Collaborative filtering and Content based recommendation system for scaling businesses

Raushan Raj
IIT2018031
B.Tech(IT)
IIIT Allahabad
Prayagraj, India

Suryasen Singh
IIT2018069
B.Tech(IT)
IIIT Allahabad
Prayagraj, India

Rahul Kumar Yadav
IIT2018071
B.Tech(IT)
IIIT Allahabad
Prayagraj, India

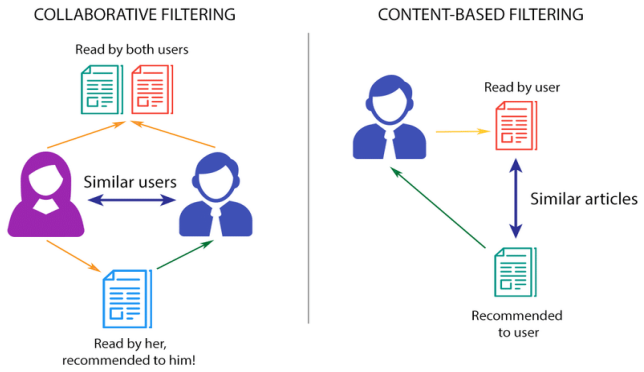
***Abstract*—In 21st century, every domain is getting automated as we are witnessing the immense growth in the application of artificial intelligence by each passing day.Apart from this, machine learning has made human life easy by enabling the systems to learn the things on their own thus reducing human effort. In recent days, people are accustomed to shopping things for their needs every day. People do spend hours in searching for the product, they wish to purchase. So this is where our recommendation system comes into picture. Our recommendation system will ease their work in such a way that consumers will spend less time in filtering the items, because our system will show them the recommended items to be sold.This recommendation system tends to analyze the things, where we can find the products which are most frequently bought and which the customer like most and wishes to buy. These types of product will be recommended for them.It increases the overall sale percentage too. Through this model, our main aim is to reduce the human effort by suggesting them the recommended items. The proposed recommendation method uses a machine learning algorithm to provide best possible suggestion.**

***Index Terms*—Product Recommendation, collaborative Filtering, ML, etc.**

I. Introduction

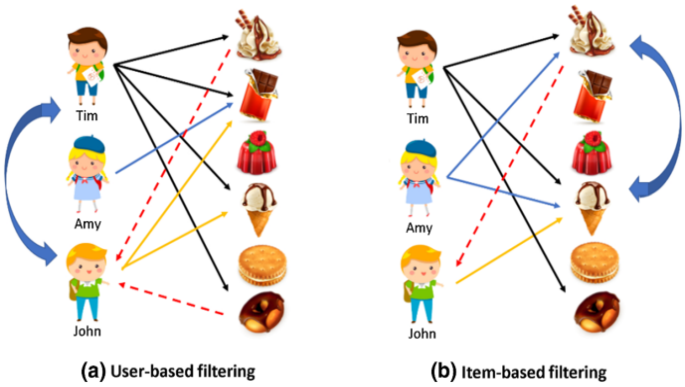
The advent of technology brings us many advanced fields like Artificial Intelligence, Machine Learning, Natural language processing, Deep Learning etc. Now every field is getting automated with the help of Artificial Intelligence and things are changing rapidly and so are the e-commerce sector. Because of E-Commerce, the life-style of people have changed completely. However, some problems remain unsolved. Suppose if the number of products are adequate and if the products are similar to each other or if the consumer doesn't know the domain very well , he/she could be lost and feel frustrated facing this large number of choices. One of the common solution to all these problem is to use the Product Recommendation Systems, which proactively suggest products to the consumers according to their specific requirements. It helps to increase customer satisfaction, which in turn to enhance brand recognition and improved market performance for the organization. Many algorithms have been used by different product recommendation system for providing accurate items to the consumers. But now a days, some recommendation system allow everyone to provide review or rating about the product , no matter whether they have purchased that particular product or not.

Recommendation Technique	Description
Collaborative Filtering	Recommends items to the user based upon the recommendations of similar users
Content Based Filtering	Recommends items to the user based on the purchase history and profile of user



Mentioned below are the techniques which has been used for the above mentioned aim

1) Collaborative based Filtering: It is a machine Learning technique which is used to find a connection between pieces of information.This type of approach is mostly used in systems where there is need of some kind of prediction or recommendation of product or service using the similarities between user data and items.for instance suppose there are two users A and B and both like an item X, additionally user B also likes item Y, then item Y could be recommended to user A by the system.



