Online Book Recommendation System by using Collaborative and Popularity based filtering

Abstract - Recommendation system is an algorithm that suggests relevant items to users. The systems that are designed to recommend things to the user based on many different factors i.e popularity, content, collaborative filtering based and hybrid. These systems predict the most likely product that the users are most likely to purchase and are of interest to by filtering a large amount of data present. This paper solves the data sparsity problem by combining the popularity-based filtering , collaborative-based filtering and associate mining to achieve better performance. The results obtained are demonstrated and the proposed recommendation algorithms perform better and solve the challenges such as data sparsity and scalability.

Keywords: Collaborative filtering, popularity-based filtering , associate mining

I. INTRODUCTION

A recommendation system seeks to predict the “rating” or “preference” a user would give to an item.It deals with a large volume of information present by filtering the most important information based on the data provided by a user and other factors that take care of the user’s preference and interest. They are primarily used in commercial applications. Examples of such applications include recommending products on Amazon, music on Spotify, book recommendations on goodreads etc. Recommendation systems use following technologies to recommend products: Popularity-based filtering , Collaborative filtering & Associate mining. Popularity-based systems check about the products which are in trend or are most popular among the users and directly recommend those. Collaborative based filtering is a method to analyze the user’s behavior by predicting the user's taste to that of similar to other users. Associate mining checks for the dependency of one data item on another data item and maps them accordingly. It tries to find some interesting relations or associations among the variables of the dataset. An association rule is a condition of the form A -> B where A and B are two data sets of a particular item. Associate mining finds correlation between A and B i.e. If we purchase A, then it looks for possibilities to purchase B.

Limitations of Popularity-based filtering:

a. Not personalized: The system would recommend the same sort of products which are solely based upon popularity to every other user.

Because of this limitation, and for making a better personalized user experience, most of the online platforms are shifting towards collaborative filtering based recommendation systems. Although collaborative based systems also have several drawbacks such as sparsity and cold start

a. Cold-start problem: It is also known as a new user problem as it is difficult to suggest any item to a new user, because no item is used by this user.

b. Sparsity: The new user and product do not have enough historical data to make it work (data sparsity)

This paper is organized as follows: Literature review will be presented in section 2. Collaborative filtering is described in detail in section 3. Section 4 represents the experimental results of the comparison. Section 6 summarizes the paper.

II. LITERATURE REVIEW

The main objective of this research paper is to summarize the recent research with comparative results between popularity based and collaborative filtering, associate mining based recommender system. Graphs have been plotted on basic datasets

to understand the concept of similarity in collaborative filtering better.

III. POPULARITY BASED FILTERING

Popularity based recommender system works on the principle of trend. It recommends the items which are popular at the moment.

For example, if there is a product which is liked by a large number of users, there is a possibility that it will be liked by a new user and therefore the system will recommend the trending item to the new user.

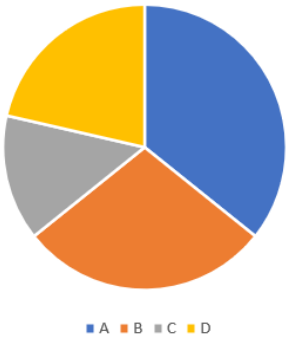

Fig 1.

In the above chart, since A and B are the most liked and trending items, the recommender system will recommend them to the new users.

Since, this system doesn't need a user's historical data it does not face the cold start problem like collaborative filtering based system.

A. Calculating the trending items

The trending items are usually calculated with the help of view count , number of likes , number of votes and average rating based on, on what platform we are.

For example, IMDB movie recommendation works on weighted average formula:

WR = [vR/(v + m)]+[mC/(v+m)]

Where,
v is the number of votes for the movie;
m is the minimum votes required to be listed in the chart;
R is the average rating of the movie; and
C is the mean vote across the whole report.

IV. COLLABORATIVE FILTERING

Collaborative filtering works on the similarity between different users and the items.Similarity is not limited to the user's preferences; similarity between different items can also be considered. This system gives better results if we have a large volume of information about the items and the users.

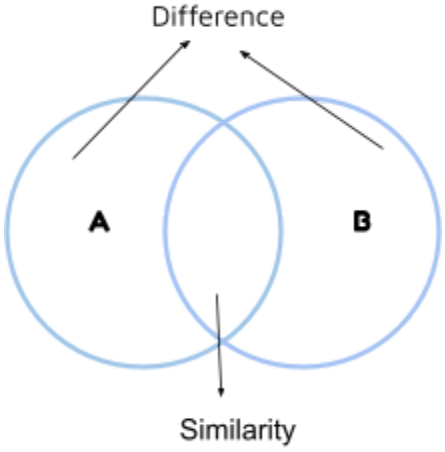

Fig 2.

