International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7393
***Article Received**: 30 April 2023 **Revised:** 28 June 2023 **Accepted:** 19 July 2023*

A Probabilistic Approach for Item Based Collaborative Filtering

D.Ganga Devi, S.Sampath[2]
[1]Department of Computer Applications,
Kalasalingam Academy of Research and Education,
Krishnankoil, Tamil Nadu, India.
icejessysarah@gmail.com
[2]Department of CS & IT,
Kalasalingam Academy of Research and Education,
Krishnankoil, Tamil Nadu, India.
Sampath1959@gmail.com

Abstract— In this era, it is essential to know the customer's necessity before they know it themselves. The Recommendation system is a sub-class of machine learning which deals with the user data to offer relevant content or product to the user based on their taste. This paper aims to develop an integrated recommendation system using statistical theory and methods. Therefore, the conventional Item Based Collaborative filtering integrated the probabilistic approach and the pseudo-probabilistic approach is proposed to update the k-NN approach. Here we synthesize the data using the Monte-Carlo approach with the binomial and the multinomial distribution. Then we examine the performance of the proposed methodologies on the synthetic data using the RMSE calculation.

Keywords- Recommendation Systems; Machine learning; Collaborative filtering; Probabilistic approach; Pseudo-probabilistic approach; k-NN;

I. INTRODUCTION

In recent days, the internet has become a medium for transaction on business, which in turn acts as a driving force for the development of the recommender system technology. The recommendation systems utilize the explicit feedback given by the users as well as the implicit feedbacks obtained on buying or browsing an item to infer customer interests. Because the past interests and inclinations are the good gauges of future choices. The Collaborative filtering methods [1] requires least information by the user to provide strong recommendation whereas the content-based recommender system uses the attribute values of the user and the items. Collaborative filtering involves three types of collaborative approaches namely, Memory-based, Model-based and Hybrid based approach [2]. Memory-based collaborative filtering depends on the ratings and the core of this algorithm is the similarity calculation [3]. In Model-based approach, a model is used to calculate the recommendations. It involves the User-based collaborative filtering and the Item-based collaborative filtering. It exploits the entire user-item interaction matrix to make the predictions. These approaches involve two challenges: scalability and quality [4]. Both are conflict to each other, though the less time in prediction results in scalability it lacks in quality. Therefore, it is important to treat them simultaneously to discover a solution. Due to its simplicity, main advantage of using the Memory-based methods is its simplicity in implementation. It provides comprehensive and easily understandable justification for the result. It does not require the costly training phase [5].

II. RELATED WORK

As in [6], the most common collaborative algorithms are the k-Nearest Neighbor (k-NN) algorithm, matrix factorization algorithm and the graph algorithms. In this work, all the merits and demerits of the personalized recommendation systems are discussed with. The k-NN algorithm is optimized by considering the fact that often the users do not interact with one another in their behavior. It reduced the cost of calculation.

The key for the personalized recommenders is the similarity computation. A weight based similarity algorithm called IR-IUF++ (Item Rating-Inverse User Frequency++) has been proposed in [7]. It considered the modified version of the Pearson correlation. The coefficient value is obtained by considering the variance of item ratings while computing the user's similarity. For example, a famous film will surely get a high rating, but not because of the likes. Thus, the variance is used to reveal it's extricate degree. Significantly, the time-aware similarity computation is introduced in [8] to give less weightage to the items that rated over a long period, since the recently evaluated items may attract the people more. The degree of correlation function is used to estimate the item's relevancy. It increases the item's weight depending on the fact that the linear curve can be used to represent the person's memory. In addition,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7393
***Article Received**: 30 April 2023 **Revised:** 28 June 2023 **Accepted:** 19 July 2023*

the covering degree function is applied, to show the strong correlation of the user's current interest with previous preferences. The item with a higher covering degree helps to predict the rating score efficiently.

The DTEC (Dual Training Error based Correction approach) [9] is used to correct the recommendations given to a user for an item. It is applied to improve the recommendation's accuracy. In [10], an improved collaborative filtering algorithm is proposed which is based on user-property matrix obtained by the SVD (Singular Value Decomposition) technique to make accurate recommendation for different kind of people. To obtain the user similarity in Collaborative filtering algorithm, the active user should be compared with all the users, which results in higher computation work. Here it adopts the clusters to reduce the amount of calculations. Thus the users belong to the active users category should only be considered. The clustering and SVD helps to obtain the similar users by considering the top-n properties. Then the corresponding items with these top-n properties are recommended with fewer calculations by examining the item-property matrix.

