A Survey on Filtering Techniques for Recommendation System

Janisa Gosbel Pereira
Computer Engineering
Mumbai, India
janisacolaco@gmail.com

Shubha Tiwari
Data Science Consultant
Coding Nest
Pune, India
shubhatwr@gmail.com

Sheetal Ajoy
Computer Engineering
Mumbai, India
antony89sheetal@gmail.com

2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC) | 978-1-7281-8880-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/iSSSC50941.2020.9358819

***Abstract*—The advancement of technology in recent years have caused a tremendous development in the ecommerce industry. With the growth in ecommerce, recommendation systems have gained importance in research. The recommendation system is based on different filtering techniques to provide appropriate personalized suggestions to the user. In this paper we focus on the analysis of the commonly used filtering techniques of recommendation system and provide a comparison of these techniques.**

Keywords— recommendation, collaborative filtering, content based filtering, hybrid filtering, similarity, rating

I. INTRODUCTION

Recommendation system is an important solution to obtain the information specific to a user from the vast amount of data available on the internet. This large amount of data requires processing and sorting to identify the useful information out of it [1]. The huge amount of data can create a lot of problems if it is not processed successfully and lead to wastage of time of the user while the user surfs the internet. Furthermore the information obtained needs to be filtered to provide suggestions to the user. The traditional recommendation systems generally use two basic and common filtering techniques which are collaborative filtering and content based filtering. The filtering techniques provide suggestion based on the user's interest, likes, dislikes, and history of the user. Collaborative filtering techniques work by collecting users rating whereas content based filtering techniques work by comparing two objects, like the content of an object is compared with the content of the object of user's interest [2]. These methods have their own pros and cons. To overcome the disadvantages of these methods a hybrid filtering approach is followed which combines the benefits of both the methods.

The recommendation systems can be used in various application areas like product recommendations, web service recommendation, and movie recommendation and so on. Recommender systems are considered useful in providing personalized and accurate timely recommendations to users. They are used by ecommerce stores to attract customers by displaying those products which match their interest thus reducing the search time and providing better results [2].

The overall flow of this paper is as follows: Description of the three commonly used filtering techniques in recommendation system is given in Section II. In Section III a comparative survey based on common parameters of these techniques is done. And Section IV provides conclusion and research directions for future.

II. FILTERING TECHNIQUES

The filtering algorithms usually can be categorized into two basic types:

A. Collaborative Filtering

Collaborative filtering is a technique that provides recommendation to the user considering the ratings given by other similar users [3].

B. Content Based Filtering

Content based filtering tends to recommend similar items considering the characteristics and content of selected items [3].

C. Hybrid Filtering

Another approach to filtering techniques is a combination of the two basic filtering techniques, content based and collaborative [9]. This technique is referred to as the hybrid filtering technique which is another popular approach to filtering. The hybrid technique combines two or more filtering methods. It thus gains an increase in the accuracy and performance.

In this paper we analyze three filtering techniques which are described below.

In Subsection A description of collaborative filtering technique is given, the Subsection B includes description of a content based filtering technique, and in the last Subsection C description on hybrid filtering technique is provided

III. REVIEW OF TECHNIQUES

A. Collaborative Filtering Recommendation Model with User Similarity Filling [5][6]

In this paper collaborative filtering recommendation algorithm combined with user similarity and the user item rating attribute is developed. The corresponding attribute dimensionality values of users are collected; also rating

978-1-7281-8880-5/20/$31.00 ©2020 IEEE

information of user's interest is used to further determine user similarity.

TABLE I. USER RATING MATRIX [6]

	$Item_1$	**$Item_2$**	**….**	**$Item_n$**
$User_1$	r_{11}	r_{12}	….	r_{1n}
$User_2$	r_{21}	r_{22}	….	r_{2n}
…..	….	….	….	….
$User_m$	r_{m1}	r_{m2}	….	r_{mn}

The flow of the algorithm is as follows [5] [7]:

Rating matix: The first step is to establish a rating matrix represented as $R = \{r_{ij}\}_{m*n}$ where m indicates user number and n is the number of item, r_{ij} is the score of user i given to item j. This rating matrix is needed to evaluate the items for users. Table 1 shows the rating matrix.

These rating values help to determine the similarity values of users. The similarity values are then sorted to determine the k neighbours.

