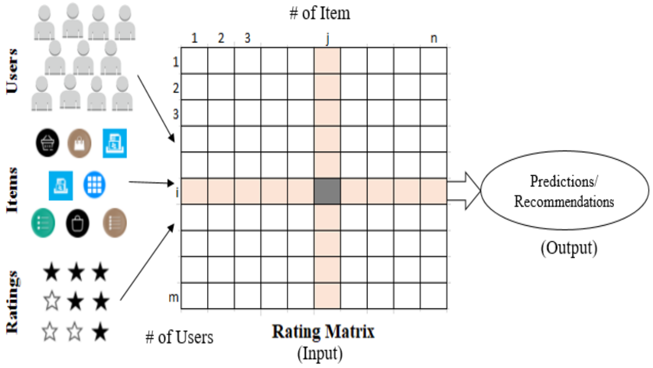
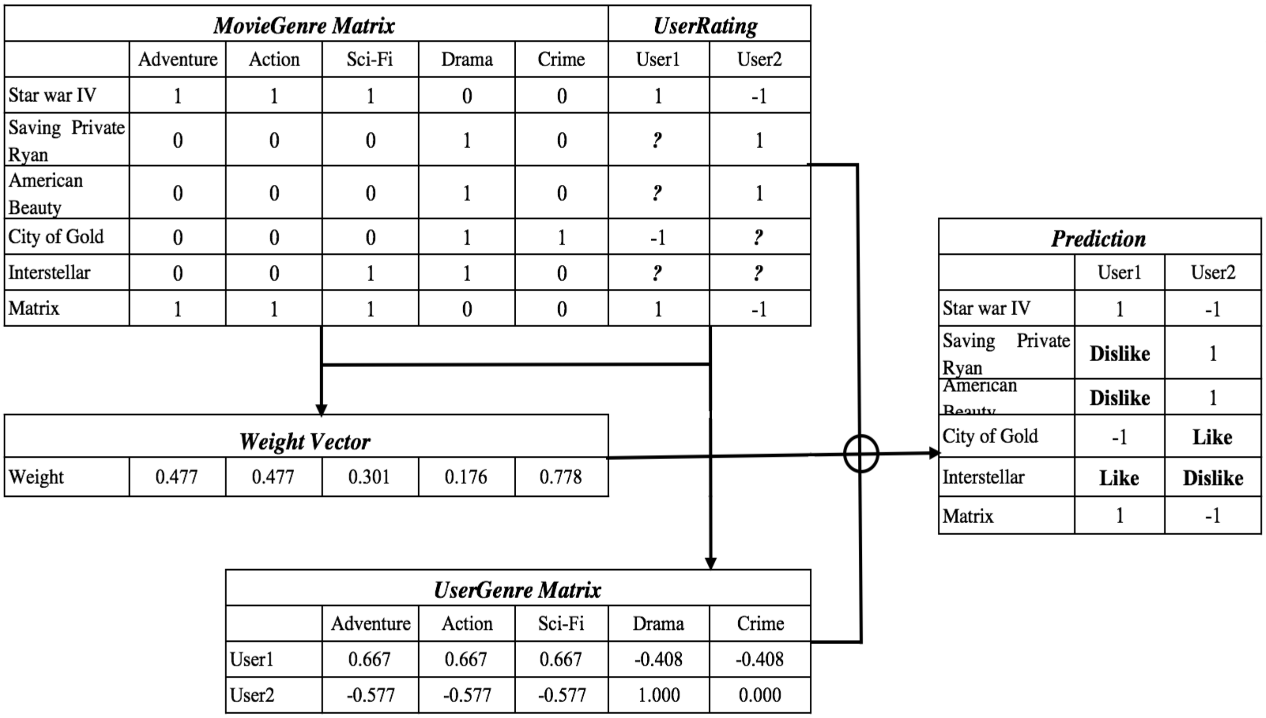


Fig 3: Framework for collaborative filtering



	MovieGenre Matrix					*UserRating*	
	Adventure	Action	Sci-Fi	Drama	Crime	User1	User2
Star war IV	1	1	1	0	0	1	-1
Saving Private Ryan	0	0	0	1	0	?	1
American Beauty	0	0	0	1	0	?	1
City of Gold	0	0	0	1	1	-1	?
Interstellar	0	0	1	1	0	?	?
Matrix	1	1	1	0	0	1	-1

	Weight Vector				
Weight	0.477	0.477	0.301	0.176	0.778

	UserGenre Matrix				
	Adventure	Action	Sci-Fi	Drama	Crime
User1	0.667	0.667	0.667	-0.408	-0.408
User2	-0.577	-0.577	-0.577	1.000	0.000

