PROTOTYPE
ONLINE BOOK RECOMMENDATION SYSTEM

INTRODUCTION

Recommendation systems are software programs that help a user to find products according to their needs and interests by using the user's rating of each item and the user's preferences. Recommendation systems are ubiquitous. If you've ever looked for movies on Netflix or glanced through posts on Instagram, you've used a recommendation system without even knowing it. With online shopping, customers have nearly countless options. Nobody has the time to browse through the infinite number of pages in order to search for a product or purchase a product. These types of algorithms lead to service enhancement and customer satisfaction which in turn, brings in more traffic onto the website. Recommendation systems play an important role in helping users find products and content they care about. The book recommendation system would help the user purchase a book, by recommending books based on collaborative filtering and association rule mining.

Collaborative filtering imitates user-to-user and item-to-item type of recommendations. It forecasts users' preferences as a linear, weighted permutation of other user's preferences. Association rule mining is a technique which intends to detect recurrently occurring patterns, correlations, or associations from datasets found in different kinds of databases such as relational databases, transactional databases, and other forms of data warehouses.

LITERATURE SURVEY

Recommender systems have become a vital research field since the emergence of the first paper on collaborative filtering in the mid- the 1990s. In general, these systems are stated as the support systems which help users to find content, products, or services (such as books, movies, music, TV programs, and websites) by gathering and examining suggestions from other users, which means reviews from various establishments, and users. These systems are broadly classified into collaborative filtering (CF) and content-based filtering (CB). CF is an information filtering practice that is based on the user's evaluation of items or previous purchases records. However, this method has been known to expose two major issues that are sparsity problem and scalability problem. CB examines a set of items rated by an individual user and then uses the content of these items, as well as the provided ratings, to deduce a profile that can be used to recommend additional items of interest. However, the syntactic nature of Content-based filtering to detect the similarity between items that share the same attributes or

features causes overspecialized recommendations that only comprise very similar items to those the user already knows.

PROBLEM STATEMENT

A recommendation system will help users who do not have enough individual knowledge to peruse through the different types of options offered by a website. It will provide the users with information to assist them to make a decision as to which items to purchase. The proposed work alters from the existing recommendation systems since the existing ones only consider one technique to recommend items to the users. They do not recommend items using two or more techniques and are not a Hybrid Recommendation System. The proposed system uses a combination of Collaborative Filtering and Association Rule Mining. Collaborative Filtering is used to predict the ratings of a particular item by calculating ratings given to similar items or ratings given by similar users. Association Rule Mining is used to extract different patterns and correlations between different items. Thus, the usage of both these techniques can help to control the data sparsity problem and cold start problem in recommendation systems.

PROBLEMS

Problems faced by recommender systems:-

A. Data Sparsity: Data sparsity arises from the phenomenon that users in general rate only a limited number of items.

B. Cold Start: Personalized recommender systems take advantage of users' past history to make predictions. The cold start problem apprehends personalized recommendations for customers with zero or negligible past history.

C. Gray Sheep: The collaborative filtering (CF) approach in recommender systems assumes that users' preferences are consistent among users. It is presumed that few of these consumers belong to a minor community of users who have uncommon likings, such users are incompliant with the CF fundamental hypothesis. They are grey sheep users.

D. Shilling Attacks: Recommender systems which are based on collaborative filtering are vulnerable to "shilling attacks" due to their open nature. Shiller's inject a few unscrupulous "shilling profiles" into the database of ratings for altering the system's recommendation, due to which some inappropriate items are recommended by the system. E. Scalability Recommender

systems play an important role in online activities by making personalized recommendations to users, as finding what users are looking for among an enormous number of items in huge databases is a tedious job. These methodologies need humongous volumes of training data, which cause scalability problems.

COLLABORATIVE FILTERING

Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person.