Based on the selected k neighbors, users rating values and item rating values are estimated. Considering S_i as the nearest neighbor of set I, and P_{ix} as the rating value of user i given to item x. P_{ix} is given as

$$P_{ix} = \overline{r_i} + \frac{\sum_{j \in s_i} sim(i,j)*(r_{jx} - \overline{r_j})}{\sum_{j \in s_i} |sim(i,j)|} \quad (1)$$

sim (i,j) is the similarity computed between user i and j. $\overline{r_i}$ and $\overline{r_j}$ denote the graded values in average of user i and j

a) Similarity Calculation [6]:

The traditional methods of calculating similarity between two objects include cosine similarity and correlation similarity. Such traditional methods lack rationality and give unreliable recommendation results [5]. To improve the recommendation accuracy and to solve the inconsistencies in the traditional method a new method of calculating similarity is used in this technique.

According to the new method the similarity between the rating of two users for the same item is computes as given

$$sim_1(i, j, x) = 2 \left(1 - \frac{1}{1+\exp(-|r_{ix} - r_{jx}|)}\right) \quad (2)$$

b) Similarity of interest tendency [6]:

Here the average rating of users is computed which determines the user's interest for a certain objective. Similarity between user i and j for the same item is thus calculated as follows

$$sim_2(i, j, x) = \left(\frac{1}{1+\exp(-(r_{ix} - \overline{r_i})(r_{jx} - \overline{r_j}))}\right) \quad (3)$$

Confidence is measured using Jacard cooeficient

$$sim_3(i, j) = \frac{|I_i \cap I_j|}{|I_i \cup I_j|} \quad (4)$$

I_i is the item set evaluated by user i

Therefore the final similarity is computed as

$sim_{score}(i , j) =$

$$\left(\frac{1}{|I_{ij}|} \sum_{x \in I_{ij}} sim_1(i, j, x) * sim_2(i, j, x) * sim_3(i, j x)\right) \quad (5)$$

Similarity of the user attribute with the rating of user is given by

$$sim(i, j) = \alpha * sim_{attr}(i, j) + \beta * sim_{score}(i, j) \quad (6)$$

where α *and* β is given by

$$\alpha = 2 * \left(1 - \frac{1}{1+\exp(-|I_i|)}\right), \quad \beta + \alpha = 1 \quad (7)$$

The overall algorithm is shown in Fig. 1 and described as follows [5]:

1) The corresponding values of attribute dimensionality of user are collected to establish attribute matrix.

2) Rating values are caluclated and similarity rating matrix is computed using the collected values of attribute rating data of users and its corresponding data values

3) Combination of two matrices is used to compute the similarity matrix of the user.

4) Further the K nearest neighbors of user i are determined by the similarity matrix, and the value of x item is calculated through the below equation [6].

$$r_{ix} = \frac{\sum_{k \in N_k} sim(i,x) * r_{kx}}{\sum_{k \in N_k} sim(i,x)} \quad (8)$$

where N_k is the most similar K neighbors of user and r_{ki} is the rating given by user k to the item i.

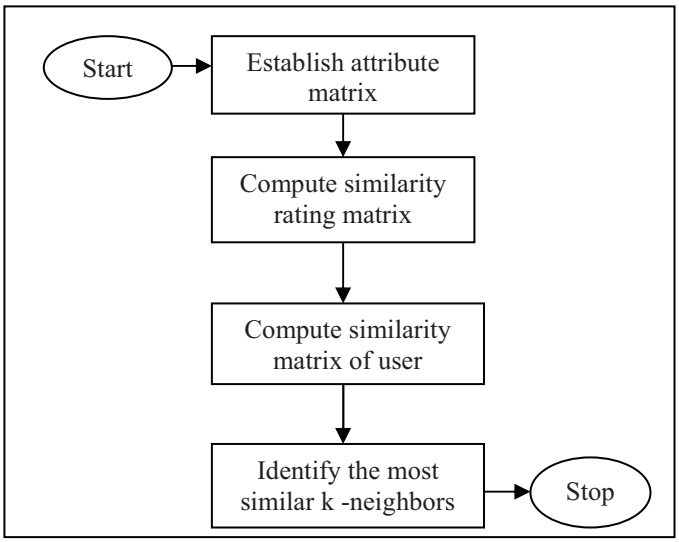

Fig. 1. Flow of collaborative filtering model [5].

Advantages:

- The quality of recommendation is increased
- User-depending on method is used

Disadvantages:

- New method for similarity calculation is required to improve rationality and reliability of recommendation results [6].

