Content Based and Rating Based Book Recommendation System

[1]Samiksha M. Pande, [2]Ashwini Gaikwad

[1]Student, Department of Computer Science and Engineering, DIEMS, Aurangabad, MH, India
[2]Assistant Professor, Department of Computer Science and Engineering, DIEMS, Aurangabad, MH, India

Abstract: On the Internet, as data available for users is expanding in an exponential rate, the overwhelming size and complexity of the information is also increasing day-by-day. With this rapid development on World Wide Web, we are moving from an environment of data scarcity towards an era of information overload. Now-a-days everyone is having their own intelligent handy devices which is turning them towards the e-commerce sites for their day-to-day chores of purchasing rather than taking out time and visiting number of physical stores. However, the enormous amount of available choices makes the customer indecisive. For the current day, the main challenging task of consumer-oriented e-commerce market is to understand the online customer's requirements and expectations. Hence this has increased the demand for Recommendation System (RS), which are information filtering system that searches and prioritize information from massive amount of dynamically generated data and recommend more efficiently the item according to user's interest, preferences and observed behaviour.
So, to reduce the impact of large volume data, a Content based and Rating based Recommendation system is proposed. This system aims at reducing the shortcoming of Collaborative Filtering and Content based methods individually, by parallelly using both methods in one System.

Index Terms - E-commerce, Recommendation system, Content based, Rating based, Collaborative Filtering.

I. INTRODUCTION

Recommendation Systems are widely used in many different domains of information retrieval for the recommendation of books, music, movies, various items available on e-commerce sites, people on social sites, partners on dating sites, trip location and hotel advisor by Online travel services and a lot more. Hence, Recommendation Systems (RSs) are techniques and an intelligent application that tries to predict items out of the large pool a user may be interested in and recommends the best option to the target user. So, to explore the large volume of data and mine useful information or knowledge for further actions, according to user interest, preferences and behaviour that are being captured from his/her previous purchase history, which they are stored and use the same for the personalized recommendation in the future[15].

Personalizing services by decreasing the searching cost of the transaction and choosing items of interest are beneficial to the customer as well as to service providers in an Online Shopping environment as it has increased revenue and helpful for selling more products by searching customer directly with the effective way of advertisements and recommendation. Two basic approaches CF and Content-Based Filtering are considered for applying various methods in RS. Collaborative Filtering(CF) is a technique used for predictions in RS which do not require any extra information about item and user or one can say it is an information filtering system which collects the choice information or preferences about the interest of a user among number of users which works for making recommendation by matching people of similar interests.

Most common forms of CF techniques implemented are user-based and item-based. User-Based CF says that users who got agreed previously i.e. in past are most probably going to agree in future.

User-Based category can shortly be described as below, (shown in fig. 1)

i) Collect likeminded users. If user u1 is rating or is shopping an item I1 find all the users purchasing or rating I1.

ii) Calculate prediction from all like-minded user for active user u1.

Item Based category can be shortly described as below, (as shown in fig. 2)

i) Relationship between a pair of items is shown as user u1 is recommended the items like what user u1 have preferred previously. Suppose u1 has bought or rated I1 then items like I1 are collected.

ii) Calculate prediction based on matching items I1 to other items in item-item matrix i.e. taste of the current user.

