

Design and Implementation of Collaborative Filtering Approach for Movie Recommendation System

Anshu Sang

Gyan Ganga Institute of Technology and Sciences
Jabalpur, India

Santosh K. Vishwakarma

Gyan Ganga Institute of Technology and Sciences
Jabalpur, India

ABSTRACT

Recommendation systems play a significant role in the user life that provides information filtering from enormous data to the user specific data for decision making. Recommendation system mainly deals with the similarity among objects (items). The collaborative filtering is a recommendation technique that contains a list of rating that the previous user has already given for an item. Using the rating and similarity among the two users, the system recommends an item to the user for the decision making. In the system, separate the movies data set into unrated and rated sample set. Then rated movies dataset is treated as training data with the help of "Item K-NN", builds a model. The model is applied to the unrated movies dataset which is treated as testing data set. At last the order list of unrated movies is created on the basis of distance Pearson and Cosine formula.

Keywords

Collaborative filtering, Recommender system, Item K-NN (Item recommendation), Item K-NN (Item Rating Prediction)

1. INTRODUCTION

Nowadays, the enormous quantity of data provided by internet suppresses the ability of information processing. The stupendous data needs more complex data processing. People get confused by number of items viewed in a websites, journals that are released every year. Today, all the work is done online. Due to this, lots of information is generated on the internet. For finding adequate information from such tremendous data is users' biggest need. So dealing with loads

of information and selection of user specific data is a cumbersome task. E commerce is an emerging field from last few decades. E commerce websites needs different technologies to refer their product to the end user; for example: food, clothing, gadgets, movies, books etc.[7] Recommendation system is one of the technology for the referral purpose to satisfy the customer need. These recommendation systems have a wide area of success in the personalized recommendation to the end user. Recommendation system facilitates decision making process through information assistance, enhance the user experience, and generate revenue for creator. In decision making process, users' experience with item of interest such as users' choice, ranking, scores, tags etc are used as an input data and are converted into ordered list and scores as an output, the list is contains rank among different items.[12] The rank are generated among items with the help of different similarity formulas . These lists are used by the system to recommend an item to the end user.

There are many ecommerce websites such as Amazon, Flipkart, Quickr, Netflix, Last.fm etc using different recommendation algorithm but the main motive behind all website is to recommend product to the end user. These recommendation systems are used in different application that works to gather information from the user.

In the past few decades, different recommendation techniques have been used. The recommendation system is broadly classified into 3 categories: 1. Collaborative filtering 2. Content based filtering 3. Hybrid filtering.

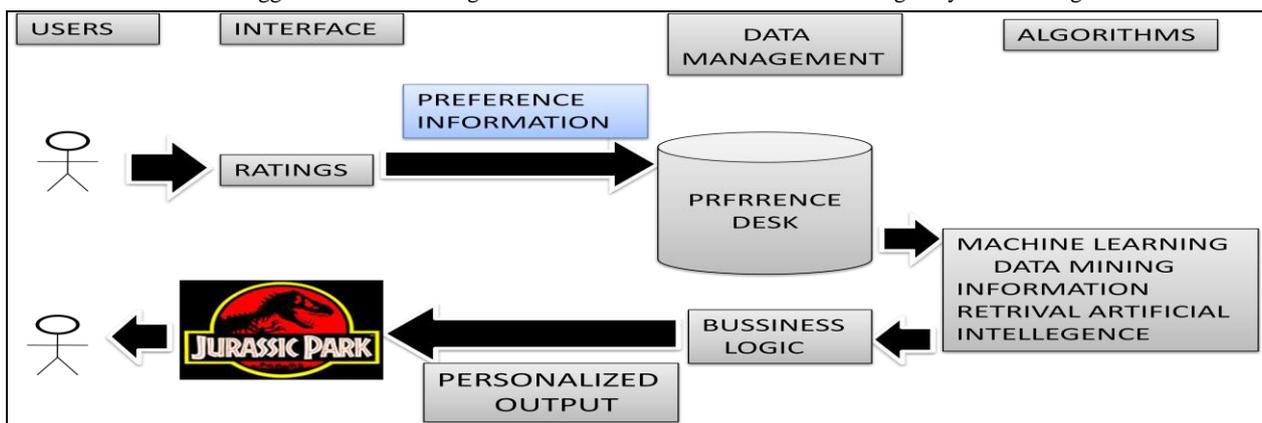


Figure 1: Workflow of recommendation system.

