Online Book Recommendation System using Collaborative Filtering (With Jaccard Similarity)

Journal of Physics: Conference Series **1362** (2019) 012130 doi:10.1088/1742-6596/1362/1/012130

Online Book Recommendation System using Collaborative Filtering (With Jaccard Similarity)

Avi Rana[1], K. Deeba[2]

[1,2]SRM Institute of Science and Technology,Chennai,India

avijrana@gmail.com[1], deeba.k@ktr.srmuniv.ac.in[2]

Abstract-- Recommendation System (RS) is software that suggests similar items to a purchaser based on his/her earlier purchases or preferences. RS examines huge data of objects and compiles a list of those objects which would fulfil the requirements of the buyer. Nowadays most ecommerce companies are using Recommendation systems to lure buyers to purchase more by offering items that the buyer is likely to prefer. Book Recommendation System is being used by Amazon, Barnes and Noble, Flipkart, Goodreads, etc. to recommend books the customer would be tempted to buy as they are matched with his/her choices. The challenges they face are to filter, set a priority and give recommendations which are accurate.

RS systems use Collaborative Filtering (CF) to generate lists of items similar to the buyer's preferences. Collaborative filtering is based on the assumption that if a user has rated two books then to a user who has read one of these books, the other book can be recommended (Collaboration). CF has difficulties in giving accurate recommendations due to problems of scalability, sparsity and cold start. Therefore this paper proposes a recommendation that uses Collaborative filtering with Jaccard Similarity (JS) to give more accurate recommendations. JS is based on an index calculated for a pair of books. It is a ratio of common users (users who have rated both books) divided by the sum of users who have rated the two books individually. Larger the number of common users higher will be the JS Index and hence better recommendations. Books with high JS index (more recommended) will appear on top of the recommended books list.

Keywords: Similarity index, filtering techniques, recommender system, Jaccard Similarity

1. INTRODUCTION

Recommendation system filters information by predicting ratings or preferences of consumers for items that the consumer would like to use [1]. It tries to recommend items to the consumer according to his/her needs and taste. RS mainly uses two methods to filter information - Content-based and Collaborative filtering. Content-based filtering involves recommending those items to a consumer which are similar in content to the items that have already been used by him/her. First, it makes a profile of the consumer, which consists of his/her taste. Taste is based on the type of books rated by the consumer. The system analyses the books that were liked by the consumer with the books he had not rated and looks for similarity. Out of these unrated books, the books with the maximum value of similarity index will be recommended to the consumer. Paul Resnick and Hal Varian were the ones who suggested Collaborative filtering algorithm in 1997. It became popular amid the various frameworks available at that time. [2] [3] [4]

A complete RS contains three main things: user resource, item resource and the recommendation algorithm. In the user model, the consumers' interests are analysed, similarly, the item model analyses the items' features. Then, the characteristics of the consumer are matched with the item characteristics to estimate which items to recommend using the recommendation algorithm. The performance of this algorithm is what affects the performance of the whole system.

In memory-based CF, the book ratings are directly used to assess unknown ratings for new books. This method can be subdivided into two ways: User-based approach and Item-based approach.

1. User-based approach: As per this method, customers with alike choices form a neighbourhood. If an item is not rated by the buyer, but it has been rated highly by other members of the

neighbourhood, then it can be endorsed to the buyer. Hence the buyer's choices can be predicted based on the neighbourhood of similar buyers. [2] [3] [4]

2. Item-based approach: In this method, similarity between the group of objects rated emphatically by the buyer and the required object is calculated. The items which are very alike are selected. Recommendation is computed by finding the weighted means of user's ratings of the alike objects. [5]

To obtain accurate and faster recommendations researchers have combined the different recommender technologies, these are known as Hybrid recommendation systems. Some Hybrid recommendation approaches are:

- Separate execution of procedures and connecting the results.
- Using some content filtering guidelines with community CF.
- Using some principles of CF in content filtering recommender
- Using both Content and Collaborative filtering in a recommender

Additionally there is on-going research on Semantic-based, Context-aware, Cross-lingual, Cross-domain and peer-to-peer methods [6].