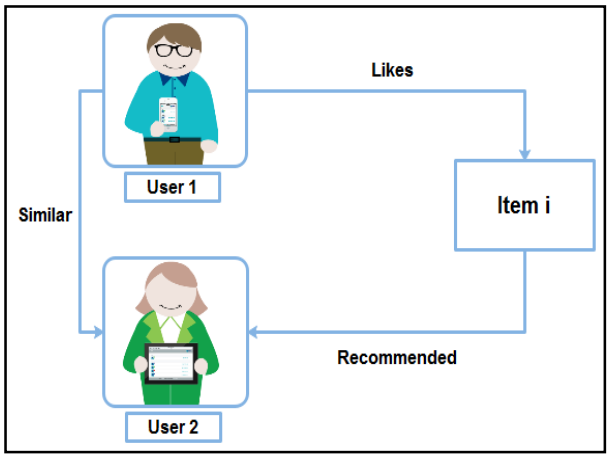

Fig -1: User-Based Recommendation

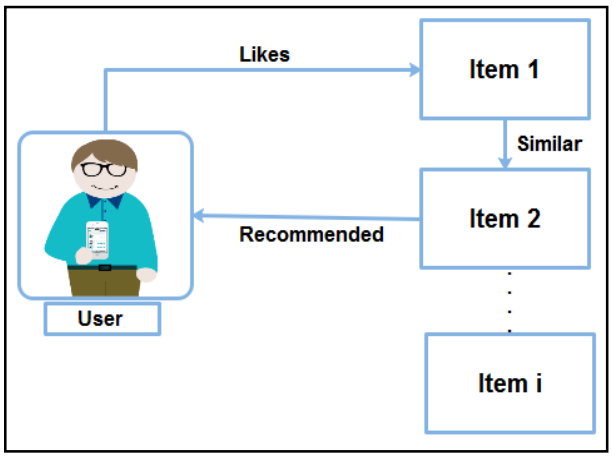

Fig -2: Item-Based Recommendation

Existing Systems like Amazon, YouTube, Netflix, Facebook, LinkedIn, Myspace, Twitter, Reddit, last.fm, Saavn, gaana.com, MakeMyTrip, magicbricks, Pinterest, OYO, Pandora and much more use RS in its different forms for different purposes. Collaborative filtering used by Facebook, LinkedIn, Myspace RS suggests friend, groups and other social connection by observing connections present in the network of the user. Twitter RS suggests whom to follows by using various signals and in-memory cell.

YouTube and Netflix are hybrid systems which combine uses Collaborative Filtering and Content-Based Filtering. Their RS works by comparing the watching and searching habits of similar users i.e., by recommending movies or videos that a user has highly rated or most watched and which also shares similar characteristics. Book or other product recommendation in Amazon is done by Item Clustering CF technique.

User-based CF technique makes suggestions by considering user choice which is mostly used by Last.fm and Reddit. Last.fm creates a station for a user considering choices of other users (based on similar user, people whom you follow) listened/like/rated to.

The content-based approach is used by Pandora which uses the content of song for the recommendation. The feature and quality of song or artist are used for tuning into stations playing similar featured music also it uses feedback from users i.e., like and dislikes of users. Pandora can start recommendation with little information.

II. LITERATURE SURVEY

[1] CLUBCF is the solution to the challenge of effectively capturing, processing and managing the service relevant Big data which is beyond the ability of traditional approaches by forming clusters of similar services and collaboratively recommending services. CLUBCF approach contains two main steps, whose first step is Clustering of services based on the description, functional similarity and Characteristic similarity computed by Jaccard Similarity Coefficient (JSC) and the weighted sum of Description Similarity and Functionality Similarity respectively, then for clustering Agglomerative Hierarchical Clustering (AHC) Algorithm is used. Collaborative Filtering (CF) is the second main step in CLUBCF, initially, Pearson correlation coefficient (PCC) is used to compute rating similarity & then clustered services are provided with a predicted rating. Lastly, all recommended services are ranked in descending order according to the predicted ratings.

In traditional CF systems with the increase in the number of thousands of participants, the amount of work is increased in rapid proportion. To provide high-quality quick Recommendation new RS technologies needs to have emerged for large scale problems. For addressing these issue, Item-Based CF techniques [2] are evaluated where the relationship between items is identified by analyzing user-item matrix; then to compute recommendation indirectly these relationships are used. Algorithms like Correlation-based similarity, Cosine-based similarity and adjusted cosine similarity are used to calculate item: item similarity and for obtaining a recommendation from them, different techniques like weighted sum and a regression model is used. Finally, results were experimentally evaluated and compared with the K-nearest neighbor approach. Here the experiment shows that a more efficient and better-quality recommendation was provided using Item-based algorithm than its User-based counterparts.