This figure shows the similarity and differences between the interests of two users A and B. Since both A and B have similar taste it is observed that, a recommender system based on collaborative filtering, suggests A's interest in B and vice-versa.

It have two approaches:

a. User-based nearest neighbor
b. Item-based nearest neighbor

1) User-based nearest neighbor:

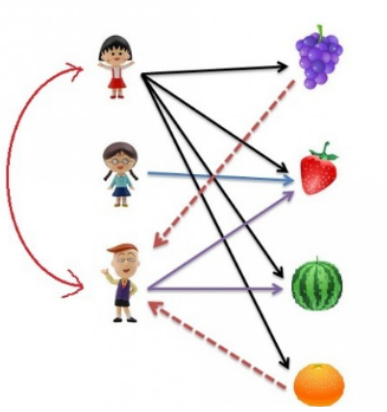

Fig 3.

This figure shows 3 users and the fruits, which they like. User based recommender system finds out the user who has similar taste and can buy the same kind of fruits. The similarity between the users is computed by using similarity measures that are then used to predict ratings and recommendations. This can be done through various algorithms which are discussed in the later part of this paper.

	Grapes	Strawberry	Watermelon	Oranges
USER 1	☑	☑	☑	☑
USER 2	☐	☑	☐	☐
USER 3	☐	☑	☑	☐

Table 1.

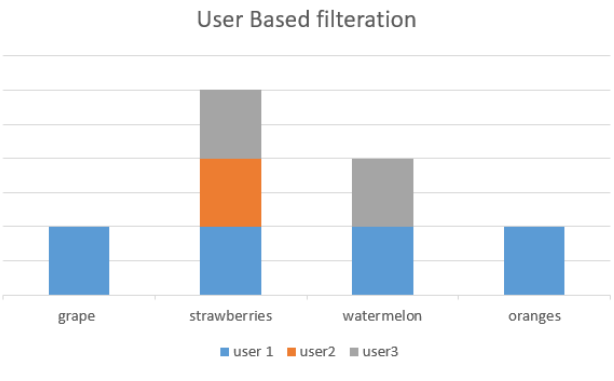

Fig 4.

Since User 3 has more similarities to that of User 1. System will recommend User 3, other likings of user 1 i.e grapes and oranges.

2) Item-based nearest neighbor

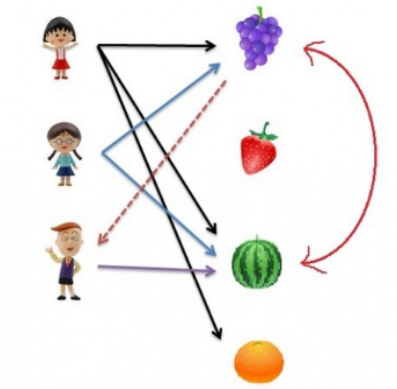

Fig 5.

In the above figure the rating of an item is predicted by the rating of user's on the neighboring items. This type of recommender system is based on the idea of item similarity.

	User 1	User 2	User 3
Grapes	☑	☑	☐
Strawberry	☐	☐	☐
Watermelon	☑	☑	☑
Oranges	☑	☐	☐

Table 2.

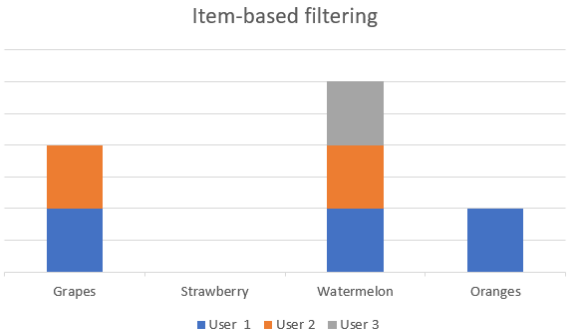

Fig 6.

Here, Watermelon is liked by each user whereas Grapes is liked by only user 1 and user 2. Therefore, the item-based recommender system will recommend Grapes to user 3.

A. Calculating similarity between the items

The similarity between items/users can be calculated through various algorithms like Euclidean algorithm , cosine similarity, Manhattan distance etc.

For calculating the similarity between the items, firstly the recommender system plots a coordinate graph on the ratings of the items by the user (based on which system is being used).