B. *Recommendation System for Property Search using Content Based Filtering [8][9]*

This technique deals with a recommendation system which is web based for choosing property which is developed using content based filtering approach. Considering the behavior of the user based on the content of advertisement previously searched by the user, the property information is provided to the user. Every time the user selects the ad to be displayed, that information gets stored for further processing to provide recommendation. In this technique content based filtering method along with Apriori algorithm and association rule mining is used to provide recommendation of property site.

The property sites generally store information only about its category i.e. whether it is a house, shop, office or boarding house. Along with category data, other information about the property can also be used which includes its size, sold/rented status, ownership letter, facilities of the house, location. The users of the property data are also of two types, one is the user who already knows the specifics and his wants and second is the user who does not know his specifics and wants. The recommendation system should satisfy the needs of both these user's.

The content based filtering method used here tends to compare the details of the advertisements that are often seen by the user with the details of the advertised property product. And it looks for similarity between the content of the recommended advertisements with the previous advertisement viewed by the user.

The details of the property advertisement viewed by the user are stored in the database. Further this stored data is assigned a weight depending on its frequency of appearing in the search process that is performed by the user.

Finally the Apriori algorithm is used for maintaining the search database depending on the content of the advertisement by analyzing the weight assigned to the data. It identifies the frequent patterns among the data. This approach requires building and maintaining user profiles based on users interest i.e. store the customers purchase transactions, the ratings, reviews comments and so on.

The overall flow of the algorithm is shown in Fig. 2 and described as follows [8]:

1) Firstly the user visits various advertisements that are displayed on the system and the content of those advertisement visited by the user is stored in the database.

2) From the data stored, certain words visited by the user are collected based on the title, address, and description of advertisement.

3) Using TF-IDF method weight of the selected words is calculated.

4) The word which obtains the highest score of TF-IDF is identified, which is seen in the ID of the property advertisement in the database.

5) The ID of the displayed ad, will be considered as the itemset for Apriori algorithm.

6) The Apriori algorithm is used to identify the frequent itemset.

7) The association rule mining method to find patterns among the data using the frequent itemset generated in the previous step.

8) The recommended advertisement details are then displayed to the user using the previous formed association rule pattern.

Advantages:-

- Considers behavioral data of system usage
- Accommodates needs of user who does not know his wants and specifics
- Provides successful results in property recommendation

Disadvantages:-

- Support & Confidence values affect the recommendation results

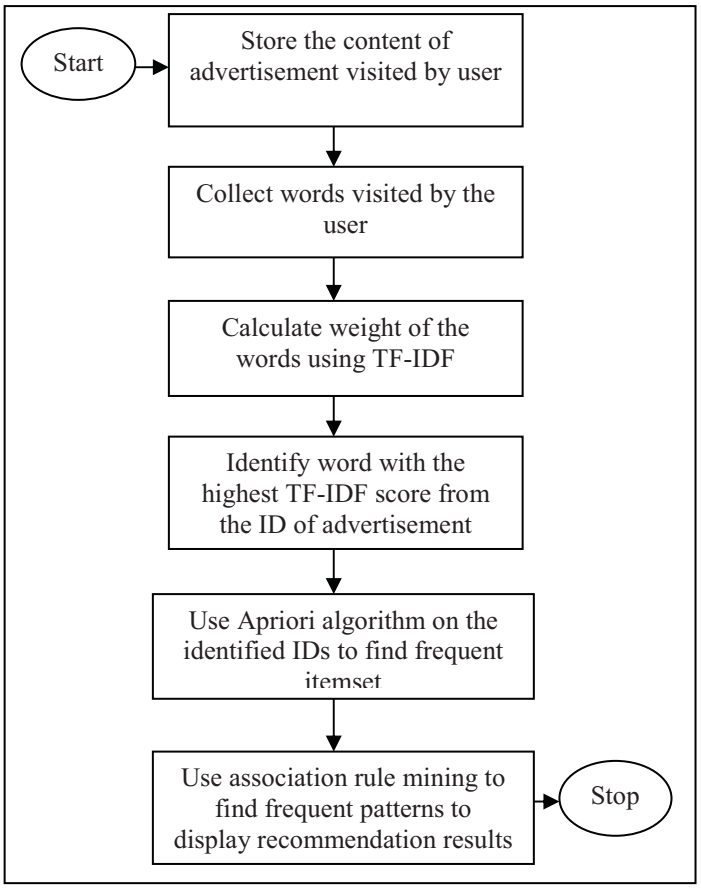

Fig. 2. Flow of content filtering model [8].