Collaborative filtering was implemented by Xerox PARC Institute called "Tapestry".[7] The collaborative filtering recommendation system uses the users' past experience to recommend a product. It's main emphasis is on users' similar behavior with other users. Collaborative methodology takes an item user matrix on the basis of user viewing and preferences as an input and generate numerical value(ordered

list) shows the like and dislike of the user .The generated list does not contain the item that user has already consumed. These generated lists are used by the system for recommending an item. Collaborative filtering technique is further divided into two parts: User based collaborative filtering, Item based collaborative filtering. User based collaborative filtering uses the rating of the item by the user

and compares it with the other user rating on similar item. This makes the system instinctive prediction of the user interest by combining rating from different user having similar taste and after that determines similarity between users. The main problem of this system is scalability. Item based collaborative filtering can resolve this problem. It follows the same steps as User based filtering; but the main difference between them is rating of the item is calculated by users who used to provide it in the past.

Content based filtering emphasis on user feedback. The method is based on specification of item and profile of user's rating. In the content based recommendation system, it emphasis on users' likes and show the item relevant to the user preference and at the same time it deemphasized the user dislike.[6] The main problem with this method is that the system is limited to the recommended content of similar type of item that user has already used. Pandora is an example of content based filtering technique.

Hybrid is combination of best feature of collaborative filtering and content based filtering. Hybrid recommendation method can be applied in many ways: by combining both the methods and form a single model, apply separately both the methods, apply first collaborative filtering then content based and vice versa. Several studies conclude that hybrid recommendation is more reliable and accurate than other recommendation technique. By this approach, some problems are resolved such as cold start and data sparsity. Netflix movie recommendation system work on hybrid recommendation system.

Rapid Miner is an open source software platform which provides an integrated environment for machine learning, data mining, text mining, predictive analytic and business analytic. Rapid Miner functionality can be extended through plugins that are available from rapid miner market place. Rapid Miner makes recommendation technique easy to understood, provides efficiency and accuracy with the recommendation extension. Rapid Miner makes the developers easy to implement the task without any overhead of program codes. Recommendation extension has 26 operators. Recommendation operator can recommend an item by applying the item recommendation operator and item rating recommendation operator on the basis of user requirement.

2. LITERATURE REVIEW

Zhihang Tang and Zhonghua Wen [12] using recommendation system for decision making process through information assistance. Personalised recommendation system uses the information as input data and converts it into ordered list and scores as an output. The list and scores are used by the creator for decision making. The technique used in this paper is collaborative filtering, in which users' past behaviour predicts the future recommendation of items and also find other user with similar taste to the active user. The user item matrix using k-NN classification to determine the similar pair of item between the 2 users or pair of item that viewed both users. The main disadvantage of using collaborative filtering is it required more allocation of space for user need and a lot of items are unexplored.

Puntheeranurak and Chaiwitooanukool [7] proposed innovative algorithm that improves efficiency for items based collaborative filtering technique. In the paper 3 methods are used out of which two already exist and one is derived from them. They are: 1. Item rating similarity method 2. Item attributes similarity method 3. Item hybrid similarity method. The author abstract data set from MovieLens website. For

calculating the accuracy of prediction it uses Mean Absolute Error to determine the rating error between predicting rating and actual rating. The lower MAE show more accurate the prediction is permitted to provide better recommendations. The result of this paper shows that the accuracy of proposed algorithm is better than traditional and demography based collaborative filtering.

Ponnam et al [6] uses the item based collaborative filtering for large scale data for determining the User Item Rating Matrix for finding the relationship between various items and use the relationship to predict the recommendation of an item to the user. Netflix data is available in Group Lens; are used in this paper. It calculates the other items similarity to the current item by computing the weight by cosine similarity, with the help of weight, the selection of similar neighbour item which helps in predicting the unknown rating which is used to recommend top N items.