Recommendation algorithm is an effective way of targeted marketing which creates a personalized shopping experience. But e-commerce recommendation algorithm often works in a challenging environment. [3] compares the three common recommendation problems solving approaches: Cluster-based, search-based and traditional CF with their proposed Item-to-Item CF algorithm. This algorithm is applied to Amazon's online shop that produces high-quality real-time recommendations which were able to scale independently of the massive amount of customer and item datasets.

[4] An Improved Collaborative Filtering Algorithm Based on User Interest algorithm combines user interest information to reduce the impact of data sparseness by User-based Collaborative Filtering algorithm which is improving it in two ways: firstly, a user-item rating is analyzed and divided in K-clusters. The target user is allocated to the most similar cluster and its nearest neighbor is generated. Then the top-M items are recommended to the target users based on their predicted rating for the item.

For predicting user interest items most of the recommender system uses either content-based filtering or CF. Individually both the methods have their own advantages and disadvantages [5] proposed an elegant and fruitful framework that combines both CF and Content-Based Filtering (CBF). In this approach, the user-movie rating database is taken which enhances existing user data by using CB predictor and then the personalized recommendation is provided through CF. Hence, from their experimental results, they clearly showed how CB CF approach is better than a pure Collaborative Filter or pure CB filter.

Item-Based(IB) and User-Based(UB) CF methods are prevailing techniques to search and retrieve the mashup services from a huge volume of service relevant data but it consumes too much time to compute similarity [6] proposes a bottom-up hierarchical clustering based CF approach which is a solution to decrease the number of services to be processed using functionalities and classifications similar services are grouped into clusters. Recommendations are made based on similar services that are grouped into clusters by their functionalities and classification using k-means clustering is one of the partitioning-based algorithms, it needs additional information for clustering process from users.

CF is mostly used in many domains with variety of algorithms in options. In spite of so many advantages, CF should have the ability to deal with data sparsity to scale with an increasing amount of items and users, scalability, noisy data, shilling attacks, cold start problem, privacy protection and to make a recommendation in a short may reduce its impact by reducing its recommendation accuracy. So as to improve the accuracy and scalability many researchers have worked on it proposing new similarity measures like to get good results H.J.Ahn(2008) in [7] has proposed Proximity Impact Popularity (PIP) a new similarity measure of CF which combines item information, content, and popularity with user-behavior data. Also, to reduce the impact of problems that affects recommendation accuracy of CF, several CF algorithms such as item-based, user-based, model-based, Content-based and so on, are proposed in [8].

[8] have performed a comparison of difference CF algorithm to observe their behavior not only in favorable conditions but also under diverse situations. Two new matrices GPIM & GIM are proposed which contributes to the evaluation of CF systems, that focuses on meaning the quality of recommendation list. In the article a novel strategy of CF is proposed which is based on differences between items and users rather than relying on its similarities and the result shows that their approach is accurate as of that to modern methods also its computational efficiency is comparatively much better.

A key approach of [9] is to find similar items or users using a user-item rating matrix. This research improves recommendation performance in Memory-Based CF algorithm by calculating the similarity between users and items. Pearson Correlation Coefficients (PCC) [10] and cosine [11], cannot capture similar users effectively, especially for the users who do not rate enough items for analyzing similar users. So, combines user ratings local context of each pair's users and its global preference. The same disadvantage of PCC [10] and Cosine [11] is analyzed by many researchers and they have proposed a new similarity measure for CF to improve the accuracy in [12] where Ahn proposed a new similarity measure called Proximity-Impact-Popularity [PIP]. But this similarity deals only with local information of rating and do not consider global preferences. The size of the common user set is not considered by traditional PCC [13] proposed a weighted PCC to solve this problem.

[14] proposed An Improved Online Book Recommender System using Collaborative Filtering Algorithm that developed a recommendation model which uses OOADM, improved CF algorithm and an efficient quicksort algorithm and adjusted cosine similarity algorithm to improve the RS. Its results show that the accuracy of rating followed by normal distribution suggests consistency and efficiency.