2. LITERATURE SURVEY

Okon et.al. (2018) [7] proposed a model that generates recommendations to buyers, through an enhanced CF algorithm, a quick sort algorithm and Object Oriented Analysis and Design Methodology (OOADM). Scalability was ensured through the implementation of Firebase SQL. This system performed well on the evaluation metrics.

Kurmashov et.al. (2015) [8] used Pearson correlation coefficient based CF to provide internet based recommendations to book readers and evaluated the system through an online survey.

Mathew et.al. (2016) [9] proposed a system that saves details of books purchased by the user. From these Book contents and ratings, a hybrid algorithm using collaborative filtering, content-based filtering and association rule generates book recommendations. Rather than Apriori, they recommended the use of Equivalence class Clustering and bottom up Lattice Transversal (ECLAT) as this algorithm is faster due to the fact that it examines the entire dataset only once.

Parvatikar et.al. (2015) [10] proposed item-based collaborative filtering and association rule mining to give recommendations. Similarity between different users was computed through Adjusted Cosine Vector Similarity function. Better recommendations were obtained as through this method data sparsity problem was removed.

Ayub et.al. (2018) [11] proposed a similarity function similar to Jaccard Similarity to locate alike items and users for the enquiring item and user in nearest neighbour based collaborative filtering. They proposed that absolute value of ratings should be taken as against the ratio of co-rated items taken in Jaccard Similarity. They also compared performance of their method with other similarity measures.

Gogna and Majumdar [12] suggested the use of buyer's demographic and item category to overcome data sparsity and cold start problems in their movie recommendation system. Latent Factor Model (LFM) was used. They developed a matrix to match the buyer and user information to get a dense user and dense item matrix. Label Consistency map, the outcome of this system was used to suggest unrated and other items to new buyers.

Chatti et.al. (2013) [13] suggested tag-based and rating based CF recommendation in technology enhanced learning (TEL) to resolve the data sparsity problem and extract relevant information from the rating database. Memory and model oriented 16 varied tag-based Collaborative filtering algorithms were evaluated for buyer satisfaction and accuracy of recommendations in Personal Learning Environments.

Choi et.al. (2010) [14] proposed RS based on HYRED, a hybrid algorithm using both content and collaborative filtering on a compact dataset (by reducing user interest items) and neighbor data. HYRED used altered Pearson Coefficient based Collaborative filtering and distance-to-boundary (DTB) Content filtering. This would result in better and faster recommendation for large amount of data.

Liu et.al. (2012) [15] added the dimension of user-interest. They proposed iExpand, a 3 tier model i.e. user, user-interest and item. This helped in overcoming the issues of overspecialization and cold-start as well as reducing computation costs.

Feng et.al. (2018) [16] proposed a RS for movies based on a similarity model constituted of factors S_1 (similarity between users), S_2 (ratio of co-rated items) and S_3 (user's rating choice weight). This RS was particularly useful for sparse datasets.

Literature survey suggested that recommendation systems are being used by large number of online marketers to increase their sales by offering products to customers which match their tastes.

These RS suffer from many problems such as data sparsity, cold start, trust, scalability and privacy. Therefore there is need for improved recommendation systems which solve these issues.

Most researchers use Adjusted Cosine Vector Similarity function to compute similarity among book ratings to recommend books, to take into account that users have different rating schemes. Some rate items high usually while others usually rate low.

3. PROPOSED WORK

According to Aggarwal, C.C., Collaborative Filtering method is used in recommendation systems to develop recommendations based on ratings provided by the other users of the system [17]. If buyer's ratings of items match, it is likely that their ratings of other items will also match, this is the basic assumption of CF. Computers cannot gauge qualitative factors such as taste or quality, therefore recommendations based on the ratings of humans who can rate on the basis of qualitative factors, i.e. collaboration, will give a better outcome.