Isinkaye et al [4] explore the different characteristics and potential of the different prediction technique in recommendation technique. In this work, they have further added different phases of recommendation to predict an item. It also includes different filtering techniques with their pros and cons and explains each technique with proper example. The paper defines a different evaluation metric for determining the accuracy of the predicting values.

Tiwari et al [1] presented blistering development of internet technologies; due to this numerous of websites are available for recommending a book that creates the competition among them. The paper explore the collaborative filtering and association mining to recommend a book on the basis of student requirement i.e., Publisher and price. In the present system, association rule is used to determine the profile of a user compare the user profile with the target user which has at least 5 years membership with the existing system. Collaborative filtering user based recommendation system is used to find the rating of the target user on the basis of prize and publication to recommend the book to the user.

3. DATASET

The Dataset used in the present system is Student Movie Rating. It is a collection of 390 rating as preferences that are given by 30 User over 13 Movies. The rating range on the scale of 1-5. In the dataset, there are 3 column names as:

1. User: This field defines the User id of different user that rate movies.
2. Movies: This field defines the name of different movies that are Rated or Unrated by the user.
3. Ratings: This field defines the rating given by the user to a particular movie on the scale of 1 to 5.

STUDENT	MOVIES	RATING
USER1	SPIDERMAN1	3
USER1	SPIDERMAN2	4
USER1	SPIDERMAN3	2
USER1	STARWAR1	1
USER1	STARWAR2	4
USER1	CONJURING1	2
USER1	CONJURING2	3
USER1	MERMAID	4
USER2	SPIDERMAN4	3
USER2	SPIDERMAN5	4
USER2	SPIDERMAN6	1
USER2	STARWAR3	3
USER2	STARWAR4	4
USER2	CONJURING1	2
USER2	CONJURING2	5
USER2	MERMAID	2
USER3	SPIDERMAN1	3
USER3	SPIDERMAN2	3
USER3	SPIDERMAN3	2
USER3	STARWAR1	4
USER3	STARWAR2	5
USER3	CONJURING1	1
USER3	CONJURING2	5
USER3	MERMAID	2

Figure2: Dataset of rated movies.

STUDENT	MOVIE	RATING
USER1	SPIDERMAN4	?
USER1	SPIDERMAN5	?
USER1	SPIDERMAN6	?
USER1	STARWAR3	?
USER1	STARWAR4	?
USER2	SPIDERMAN1	?
USER2	SPIDERMAN2	?
USER2	SPIDERMAN3	?
USER2	STARWAR1	?
USER2	STARWAR2	?
USER3	STARWAR1	?
USER3	STARWAR2	?
USER3	CONJURING1	?
USER3	CONJURING2	?
USER3	MERMAID	?
USER4	SPIDERMAN2	?
USER4	STARWAR1	?
USER4	STARWAR2	?
USER4	CONJURING1	?
USER4	SPIDERMAN6	?
USER5	SPIDERMAN3	?
USER5	STARWAR1	?
USER5	STARWAR2	?
USER5	CONJURING1	?

Figure 3: Dataset the Unrated movies

4. METHODOLOGY

In the present system, enumerated the theoretical analysis of the method that is: collaborative item based recommendation .In this work, Rapid Miner ver 5.0 with extension of text mining and recommendation system are used . Rapid miner workflow approach entices systematic research and facilitates its implementation in combination with Rapid Analytics. In Rapid Miner, the recommendation extension has 26 recommendation operators. These operators are categorized into the following groups: Item Recommendation, Item Rating Prediction and Recommendation performance.

In the present system the method used is Item k-NN with Item Recommendation and Item Rating Prediction.

The two most popular similarity measures are Pearson correlated-based and cosine-based. Pearson correlation coefficient is used to measure the extent to which two variables linearly relate with each other and is defined as

$$S(a, u) = \frac{\sum_{i=1}^n (r_{ai} - \bar{r}_a)(r_{ui} - \bar{r}_u)}{\sqrt{\sum_{i=1}^n (r_{ai} - \bar{r}_a)^2} \sqrt{\sum_{i=1}^n (r_{ui} - \bar{r}_u)^2}} \quad (1)$$

From the above equation, S(a,u) denotes the similarity between two users a and u, r_{ai} is the rating given to the item i by user a. r_{ui} is the rating given to item i by user u. \bar{r}_a is the mean rating given by user a while n is the total number of items in the user-item space.