	Prediction	
	User1	User2
Star war IV	1	-1
Saving Private Ryan	**Dislike**	1
American Beauty	**Dislike**	1
City of Gold	-1	**Like**
Interstellar	**Like**	**Dislike**
Matrix	1	-1

Fig 4: Content-Based Movie Recommendation System

This paper is limited to discuss only memory-based techniques. Memory-based techniques further classify as user based and item based collaborative filtering. User-Based Filtering technique computes the similarity between users by comparing their preference over the same item and calculates the predicted preference for items for the target user whereas Item Based Filtering techniques compute predictions using similarity between items. The technique works by retrieving all the items rated by a target user and determine similarities of retrieved items with target item. [9] Let the model is defined as the rating matrix of order m×n with m users (u_1, u_2, …u_m) and n movies(i_1, i_2,…..i_n). Each cell $r_{u,i}$ of matrix ins the rating specified by user u over movie i in the range of let 0 to k. Each user has a list of movies m for which he/she has shown interest by specifying a rating. Let the following table shows a snapshot of user movie rating matrix.

TABLE 1: User Movie Rating Matrix

	m_1	m_2	m_3	m_4	m_5	m_6
u_1	4	**?**	3	5	**?**	4
u_2	1	2	**?**	2	1	**?**
u_3	3	5	**?**	4	**?**	5
u_4	4	3	2	**?**	4	5
u_5	1	**?**	3	1	2	2

Collaborative filtering steps may be summarized as:

1) *Compute Similarity Metric:* This is the first step of neighbourhood-based filtering, in order to identify the neighbours of target user his/her similarity to all other users is computed. Various similarity metrics are used to define similarity among users. This is a tricky step as different users have different taste of preferences and the may have rated different set of movies. The two famous similarity metrics used in neighbourhood-based filtering approach are:

Pearson Correlation Similarity: Pearson correlation defines the linear correlation between two vectors and has a value between -1 to 1. The similarity between the two vectors u and v is defined as:[1]

$$s_{Pearson}(u,v) = \frac{\sum_{i=1}^{n}(r_{u,i}-\bar{r_u})\times(r_{v,i}-\bar{r_v})}{\sqrt{\sum_{i=1}^{n}(r_{u,i}-\bar{r_u})^2\times\sum_{i=1}^{n}(r_{v,i}-\bar{r_v})^2}}$$

Where r_u is mean rating of the user

Cosine Similarity: Cosine is one of the most popular methods of statistics to find similarity between two non-zero real values vectors. It looked for an angle between two vectors in n-dimensional space and defined as:[1]

$$s_{Cosine}(u,v) = \frac{\sum_{i=1}^{n}(r_{u,i}).(r_{v,i})}{\sqrt{\sum_{i=1}^{n}(r_{u,i})^2\times\sum_{i=1}^{n}(r_{v,i})^2}}$$

For Example, Table2 shows the Pearson similarity measure for rating matrix shown in Table 1.

TABLE 2: Pearson Similarity Metric

	u_1	u_2	u_3	u_4	u_5
u_1	1.000				
u_2	0.707	1.000			
u_3	-0.169	0.683	1.000		
u_4	0.739	-0.980	0.038	1.000	
u_5	-0.851	-0.101	0.794	-0.581	1.000

2) *Selection of K neighbours*: This is the second step of Collaborative filtering. In this step, a subset of users is selected with higher similarity to the target user. For this, the similarity metric is rearranged in descending order to choose top K neighbour of the target user. For example, from above metric top 2 neighbours of user u_1 will as shown in the following table.

TABLE 3: Selecting top-2 Neighbours

	u_4	u_2	u_3	u_5
u_1	0.739	0.707	-0.169	-0.851

3) *Prediction and Recommendation:* Prediction or recommendation is the final step where an active user might have appraised an item not rated yet or recommended by some top k items, which might be liked by the active user. The prediction can be calculated by the following equation.

$$P_{a,i} = \bar{r_a} + \frac{\sum_{u=1}^{n} S(a,u)\times(r_{u,i}-\bar{r_u})}{\sum_{u=1}^{n}|S(a,u)|}$$

Where S (a, u) is similarity of active user to user u. Table 4 shows the prediction of users over items. High rating value items may be recommended to respective users.

TABLE 4 Prediction Metric

	m_1	m_2	m_3	m_4	m_5	m_6
u_1	4	**3.94**	3	5	**3.96**	4
u_2	1	2	**0.99**	2	1	**1.86**
u_3	3	5	**4.89**	4	**4.12**	5
u_4	4	3	2	**4.53**	4	5
u_5	1	**2.40**	3	1	2	2

C. Content-Based Vs Collaborative Based:

Collaborative filtering methods are completely based on user interest and do not require any other domain knowledge whereas content-based methods make use of item attribute similarity to predict preference of a user over the item. Both techniques have their pros and cons in different situations and are summarized in table 5.