Soanpet Sree Lakshmi and Dr. T. Adi Lakshmi in their work have described in detail the problems like overspecialization, data sparsity, cold start, scalability, ranking of the recommendation, etc. that are related to the CF recommendation system [6]. Buyers do not like to spend too much time on rating items online. Hence rating data is usually sparse, which in turn reduces the recommendation quality. As new users have not made any explicit or implicit ratings, therefore it is difficult to find similarity and hence provide recommendations. This is known as cold start problem.

This paper proposes the use of Jaccard Similarity [11]. Jaccard Similarity for item-based recommendation calculates the similarity between two books by taking the number of users who have rated both books in the numerator and the number of users who have rated either of the two books in the denominator. This does not consider the absolute ratings rather it considers number of items rated. Two items will be more similar, when they have been rated by more common users. Jaccard Similarity does not consider the actual rating given by the user to a book, but only goes by the number of users who have rated the books. Incorporating a weightage to the similarity index based on the rating given by user, overcomes this issue. For the item-based algorithm, Jaccard Similarity is given as

$$jaccard\ sim(A,B) = \frac{|n_A \cap n_B|}{|n_A \cup n_B|}$$

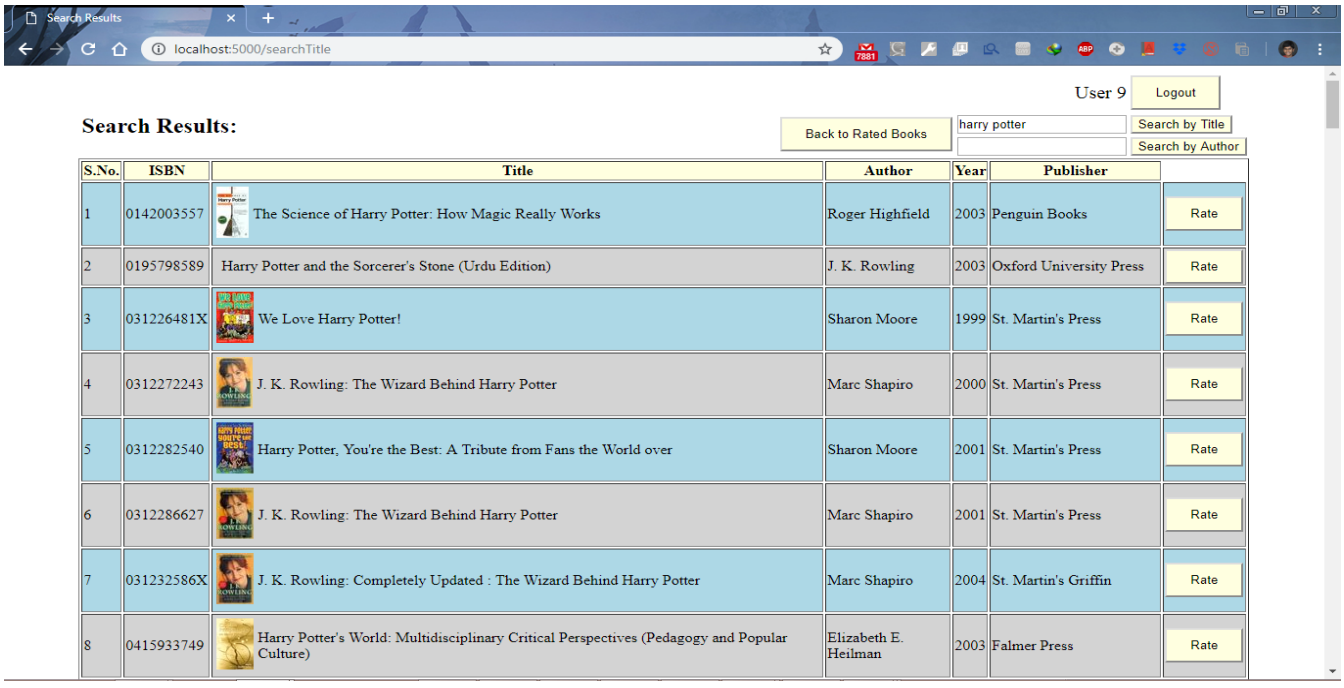
Where n_A is the set of users who have rated item A and n_B is the set of users who have rated item B.

