A Hybrid Recommendation System for E-Commerce

Shefali Gupta
Computer Science
Jagannath University
Jaipur
shefali.gupta2108@gmail.com

Dr. Meenu Dave
Computer Science
Jagannath University
Jaipur
meenu.s.dave@gmail.com

***Abstract*— There is a huge dependency on e-commerce for goods, electronics, clothes, etc., nowadays. People are searching for e-commerce websites frequently for their daily needs. There are various techniques to filter out relevant data and make it available to the user but it remains difficult to provide da-ta that fits user preferences because of which recommendation system came into play. Recommendation system filters data based on users preferences and refers this filtered data to the user. These systems use machine learning, data mining and various other algorithms to accomplish this task. Collaborative and content-based filtering are amongst the most popular recommendation system techniques. Both these techniques have numerous limitations hence this paper aims to provide a hybrid algorithm that combines collaborative filtering as well as content-based filtering to levy the benefits of both. A comparative analysis of the proposed hybrid algorithm against the content and collaborative filtering methods has also been carried out.**

Keywords*— Recommendation system, Content-based filtering, Collaborative filtering, Hybrid Algorithm

I. INTRODUCTION (*HEADING 1*)

Every year, e-commerce websites introduce a vast amount of items on their websites. This huge amount of information urges the need to develop a system that filters out relevant information and provide recommendations that fits the user's interest [20]. Over the past decade, many recommendation systems have been developed in various areas. These systems aims at providing personalized recommendation to the users by understanding taste of each user and finding items that might be desirable by them [2,3,8,15]. Personalized recommendation system uses various methods such as content based filtering, collaborative filtering, hybrid filtering, and knowledge based filtering and so on [5].

Content based filtering looks at features of items and users taste in order to recommend similar items to the user [18]. On the other hand collaborative filtering uses user's behavior to create a neighborhood of similar users and then provide recommendations [16]. Both these filtering techniques have advantages as well as disadvantages with cold start, diversity, scalability and data sparsity being some of the challenges faced by these techniques [11]. To bring synergy between them, hybrid algorithms are developed that uses combinations of various recommendation techniques. Numerous researchers have combined Collaborative and Content-based filtering together to levy the advantages of both.

Burke [4] divided hybrid techniques in seven different classes:-

1. Weighted: It linearly combines recommendations obtained from different recommendation techniques
2. Switching: This technique switches from one technique to another based in certain defined criteria
3. Mixed: This technique provides recommendations from several techniques at the same time
4. Feature combination: This technique combines features from different recommendation techniques and input it in a single recommendation technique
5. Cascade: Output from one technique is used as an input to other technique to improve the results
6. Feature Augmentation: Output from one technique is injected as an input feature to other technique
7. Meta-level: Model learned on one recommendation technique is used on other technique

In particular, the main contribution of this paper is as follows:

1. A hybrid algorithm that combines content-based and collaborative based filtering using a cascade method has been proposed
2. This algorithm has been implemented on mobile data that consists of mobile ratings datasets and mobile attributes datasets
3. The performance of the above algorithm is measured using the recall parameter and comparative analysis is presented with other recommendation techniques

The rest of the paper is arranged as follows. Section 2 provides background knowledge about various hybrid algorithms developed so far. The Proposed hybrid algorithm is explained in detail in Section 3. In Section 4, experimental results obtained from the above developed hybrid algorithm are displayed along with comparative analysis of results obtained from other recommendation techniques. Finally, In Section 5, the conclusion of the paper and future research work is presented.

II. BACKGROUND

To take advantage of recommendation system techniques, various algorithms have been proposed by researchers that uses a mixture of collaborative filtering or content-based filtering or any other recommendation techniques.

Jain, Kartik & Kumar, Vikrant & Kumar, Praveen & Choudhury, Tanupriya [1] developed a hybrid algorithm that calculates the similarity of users among others grouped around various genres. This algorithm uses user's preference of movies in terms of genres as the deciding factor for providing recommendations of movies to the user.

