DS JULY 2022 Batch
Module 18 – Recommender System

Topics

- Recommender System
- Market Basket Analysis
- Apriori

Recommendation System

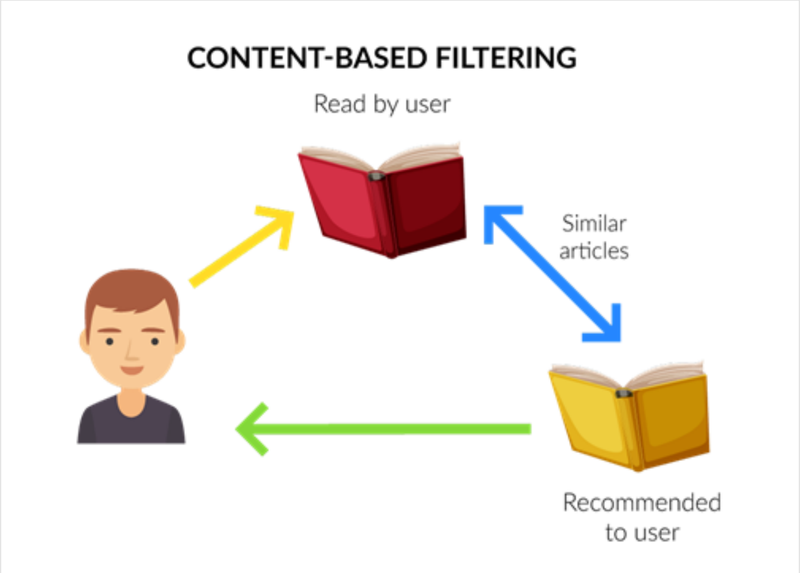
Recommendation System

- A recommendation system is a subclass of Information filtering Systems that seeks to predict the rating or the preference a user might give to an item. In simple words, it is an algorithm that suggests relevant items to users. Eg: In the case of Netflix which movie to watch, In the case of e-commerce which product to buy, or In the case of kindle which book to read, etc.

- There are many use-cases of it. Some are
  - A. Personalized Content: Helps to Improve the on-site experience by creating dynamic recommendations for different kinds of audiences like Netflix does.
  - B. Better Product search experience: Helps to categories the product based on their features. Eg: Material, Season, etc.

Types of Recommendation System

- **Content-Based Filtering**



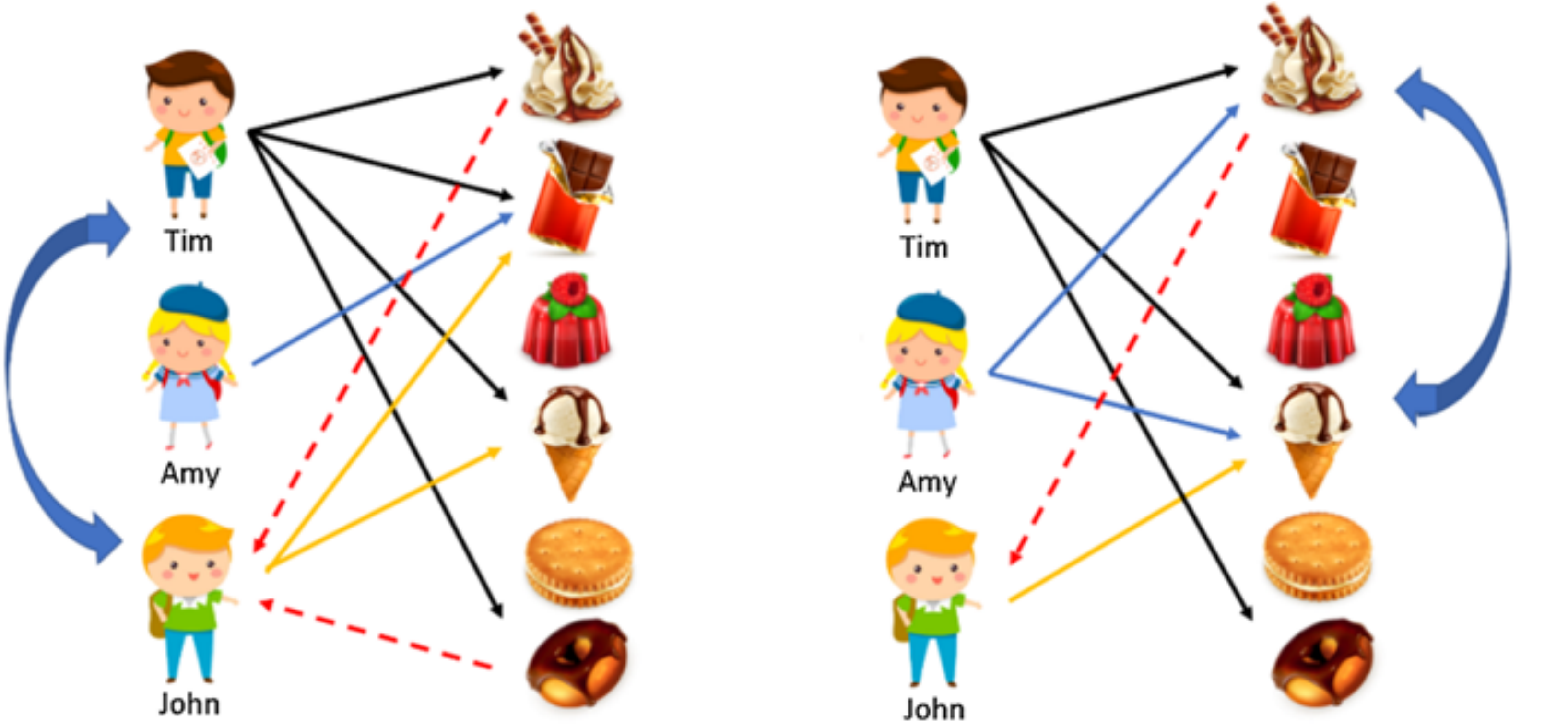
Source: *https://recosenselabs.com/blog/content-recommendations-media-websites*