The cosine-based measure is widely used in the fields of information retrieval and text mining to compare two text documents. In this case, documents are represented as vectors of terms. The similarity between two items u and v can be defined as:

$$S(u, v) = \frac{\bar{u} \cdot \bar{v}}{|\bar{u}| |\bar{v}|} = \frac{\sum_i r_{ui} r_{vi}}{\sqrt{\sum_i r_{ui}^2} \times \sqrt{\sum_i r_{vi}^2}} \quad (2)$$

From the above equation, S(u,v) is the similarity between two item u, v. r_{ui} is the rating given by item u. r_{vi} is the rating given by item v.

In the present system, the following steps are taken for Item Recommendation .They are:

1. Separate the dataset into rated and unrated examples.
2. Build the model using Item k-nn (item recommendation) on rated subset.
3. Apply the model on the unrated subset.

The following operators are used in the system such as Retrieve, Set Role, Item K-NN, and Apply Model.

In the Item Rating Prediction, same above process is applied except Item k-nn(item rating prediction) is used in place of Item k-nn (item recommendation).

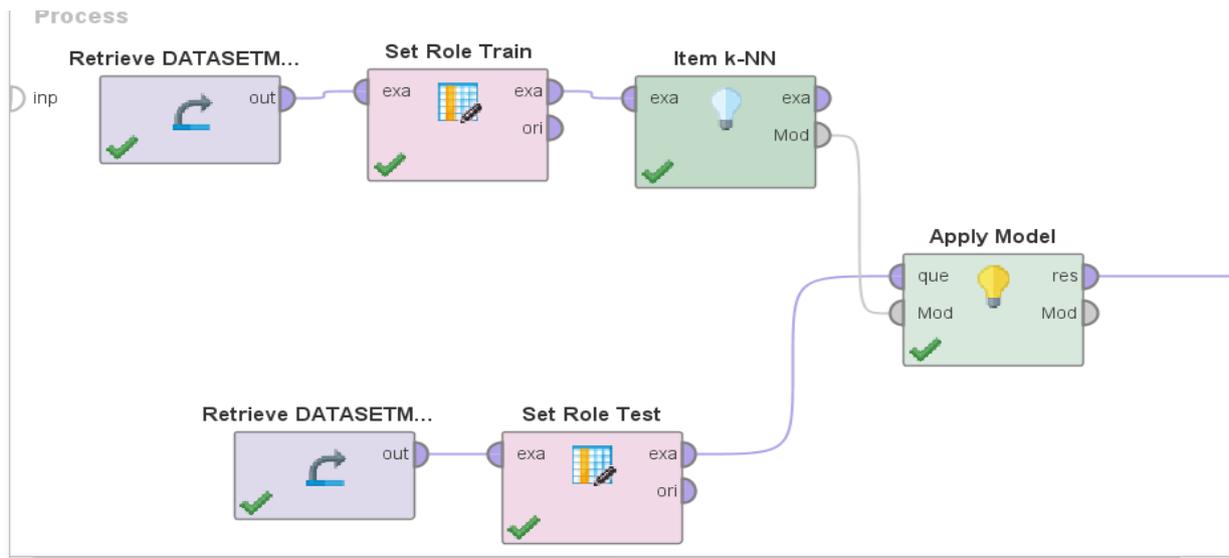


Figure4: Shows workflow Item k-nn (Item Recommendation)

The figure 4 shows Item k-nn (item recommendation): Unweighted k-nearest neighbor item-based collaborative filtering to predict unrated movies on the basis of rated movies using cosine similarity.