Doke N., Joshi D [7] combined user and item-based collaborative filtering model for providing suggestions to the user. Ming Li, Ying Li, Wangqin Lou, Lisheng Chen [13] developed a hybrid algorithm for Q&A documents in which they combined collaborative filtering, content-based filtering, and a complementarity-based recommendation method to find documents that match user needs. Paul, Sutanu & Das, Dipankar [14] used item-based and user-based filtering algorithms along with factor-based hybrid models to build an effective recommendation system for an artist.

Researchers like Shan Liu, Yao Dong, and Jian ping Chai [17] introduced a hybrid algorithm for the news recommendation system. They improved the correlation coefficient formula by adding news as an important parameter while calculating the similarity of users in order to provide better recommendations to the users.

These researches show that a combination of two or more recommendation techniques result in a more robust model as limitations of one technique is overcome by the strengths of other technique.

III. PRELIMINARIES

This section describes the structure of the collected data, the statistics developed on the data, and the preprocessing needed to be done before using the data. Section III-A explains the structure of the data. Section III-B presents the insights obtained from the data. Finally, in Section III-C, a detailed description of the preprocessing needed to be done on the collected data.

A. Structure of data

The dataset has been collected from the Kaggle website and it depicts contextual data on smartphones and its reviews on Amazon. This dataset has been widely used in a variety of recommendation systems as it contains both feature-related data as well as ratings given by users to different items.

This dataset contains of two files: one is the attribute dataset and the other is reviews dataset. Table 1 Depicts the attributes of the smartphones used in research divided under various categories. Review dataset consists of three columns: UserID, ItemID, rating.

CATEGORY	ATTRIBUTES
NETWORK	• Memory card slot • SIM
BODY	• Edge • Size • Weight
DISPLAY	Multitouch
PLATFORM	• OS • CPU
MEMORY	• Internal memory • RAM
MAIN CAMERA	Primary camera
SOUND	• Loudspeaker • 3.5mm jack
BATTERY	• Standby
MISCELLEANEOUS	• Alarm • Alert types • Brand • Price

Table 1: Various contextual features of smartphones

B. Explored statistics

This section gives an overview of descriptive statistics on entire dataset. Table 2 describe some statistics obtained from the above datasets.

# of unique users	45830
# of unique items	720
# of transactions	67987
Maximum number of items purchased by a single user	6514
Minimum number of items purchased by a single user	1
Average number of items purchased by a single user	1.5

Table 2. Overall statistics

Fig 1. Shows the top 10 items purchased by the user. Item54 has been purchased 981 items followed by item29 which is purchased 925 times by the users. These items are the ones that has been purchased maximum number of items by the user.

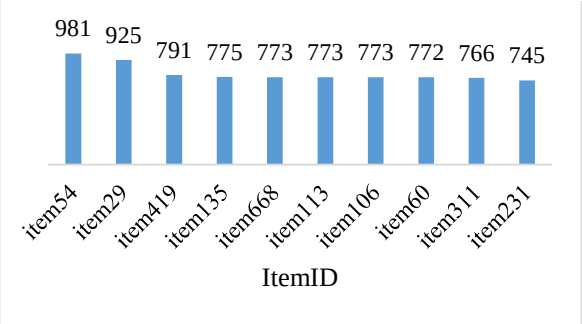


Fig 1. Top 10 popular items

In table 3, count of each rating has been presented. 37752 items have been given rating 5 by the users followed by rating 4 scored by 8824 items. On contrary, 12743 items have been rated with rating 1 by the users.

Rating	Count of ratings
1	12743
2	3915
3	4752
4	8824
5	37752

Table 3. Count of each type of ratings

C. Preprocessing data

From the data it is observed that the userId column, as well as the movieId column, consist of integers. Furthermore, each attribute needs to be converted into a more usable format to be used by the users. In order to do so, a one-hot encoding is used to create a matrix that comprises of corresponding attribute for each of the items.

This encoding is needed for supporting categorical data. Each different attribute is stored in columns that contain either 1 or 0. 1 shows that an item has the attribute present and 0 shows that it doesn't. Figure 2 represents one hot encoding for the above dataset.