(a) User-based filtering (b) Item-based filtering

There are two ways to perform collaborative filtering.

1) User to user based collaborative filtering.
2) Item to item based collaborative filtering

2) Content based filtering: Content-based filtering uses features of the items to recommend other items similar to what the user likes, based on their latest action. It doesn't need any data about other users, since the recommendations are specific to this user.

There are various techniques which fall under content based filtering some of them are listed below

1) Cosine similarity.
2) Manhattan distance.
3) Euclidean distance.

II. Problem Statement

The fundamental problem of our project is the recommendation of online grocery items or products based on customer's purchase history and ratings provided by other users who bought similar items.This system is based on the journey of a new customer from the time he/she lands on the business's website for the first time to when he/she makes repeat purchases.Since the system is majorly based on collaborative filtering technique, it computes the connection among multiple users and based upon their ratings, it helps in predicting products for a particular user by identifying patterns based on preferences from multiple user data.

III. Literature Review

We have got some useful articles related to recommendation system by doing a comprehensive online search. The below mentioned papers have also focused on recommendation system but they differ from our system either in the strategy or the aim.

Kumar et al. [1] proposed the movie recommendation system called "MOVREC" which is based on collaborative filtering technique. Mentioned technique utilizes the information given by the customer. That information is analyzed and a movie is recommended to the customer which are arranged with the movie with highest rating first.

AnuPrabha et al. [2] proposed the music recommendation system based on collaborative filtering technique.Collaborative Filtering algorithm is one of the popular successful techniques of recommendation system, which aims to find users closely similar to the active one in order to recommend items. Collaborative filtering with alternating least squares algorithm is the most imperative techniques which are used for building a music recommendation engine

Mittra et al. [3] proposed Book Recommendation System using Machine learning. The proposed system used the K-means Cosine Distance function to measure distance and Cosine Similarity function to find Similarity between the book clusters. Sensitivity, Specificity, and F Score were calculated for ten different datasets. The average Specificity was higher than sensitivity, which means that the classifier could re-move boring books from the reader's list.

Sharma et al. [5] presented Movie Recommendation System Using Item Based Collaborative Filtering. It aims to provide users with accurate movie recommendations.The objective of this system is to recommend movies to our users based on their viewing history and ratings they provide

Tayade et al . [6] in this paper presented a deep learning based product recommendation system.It uses transfer learning to elicit the rich information from the product images, and the use of cosine similarity approach, the user is provided with eclectic recommended products depending on their choices.

IV. Methodology

Even though the system is majorly based on collaborative filtering techniques but the system has been designed in three phases and each phase uses different kind of strategy/techniques as their role of work is different for different situations. Suppose when new customers, who have no previous purchase history with the system, arrive for the first time, they would be recommended the most popular items sold on the website by the system and Once, they purchase anything on the website, the system updates and thus by using collaborative filtering techniques it would recommend other products based on the purchase history and ratings provided by other users on the website.

For the above discussed purposes two types of data-set has been used, they are

1) Amazon product rating dataset.
2) Flipkart product description dataset.

As stated above the recommendation system has been designed in three phases/parts:

1) Popularity based recommendation system: It is one of the best and easiest way to target the new customers who have no purchase history with the system.The reverse sorting technique on the "Amazon product rating" dataset provides

us the product which has been rated the most number of times.

2) Content based recommendation system: This section focuses on recommending items to users based on the features of the items he/she had recently purchased. Unlike collaborative based filtering it does not depend on the other users. It extracts information about the product from the description and finds other items based on that information.

3) Item to item based collaborative filtering system: This section focuses on recommending items to users based on their purchase history. This item to item based collaborative filtering focuses on the relationship between items.

V. Methods/Techniques used:

A. Sorting:

Sorting a process of arranging item or values in a systematic manner. a set of values can either be arranged in an ascending order or descending order.In our model, reverse sorting (i.e decreasing order) has been performed at almost every level to obtained the top products.

B. Text/word vectorization:

word vectorization also known as word embeddings is a methodology in natural language processing to map words or phrases from vocabulary to a corresponding vector of real numbers. Below are some of the most commonly used word embedding techniques:

1) Count Vectorizer
2) TF-IDF Vectorizer
3) Hashing Vectorizer
4) Word2Vec

Among all of the above mention vectorizer TF-IDF vectorizer is a better choice as it not only focuses on the frequency of words present in the documents but also provides the significance of the words. It reduces the complexity of model building by removing words that are less significant for the analysis and thus reducing the input dimensions.

TF-IDF is a product of two parts:

1) TF (Term Frequency) : Frequency of occurrence of a particular word in a sentence.