Collaborative filtering (CF) such as user-based and item-based methods are popular techniques to retrieve the services from overwhelming services, but it consumes a lot of time. Clustering techniques are the solution to decrease the data size of service. Bottom-up hierarchical Clustering based collaborative filtering approach is used in [4]. In this approach, similar services are grouped into clusters by their functionalities and classifications, and recommendations are made based on the similar services under the same cluster where the K-means clustering algorithm issued for recommending. It is one of the partition-based clustering algorithms. It was applied to the partition of services based on the user's preference. Even though K-means is one of the partition-based clustering algorithms, but for clustering process, it requires additional information from users.

Some researches like [16] use extra information like user interest, user activity, user location for increasing the performance of CF algorithm. User similarity is calculated, and user interest based on item relationship is expanded by Liu. Q & et. Al (2012) [17]. Yehuda Koren (2009) proposed a CF algorithm in [18] where according to target items dynamic adjustments are done on the weight of neighboring user's set based on user activity. All the above studies have reduced the impact of data sparseness and improved the accuracy and scalability of the CF algorithms up to some extent.

III. SYSTEM DEVELOPMENT

In the proposed system we have two main modules,

1) Rating based Recommendation
2) Content-based Recommendation

As we have seen in our studies that both CF (rating based) and CBF have their own advantages and disadvantages individually. In this paper, a Book Recommendation System is proposed which combines the use of Content-Based and Rating based CF techniques which up to an extent solves the problem of data sparsity, cold start, popularity biased and the problem of quality judgment from other users by providing both the approaches parallelly.

The architectural design of proposed Book RS is illustrated in fig. 1 and description of the terms are as below.

1) Login Page: Can log in to the system by entering the credentials that a user has filled in the registration form.
2) Registration Page: Details to be filled by a user which are validated for user account maintenance.
3) Book Recommendation Home Page: This is a Home page of the system which contains all the Books from amazon dataset with images. Logged In users can rate books on this page.
4) User Rating: Ranging from 1 to 5 users are supposed to give a rating to the book according to their preference.
5) Rating based Recommendation Page: User rated books are taken as input; users are searched for similar rated books using PCC and Euclidean distance score for recommending books.
6) Content-based Recommendation Page: Generates a book recommendation list for the logged-in user by matching the content of the book user has clicked in this module. Using the book genre as content and implementing the k-means clustering algorithm on its book is recommended.

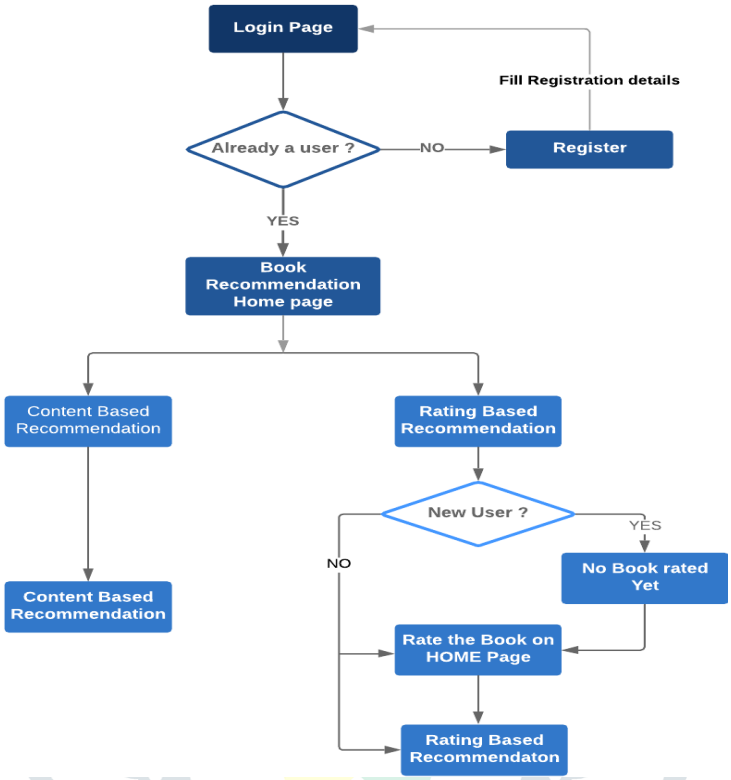

Fig -3: System Flow

Book Recommendation architecture in fig.3 can be explored in steps as,

1) Initially running the python program Login page is generated.
   - If a user is new user, she/he is asked to get registered.
   - If already registered Login credentials are filled.
2) Book Recommendation home page is displayed with the large number of books options to go through.
3) From such a huge volume data which book should a user choose? Book RS provides two options,