Item	Brand	SIM	OS	CPU
item40	Apple	No	iOS	Dual-core
item41	Ericsson	Mini-SIM	Android	Dual-core

One-hot encoding

itemID	Brand.Apple	Brand.Ericsson	SIM.Mini.SIM	SIM.No	OS.Android	OS.iOS.	OS.symbian	CPU.Dual.core
item40	1	0	0	1	0	1	0	1
item41	0	1	1	0	1	0	0	1

Fig 2. One hot encoding

Each missing attribute has been replaced with value zero indicating it is not present. Also, for reviews dataset, all items that are not rated by the user has been replaced by calculating the mean rating for each user and subtracting it from each rating of a user to calculate the adjusted rating.

To improve the quality and speed of the recommendation system, data dimensionality reduction is required as shown in fig 3. Data has a very high dimensional space but it is preferred not to. Features that barely express similarity between articles, can be removed from the feature set to significantly reduce the number of features and to improve the quality of the recommendations. Additionally, a smaller dataset improves the speed of the recommendation system.

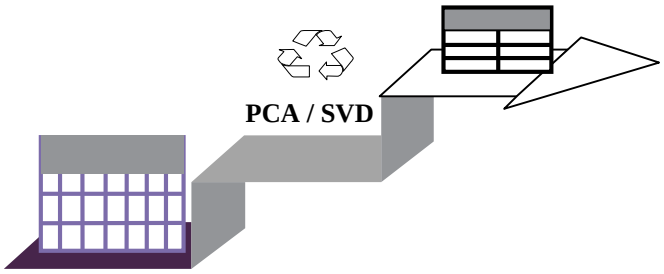


Fig 3. Dimensionality reduction

The problem of dimensionality reduction can be represented as follows. Consider a dataset **X** represented as having n rows and D columns consisting of n data vectors $\mathbf{x_i}$ ($i \in \{1, 2, \ldots, n\}$) with dimensionality D, reduced to a representation $U^{n \times d}$ where d<<D, to keep as much information as possible. In this case, two dimensionality reduction techniques have been used described as below:

1. **Principal Component Analysis (PCA)**

PCA allows to obtain an ordered list of components that account for the largest amount of the variance from the data in terms of least square errors. It helps in obtaining important features from a large dataset with 3 or more dimensions. Reducing the dimensions of dataset helps in visualizing it a better way as shown in fig 4.

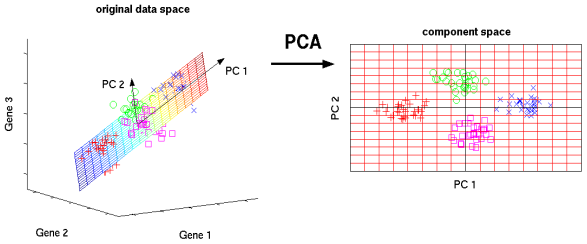


Fig 4. Principal Component Analysis

1. Singular value decomposition

IV. Proposed Hybrid Filtering Algorithm

In this section, proposed hybrid algorithm has been discussed which is developed by combining content based filtering and collaborative filtering using cascade method. Cascade method helps in refining the recommendations obtained by applying first technique.

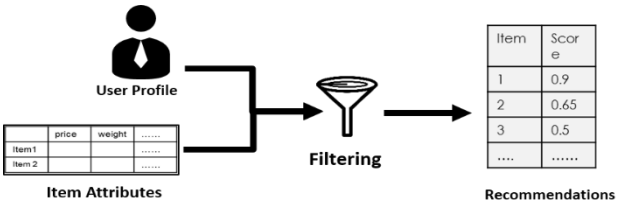


Fig. 1. Content Based Filtering

Fig. 1 depicts the flow of content based filtering. Content based filtering takes user profile and item attributes as input and suggest items to the user based on their interest. For example, if a user likes movies starring Robert Downey Jr, the system tracks its choice and recommends movies starring Robert Downey Jr.