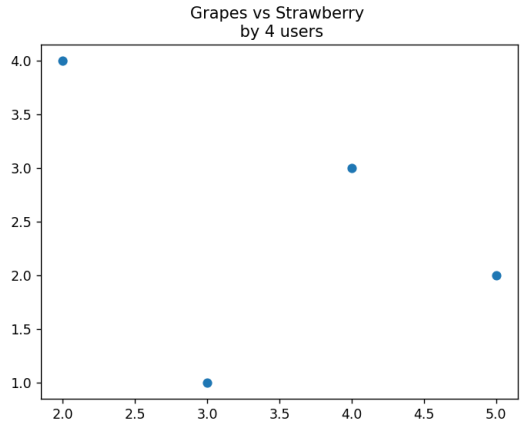

Fig 7.

The above graph plots the ratings given by 4 users on two fruits i.e grapes and strawberries on x and y axis respectively. (item-based system)

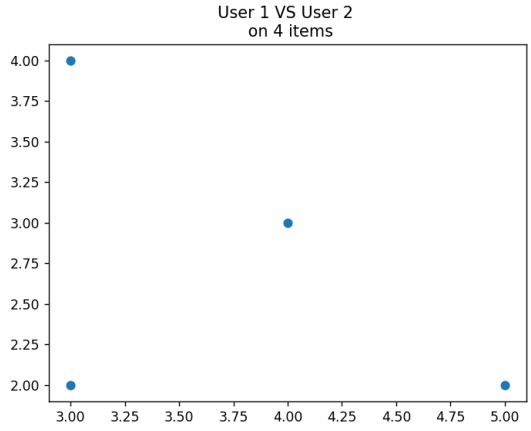

Fig 8.

The above graph plots the ratings of 4 fruits given by 2 users on the x and y axis.
Similarly, the system plots the rating coordinates, multi-dimensional graphs between various items.

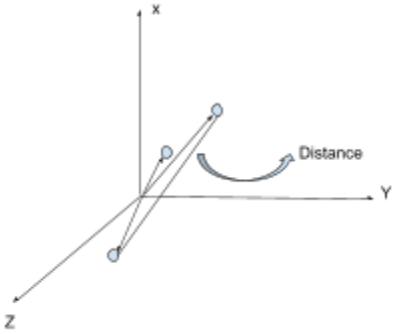

Fig 9.

The system then counts the distance between two points using the algorithms mentioned earlier and calculates the ratings that are closest to a particular point and give it as a recommendation.

In this project we have used the concept of cosine similarity which works on the angles and distances between two points, about which we will learn in the following sections.

V. METHODOLOGY

A. Data collection and dataset

The dataset being used is Book Recommendation Dataset from Kaggle. It contains the following datasets : books, users and ratings.

1) Books Dataset: The shape of this data is (271360, 8). The figure below shows the information about this dataset.

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher
0	0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press
1	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada
2	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial
3	0374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux
4	0393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton & Company

Table 3.

2) Users Dataset: Shape of this dataset is (278858, 3). The figure below shows the information about this dataset.

	User-ID	Location	Age
0	1	nyc, new york, usa	NaN
1	2	stockton, california, usa	18.0
2	3	moscow, yukon territory, russia	NaN
3	4	porto, v.n.gaia, portugal	17.0
4	5	farnborough, hants, united kingdom	NaN

Table 4.

3) Ratings dataset: Shape of this dataset is (1149780, 3). The figure below shows the information about this dataset.

	User-ID	ISBN	Book-Rating
0	276725	034545104X	0
1	276726	0155061224	5
2	276727	0446520802	0
3	276729	052165615X	3
4	276729	0521795028	6

Table 5.

B. Data preparation and preprocessing

To work with these three datasets, each dataset is further cleaned to remove the unwanted and noisy data. Data is reduced to make the analysis and understanding easier.

C. Algorithms

- For popularity based recommender system we have sorted the data on the basis of number of ratings and average ratings. To make the system more precise, we are considering books which have more than 250 ratings.
- For calculating the similarity in collaborative filtering system we are using the cosine similarity algorithm, among the various available options.
  In figure 9, if you notice if the angle between the lines is increased, then the similarity decreases. To calculate similarity, we need a function that returns a higher similarity or smaller distance for a lower angle and a lower similarity or larger distance for a higher angle.
  The cosine of an angle is a function that decreases from 1 to -1 as the angle increases from 0 to 180. The higher the angle, the lower will be the cosine and thus, the lower will be the similarity of the users.

Cosine similarity means the similarity between two vectors of inner product space, It is measured by the cosine of the angle between two vectors.