In this approach, the recommendation system will calculate the similarity between books that have been rated by the user already and the books available in the Book Crossing dataset. This dataset was compiled by Cai-Nicolas Ziegler, it contains data of 2,71,379 books, along with 11,49,780 user ratings for these books, given by 2,78,858 users. Datasets are available in CSV (Comma Separated Values)

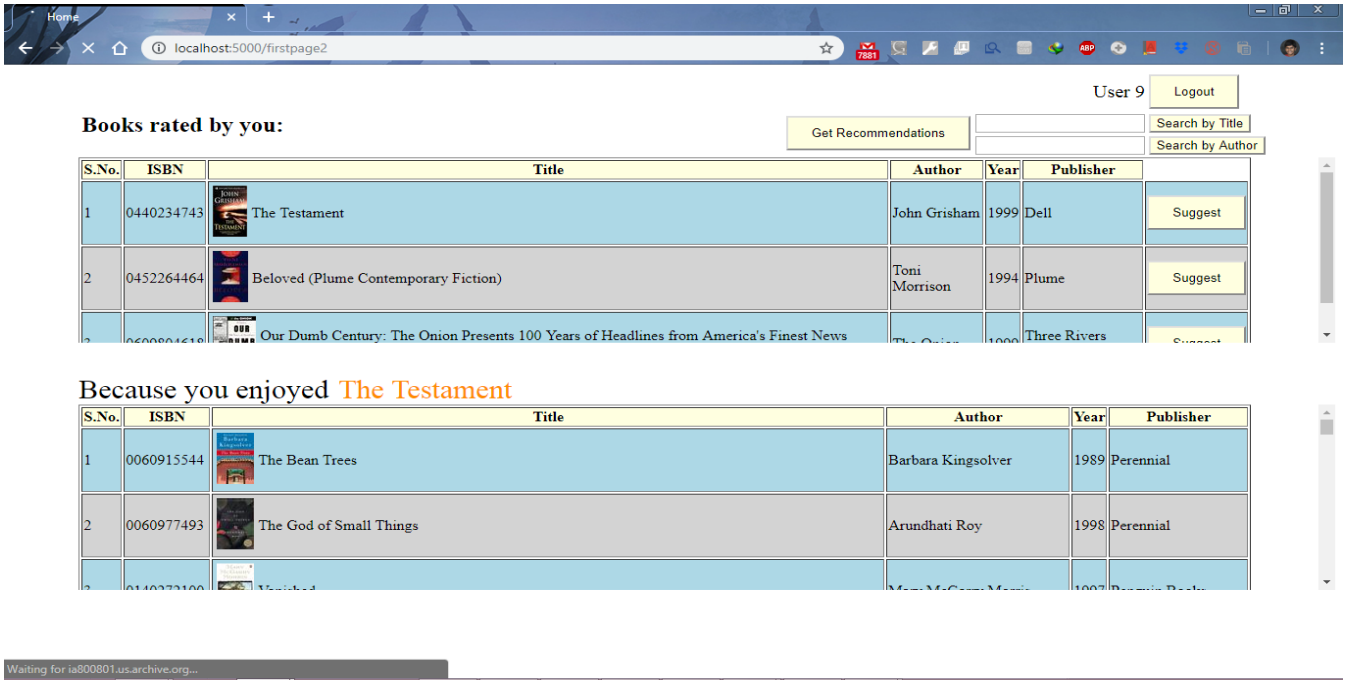
and SQL formats. Since large amount of data is to be handled, SQL database was used for faster and more efficient processing.