   I. **Rating Based Recommendation (RBR)**: recommends the logged-in user similar books to what she/he has rated previously. In the case, a user is a new user and haven't rated a single book yet or enough books he/she will not get the recommendation based on rating.

   Rating Based Recommendation matches the rating by calculating similarity score as below.

   i) Euclidean Distance is calculated (Returns ratio Euclidean distance score of user1 and user2)
      - firstly, both rated books by user1 and user2 are taken
      - condition is checked if both have any common rated book

   ii) Pearson Correlation Coefficient is calculated (Get rated items)
      - check the number of ratings in common
      - add up all the preferences of each user
      - sum up the squares of preferences of each user
      - sum up the product value of both preferences for each user

- calculate the Pearson score
- return the number of similar users for a given specific user
- sort the similar users so that highest score user will appear first
- get a recommendation by using a weighted average of every other user's ranking
- the normalized list is created
- Books are recommended.

Pearson's correlation

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Here,

- r_{xy} is sample Pearson coefficient or sample correlation coefficient when applied to a sample.
- $X_i = X_1 \ldots\ldots X_n$;
- $Y_i = Y_1 \ldots\ldots Y_n$;
- n is sample size;
- Xi, Y_i are the individual sample points indexed with I,
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean of x and analogously for y.

Rearranging the formula, we get r_{xy} as,

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

II. **Content-Based Recommendation (CBR):** recommends user the similar book to what she/he has visited recently by matching its content. In the proposed Book RS genre of the book is matched content using the k-means clustering algorithm book is recommended.

K-Means Clustering Algorithm

It is an unsupervised learning algorithm which means it does not require training data and is iterative based algorithm meaning it calculates cluster centroid repeatedly. It takes 'n' point dataset as input and 'k' an integer parameter for specifying the number of clusters to create. Its output is a set of k clusters centroids mapping each of the data points to a unique cluster. Here in Book RS genre of the book is tried to group in similar clusters where no prediction is involved, k-means clustering algorithm works as below:

i) Handle Data:

- Clean the file, normalize the parameters, given numeric values to non-numeric attributes.
- Read data from the file and split the data for cross-validation.

ii) Find Initial Centroids: Choose k centroids in random. In Book RS k=25, as we have 25 types of genres.

iii) Distance Calculation: Euclidean distance is used to find the distance between each of the data points with each of the centroids.

[19] In Cartesian coordinates, if **p** = (p1, p2,…, pn) and **q** = (q1, q2,…, qn) are two points, then the distance (d) from **p** to **q** , or from **q** to**p,** in Euclidean space is given by,

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$

iv) Clustering: Using the calculated distances between the points we will decide which data point to be in which, among k clusters.

v) Re-calculating the centroids: Again, find the new values for centroid.

vi) Stop the iteration: When there is negligible difference between the new and the old centroids, Stop the algorithm.

IV. PERFORMANCE ANALYSIS

Showing the Rating similarity calculation of the Book RS and comparing the results from the database.

Table 1: Books and their rating by users.

Users → Items(Books)	Samiksha	Ashwini	Yash	LisaRose	Sankalp
The Alchemist(Book1)	X	X	X	X	X
Brave New World(Book2)	X	X	-	X	X
The Tao of Pooh(Book3)	-	X	-	X	-
The Zahir: A Novel of Obsession (Book4)	-	-	X	-	-
The Pilgrimage (Plus) (Book5)	-	-	X	-	X
A Year with Rumi: Daily Readings (Book6)	-	-	X	-	-
The Tibetan Book of Living and Dying: The Spiritual Classic (Book7)	X	-	-	-	-

Table 2: Recommended books based on similar user rating

	Similar Books	Recommended Books
(Samiksha, Ashwini)	Book1,Book2	Book3
(Samiksha, Yash)	Book1	Book4, Book5, Book6
(Samiksha, LisaRose)	Book1,Book2	Book3
(Samiksha, Sankalp)	Book1, Book5	Book5
….		