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

For finding similarity among the books wherein the difference in each users rating is taken into consideration is:

$$sim_{(t,r)} = \frac{\sum_{i=1}^{m}(R_{it}-A_i)(R_{ir}-A_i)}{\sqrt{\sum_{i=1}^{m}(R_{it}-A_i)^2(R_{ir}-A_i)^2}}$$

where Rit is the rating of target item t by user i,Rir is the rating of remaining item r by user I,Ai is the average rating of user i for all co-rated items ,and m is the number of all rating users to item t and item r .

```
similarity_scores[0]
```

```
array([1.        , 0.10255025, 0.01220856, 0.        , 0.05367224,
       0.02774901, 0.08216491, 0.13732869, 0.03261686, 0.03667591,
       0.02322418, 0.06766487, 0.02083978, 0.09673735, 0.13388865,
       0.08303112, 0.11153543, 0.05100411, 0.02517784, 0.11706383,
       0.        , 0.14333793, 0.07847534, 0.06150451, 0.08723968,
       0.        , 0.07009814, 0.13658681, 0.07600328, 0.12167134,
       0.00768046, 0.01473221, 0.        , 0.07965814, 0.04522617,
       0.01556271, 0.09495938, 0.0182307 , 0.02610465, 0.07984012,
       0.11679969, 0.0569124 , 0.08354155, 0.08471898, 0.08785938,
       0.05491435, 0.0548505 , 0.27026514, 0.09779123, 0.06016046,
       0.08958835, 0.06748675, 0.        , 0.04468098, 0.01920872,
       0.        , 0.05629067, 0.00557964, 0.07877059, 0.05219479,
```

Fig 10.

Here, the similarity score of the book at index number 0 is calculated to that of other books in the dataset. We notice that it's similarity score with itself is 1 as the angle will be 0.

The rating of target user u to target item t can be predicted as:

$$P_{ut} = \frac{\sum_{i=1}^{c} R_{ui} \times sim(t,i)}{\sum_{i=1}^{c} sim(t,i)}$$

Where Rui is the rating of target user u to the neighbor item i, sim (t, i) is the similarity of target item t and neighbor item i and c is no. of neighbors.

For precise results we have considered books which have more than 50 ratings and users who have voted on more than 200 books.

VI. EXPERIMENTAL RESULTS

- The popularity based recommender system in our project gives the top 50 most voted books.

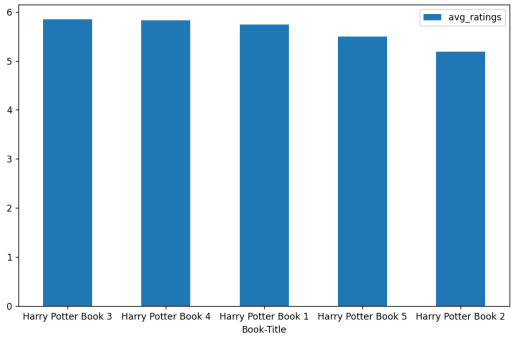

Fig 11.

The above figures give the top 5 books on the basis of popularity.

- Collaborative filtering based system gives the recommendation to the user on the basis of the book he/she have liked.

```
recommend('1984')
```

```
Animal Farm
The Handmaid's Tale
Brave New World
The Vampire Lestat (Vampire Chronicles, Book I
The Hours : A Novel
```

Fig 12.

The above figure gives the top 5 books to the user who has liked the book ‘1984’.

VII. CONCLUSION

The source of information transforming to online mode has led to invention of new technologies based on user interest for the upcoming generation. This paper gives a glimpse at both popularity as well as collaborative filtering based recommender systems. We recommend using both the systems together to overcome the problem of data sparsity, cold start and personalisation. The results give good performance at accuracy.

REFERENCES

1. https://www.analyticsvidhya.com/blog/2022/02/introduction-to-collaborative-filtering/
2. http://www.salemmarafi.com/code/collaborative-filtering-with-python/
3. https://www.analyticsvidhya.com/blog/2021/07/recommendation-system-understanding-the-basic-concepts/
4. https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa#:~:text=Many%20algorithms%2C%20whether%20supervised%20or,%2C%20UMAP%2C%20HDBSCAN%2C%20etc.
5. https://www.machinelearningplus.com/nlp/cosine-similarity/
6. https://medium.com/the-owl/recommender-systems-f62ad843f70c
7. https://www.analyticssteps.com/blogs/what-are-recommendation-systems-machine-learning
8. https://realpython.com/build-recommendation-engine-collaborative-filtering/
9. https://ieeexplore.ieee.org/document/7435717