The work cited in [11] introduces the Tag matrix to calculate the comprehensive similarity. The TF (Term Frequency) calculates the proportion of the frequency of the occurrences of specific words in a document. The IDF (Inverse Document Frequency) measures the importance of a word in the appropriate documents. The TF - IDF (Term Frequency - Inverse Document Frequency) weight describes the degrees of the importance of the Tag in the item set. The Item-Tag matrix is used along with the traditional Item-Score matrix to calculate the item comprehensive similarity for prediction. A niche approach is presented in [12] for selecting the suitable neighborhood for a new item by utilizing the item attributes in IBCF. To solve the cold-start problem of IBCF, the different attributes of the new item are used to obtain the interrelated attributes. A hybrid HyCov algorithm is presented in [13] to improve the coverage of the predictions. If the active user does not rate any items, then the conventional algorithm will not generate any predictions. However, in HyCov algorithm, the neighborhood of similar users to the active user are identified. Then the ratings for the active user is predicted by utilizing the required rating of the neighbors.

III. PROBLEM OBJECTIVE

The objective is to improve the existing Item Based Collaborative filtering approach by integrating a probabilistic and pseudo-probabilistic approach to update the k-NN approach. The paper aims to synthesize data using the Monte-Carlo approach with the binomial and the multinomial distribution and then examine the performance of the proposed methodologies on synthetic data using the RMSE calculation. The ultimate goal is to enhance the accuracy and effectiveness of recommendation systems in providing personalized recommendations to users based on their preferences.

IV. PROPOSED WORK

This proposed method paved the way to investigate the application of statistical theory and methods in building the recommendation system. The probabilistic k-NN method & pseudo-probabilistic approach are introduced to develop an integrated recommendation system.

A. Item-Based Collaborative Filtering

The similar items receive similar ratings in the Item-Based Collaborative filtering [14]. Here the similarity among the items are used to define the neighborhood to predict the ratings. The User-based collaborative filtering fades out since even the active users can't rate minimum threshold of item sets.

Here the ratings matrix R should be converted into the mean-centered matrix S [17], by subtracting each rating with the average rating of each items as shown in equation (1).

$$S_{uj} = r_{uj} - \mu_u \forall\, u\epsilon\{1,2,\ldots max.no\ of\ users\} \quad (1)$$

where r_{uj} be the ratings of user u on item j and μ_u represents the mean ratings of user u.

Here the rating matrix R is generated using the Monte Carlo approach. The Monte Carlo methods [8, 16] are a class of computational algorithms that rely on repeated random sampling to generate simulated data that mimics real-world data. One approach of generating simulated data using Monte Carlo methods is to use the binomial distribution or the multinomial distribution. It randomly samples from the distribution using a probability parameter, which is based on real-world data. This process is repeated multiple times to generate a large sample of simulated data that can be used for statistical analysis and modeling.

B. k-NN Method

The k-NN (k-Nearest Neighbors) is a supervised learning algorithm relies on item feature similarity. The k-NN calculate the distance between the target item and every other items. Then rank its distances to return the top k neighbors, which are nearer to the target item.

The k- approach is used to find the clusters of k similar users having nearby similarity in ratings to the item t. Let we consider the set of users who have provided the ratings for the item i as Ui and the set of users who have rated the item j can be denoted as Uj[17]. The adjusted cosine similarity is one of the common way to determine the similarities between the items as shown in the equation (2).

$$\text{Adj. Cos(i, j)} = \frac{\sum_{u\in U_i\cap U_j} s_{ui}\cdot s_{uj}}{\sqrt{\sum_{u\in U_i\cap U_j} {s_{ui}}^2}\cdot\sqrt{\sum_{u\in U_i\cap U_j} {s_{uj}}^2}} \quad (2)$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7393
***Article Received**: 30 April 2023 **Revised:** 28 June 2023 **Accepted:** 19 July 2023*

The distance matrix with only non-negative values is calculated from the obtained similarity matrix. Let Qt(u) can be used to denote the top-k matching items to the considered item t.