Login module for the proposed system:

Once the similarities have been computed for all the books, they are sorted and the books most similar to the books rated by the user are recommended to him. Recommendations to the user can also be based on a particular book chosen by the user. In this case, the similarity calculation will be done only for this book and the most similar books will be recommended to him. Books in the entire dataset can be searched by author or by title, and the user can give ratings to these books for more accurate results.



S.No.	ISBN	Title	Author	Year	Publisher	
1	0142003557	The Science of Harry Potter: How Magic Really Works	Roger Highfield	2003	Penguin Books	Rate
2	0195798589	Harry Potter and the Sorcerer's Stone (Urdu Edition)	J. K. Rowling	2003	Oxford University Press	Rate
3	031226481X	We Love Harry Potter!	Sharon Moore	1999	St. Martin's Press	Rate
4	0312272243	J. K. Rowling: The Wizard Behind Harry Potter	Marc Shapiro	2000	St. Martin's Press	Rate
5	0312282540	Harry Potter, You're the Best: A Tribute from Fans the World over	Sharon Moore	2001	St. Martin's Press	Rate
6	0312286627	J. K. Rowling: The Wizard Behind Harry Potter	Marc Shapiro	2001	St. Martin's Press	Rate
7	031232586X	J. K. Rowling: Completely Updated : The Wizard Behind Harry Potter	Marc Shapiro	2004	St. Martin's Griffin	Rate
8	0415933749	Harry Potter's World: Multidisciplinary Critical Perspectives (Pedagogy and Popular Culture)	Elizabeth E. Heilman	2003	Falmer Press	Rate



The interaction between the application server and the user is shown in fig 1. Fig. 2 displays the architecture of the proposed system