- In this type of recommendation system, relevant items are shown using the content of the previously searched items by the users. Here content refers to the attribute/tag of the product that the user like. In this type of system, products are tagged using certain keywords, then the system tries to understand what the user wants and it looks in its database and finally tries to recommend different products that the user wants.

Types of Recommendation System

- **Collaborative Based Filtering**
- Recommending the new items to users based on the interest and preference of other similar users is basically collaborative-based filtering. For eg:- When we shop on Amazon it recommends new products saying *"Customer who brought this also brought"* .
- This overcomes the disadvantage of content-based filtering as it will use the user Interaction instead of content from the items used by the users. For this, it only needs the historical performance of the users. Based on the historical data, with the assumption that user who has agreed in past tends to also agree in future.
- There are 2 types of collaborative filtering:-
  - **a) User-Based Collaborative Filtering :** Rating of the item is done using the rating of neighbouring users. In simple words, It is based on the notion of users' similarity.
  - **b) Item-Based Collaborative Filtering :** The rating of the item is predicted using the user's own rating on neighbouring items. In simple words, it is based on the notion of item similarity.

Types of Recommendation System

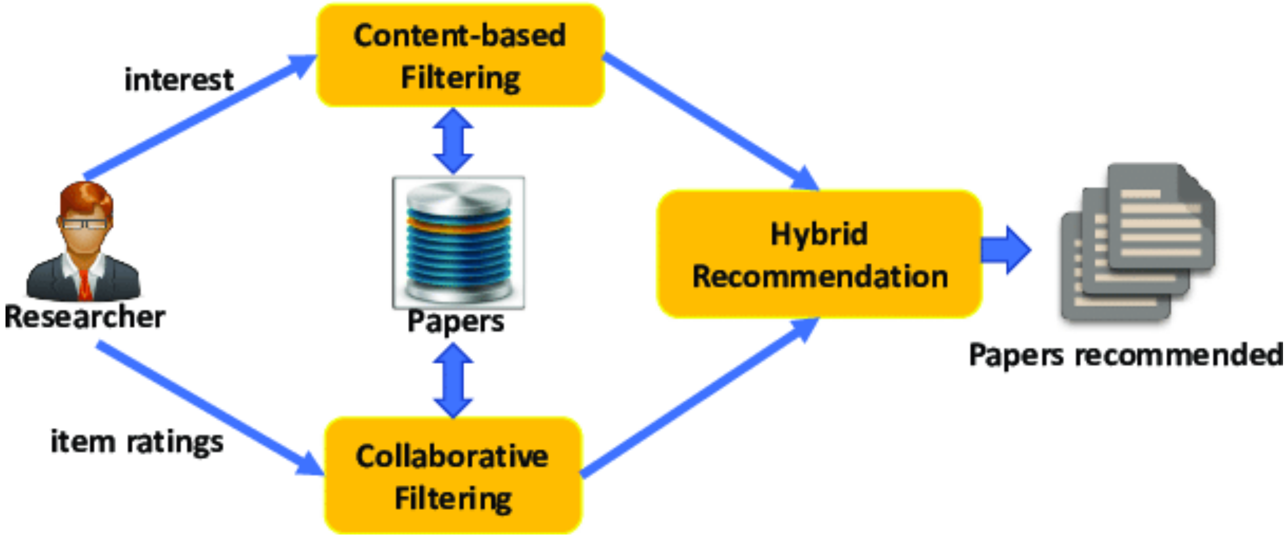


(a) User-based filtering

(b) Item-based filtering

Types of Recommendation System

- **Hybrid**