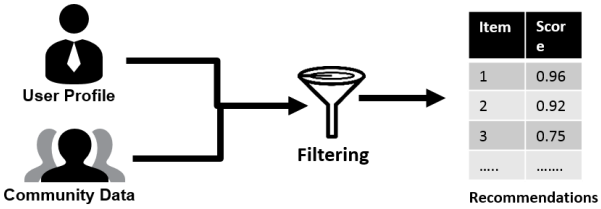
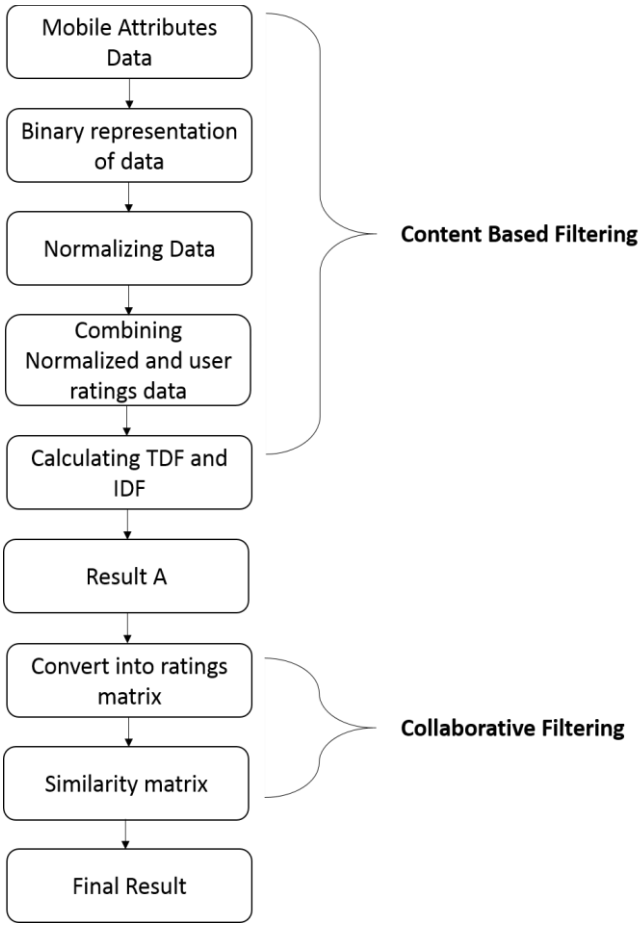


Fig. 2. Collaborative Filtering

Fig. 2 depicts the working of collaborative filtering. Collaborative filtering takes user as well as community data as input and recommends items to the user based on items liked by similar users [12]. For example, if a user A and user B has same preferences in movies and user B has liked some

movies different than user A, system will recommend other movies liked by user B to user A.



.

Fig. 3. Proposed Hybrid Filtering Algorithm

Proposed hybrid algorithm combines content based filtering and collaborative filtering. Fig. 3 depicts general procedures of hybrid algorithm which will be described in detail as follows:

- **Step 1**: Content based filtering works by recommending items to the user on the basis of its profile. In order to do it, item attribute data is first converted into binary representation shown in Fig. 4.

Items	Brand	SIM
Item9	Apple	Nano
Item10	Apple	Nano
Item11	Apple	Nano
Item12	Apple	Nano
Item13	Apple	Nano

Items	Brand.Apple	Brand.Ericsson	SIM.Micro	SIM.Mini	SIM.Nano
Item9	1	0	0	0	1
Item10	1	0	0	0	1
Item11	1	0	0	0	1
Item12	0	1	0	1	0
Item13	0	1	0	1	0

Fig. 4. Binary representation of data

- **Step 2:** Each data vector is then normalized before any similarity calculations is to be done so that ratings lie between 0 or 1 instead of getting 0 or 1.