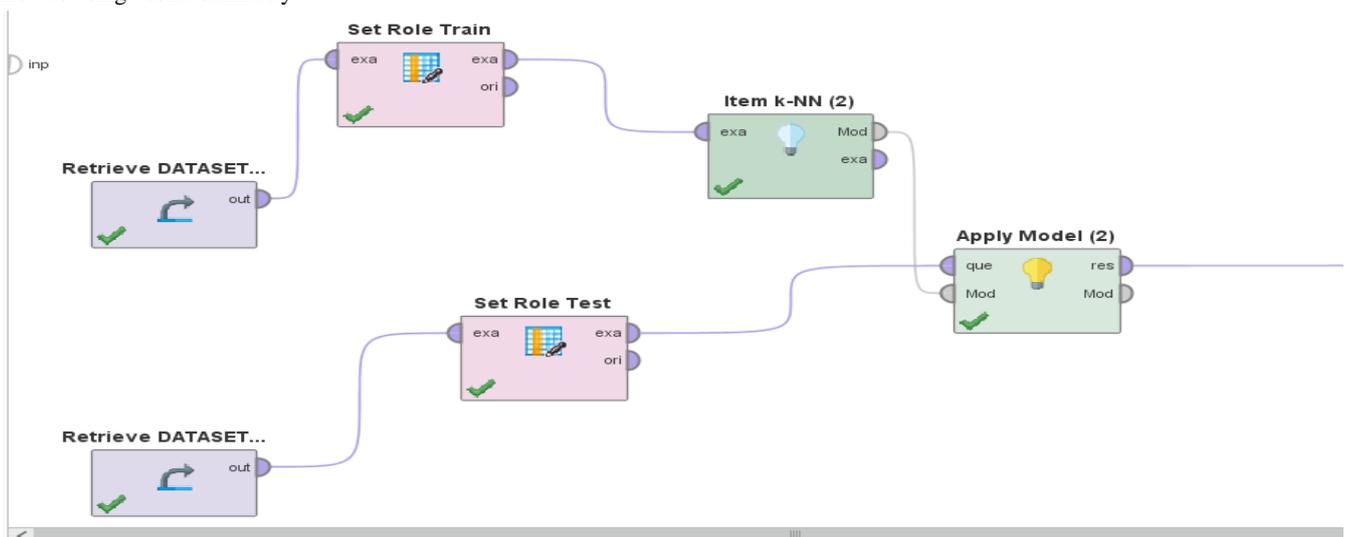


Figure 5: Shows the workflow of Item k-nn(item rating prediction).

In the figure 5 shows Item k-nn(item rating prediction) weighted item-based KNN is used to predict the rating of unrated movies with cosine/Pearson similarity. The ranking is provide to the unrated movie with help of predicted value which is generated in this process.

5. RESULT ANALYSIS

In order to demonstrate the proposed method, implementation of collaborative filtering item

recommendation and collaborative filtering item rating prediction. In this work, User Rating Dataset that contains Rated and Unrated Movies. The rated movies are treated as training set and an Unrated as testing set. In item recommendation, the ranking is given to the different item on the based on user id and item id. In item rating prediction, the value generated by the system which can be used to predict the rating of different items.

Row No.	user_id	item_id	rank
1	0	12	1
2	0	11	2
3	0	10	3
4	0	9	4
5	0	8	5
6	1	4	1
7	1	3	2
8	1	2	3
9	1	1	4
10	1	0	5
11	2	12	1
12	2	11	2
13	2	10	3
14	2	9	4
15	2	8	5
16	3	10	1
17	3	5	2

Figure 6: Shows the result of Item recommendation provides ranking of the items

Figure 6 shows the result of item recommendation. There are 3 columns: User_id defines user from User1 to User30; contains 30 users such as: User1 is shown by 0, User2 is shown by 1 and so on. Item_id defines Movie name from

Spiderman1 to Mermaid contain 13 movies such as: Spiderman1 is shown by 0; Spiderman2 is shown by 1 and so on. The rank defines rating given to the Unrated Movies on the range 1 to 5.