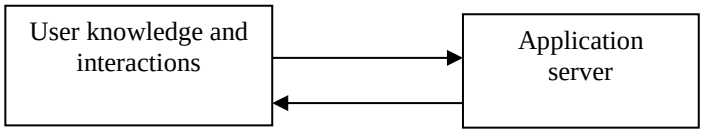


Fig1: Block diagram

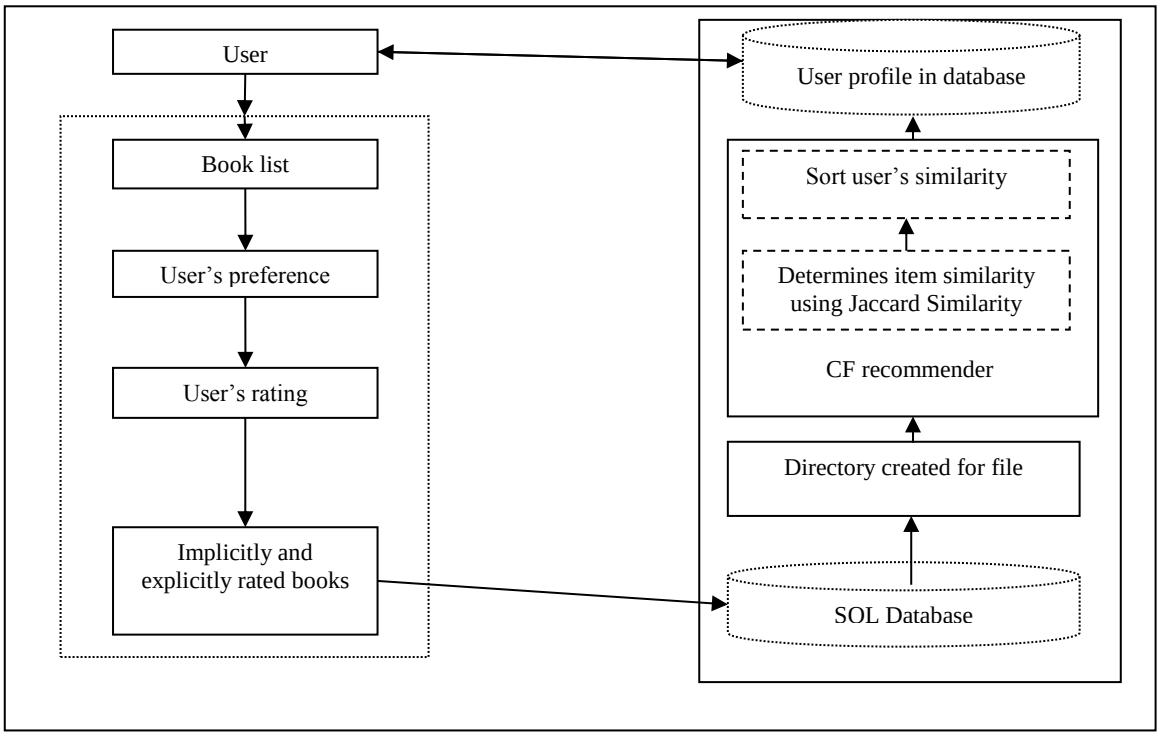


Fig. 2: Architecture of proposed system

4. EXPERIMENTAL RESULTS

This paper used Root Mean Square Error (RMSE) statistical accuracy metrics to evaluate the online book recommendation system. RMSE measures the variation between the predicted value and actual user rating. The predicted rating is calculated as follows:

$$P_{u,i} = \frac{\sum_{all\ similar\ items,N}(s_{i,N} * R_{u,N})}{\sum_{all\ similar\ items,N}(|s_{i,N}|)}$$

Hence, smaller the value of RMSE, better the recommendation system. The computed value of RMSE for this system is 1.504.

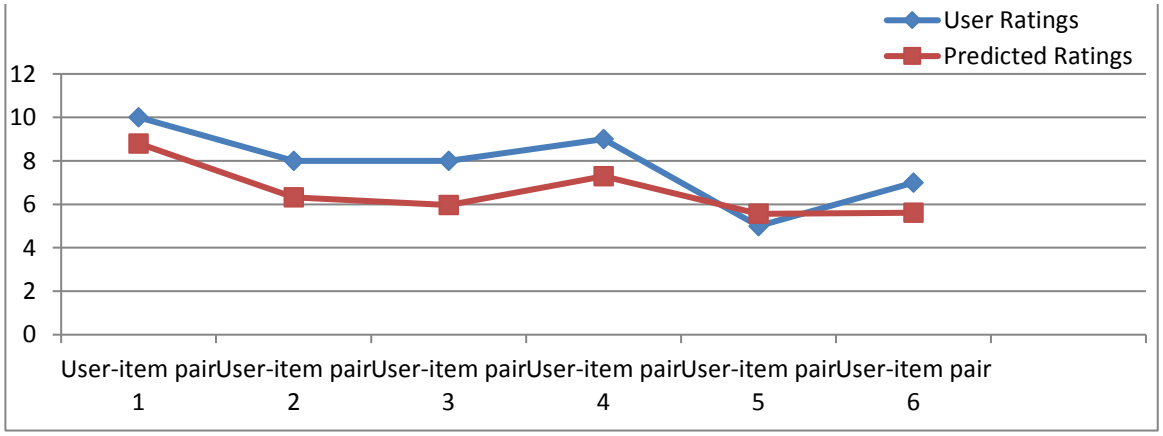


Fig. 3: Comparison of Ratings and Predictions for user-item pairs

5. CONCLUSION

Even with the adaptation of a fitting algorithm for recommendation, the RS faces an obstacle because of large quantity of data that needs to be handled. According to the experimental results, the proposed algorithm with compact dataset was more accurate than existing algorithms with full datasets. In addition, JS uses the number of common users as a basis for measuring similarity, rather than the absolute ratings, as used by most existing algorithms, which gives a more accurate result.

6. FUTURE WORK

The recommendation system proposed here takes the number of users who have rated the books into account, without factoring in the absolute rating. Due to this, a recommendation might arise from a book that a user has given low rating to, in which case a book might be recommended from a genre that the user dislikes.