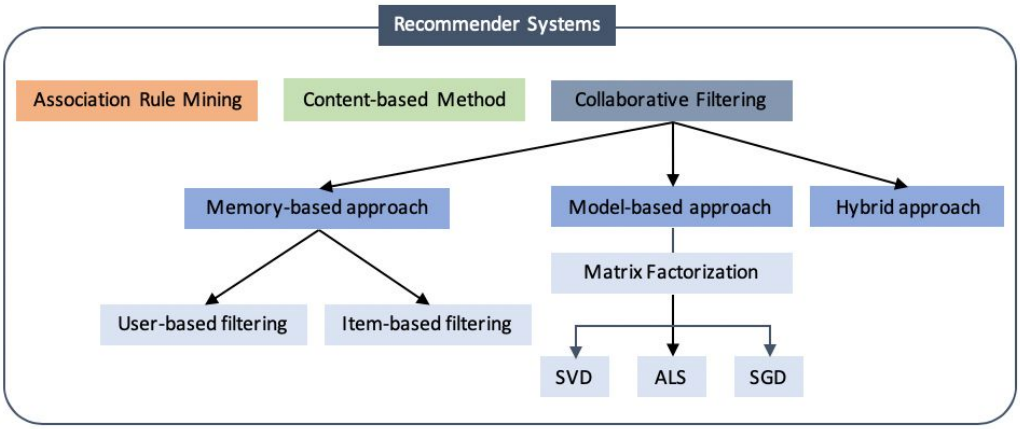


Fig. Types of Collaborative Filtering

Collaborative Filtering uses two types of approaches:

A. User to user Collaborative Filtering: User to user or user-based collaborative filtering finds a neighbourhood of users which are similar or have similar interests as the active user and predicts the rating the active user might give to the product. Based on these ratings the system recommends the active user, the items reviewed or rated by the users in that neighbourhood.

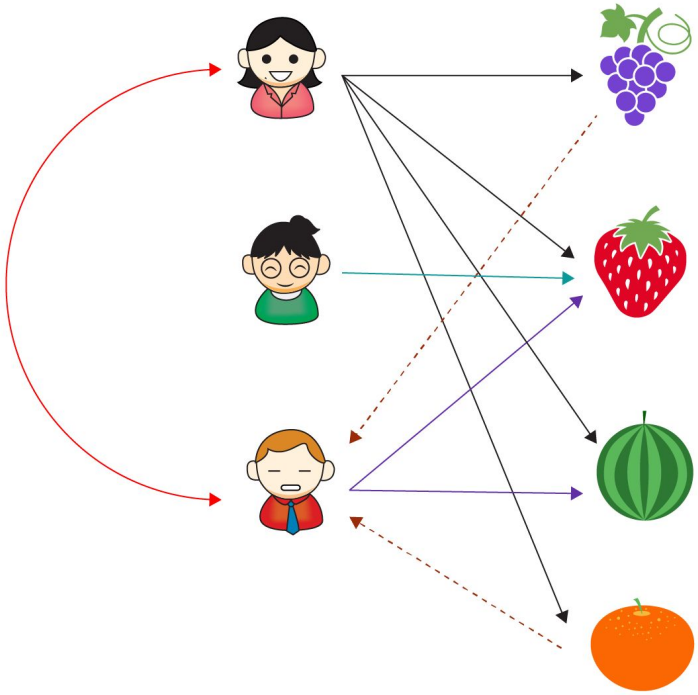
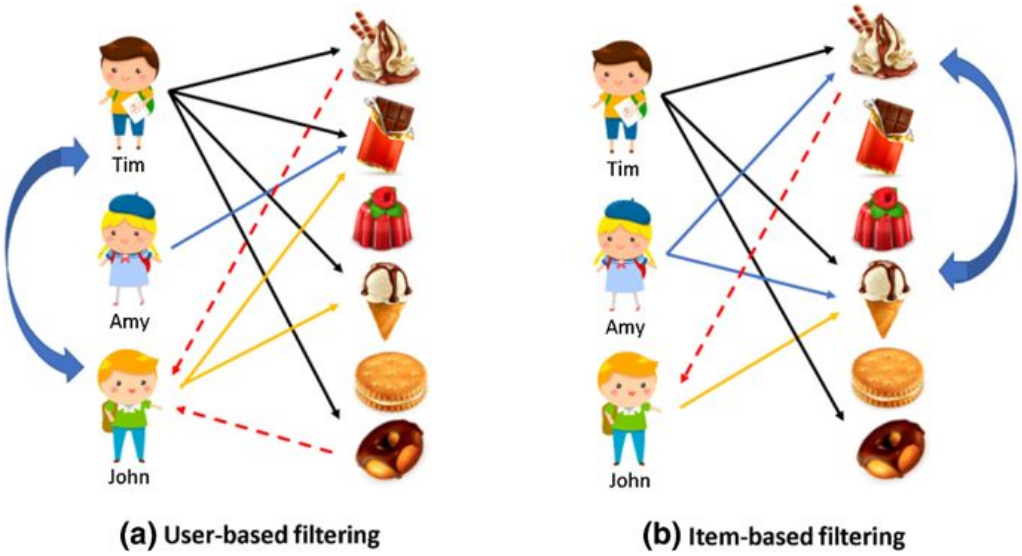