C. *Hybrid Recommender System for Learning Material using Content Based Filtering and Collaborative Filtering with Good Learners Rating [10]*

It is an improved and hybrid method for recommending learning material with the combination of content based and collaborative filtering [11]. Only the rating of good learners having certain amount of similarity with the active learners is used in this method.

The proposed model is as follows [10] and shown in Fig. 3:

1) Domain model: It involves developing the following model

Learning material model described as $d = (t_1, t_2, t_3, \ldots..t_n)$

Where t_i is the no of tags i in the learning material d and n is the number of terms considered as 9

2) User model: The user model is deleveloped to determine user competence for every item

Described as $u = (c_1, c_2 \ldots c_9)$

c_i is the competence of user i associated with term I, considering user vector is 9

3) User rating model: This denotes the ratings given by the user

$s = (r_1, r_2 \ldots r_n)$

r_i is the rating value given by user s to learning material I and n denotes the count of learning material

4) Learning Material Weighting

This uses the documents dataset and the tagging dataset. TF-IDF method is used for calculating weights

$$W_{i,j} = \frac{f_{ij}}{\max_z f_{z,j}} * \log \frac{D}{d_i} \quad (9)$$

$F_{i,j}$ is the occurrence frequency for term i in the learning material j. $\max_z f_{z,j}$ denotes the max frequency for the term in the learning material j, d_i determines the count of learning material that contains term i. D denotes the total count of learning material.

5) Content based filtering [11]

The similarity of the learner profiles with the learning material is computed using the method of cosine similarity. Learner's competence determines the first vector and the second vector is computed from the learning materials weight matrix

$$\text{Cos}\left(\underset{u}{\rightarrow}, \underset{v}{\rightarrow}\right) = \frac{\vec{u}\,\vec{v}}{\left\|\underset{u}{\rightarrow}\right\| * \left\|\underset{v}{\rightarrow}\right\|} \quad (10)$$

where $\underset{u}{\rightarrow}$ denotes the competence of user vector and $\underset{v}{\rightarrow}$ denotes the vector of weights of learning material.

6) Collaborative filtering with good learners rating

In this step the good learners with rating pattern similar to that of active learners are identified using the cosine similarity measure. The rating dataset based on user id and the ratings given by the user to the learning material is used. Pearson correlation coefficient method is used to determine similarity. The score of similarity between the active learners and good learners are obtained as the output considering the equation below.

$$\text{sim}(a, b) = \frac{\sum_{p \in P}\left(r_{a,p} - \overline{r_a}\right)\left(r_{b,p} - \overline{r_b}\right)}{\sqrt{\sum_{p \in P}\left(r_{a,p} - \overline{r_a}\right)^2}\sqrt{\sum_{p \in P}\left(r_{b,p} - \overline{r_b}\right)^2}} \quad (11)$$

where $r_{a,p}$ is the rating given to learning material p by learner a, similarly for $r_{b,p}$. $\overline{r_a}$ is the average rating provided by learner a and $\overline{r_b}$ is the average rating provided by learner b.

7) Recommendation [11]

It includes selecting the top N items with the following formula

$$\text{Pred}(a,p) = \overline{r_a} + \frac{\sum_{b \in N} \text{sim}(a,b)\left(r_{b,p} - \overline{r_b}\right)}{\sqrt{\sum_{b \in N} \text{sim}(a,b)}} \quad (12)$$

N determines the number of learners, pred(a,p) is the predicted rating given to learning material p by learner a. The rating given to the learning material p by a good learner b is denoted by $r_{b,p}$. sim (a, b) is the similarity between good learner b and active learner a.

MAE score is the evaluation metric used to determine the accuracy of ratings predicted or recommended [12].

Advantages:-

- Considers inclusion of good learners rating in collaborative filtering method along with content based filtering method
- Improved recommendation accuracy