Row No.	RATING	STUDENT	MOVIE	prediction
1	?	USER1	SPIDERMAN4	2.806
2	?	USER1	SPIDERMAN5	2.719
3	?	USER1	SPIDERMAN6	3.049
4	?	USER1	STARWAR3	3.572
5	?	USER1	STARWAR4	3.074
6	?	USER2	SPIDERMAN1	2.811
7	?	USER2	SPIDERMAN2	2.814
8	?	USER2	SPIDERMAN3	3.321
9	?	USER2	STARWAR1	3.391
10	?	USER2	STARWAR2	3.341
11	?	USER3	STARWAR1	3.060
12	?	USER3	STARWAR2	3.314
13	?	USER3	CONJURING1	2.929
14	?	USER3	CONJURING2	2.781
15	?	USER3	MERMAID	2.633
16	?	USER4	SPIDERMAN2	3.324
17	?	USER4	STARWAR1	3.516

Figure 7: Shows the result of the item rating prediction using the cosine formula.

Figure 7: shows the result of Item Rating Prediction of different unrated movies using Cosine distance formula. From this result, we can rate the items on the basis of their predicted values in the range of 1 to 5. In Figure 9, the rating is

provided to the unrated movies with the help of predictive values. For ex – The predicted value of User1 where value is higher, is rated as 1 and the lower value is rated as 5. The other value lies between range of 1-5.

Row No.	RATING	STUDENT	MOVIE	prediction
1	?	USER1	SPIDERMAN4	3.039
2	?	USER1	SPIDERMAN5	3.063
3	?	USER1	SPIDERMAN6	1.442
4	?	USER1	STARWAR3	2.946
5	?	USER1	STARWAR4	3.749
6	?	USER2	SPIDERMAN1	1.622
7	?	USER2	SPIDERMAN2	4.397
8	?	USER2	SPIDERMAN3	2.613
9	?	USER2	STARWAR1	3.152
10	?	USER2	STARWAR2	3.148
11	?	USER3	STARWAR1	2.988
12	?	USER3	STARWAR2	2.442
13	?	USER3	CONJURING1	2.295
14	?	USER3	CONJURING2	3.110
15	?	USER3	MERMAID	3.948
16	?	USER4	SPIDERMAN2	2.812
17	?	USER4	STARWAR1	3.112

Figure 8: Shows the result of item rating prediction using Pearson formula.

Figure 8: shows the result of Item Rating Prediction of different unrated movies using Pearson distance formula. From this result, we can rate the item on the basis of their predicted values in the range of 1 to 5. In Figure 9, the rating

is provided to the unrated movies with the help of predictive values. For ex – The predicted value of User1 is higher rated 1 and the lower value rated as 5. The other value lies between range of 1-5.

PEARSON					COSINE				
ROW NO	STUDENT	MOVIE	PREDICTION	RATING	ROW NO	STUDENT	MOVIE	PREDICTION	RATING
1	USER1	SPIDERMAN4	3.0391356	3	1	USER1	SPIDERMAN4	2.806442902	4
2	USER1	SPIDERMAN5	3.06290672	2	2	USER1	SPIDERMAN5	2.719084778	5
3	USER1	SPIDERMAN6	1.44154165	5	3	USER1	SPIDERMAN6	3.048537992	3
4	USER1	STARWAR3	2.94596722	4	4	USER1	STARWAR3	3.571581647	1
5	USER1	STARWAR4	3.74914688	1	5	USER1	STARWAR4	3.07405685	2
6	USER2	SPIDERMAN1	1.62167091	5	6	USER2	SPIDERMAN1	2.81131028	5
7	USER2	SPIDERMAN2	4.39694434	1	7	USER2	SPIDERMAN2	2.813557725	4
8	USER2	SPIDERMAN3	2.61309755	4	8	USER2	SPIDERMAN3	3.320826269	3
9	USER2	STARWAR1	3.15173662	2	9	USER2	STARWAR1	3.391165644	1
10	USER2	STARWAR2	3.14826354	3	10	USER2	STARWAR2	3.341068746	2
11	USER3	STARWAR1	2.9884095	3	11	USER3	STARWAR1	3.060130238	2
12	USER3	STARWAR2	2.44186309	4	12	USER3	STARWAR2	3.314198027	1
13	USER3	CONJURING1	2.29470259	5	13	USER3	CONJURING1	2.928999479	3
14	USER3	CONJURING2	3.11041039	2	14	USER3	CONJURING2	2.781005701	4
15	USER3	MERMAID	3.94761629	1	15	USER3	MERMAID	2.632783099	5
16	USER4	SPIDERMAN2	2.81158706	5	16	USER4	SPIDERMAN2	3.324186459	5
17	USER4	STARWAR1	3.11151011	4	17	USER4	STARWAR1	3.515767419	2
18	USER4	STARWAR2	4.8432233	1	18	USER4	STARWAR2	3.497395685	3
19	USER4	CONJURING1	3.78143343	3	19	USER4	CONJURING1	3.396837975	4
20	USER4	SPIDERMAN6	3.96080222	2	20	USER4	SPIDERMAN6	3.518210816	1
21	USER5	SPIDERMAN3	3.85672704	3	21	USER5	SPIDERMAN3	3.180365589	4