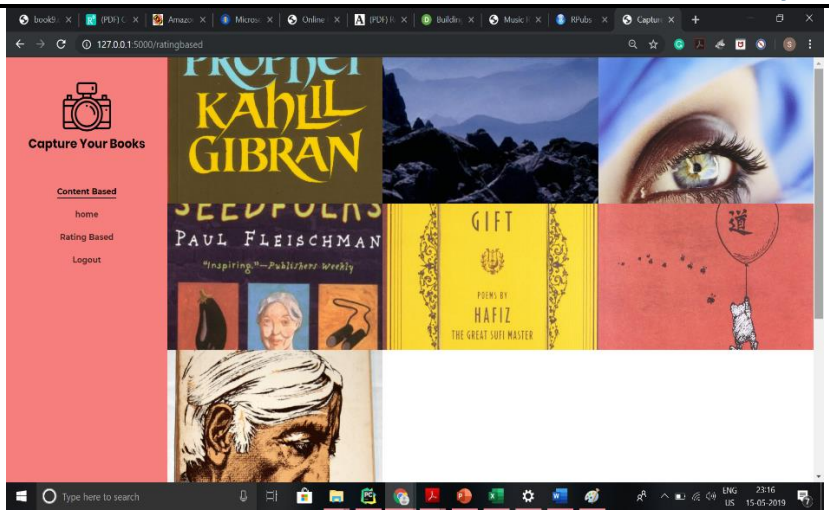

Fig - 4: Output showing Results for recommended books based on similar user rating

- In the analysis Table 1 shows user-book rating matrix. For demonstration purpose, only some of the users are taken from numbers of users and only 7 items from large volume dataset are considered for calculation where Rating Based recommendation with output is shown. So, total Books Recommended to User Samiksha are Book3, Book4, Book5 and Book6. (and more, by comparing with other users not shown in table)
- Likewise, we can recommend books to all other users based on Rating using this concept of matching similarity & calculating distance.
- The performance of the K-means algorithm is evaluated for recognition rate with different numbers of clusters with distance metric mentioned in given in steps of k-means algorithm. Table 3 shows the performance evaluation of k-means clustering algorithm for our dataset where in case of Book Recommendation dataset it has k as 25 i.e. classes of Genre.
- Table 3 and its corresponding graph shows that as the number of clusters increased dataset recognition rate also increases as uniqueness is there and cluster contains almost all similar items whose centroid is not going to change further.

Table 3: Performance evaluation of k-means clustering algorithm at different k values.

	Number of Clusters formed				
	5	10	15	20	25
Dataset Recognition Rate(%)	69.60	72.64	74	75.87	99.33