Now the item-based collaborative filtering can predict the missed ratings by finding the weighted average of the most similar items as shown in the equation (3).

$$\hat{r}^{ut} = \frac{\sum_{j \in Q_t(u)} AdjustedCosine(j,t).r^{uj}}{\sum_{j \in Q_t(u)} |AdjustedCosine(j,t)|} \quad (3)$$

Here the user's own ratings on similar items are used as a leverage to predict the missed ratings.

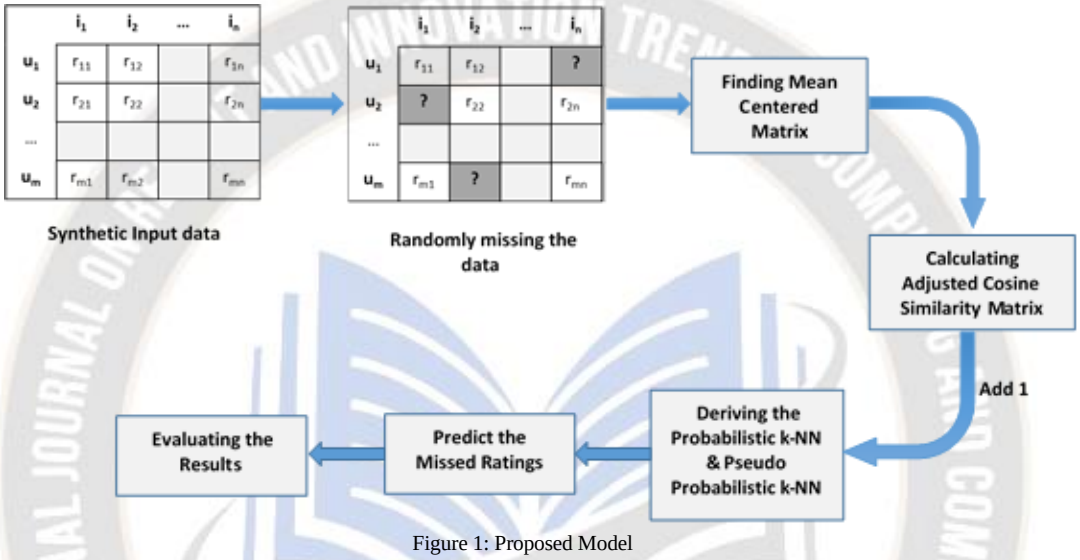

Figure 1: Proposed Model

Step 1: The rating matrix R is created using the Monte-Carlo approach.

Step 2: From the synthetic data, the ratings are missed randomly.

Step 3: The mean-centered matrix S, is calculated as shown in the equation (1).

Step 4: The adjusted cosine similarity is calculated using the equation (2) to show the similarity between the items. The obtained similarity matrix is being modified by adding 1.

Step 5: Find the probability distribution matrix.

C. Probabilistic k-NN Method

In traditional k-NN method, the top-k most similar items are considered rigidly. The most significant innovation in the proposed method namely, the probabilistic-k-NN method is that the qualified similar items are also given a chance while deciding the neighboring objects. In this research, we implement a flexible approach in k-NN classifier by making the qualified similar items to involve in the prediction by giving a chance. The workflow of the predicted model is expressed in the figure 1.

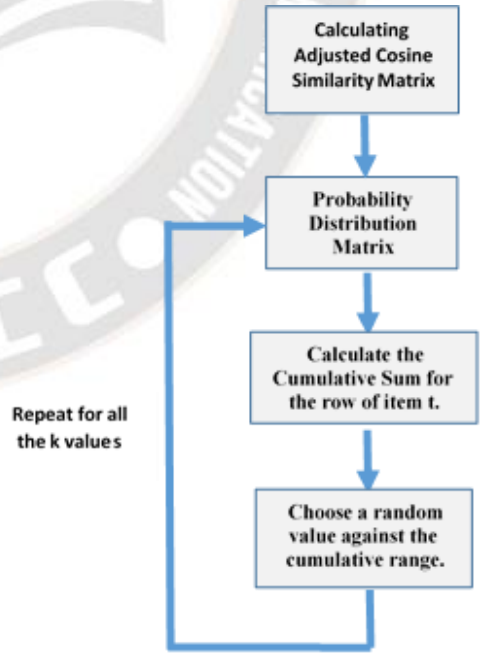

Figure 2: Workflow of probabilistic k-NN

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7393
***Article Received**: 30 April 2023 **Revised:** 28 June 2023 **Accepted:** 19 July 2023*