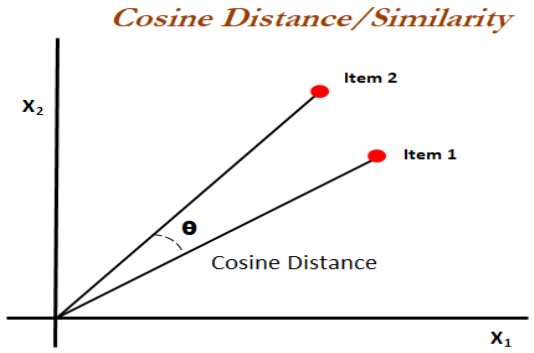
$$TF(w,d) = \frac{occurences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

2) IDF (Inverse Document Frequency) : Natural log of ratio of total number of sentences in a document to the number of sentences containing that particular word.

$$IDF(w,D) = \ln\left(\frac{Total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w}\right)$$

C. Cosine similarity:

It is a mathematical technique which is used to measure the similarity between two sequences of numbers (also called vectors).It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

D. Correlation coefficient:

It is used to determine the relationship strength of the two variables. There are many techniques to achieve the mentioned aim but the most famous and popularly used is Pearson correlation. Pearson correlation generates values ranging in the closed interval of -1 to +1 where -1 denotes most negative correlation, 0 denotes no correlation and 1 denotes most positive correlation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

another.

Step 6: A reverse map of indices and product name is constructed.

Step 7: This is the last step where top 10 most similar products are recommended based on the similarity score generated previously.

VI. Implementation:

A. *Popularity based recommendation system:*

Step 1: It involves loading the dataset and importing the required libraries.

Step 2 : At this step, a dataframe is created.

Step 3: After the creation of dataframe, the presence of null values are checked and if present then removed.

Step 4. At this step, reverse sorting is done based on the number of times a product has been rated so that the product which has been rated the most occurs at the top and hence top 10 such results are displayed.

B. *Content based recommendation system:*

Step 1: It involves loading the dataset and importing the required libraries

Step 2 : At this step, a dataframe is created and only the required fields such as product name and description are taken.

Step 3: At this step, all the null values are removed and a document is formed which consists only of product description.

Step 4: This step involves word vectorization i.e conversion of words into vectors.TF-IDF vectorizer has been used for this purpose. After passing the description document to TF-IDF vectorizer, a TF-IDF matrix is returned

Step 5:At this step cosine similarity is performed on the TF-IDF matrix which gives a similarity matrix containing values representing the similarity of all items with one

C. *Item to item based collaborative filtering:*

Step 1: It involves loading the dataset and importing the required libraries

Step 2 : At this step, a utility matrix is prepared, a utility matrix is a NXM userId to productId matrix and the value inside cell is the rating that the customer has given to the corresponding product.

Step 3: At this step, products which have been rated less than 10 times are dropped from the utility matrix and cells containing null value are filled with 0.

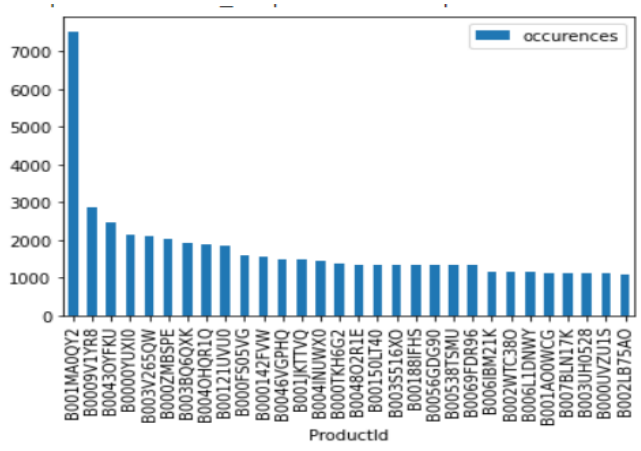
Step 4: It involves preparation of correlation matrix, correlation of all items with one another.

Step 5: at this step, a user purchase history is prepared and passed to a function one by one, the function returns a set of values which are actually the similarity score of the purchased product with the other products.

Step 6: At last top 10 products with the highest similarity scores are displayed as the final result.