- A hybrid recommendation system is a special type of recommendation system which can be considered as the combination of the content and collaborative filtering method. Combining collaborative and content-based filtering together may help in overcoming the shortcoming we are facing at using them separately and also can be more effective in some cases.

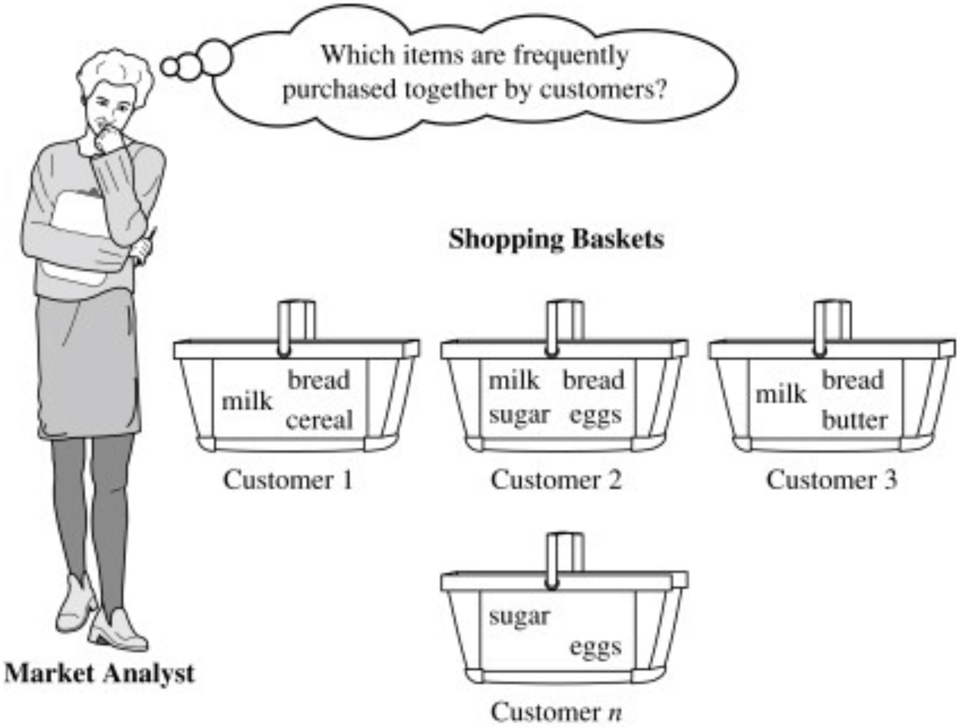


Market Basket Analysis

Market Basket Analysis

- Frequent itemset mining leads to the discovery of associations and correlations between items in huge transactional or relational datasets.
- With vast amounts of data continuously being collected and stored, many industries are becoming interested in data mining to find patterns in their databases.
- The disclosure of “Correlation Relationships” among huge amounts of transaction records can help in many decision-making processes, such as the design of catalogues , cross-marketing, and customer shopping Analysis. These mining algorithms are used for recommender systems.
- **A popular example of frequent itemset mining is Market Basket Analysis.** This process identifies customer buying habits by finding associations between the different items that customers place in their “shopping baskets,” as you can see in the following figure.

Market Basket Analysis



- The discovery of this kind of association will be helpful for retailers or marketers to develop marketing strategies by gaining insight into which items are frequently bought together by customers, whether from a grocery store or online retail.

Apriori Algorithm

What is Apriori

- Apriori algorithm is one of the data mining techniques used for mining different types of patterns in data.
- But before directly jumping to the algorithms we must have a clear understanding of ‘Association Rule Mining’.