Step 6: To predict the missing value r_{ut} (where user u, item t), the probabilistic k-NN is obtained as shown in the figure 2. The steps are explained as follows:

1. Get the data row of the missing item t of probability distribution matrix. Find the cumulative sum of that row.
2. Now generate a random number. Check the random number against the cumulative range under which it falls.
3. Now remove that index value from the probability distribution array and recalculate the probability.
4. Repeat the steps until the k-NN range.
5. Use the obtained index positions as the probabilistic k-nearest neighbour and predict the missed ratings by using the equation (3).

D. Psuedo-Probabilistic k-NN Method

It involves the combination of conventional k-NN and the probabilistic k-NN approach as shown in the figure 3. Here we take the 1/3rd ratio of the probabilistic k-NN is chosen for the experiment purpose.

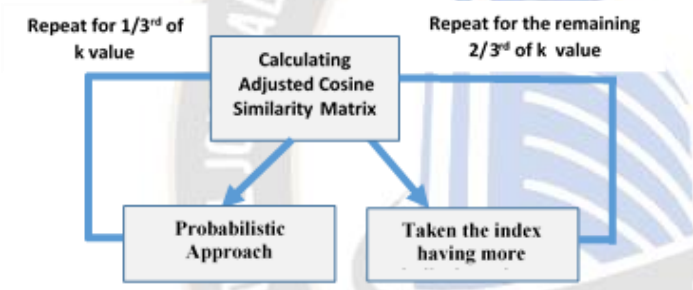

Figure 3: Workflow of pseudo-probabilistic k-NN

V. EXPERIMENTAL STUDY

A. Evaluation

The recommender system field involves many metrics to evaluate the prediction. Generally, two methods are used to evaluate the prediction error [15]. The Mean Absolute Error (MAE) is used to find the difference between the actual ratings and the predicted ratings. The better recommenders need a lower value. It may be calculated using the formula (4)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

where it calculates the sum of the absolute value of residuals and divide it by the total number of data points.

In the RMSE (Root Mean Squared Error), the absolute value of the residuals is being squared and the square root of the whole term is taken for comparison as shown in the equation (5).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (5)$$

On comparing with the RMSE will penalize more when the error is high.

B. Experimental Setup

For our research purpose, we are using the Synthetic data set with 10K users [8, 16]. We generate the data using the Monte-Carlo approach. The binomial probabilities are calculated using the discrete probability distribution, which gives either Success or Failure as the result. The equation (6) is used to calculate the binomial distribution.

$$P(x) = nCx\, p^x\, (1-p)^{n-x} \quad (6)$$

Here n represents the total number of items and x represents the number of items chosen at a time. It involves the combination of n trials of independent binomial experiment to get x successes in calculating the probability. The user-item interaction matrix is populated using this binomial distribution approach against p=0.4, 0.6 and 0.8.

In synthesizing the dataset using the multinomial distribution, the probability distribution matrix is generated with the specified k possible outcomes in each trials. The probability of each outcome should be given by the vector $p = (p_1, p_2, \ldots p_k)$ where p_i is the probability of i^{th} outcome. The rating matrix is synthesized using binomial and multinomial distributions are used for further investigation.

C. Experimental Results

Here we compare the conventional k-NN approach based IBCF and the probabilistic k-NN and pseudo-probabilistic k-NN approach based IBCF with RMSE evaluation.

The comparison performs based on the chosen p value of binomial distribution value like 0.4, 0.6 and 0.8 respectively as shown in figure 4, 5 and 6. The x-axis stands for the number of neighbors ranging from 5 to 40.