TABLE 5: Comparison between Content-Based and Collaborative Filtering

Method	Advantages	Disadvantages
Content-Based Filtering	• Provide user independence with exclusive ratings.	• Limited content Analysis. • Difficult to generate the attribute for items in

	• Adequate to new items. • Provide explanation how recommendation works. • Quality improves over time.	certain cases. • New user problem
Collaborative Filtering	• Easy to implement. • No domain knowledge required. • New data can be added easily. • Scalable with co-rated data. • Model based techniques improve prediction accuracy.	• Cold start- User preferences are needed. • Scalability: Huge amount power is required to process billions of data values. • Sparsity-Performance decreases with sparsity.

III. EXPERIMENT MODEL AND EVALUATION

The input of any recommendation system is rating matrix or the user and item profile data, and the task is to predict missing data in rating matrix. Our experimental model is shown in figure 5. We first pre-process the data set as per the need for filtering techniques and divide it to in training data and testing data. Then after recommendation algorithm is applied to the training data to get a learned recommendation model. The performance of the learned model is tested with the help of testing data. To validate our implemented model we used MovieLens 100K dataset[11], which is further divided into training and a testing dataset of 75% and 25% of each respectively. We have implemented Content-based and Neighborhood-based Collaborative Filtering (user-user based and item-item based) for recommendation algorithms.

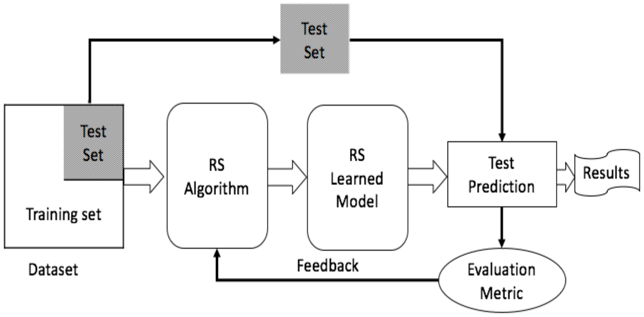


Fig 5: Experimental Model

There are different metrics used to test the performance of recommendation system. In this paper, we use Mean Absolute Error(MAE) and Root Mean Square Error(RMSE) which are the most popular evaluation metric after the Netflix prize competition.[7] These metrics are defined as:

MAE: Mean Absolute Error is the average of the absolute difference between the predictions and actual values.

$$MAE = \frac{1}{N}\sum_{i=1,}^{m}\sum_{j=1}^{n}\left|r_{i,j}-\widehat{r_{i,j}}\right|$$

RMSE: Root Mean Square Error computed by the square root of the average of the difference between predictions and actual values. Lower the RMSE is better the recommendation.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1,}^{m}\sum_{j=1}^{n}(r_{i,j}-\widehat{r_{i,j}})^{2}}$$

Results: This paper implements the two fundamental approach to information filtering for movie recommendation named as content-based filtering and collaborative filtering. We used MovieLens 100K dataset to perform the experiment and evaluate the performance of two filtering techniques. This dataset consists a set of 943 users over 1682 movies. Users have rated movies on a scale of 1 to 5. Dataset consists of 100000 rating with every user must rate at least 20 movies. We followed the experiment model as shown in figure 5. The researcher has proposed various similarities for neighbourhood-based collaborative filtering, but this paper is only limited to Cosine and Pearson coefficient similarity. The following table shows the summary of results obtained for the experiment over evaluation parameters.

TABLE 6: Results

Method		Similarity	MAE	RMSE
Content based Filtering(CBF)		Weighted	0.987	1.253
Collaborative Filtering(CF)	User-User based	Cosine	0.818	1.023
		Pearson	0.805	1.009
	Item-Item Based	Cosine	0.792	0.992
		Pearson	0.786	0.985

Following figure 6 shows the MAE values obtained for each method, it is clear from the figure the best MAE value is 0.786 has been obtained for Item-Item based collaborative filtering using Pearson similarity. The Second best MAE value is 0.792 is also for Item-Item based Collaborative filtering using cosine similarity. With these result, one can say that Item-Item neighbourhood-based collaborative filtering shows better performance among the three implemented techniques.

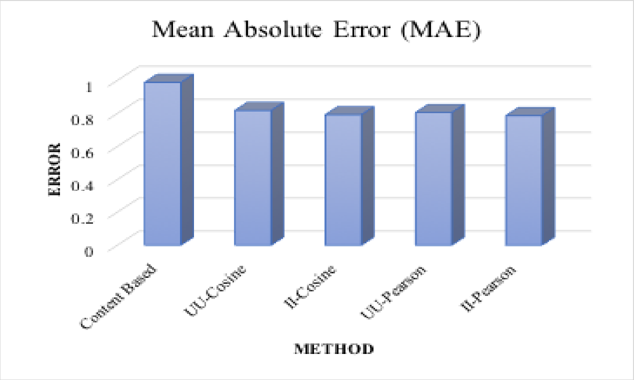


Fig 6: Comparison of Mean Absolute Error

Figure 7 Shows the RMSE values obtained for each implemented method in the experiment. From the figure, it is clear that Item-Item based Neighborhood collaborative filtering outperformed the other two implemented techniques of filtering by achieving minimum RMSE among all implemented methods.