This recommendation system relies on the ratings given by users. So, trust is a major issue, like whether the feedback and rating given by the user is genuine or not. This recommendation system does not solve the trust issue.

Therefore future research should focus on resolving both these issues.

7. REFERENCES

[1] Elgohary, A., Nomir, H., Sabek, I., Samir, M., Badawy, M. and Yousri, N.A. (2010). Wiki-rec: A semantic-based recommendation system using Wikipedia as ontology. In 10th International Conference on Intelligent Systems Design and Applications (ISDA).

[2] Oku, K., Nakajima, S., Miyazaki, J. and Uemura, S. (2006). Context-aware SVM for context-dependent information recommendation. In MDM'06 proceedings of International Conference on Mobile Data Management.

[3] Ghani, R. and Fano, A. (2002). Building recommender systems using a knowledge base of product semantics. In proceedings of 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems.

[4] Miyahara, K. and Pazzani, M.J. (2000) Collaborative filtering with the simple Bayesian classifier. In proceedings of Pacific Rim International Conference on Artificial Intelligence.

[5] Asanov D. (2011) Algorithms and Methods in Recommender Systems. Berlin Institute of Technology Berlin, Germany

[6] Lakshmi, S.S. and Lakshmi, T.A. (2014). Recommendation Systems: Issues and challenges. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5771-5772

[7] Okon, E.U., Eke, B.O. and Asagba, P.O. (2018). An improved online book recommender system using collaborative filtering algorithm. International Journal of Computer Applications(0975-8887) Volume 179-No.46, June 2018

[8] Kurmashov, N., Konstantin, L., Nussipbekov, A. (2015). Online book recommendation System. Proceedings of Twelve International Conference on Electronics Computer and Computation (ICECC)

[9] Mathew, P., Kuriakose, B. And Hegde, V. (2016). Book Recommendation System through content based and collaborative filtering method. Proceedings of International Conference on Data Mining and Advanced Computing (SAPIENCE)

[10] Parvitikar, S. and Dr. Joshi, B. (2015). Online book recommendation system by using collaborative filtering and association mining. Proceedings of IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)

[11] Ayub, M., Ghazanfar, M.A., Maqsood, M. and Saleem, A. (2018). A Jaccard base similarity measure to improve performance of CF based recommendation system. Proceedings of International Conference on Information Networking (ICOIN)
[12] Gogna, A., Majumdar, A. (2015). A Comprehensive Recommender System Model: Improving Accuracy for Both Warm and Cold Start Users. IEEE Access Vol. 3, 2803-2813, 2015
[13] Chatti, M.A., Dakova, S., Thus, H. and Schroeder, U. (2013). Tag-Based Collaborative Filtering Recommendation in Personal Learning Environments. IEEE Transactions on Learning Technologies, Vol. 6, No. 4, October-December 2013
[14] Choi, S.H., Jeong, Y.S. and Jeong, M.K. (2010). A Hybrid Recommendation Method with Reduced Data for Large-Scale Application. IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews, Vol. 40, No.5, September 2010.
[15] Liu, Q., Chen, E., Xiong, H., Ding, C.H.Q. and Chen, J. (2012). Enhancing Collaborative Filtering by User Interest Expansion via Personalised Ranking. IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics, Vol. 42, No.1, February 2012.
[16] Feng, J., Fengs, X., Zhang, N. and Peng, J. (2018). An improved collaborative filtering method based on similarity. PLOS ONE 13 (9): e0204003, September 2018.
[17] Aggarwal, C.C. (2016). Recommendation System: The Textbook. XXI, 29 p. ISBN 978-3-319-29657-9
[18] NHK K. ISMAIL*,"Estimation Of Reliability Of D Flip-Flops Using Mc Analysis", Journal of VLSI Circuits And Systems 1 (01), 10-12,2019.
[19] MN BORHAN,"Design Of The High Speed And Reliable Source Coupled Logic Multiplexer",Journal of VLSI Circuits And Systems 1 (01), 18-22,2019