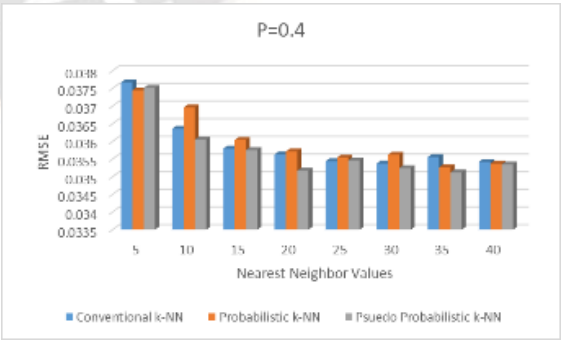

Figure 4. Comparing the probabilistic and pseudo-probabilistic based IBCF with conventional IBCF on the data synthesized using binomial distribution with p=0.4

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7393
***Article Received**: 30 April 2023 **Revised:** 28 June 2023 **Accepted:** 19 July 2023*

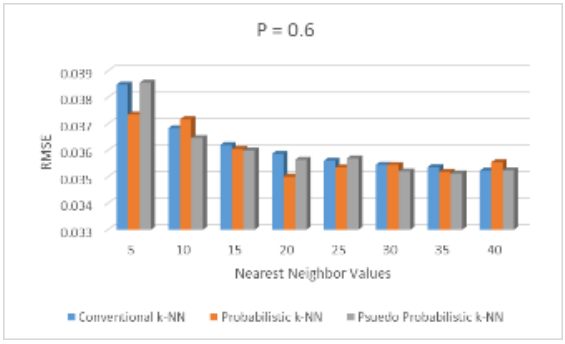

Figure 5. Comparing the probabilistic and pseudo-probabilistic based IBCF with conventional IBCF on the data synthesized using binomial distribution with p=0.6

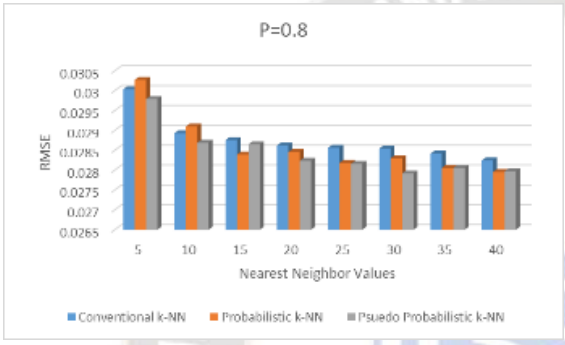

Figure 6. Comparing the probabilistic and pseudo-probabilistic based IBCF with conventional IBCF on the data synthesized using binomial distribution with p=0.8.

In the case of binomial distribution, the investigation shows that when p=0.4, the pseudo-probabilistic approach works the best as shown in figure 4. When p=0.6, both probabilistic and pseudo-probabilistic approach works better in a non-linear manner as shown in figure 5. When p=0.8, on comparing with probabilistic based k-NN, pseudo-probabilistic approach works better.

However, on using the proposed method on the synthetic data through multinomial distribution, it does not favor us when k-NN>20 as shown in the figure 7.

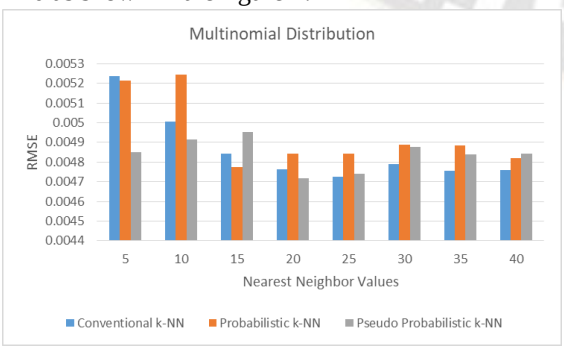

Figure 7. Comparing the probabilistic and pseudo-probabilistic based IBCF with conventional IBCF on the data synthesized using multinomial distribution

VI. CONCLUSION

The Recommendation Systems have become a buzzword in this era, owing to its contribution to a product's success. It intends to recommend a high-quality suggestion to users to help them in decision-making. Through this paper, we have implemented the integrated Item Based Collaborative filtering using the probabilistic approach and pseudo-probabilistic approach to investigate the result. The experimental results shows that the pseudo-probabilistic approach improved the prediction over probabilistic approach in the case of synthetic data generated using binomial distribution when p=0.8. We have improved the performance of both over the conventional approach to some acceptable amount when p=0.6. It also shows that the improved probabilistic and pseudo-probabilistic k-NN approaches outperformed the conventional k-NN approach for the synthetic data set obtained through binomial distribution when p=0.4.