VII. Results:

Here are the results of all the 3 parts in the order mentioned above:



Below are the recommendations for the product - HRS ULTIMATE MEN Chest Pads

	Product name	Similarity score
0	HRS ULTIMATE MEN Chest Pads	1.000000
1	HRS ULTIMATE BOY Chest Pads	0.891930
2	HRS ULTIMATE BOY Elbow Pads	0.740572
3	HRS CLUB BOY Thigh Pads	0.721257
4	Parth Collection English Y-pad	0.171612
5	India Inc Women's Solid Casual Shirt	0.159057
6	TeeMoods Casual Full Sleeve Striped Women's Top	0.154496
7	TeeMoods Casual Full Sleeve Striped Women's Top	0.154496
8	TeeMoods Casual Full Sleeve Striped Women's Top	0.154496
9	TeeMoods Casual Full Sleeve Striped Women's Top	0.154496
10	TeeMoods Casual Full Sleeve Striped Women's Top	0.154496

Below are the recommendation based on the User history - [('1403790965', 4), ('B0002VHBTU', 5), ('535795531X', 2)]

```
B0002VHBTU    2.499806
1403790965    1.499566
B000052YKM    0.075251
B000280SI0    0.041307
B00028WLG6    0.034251
B000140MQE    0.031249
B0000C0XL8    0.030753
B000142P12    0.023551
B0000D0BWH   -0.000462
B0000589YB   -0.000480
dtype: float64
```

VIII. Conclusion:

Our proposed model is very much useful for recommending products to both the type of customers i.e new as well as the existing customers. When a new customer who hasn't bought anything from the system yet, then the recommendations will be based on the popularity i.e the product which has been rated the most by now.The content based recommendation mainly focuses on the customer who has just performed the order or has no to limited number of purchase history as because collaborative filtering can't be performed on such cases.The content based system uses the product description for feature extraction and the information extracted is used in finding closest related products.The collaborative based filtering uses customer's rating for it's completion.There were two ways to do collaborative filtering i.e 1) user to user based and 2) item to item based. The best among these two ways is item to item based filtering as gives better result and also it is fast as compared to user to user based. The item to item based filtering uses Pearson's correlation to find the similarity between the product based on the rating which the user or customer has given. With the increase in the use of internet, almost every kind of purchase is done through an application or website so such scenario gives us a validation that it has a great future ahead. This model's approach can be used for various other purposes such as movie recommendation, book recommendation, song recommendation and many more.

References

[1] Manoj Kumar, D.K Yadav, Ankur Singh, Vijay Kr. Gupta, "A Movie Recommendation System: MOVREC" International Journal of Computer Applications (0975 –8887) Volume 124 – No.3, August 2015.

[2] AnuPrabha P S, HarsithaN, Vaishnavi K, Dr.P.Velvadivu and Dr.M.Sujithra, "A Music Recommendation System" International Journal of Advances in Engineering and Management (IJAEM) Volume 2, Issue 8, pp: 635-639

[3] Dhiman Sarma, Tanni Mittra, Mohammad Shahadat Hossain, "Book Recommendation System using Machine learning" International Journal of Advanced Computer Science and Applications, Vol. 12, No. 1, 2021

[4] Yading Song, Simon Dixon, and Marcus Pearce, "A Survey of Music Recommendation Systemsand Future Perspectives" 9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)

[5] Poonam Sharma, Lokesh Yadav,"A Movie Recommendation System" International Journal of Innovative Research in Computer Science Technology (IJIRCST) ISSN: 2347-5552, Volume 8, Issue 4, July 2020

[6] Akshit Tayade, Vidhi Sejpal, Ankit Khivasara, "Deep Learning Based Product Recommendation System and its Applications" International Research Journal of Engineering and Technology (IRJET) Volume 8, Issue 4, Apr 2021

[7] Sheela Kathavate, "Music Recommendation System using Content and Collaborative Filtering Methods" International Journal of Engineering Research Technology (IJERT) ISSN: 2278-0181, Vol. 10 Issue 02, February-2021

[8] Debashis Das, Laxman Sahoo, Sujoy Datta, "A Survey on Recommendation System" International Journal of Computer Applications (0975 – 8887) Volume 160 – No 7, February 2017

[9] Sushama Rajpurkar, Pooja Malhotra, Darshana Bhatt, "Book recommendation system using Content and Collaborative Filtering Methods" IJIRST –International Journal for Innovative Research in Science Technology— Volume 1 — Issue 11 — April 2015

[10] Kunal Shah, Akshay Kumar Salunke, Saurabh Dongare, Kisandas Antala, “Recommender Systems: An overview of different approaches to recommendations”,Dept. Computer Science, Sinhgad Institute of Technology, Lonavala, India International Conference on Innovations in information Embedded and Communication Systems (ICIIECS), 2017.