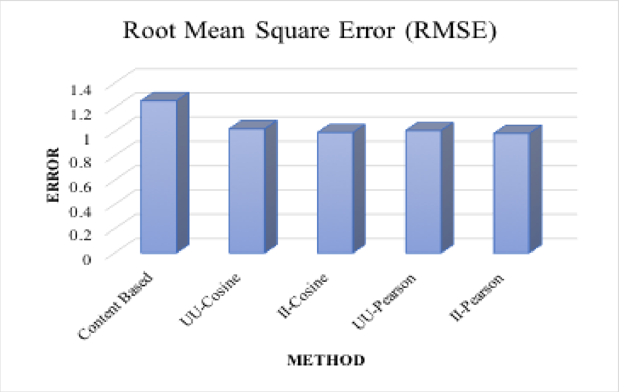


Fig 7: Comparison of Root Mean Square Error

IV. Conclusion and Future Scope

The study in this paper aimed to show the use of filtering techniques in movie recommendation. This paper demonstrates the implementation of two fundamental techniques of information filtering. The experiment was conducted using MovieLens 100K dataset, where the obtained results are showing that Item-Item neighbourhood techniques are better as compared to User-User based and Content-based filtering method.

Researchers aim to build a recommendation system with higher prediction accuracy. This paper has studied the basic techniques of filtering for movie recommendation with their pros and cons. A new technique can be suggested by using a hybrid version of studied two techniques with some explicit and implicit improvement in the system.

References

[1] G. Adomavicius and a Tuzhilin, “Toward the Next Generation of Recommender Systems: a Survey of the State of the Art and Possible Extensions,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[2] G. Linden, B. Smith, and J. York, “Amazon.com recommendations: Item-to-item collaborative filtering,” *IEEE Internet Comput.*, vol. 7, no. 1, pp. 76–80, 2003.

[3] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, “Item-based collaborative filtering recommendation algorithms,” *Proc. tenth Int. Conf. World Wide Web - WWW ’01*, pp. 285–295, 2001.

[4] J. Zhang, Y. Lin, M. Lin, and J. Liu, “An effective collaborative filtering algorithm based on user preference clustering,” *Appl. Intell.*, vol. 45, no. 2, pp. 230–240, 2016.

[5] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl, *Application of Dimensionality Reduction in Recommender System - A Case Study*, vol. 1625. 2000.

[6] C. C. Aggarwal, “Content-Based Recommender Systems,” in *Recommender Systems*, 2016, pp. 139–166.

[7] F. Cacheda, V. Carneiro, D. Fernández, and V. Formoso, “Comparison of collaborative filtering algorithms,” *ACM Trans. Web*, vol. 5, no. 1, pp. 1–33, 2011.

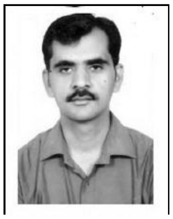
[8] J. Bobadilla, F. Ortega, and A. Hernando, “A collaborative filtering similarity measure based on singularities,” *Inf. Process. Manag.*, vol. 48, no. 2, pp. 204–217, 2012.

[9] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, “Recommendation systems: Principles, methods and evaluation,” *Egypt. Informatics J.*, vol. 16, no. 3, pp. 261–273, 2015.

[10] X. Su and T. M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” *Adv. Artif. Intell.*, vol. 2009, no. Section 3, pp. 1–19, 2009.

[11] F. M. Harper and J. A. Konstan, “The MovieLens Datasets: History and Context,” *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, p. 19:1--19:19, 2015.

Authors Profile



Sandeep K. Raghuwanshi is Phd Scholar in CSE department at Maulana Azad National Institute of Technology Bhopal. He is also assistant professor in CSE department at Samarat Ashok Technological Institute Vidisha and has teaching experience of 13 years. He has published more than 10 papers in international journals and conferences in the area of cloud computing, information security and machine learning. His research field is primarily concentrated on information Retrieval and Machine learning techniques.

R. K. Pateriya is an associate professor in CSE department at Maulana Azad National Institute of Technology Bhopal. He received PhD in year of 2011 and has 24 years of teaching experience in field of computer science at Institute of National Importance. His research work has been published in various reputed journals and conferences which include IEEE, Scopus and Web of Science Index. His current research includes information security, cloud computing data mining and information retrieval.