REFERENCES

[1] J. Ben Schafer, Dan Frankowski, 1Jon Herlocker & Shilad Sen. (2007). Collaborative Filtering Recommender Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds) The Adaptive Web. Lecture Notes in Computer Science, vol 4321. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_9

[2] Najdt Mustafa, Ashraf Osman Ibrahim, Ali Ahmed & Afnizanfaizal Abdullah. (2017). Collaborative filtering: Techniques and applications. IEEE International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE).

[3] Ruonan Ji, Yi Tian & Mengdi Ma. (2020). Collaborative Filtering Recommendation Algorithm Based on User Characteristics. IEEE 5th International Conference on Control, Robotics and Cybernetics (CRC), Pp.56-60.

[4] Badrul Sarwar, George Karypis, Joseph Konstan & John Riedl. (2001). Item-based collaborative filtering recommendation algorithms. WWW '01: Proceedings of the 10th international conference on World Wide Web, pp.285–295.

[5] Kunal Shah, Akshaykumar Salunke, Saurabh Dongare & Kisandas Antala. (2017). Recommender systems: An overview of different approaches to recommendations. IEEE International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).

[6] Hua-Ming Wang & Ge Yu. (2015). Personalized recommendation system K-neighbor algorithm optimization. International Conference on Information Technologies in Education and Learning (ICITEL).

[7] Bin Wang, Qing Liao & Chunhong Zhang. (2013). Weight Based KNN Recommender System. IEEE 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, pp.449-452.

[8] Zhipeng Zhang, Yasuo Kudo & Tesuya Murai. (2016). Improvement of item-based collaborative filtering by adding

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7393
***Article Received**: 30 April 2023 **Revised:** 28 June 2023 **Accepted:** 19 July 2023*

time factor and covering degree. IEEE Joint 8th international conference on soft Computing and Intelligent Systems and 17th International symposium on advanced Intelligent Systems, pp.543-547.

[9] Costas Panagiotakis, Harris Papadakis, Antonis Papagrigoriou & Paraskevi Fragopoulou. (2021). Improving recommender system via a Dual Training Error based Correction. Expert Systems with applications, Vol:9.

[10] Wang Hong-xia. (2019). An Improved Collaborative Filtering Recommendation Algorithm. IEEE 4th International Conference on Big Data Analytics (ICBDA), pp.431-435.

[11] Qiuyue Xu, Shangzhi Zheng & Min Cai. (2013). IBCF Improved Algorithm Based on the Tag. Proceedings of the International Conference on Computer, Networks and Communication Engineering (ICCNCE), pp.265-268.

[12] Shanthi, D. N. ., & J, S. . (2022). Social Network Based Privacy Data Optimization Using Ensemble Deep Learning Architectures. Research Journal of Computer Systems and Engineering, 3(1), 62–66. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/43

[13] Morzelona, R. (2021). Human Visual System Quality Assessment in The Images Using the IQA Model Integrated with Automated Machine Learning Model . Machine Learning Applications in Engineering Education and Management, 1(1), 13–18. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/5

[14] Zhi-Peng Zhang,Yasuo Kudo,Tetsuya Murai & Yong-Gong Ren. (2019). Addressing Complete New Item Cold-Start Recommendation: A Niche Item-Based Collaborative Filtering via Interrelationship Mining. MDPI Journal on Applied Sciences, 9(9).

[15] Manolis G. Vozalis, Angelos I. Markos & Konstantinos G. Margaritis. (2009). A Hybrid Approach for Improving Prediction Coverage of Collaborative Filtering. Proceedings of the 5TH IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI).

[16] Basit Mehmood Khan, Asim Mansha, Farhan Hassan Khan & Saba Bashir. (2017). Collaborative filtering based online recommendation systems: A survey. IEEE International Conference on Information and Communication Technologies (ICICT).

[17] Mahamudul Hasan, Shibbir Ahmed, Md. Ariful Islam Malik & Shabbir Ahmed. (2016). A comprehensive approach towards user-based collaborative filtering recommender system. IEEE International Workshop on Computational Intelligence (IWCI), pp.159-164.

[18] S.Sampath & S.Suresh. (2019). Quantiles based Neighborhood Method of Classification. International Journal of Computational and Theoretical Statistics, 6(1).

[19] Charu C.Aggarwal. (2016). Recommender Systems (pp.40-42). Springer International Publishing. 10.1007/978-3-319-29659-3