  This is done by calculating the magnitude of all the items by taking the square root of the sum of the squares of all the item attributes for each data vector.

$$Magnitude = \sqrt{x^2 + y^2 + z^2 + \cdots} \quad (1)$$

  New data vector is then created by dividing item's attribute value with the magnitude value obtained in (1).

$$Data\ vector = \frac{x}{magnitude}, \frac{y}{magnitude}, \ldots. \quad (2)$$

- **Step 3:** User taste matrix is obtained by taking product of ratings for each item with the data vector obtained in (2).
- **Step 4:** In order to calculate similarity between items, TF (Term Frequency) and IDF (In-verse Term Frequency) are used [7].

$$TF = \frac{\text{Frequency of the word in the document}}{\text{total number of words in document}} \quad (3)$$

$$IDF = log10(\frac{\text{Total number of document}}{\text{Number of documents having the term}}) \quad (4)$$

  For example, in case word apple appears 10 times in the dataset with 100 total words, it will represent a term frequency of 10 and inverse document frequency of 1.

  These measures help in evaluating the importance of a word in the document. As content based filtering depends largely on contextual information about items, this helps in providing more information about item in order to distinguish between them.

- **Step 5:** Final scores are obtained by taking sum product of user taste matrix and IDF matrix calculated above.
- **Step 6:** Collaborative filtering algorithm takes final scores matrix as input and convert it into ratings matrix for applying similarity measure [19].
- **Step 7:** Similarity measure is used to build a user-user similarity matrix through which neighbors for the current users can be computed.
- **Step 8:** Similarity matrix thus obtained is then merged with the actual user ratings data by applying various aggregation functions resulting in final dataset of items that is recommended to the user.

V. Experimental Evaluation

The evaluation of results of any model is very important, therefore recommendation system need these metrics to compare the accuracy of the model. There are several types of evaluation metrics used for comparing model such as MAE, MAPE Precision, and Recall etc. [10]. In this section, recall metric will be used for evaluating the performance of the model [9]. Recall also known as sensitivity is defined as:

$$Recall = \frac{True\ psitive}{true\ postive + false\ negative} \quad (5)$$

Complete dataset is divided into two parts: train dataset (80%) and test dataset (20%). Recall measures how accurately the predictions made from train data matches with test data. Following chart compares various recommendation techniques against the proposed hybrid model.

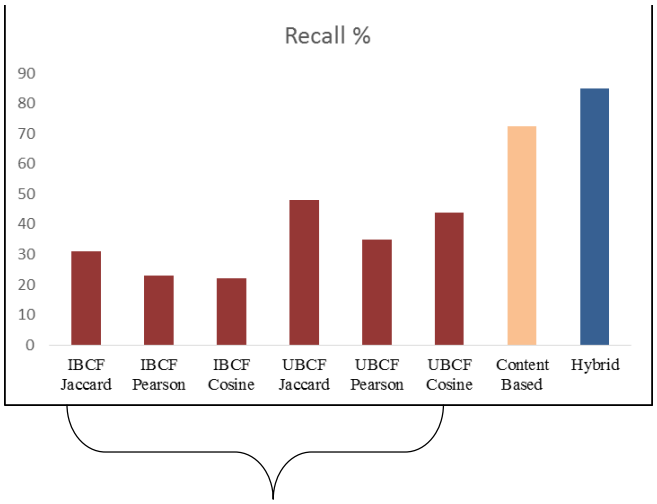


Fig. 5. Comparison of Recall among different recommendation techniques

As shown in Fig. 5, the predictive accuracy of item based collaborative filtering is the lowest. The user-based CF is slightly better than the item based algorithm on account of the more accurate neighbor selection. Content based filtering performs better that collaborative filtering with 72% recall measure. The proposed hybrid algorithm works best compared to all recommendation techniques mentioned here. Therefore, the proposed hybrid model is proved to improve the quality and accuracy of recommendation.

VI. Conclusion and Future Work

The main contribution of this paper was to describe the framework of proposed hy-brid algorithm. This method combines content based and collaborative filtering to leverage the advantages of both the techniques. The experiment result on Mobile data set confirms that the proposed model outperforms other recommendation tech-niques.

In the future work, Aim is to do following things:-

1. Above mentioned algorithm will be applied on various other datasets in order to verify the feasibility.
2. Modifying it to be able to solve cold start problem.

REFERENCES

[1] Aggarwal, C. C. (2016). An introduction to recommender systems. Recommender systems. Cham: Springer International Publishing1-28. https://doi.org/10.1007/978-3-319-29659-3_1.