Association rule

- Association rule mining is a way to find the pattern in data. The main aim of association rule mining is you have to come out with a rule through which we can predict the occurrence of an item based on the occurrence of some other items in a transaction. Defining Association rule to the point: Given a set of transactions, find rules to predict the occurrence of an item based on the occurrence of other items in the 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

Steps to Find Association Rules

Step 1: The first step towards discovering the association rule is to detect ‘frequent itemset’.

- **Itemset:** Itemset is nothing but a collection of items and the itemset which contain k items is called k – itemset like {Milk, Bread} is 2 – itemset.
- **Support Count (σ):** This is the frequency of occurrence of an itemset. For example, σ({Milk, Bread, Diaper}) = 2 as it is occurring in the transaction number (TID) 4 and 5 in the table above.
- **Support:** Fractions of transactions that contain an itemset. For example, s({Milk, Bread, Diaper}) = 2/5, 5 is a total number of transaction and 2 is the support count. We can also express it in percentage and s({Milk, Bread, Diaper}) = 40%.
- **Lift:** Lift is nothing but the likelihood of the item on right-hand side being purchased when the item on the left-hand side is sold.
- **Frequent itemset:** An itemset whose support is greater than or equal to a ‘minsup’ threshold. For the given example if the minsup threshold value is 30% then {Milk, Bread, Diaper} would be considered as the frequent itemsets.

Steps to Find Association Rules

Step 2: Defining association rule

- An implication expression of the form X → Y, where X and Y are the item sets. Example, {Milk, Diaper} → {Beer}
- For the above association rule (X⊆Y) to be valid, two conditions need to be satisfied - first is that both the item (X & Y) has to be frequent itemset, this condition is called the support (s) condition, the second, the confidence (c) condition which says that in most of the transaction where X appears Y should also appear.
- Suppose if the confidence threshold is 90%, then in 90% of the transactions where X appears Y should also appear.
- Example: {Milk, Diaper} → {Beer}

  s = $(\sigma\ (Milk,\ \ Diaper, Beer))/(|T|)$ = 2/5=0.4
- There is a total of 5 transactions and in two out of 5 {Milk, Diaper, Beer} appear together two times i.e. in transaction 3 and 4.
- The support of X and Y both taken together should be greater than the threshold value. In this case, the support of {Milk, Diaper, Beer} should be greater than the threshold value.

Steps to Find Association Rules

- **c = $(\sigma\ (Milk, Diaper, Beer))/(\sigma\ (Milk, Diaper))$= 2/3=0.67**

- The confidence factor is that out of how many times the left-hand side appear what fraction of them the right-hand side also appear.

- {Milk, Diaper} appears 3 times in the above transactions (i.e. TID 3,4 and 5), and, {Milk, Diaper} and {Beer} together appears 2 times in the transactions (i.e. TID 3 and 4).

- So, it is the ratio of the support count of all three together {Milk, Diaper, Beer} (i.e. left-hand side + right – hand side) and the support count of {Milk, Diaper} (i.e. left – hand side). The confidence factor here is 0.67.

- Lift = $(c\ (\{Milk, diaper\}{=}{=}{>}\{Beer\}))/(s(\{Beer\}))$=(2/3)/(3/5)= 10/9=1.11

Steps to Find Association Rules

- **Step 3:** Now we will call a rule a valid association rule if it satisfies that both support and confidence values are greater than some user-defined threshold value, i.e.,

  Support >= minsup threshold

  Confidence >= minconf threshold

- So suppose our minsup threshold is 30% and the minconf threshold is 90% then the rule above where s = 0.4 and c = 0.67 would qualify as an association rule.

How do we discover the rules ?

- List all possible association rules, i.e. take all possible combination of items.

- Find how many times the left – hand side and right – hand side appears

- After finding the counts compute the support and confidence for each rule.