Disadvantages:-

- Identification of good learners is required

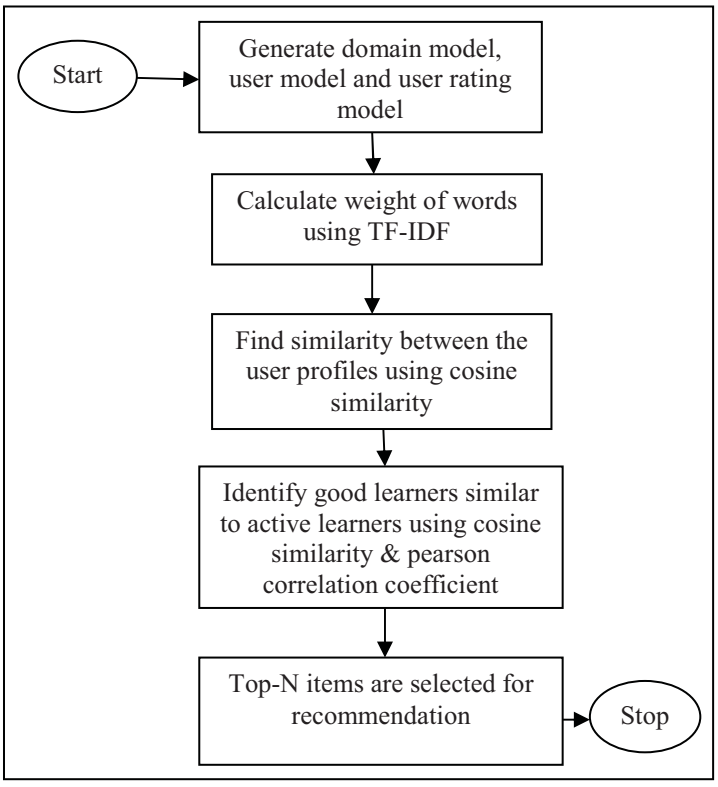

Fig. 3. Flow of hybrid filtering model [10].

IV. COMPARISON

TABLE II. COMPARATIVE ANALYSIS [5][8][10]

Technique / Parameters	**Collaborative filtering [5][6]**	**Content based filtering [8][9]**	**Hybrid Recommender System [10] [11]**
Recommendation Application area	Movie/ Books	Property search	Learning Material
Filtering Technique [12][13]	Considers user's ratings	Considers features of items	Collaborative & content based filtering combined
Input parameters	1) Attribute rating data of users 2) Attribute dimensions	1) Advertisement data 2) User profile data	1) Learning material data 2) Ratings of learning material given by user 3) Good learners ratings
Process/ Algorithms used	Selection of similar k neighbors	1) Aprioi algorithm with association rule mining 2) TF-IDF weighting method	1) Selection of top N items 2) Similarity between active user & good learner 3) TF-IDF weighting method
Similarity Calculation	Jacard coefficient & combination of similarity using user attribute & user rating	TF-IDF weighting method	Cosine similarity & Pearson correlation coefficient
Advantages	1) Considers user depending on method 2) Improved quality of recommend-ation	1) Considers behavioral data of system usage 2) Accomm-odates needs of user who does not know his wants & specifics 3) Successful recommenda-tions	1) Considers inclusion of collaborative filtering with good learner rating and content based filtering 2) Improved accuracy
Limitations	New method for similarity calculation is needed to improve rationality & efficiency of recommend-ation	Support & confidence values affect the recommendation results	Identification of good learners is required
Performance Metric	Lower MAE score Average MAE score is 0.68	Support & Confidence if smaller, more recommendation values are displayed & vice versa	Lower MAE score Average MAE score is 0.447
Accuracy [13]	Improved recommend-ation accuracy	Provides successful recommendation with improved accuracy	Better accuracy of recommend-ation

V. CONCLUSION AND FUTURE SCOPE

To summarize the analysis of the techniques described in Table II: Recommendation system using collaborative filtering [5] [6] has better recommendation accuracy as it considers users interest into account and an improved method for similarity calculation. User's interest tends to increase the quality of recommendation. Recommendation system using content based filtering [8] [9] also provides successful recommendation as it considers behavior of the user in the system and provides recommendation for users who are unaware of their wants & specifics. It also helps in providing recommendations with user preference of multiple keywords. Recommendation system using hybrid filtering [10] [11] provides improved recommendation accuracy due to the inclusion of good learners rating in the system. The users generally prefer recommendations provided by good learners. So considering the ratings of good learners in this technique provides a positive influence on the recommendations.

The advantages & limitations of each of the technique are also listed in the above sections. Each of the technique has its own pros & cons [13]. The comparative survey provided analyzes these techniques with respect to some common characteristics among them. They help in providing improved recommendation results. These techniques have proved to be beneficial in different application areas for recommendation.