[2] A.H. Celdrán, M.G. Pérez, F.J. García Clemente, G.M. Pérez, Design of a recommender system based on users' behavior and collaborative location and tracking, J. Comput. Sci. 12 (2016) 83-94.

[3] Batmaz, Z., Yurekli, A., Bilge, A., & Kaleli, C. (2019). A review on deep learning for recommender systems: Challenges and remedies. Artificial Intelligence Review, 52(1), 1-37. https://doi.org/10.1007/s10462-018-9654-y.

[4] Burke R (2002) Hybrid recommender systems: survey and experiments. User Model User Adapt Interact 12:331-370

[5] Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. Intelligent Data Analysis, 21(6), 1487-1524. https://doi.org/10.3233/IDA-163209.

[6] Das, N., Borra, S., Dey, N., & Borah, S. (2018). Social Networking in Web Based Movie Recommendation System. Social Networks Science: Design, Implementation, Security, and Challenges, 25-45. doi:10.1007/978-3-319-90059-9_2

[7] Doke N., Joshi D. (2020) Song Recommendation System Using Hybrid Approach. In: Bhalla S., Kwan P., Bedekar M., Phalnikar R., Sirsikar S. (eds) Proceeding of International Conference on Computational Science and Applications. Algorithms for Intelligent Systems. Springer, Singapore

[8] F. Ricci, L. Rokach, B. Shapira, Recommender systems: introduction and challenges, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, Boston, MA, 2015, pp. 1-34.

[9] García-Sánchez, F., Colomo-Palacios, R., & Valencia-García, R. (2020). A social-semantic recommender system for advertisements. Information Processing & Management, 57(2), 102153. doi:10.1016/j.ipm.2019.102153

[10] Isinkaye F, Folajimi Y, Ojokoh B (2015) Recommendation systems: principles, methods and evaluation. Egypt Inform J 16(3):261-273. https://doi.org/10.1016/j.eij.2015.06.005

[11] Jain, Kartik & Kumar, Vikrant & Kumar, Praveen & Choudhury, Tanupriya. (2018). Movie Recommendation System: Hybrid Information Filtering System. 10.1007/978-981-10-7245-1_66.

[12] Logesh, R., Subramaniyaswamy, V., Malathi, D., Sivaramakrishnan, N., & Vijayakumar, V. (2018). Enhancing recommendation stability of collaborative filtering recommender system through bio-inspired clustering ensemble method. Neural Computing and Applications. doi:10.1007/s00521-018-3891-5

[13] Ming Li, Ying Li, Wangqin Lou, Lisheng Chen. (2019) A hybrid recommendation system for Q&A document. In; Expert Systems with Applications, Volume 144, 15 April 2020, 113088.

[14] Paul, Sutanu & Das, Dipankar. (2020). User-Item-Based Hybrid Recommendation System by Employing Mahout Framework. 10.1007/978-981-13-7403-6_32.

[15] P.G. Campos, F. Díez, I. Cantador, Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols, User Model. UserAdap. Inter. 24 (2014) 67-119.

[16] Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web. ACM, pp 285-295

[17] Shan Liu, Yao Dong and Jianping Chai, "Research of personalized news recommendation system based on hybrid collaborative filtering algorithm," 2016 2nd IEEE International Conference on Computer and Communications (ICCC), Chengdu, 2016, pp. 865-869

[18] Son, J., & Kim, S. B. (2017). Content-based filtering for recommendation systems using multiattribute networks. Expert Systems with Applications, 89, 404-412. doi:10.1016/j.eswa.2017.08.008

[19] X. Yang, Y. Guo, Y. Liu, H. Steck, A survey of collaborative filtering based social recommender systems, Comput. Commun. 41 (2014) 1-10.

[20] Zhang, T., Li, W., Wang, L., & Yang, J. (2019). Social recommendation algorithm based on stochastic gradient matrix decomposition in social network. Journal of Ambient Intelligence and Humanized Computing. doi:10.1007/s12652-018-1167-7