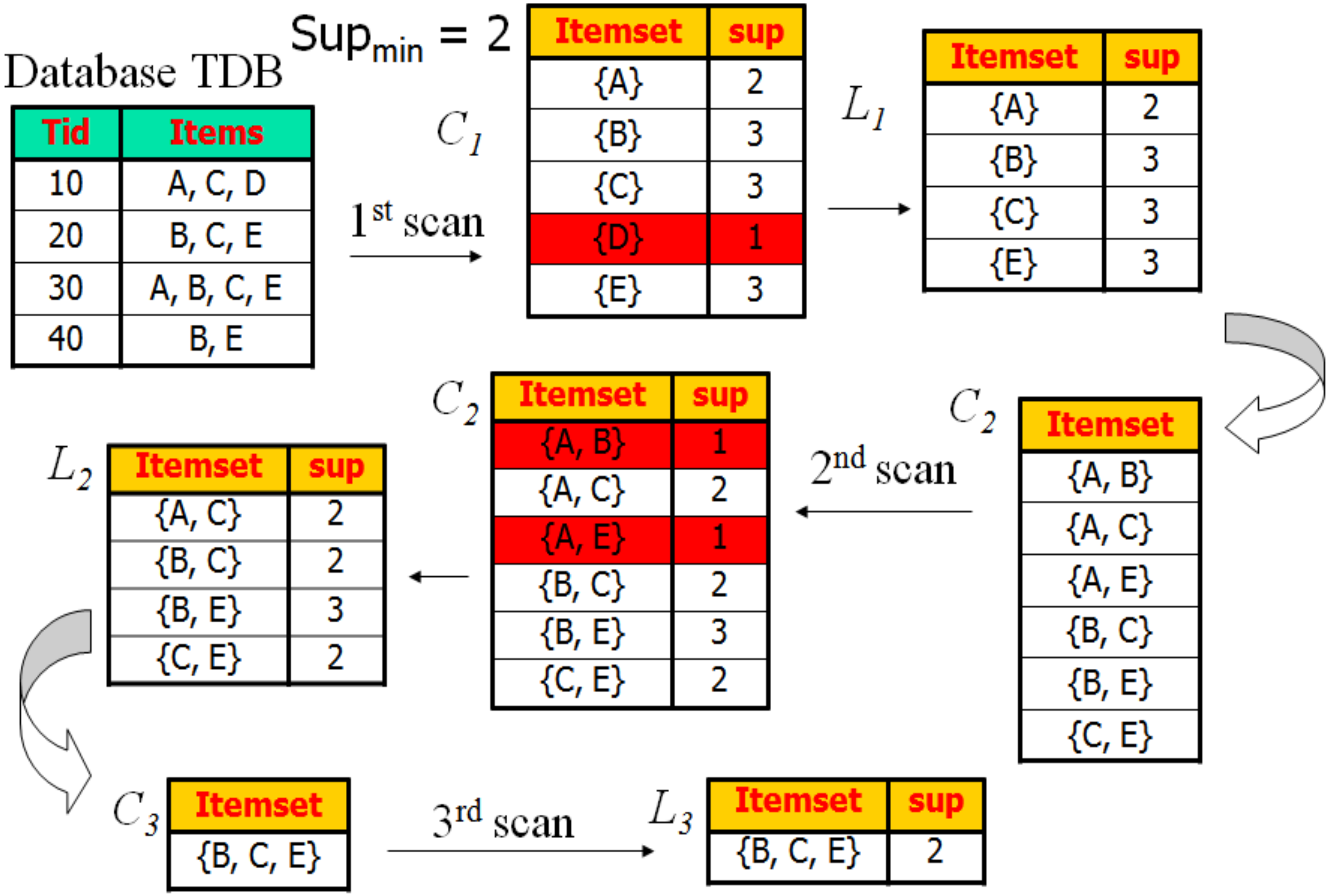
- Prune rules that fail the minsup and minconf threshold

- If the rule satisfies the two conditions of minsup and minconf, i.e. their support and confidence values are greater than minsup and minconf threshold respectively then they are the association rule.

Apriori Algorithm

- Apriori algorithm is a data mining algorithm for learning association rule over the transactional databases.
- This algorithm uses the prior knowledge of frequent itemset properties and hence is the name of this algorithm.
- The apriori property is that all the subsets of a frequent itemset must be frequent.
- Assume K = 1
    - Start generating the frequent itemset of unit length
    - Repeat until and unless no new frequent itemset are identified
    - From length K frequent itemset generate length (K+1) candidate itemset
    - Remove candidate itemset that contain subsets of length K which are not frequent
    - Count the support of each candidate by scanning the Database
    - Eliminate candidates that are not frequent, leaving only the frequent one.

Apriori Algorithm



Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

$Sup_{min} = 2$

1st scan

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

Application

- From market basket analysis we can come to know that a customer often purchases milk and bread, butter and bread together, shampoo and conditioner together and so on.
- So through the market basket analysis, a retailer can understand the purchase behaviour of the customer.
- This algorithm is also used for stock analysis like when the price of one stock increases the other also increases or decreases.

Thank you!