Figure 9: Shows the rating given to the unrated movies with the help of Pearson and Cosine formula on Item Rating Prediction.

6. CONCLUSION:

Item based collaborative filtering find similar items with the help of target item and predicts the preference rating. It finds the group of items that are close by using the rating data to

each item. The collaborative filtering has many potential problems but in this proposed the main focus on two main problem i. e., cold start, +and data sparsity. The cold start is a situation where a recommender does not have adequate information about a user or an item in order to make relevant

predictions. The proposed work can make prediction for the new user who has not rated any movies and new to the system. The data sparsity problem occurs when only a few of the total number of items available in a database are rated by the users. This leads to the sparse user item matrix. The proposed system deals with a large number of sparse data because dataset has huge quantity of unrated movies. The future scope of this work is to recommendation of movies to the different user according user's choice who rated a movie as well as new user who has not rated any movie in the system.

This paper presented Movie Recommendation System with the help of rapid miner tool can be used for recommendation of different movies to the end user for entertainment purpose. The proposed work can be extended to a large number of user. The result shows that the model can be used for large dataset. The performance of the collaborative filtering is affected by Synonymy, Shriller attack for new users and hence it is a great chance of future research on this problem

7. REFERENCES

- [1] Anand Shanker Tewari and Kumari Priyanka” Book Recommendation System Based on Collaborative Filtering and Association Rule Mining for College Students” 2014, IEEE Page(s):135 - 138
- [2] Case study on Collaborative filtering using Rapid Miner: item based recommender's source: Simafore (internet).
- [3] Case study on Collaborative filtering using Rapid Miner: user based recommender's source: Simafore (internet).
- [4] F.O.Isinkaye,Y.O.Folajimi,B.A.Ojokoh “Recommendation systems: Principles, methods and evaluation” Egyptian Informatics Journal Vol 16, Issue 3, November 2015, Pages 261–273
- [5] M. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre, “Collaborative recommendation: a robustness analysis,”ACM Transactions on 01/Internet Technology, 45v20ol. 4, no. 4, pp. 344–377, 2004.
- [6] Ponnampunyasamudram,Nallagulla,Yellamati”Movie Recommender System Using Item Based Collaborative Filtering Technique” 2016, IEEE 978-1-4673-6725-7/16/
- [7] Puntheeranurak and Chaiwitooanukool,” An Item-based Collaborative Filtering Method using item-based Hybrid Similarity” 2011, IEEE 978-1-4244-9698-3/11
- [8] RapidminerStudioDocumentation<http://docs.rapidminer.com/studio/>
- [9] Shahab Saquib Sohail, Jamshed Siddiqui, Rashid Ali, “Book recommendation system using opinion mining technique”,International Conference on Advances in Computing, Communications and Informatics (ICACCI),2013.
- [10] X. Su, T.M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques, Advances in Artificial Intelligence (AAI)”, vol. 2009, Article ID 421425, 19 pages, 2009. doi:10.1155/2009/421425
- [11] Y. Huang, "An Item Based Collaborative Filtering Using Item Clustering Prediction," ISECS International Colloquium on Computing, Communication, Control, and Management, pp54-56, 2009.
- [12] Zhihang Tang and Zhonghua Wen: “Recommendation system based on collaborative filtering in Rapid Miner” Computer Modelling & New Technologies 18(11) 1004-1008