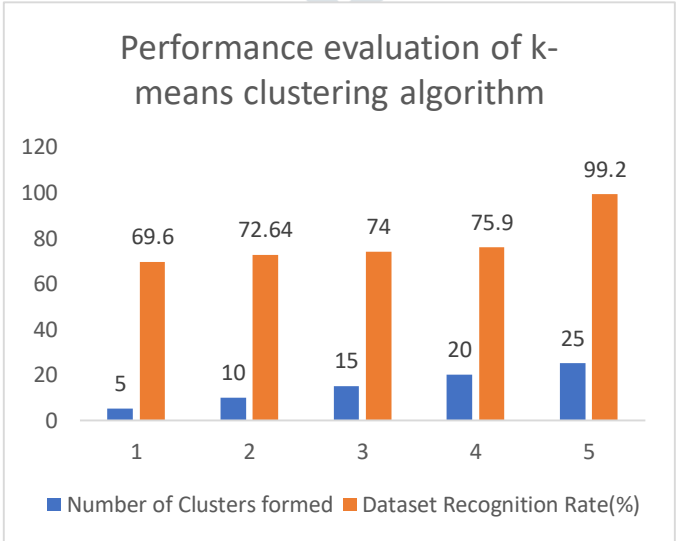

Chart - 1: Performance evaluation of k means clustering algorithm

V. CONCLUSION

Recommendation and personalization are important approaches to reduce information over-load. Various methods of Recommender System are implemented considering one of the either CF or Content-based filtering. In proposed Book Recommendation System both approaches i.e. Collaborative filtering & Content-based filtering are used parallelly which address the problems of data sparsity, cold start problem, popularity biased problem and problem of quality judgement from other users.

REFERENCES

[1] RONG HU, (Member, IEEE), WANCHUN DOU, (Member, IEEE), and JIANXUN LIU, (Member, IEEE): ClubCF: "A Clustering-Based collaborative Filtering Approach for Big Data Application," Digital Object Identi_er 10.1109/ TET.2014. 2310485.

[2] Sarwar, B., Karypis, G., Konstan, J., and Reidi, J. (2001). Item-based collaborative filtering algorithms. GroupLens research group, Army HPC research center, University of Minnesota, Minneapolis, 1-11.

[3] Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1), 76–80.

[4] Li Zhang, Tao Qin PiQiang Teng: An Improved Collaborative Filtering Algorithm Based on User Interest, JOURNAL OF SOFTWARE, VOL. 9, NO. 4, APRIL 2014, doi:10.4304/jsw.9.4.999-1006

[5] Melville, P., Mooney, R. J., and Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. In Proceedings of the eighteenth national conference on artificial intelligence (AAAI-02), Edmonton, Alberta, 187-192.

[6] V. Devi, R. Kanaga Selvi: "An Innovative Approach to Endorse the Best Web Services to Craft Mashup Web Applications Using Bigtable," DOI 10.4010/2014.268 ISSN-2321-3361 © 2014 IJESC.

[7] H. J. Ahn. "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," Information Sciences 178, pp.37-51, 2008.

[8] FIDEL C., V´ICTOR C., DIEGO F.& VREIXO O. Comparison of Collaborative Filtering Algorithms: Limitations of Current Techniques and Proposals for Scalable, High-Performance Recommender Systems [J]. ACM Transactions on the Web, Vol. 5(1), pp.2-33, 2011.

[9] Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, Xuzhen Zhu: "A new user similarity model to improve the accuracy of collaborative filtering," Knowledge-Based Systems 56 (2014) 156–166.

[10] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in: Proceeding of the ACM Conference on Computer Supported Cooperative Work, 1994, pp. 175–186.

[11] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 734–749.

[12] J.L. Herlocker, J.A. Konstan, A. Borchers, J. Riedl, "An algorithmic framework for performing collaborative filtering", in: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 230–237.

[13] Qiu T, Chen G, Zhang Z K, et al. An item-oriented recommendation algorithm on cold-start problem [J]. EPL (Europhysics Letters), 95(5): 58003, 2011.

[14] E. Uko Okon, B. O. Eke,P. O. Asagba. "An Improved Online Book Recommender System using Collaborative Filtering Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 179 – No.46, June 2018.

[15] Samiksha Pande, A. Gaikwad, "A Novel Approach for Smart Shopping Using Clustering-Based Collaborative Filtering", International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, p-ISSN: 2395-0072, Volume: 05 Issue: 02 Feb-2018.

[16] Akihiro Y., Hidenori K.&Keiji Suzuki. "Adaptive Fusion Method for User-Based and Item-Based Collaborative Filtering" [J]. Advances in Complex Systems, 14(2), pp.133- 149, 2011.

[17] Liu, Q., Chen, E., Xiong, H., Ding, C. H., & Chen, J. Enhancing collaborative filtering by user interest expansion via personalized ranking. Systems, Man, and Cybernetics [J]. Part B: Cybernetics, IEEE Transactions on, 42(1), pp.218-233,2012.

[**18**] Yehuda Koren. Collaborative filtering with temporal dynamics. 15th ACM SIGKDD'2009, pp. 447-456, 2009.