Fig User-based Collaborative Filtering

B. Item to item Collaborative Filtering: This approach is similar to the user based collaborative filtering but instead of finding a neighbourhood of users it finds a neighbourhood of items which are correlated that the active user has highly rated. Based on these ratings the system will predict the rating of the item in the neighbourhood not yet rated by the active user. Pearson correlation, adjusted cosine vector similarity, Cosine similarity, Jaccard Distance etc. are different methods to find a similarity between two items. We have used the Pearson correlation in our paper. The Pearson correlation coefficient also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC) or the bivariate correlation, is a measure of the linear correlation between two variables X and Y. It has a value between +1 and −1, where 1 is a total positive linear correlation, 0 is no linear correlation, and −1 is a total negative linear correlation.



ASSOCIATION MINING

Association Rule Mining is one of the techniques used in Data Mining. Association Rule Mining is used to find out different types of patterns, correlations and relationships between different types of data in a large data set. It makes use of different Association rules. These rules are nothing but certain IF/THEN statements. Most machine learning algorithms work with numeric data and hence they are mathematical. However, association rule mining can only be effectively used for non-numeric data and requires much more than simply counting. Association Rule Mining can also be referred to as “Market Basket Analysis”, where the study is done based on various types of transactions. Association Rule extracts the patterns from the database based on two measures that are Minimum Support and Minimum Confidence. Support indicates the number of times an item appears in a transaction. Confidence: Confidence indicates the probability of a transaction containing item A along with item B.

Market-Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

TOOLS/LIBRARIES NEEDED

- PANDAS
- MATPLOTLIB
  - PYPLOT
- SKLEARN
  - METRICS
  - PAIRWISE
- NUMPY
- SCIPY
  - CORRELATION
- IPYTHON
  - DISPLAY
- CONTEXTLIB
- WARNINGS
- OS
- SYS
- RE
- SEABORN

DATASET [Link]

The Book-Crossing dataset comprises 3 tables.

BX-Users: It contains the users. Note that user IDs (`*User-ID*`) have been anonymized and map to integers. Demographic data is provided (`*Location*`, `*Age*`) if available. Otherwise, these fields contain *NULL*-values.

BX-Books: The books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (`Book-Title`, `Book-Author`, `Year-Of-Publication`, `Publisher`), obtained from Amazon Web Services. Note that incase of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavours (`Image-URL-S`, `Image-URL-M`, `Image-URL-L`), i.e., small, medium, large. These URLs point to the Amazon web site.

BX-Book-Ratings: It contains the book rating information. Ratings (`*Book-Rating*`) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

CONCLUSION

The escalating demands of Online Data have led to the invention of new techniques for presenting or rather recommending different products to users. This project will use both Collaborative filtering as well as Association Rule Mining in order to recommend different types of books to the users. Both these techniques will make a good Hybrid Recommendation System and also get rid of the Data Sparsity and Cold Start problem. Finally, the calculations of both algorithms will be checked corresponding to the individual results given by each collaborative method in order to validate this Hybrid-method.