The comparison helps us to understand the benefits of these techniques which if integrated into a new method can provide beneficial and good quality recommendation results. A novel hybrid approach of content and collaborative filtering can be proposed with the new similarity method used in [5] [6] and integrating it with the recommendations of good learners along with Apriori algorithm [8] [9] in the hybrid approach discussed in [10] [11]. This new technique proposed by combining the advantages of all three discussed filtering methods can then be developed and tested and compared with all the three discussed techniques to see the improvement in

results. It thus helps to establish a new approach to filtering techniques for recommendation as compared to the traditional techniques.

In Future these techniques can be implemented on a same application area and compared so that their performance can be measured in terms of prediction accuracy, precision; MAE and RMSE score in order to find the best one among them; which will provide us with an in-depth analysis of the techniques. Further research can be carried out to design a new hybrid recommendation algorithm combined with Apriori and association rule and an improved method of similarity calculation. Naïve Bayes algorithm can be used instead of Apriori to check the accuracy and speed of recommendation results compared to Apriori [14]. Such an algorithm can be integrated with ranking algorithms which will be beneficial in providing accurate personalized search results. Another research direction could be to use filtering techniques along with the concepts of deep learning, neural networks and check its performance on various applications [15].

REFERENCES

[1] Johan Eko Purnomo, Sukhmavati Endah, "Rating Prediction on Movie Recommendation System: Collaborative Filtering Algorithm (CFA) vs. Dissymetrical Percentage Collaborative Filtering Algorithm (DSPCFA)" 3rd International Conference On Informatics and Computational Sciences, IEEE, 2019.

[2] Prem Melville, Raymond J. Mooney, Ramadas Nagarajan, "Content Boosted Collaborative Filtering for Improved Recommendations", Proceedings of 18th National Conference on Artificial Intelligence, 2002.

[3] Ratsameetip Wita, Kawinwit Bubphachuen, Jakarin Chawachat, "Content Based Filtering Recommendation in Abstract Search using Neo4j" 21st International Computer Science and Engineering Conference, IEEE, 2017.

[4] Maddala Lakshmi Bai, Rajendra Pamula, Praphul Kumar Jain, "Tourist Recommender System using Hybrid Filtering", 4th International Conference on Information Systems and Computer Networks, IEEE, 2019.

[5] Bo Guo, Shiliang Xu, Dongdong Liu, Lei Niu, Fuxiao Tan, and Yan Zhang , "Collaborative Filtering Recommendation Model with User Similarity Filling" 3rd Information Technology and Mechatronics Engineering Conference, IEEE, 2017.

[6] Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, Xuzhen Zhu, **"**A New User Similarity Model to Improve Accuracy of Collaborative Filtering" , Knowledge Based System, Elseveir, 2014.

[7] Zhu Qianqian, Wang Dongliang, " A SVM Recommendation IoT model based on similarity evaluation and collaborative filtering of multiangle knowledge units", International Journal of Computers and Applications, 2018.

[8] Tessy Badriyah, Sefryan Azvy, Wiratmoko Yuwono, Iwan Syarif "Recommendation System for Property Search Using Content Based Filtering Method", International Conference on Information and Communications Technology, IEEE, 2018.

[9] Nishigandha Karbhari, Asmita Deskhmukh, Vinayak Shinde, "Recommendation System using Content Filtering", International Conference on Energy, Communication, Data Analytics and Soft Computing, 2017

[10] Rudolf Turnip, Dade Nurjanah, Dana Sulistyo Kusumo "Hybrid Recommender System for Learning Material using Content Based Filtering and Collaborative Filtering with Good Learners Rating", Conference on eLearning, eManagement and eServices, IEEE, 2017.

[11] Geetha G, Safa M, Fancy C, Saranya D. "A Hybrid Approach using Collaborative filtering and Content based filtering for Recommender Systems", National Conference on Mathematical Techniques and its Applications, 2018.

[12] Nanthaphat Koetphrom, Panachai Charusangvittaya, Daricha Sutivong "Comparing Filtering Techniques in Restaurant Recommendation System", 2nd International Conference on Engineering Innovation, IEEE, 2018.

[13] Iateilang Ryngksai, L. Chameikho, "Recommender Systems: Types of Filtering Techniques", International Journal of Engineering Research and Technology, IJERT, 2014.

[14] Manoharan, Samuel. "Patient Diet Recommendation System Using K Clique and Deep learning Classifiers." Journal of Artificial Intelligence 2, no. 02: 121-130, 2020.

[15] Bashar, A. "Survey on evolving Deep Learning Neural Network Architectures", Journal of Artificial Intelligence, 1(02), 73-82, 2019.