**Assignment-2**

**Q.1) What are Association Rules, Brief various parameters of Association Rules?**

* **Association Rules:** Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.
* Association Rule Mining was defined in the 1990s by computer scientists Rakesh Agrawal, Tomasz Imieliński and Arun Swami.
* They developed an algorithm-based way to find relationships between items using point-of-sale (POS) systems.
* Applying the algorithms to supermarkets, the scientists were able to discover links between different items purchased, called Association Rules.
* Association Rules for this information to predict the likelihood of different products being purchased together.
* For retailers, association rule mining offered a way to better understand customer purchase behaviors. Because of its retail origins, Association Rule Mining is often referred to as Market Basket Analysis.
* Association rules are "if-then" statements, that help to show the probability of relationships between data items, within large data sets in various types of databases.
* Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets.
* **Various parameters of Association Rules:** The importance of an Association Rules can be determined by 3 parameters that are used to identify the strength of the algorithm. Namely:
* **Support :** This measures the frequency of an itemset in the dataset. It is calculated as the proportion of transactions containing the itemset.
* **Confidence :** This measures the reliability of the association between two items. It is calculated as the proportion of transactions containing the antecedent itemset that also contain the consequent itemset.
* **Lift :** It is the probability of all items occurring together divided by the product of antecedent and consequent occurring as if they are independent of each other.

|  |
| --- |
|  |

**Q.2)With suitable example of a dataset, explain calculation of parameters of Association Rules?**

1. **SUPPORT:**

* Support is an indication of how frequently the itemset appears in the dataset.

|  |
| --- |
| https://miro.medium.com/v2/resize:fit:382/1*ft3eXkuS6I9pX6uebZMHww.png |

* Using the example dataset, the itemset X = {Honey, Diaper} has support of 1/5 = 0.2 since it occurs 20 % of all transactions (1 out of 5 transactions)
* if milk is bought, then bread is bought has a support of 0.4 or 40%

1. **CONFIDENCE:**

|  |
| --- |
| https://miro.medium.com/v2/resize:fit:456/1*FbQIxCG1c8Jw5pgtEjDh6w.png |

* Confidence is the percentage of all 2. CONFIDENCE transactions satisfying X that also satisfy Y.
* With respect to T, the confidence value of an association rule, often denoted as X=>Y.
* Which is the ratio of transactions containing both X and Y to the total amount of X values present, where X is the Antecedent and Y is the Consequent.
* Confidence can be computed by calculating 2. CONFIDENCE the co-occurrence of transactions X and Y within the dataset in ratio to transactions containing only X.
* This means that the number of transactions in both X and Y is divided by those just in X As in example dataset the rule, {butter, bread} => milk, has confidence of (1/5)/(1/5)=0.2/0.2=1.0
* This denotes that every time a customer buys butter and bread, they also buy milk.
* This example demonstrates the rule being correct 100% of the time for transactions containing both butter and Bread.

|  |
| --- |
| https://miro.medium.com/v2/resize:fit:700/1*afhqaai-Z63T2hy6_17ayA.png |

1. **LIFT:**

* The lift of rule is defined as the ratio of the observed 3. LIFT support to that expected if X and Y were independent.
* E.g. the rule {milk, bread} =>{butter} has a lift of 0.2/(0.4 ×0.4)=1.25
* If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other, When two events are independent of each other, no rule can be drawn involving those two events.
* If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets
* If the lift is < 1, that lets us know the items are substitute to each other. This means that presence of one item has negative effect on presence of another item and vice Versa.

**Q.3) Write a Short Note on:**

**a) Content based filtering:**

* Content-based filtering works on the principle of describing a product and a profile of the user’s desired choices. It assumes that you will also like this other item if you like a particular item. Products are defined using keywords (genre, product type, color, word length) to make recommendations. A user profile is created to describe the kind of item this user enjoys. Then the algorithm evaluates the similarity of items using cosine and Euclidean distances.
* **Diagram:**

|  |
| --- |
|  |

* **Advantages:**
* **No cold-start problem:** Content-based filtering can recommend items to new users who have not yet provided any feedback or ratings. This is because it only relies on the attributes of the items, which are available even for new items.
* **Transparency:** Content-based filtering is transparent in the sense that the recommendations are based on the features of the items and the user's preferences. This can help users understand why certain items are recommended to them.
* **User independence:** Content-based filtering does not require information about other users or their preferences, which can be useful in situations where user information is not available or unreliable.
* **Diversity:** Content-based filtering can recommend diverse items based on the user's preferences. This is because it focuses on the attributes of the items, which can capture different aspects of the items.
* **Disadvantages:**
* **Limited content analysis:** Content-based filtering only considers the features or attributes of items and does not take into account other factors that may influence a user's preferences, such as social context or current trends. This can result in recommendations that may not be relevant or interesting to the user.
* **Limited personalization:** Content-based filtering may not capture the user's evolving preferences over time. This can result in recommendations that are not personalized or relevant to the user's current interests.
* **Limited diversity:** Content-based filtering may not recommend items that are outside the user's comfort zone or that the user may not have considered before. This can result in recommendations that are limited in scope and may not expose the user to new experiences or perspectives.
* **Cold-start problem for items:** Content-based filtering may not be effective for new items that do not have any features or attributes yet. This can result in new items not being recommended until they have sufficient user feedback or ratings.
* **Examples/Applications:**
* **Example**: Suppose a user X likes to watch action movies like Spider-man. In that case, this recommender engine technique only recommends movies of the action genre or films describing Tom Holland.

**b) Collaborative Filtering:**

* The collaborative filtering method collects and analyzes data on user behavior, online activities, and preferences to predict what they will like based on the similarity with other users. It uses a matrix-style formula to plot and calculates these similarities
* **Diagram:**

|  |
| --- |
|  |

* **Advantages:**
* **Personalization:** Collaborative filtering can provide personalized recommendations by leveraging the preferences and behaviors of similar users. This can result in recommendations that are more relevant and interesting to the user.
* **Serendipity:** Collaborative filtering can recommend items that the user may not have considered before but that are popular or highly rated among similar users. This can result in recommendations that expose the user to new experiences or perspectives.
* **Scalability:** Collaborative filtering can handle large amounts of data and users, making it scalable for large-scale recommender systems.
* **Cold-start problem for users:** Collaborative filtering can recommend items to new users who have not yet provided any feedback or ratings, by leveraging the preferences of similar users.
* **Disadvantages:**
* **Cold-start problem for items:** Collaborative filtering may not be effective for new items that do not have any user feedback or ratings yet. This can result in new items not being recommended until they have sufficient user feedback or ratings.
* **Data sparsity:** Collaborative filtering can suffer from data sparsity, where some users may have very few ratings or interactions with items. This can result in recommendations that may not be accurate or relevant for these users.
* **Popularity bias:** Collaborative filtering can suffer from popularity bias, where popular items are recommended more often than less popular items, even if they may not be relevant to the user's preferences.
* **Privacy concerns:** Collaborative filtering requires access to user data, which can raise privacy concerns for some users.
* **Examples/Applications:**
* **Example:** If user X likes Book A, Book B, and Book C while user Y likes Book A, Book B, and Book D, they have similar interests. So, it is favorably possible that user X would select Book D and user Y would enjoy reading Bood C.

**c) Hybrid Filtering:**

* In hybrid recommendation systems, both the meta (collaborative) data and the transactional (content-based) data are used simultaneously to suggest a broader range of items to the users. In this technique, natural language processing tags can be allocated for each object (movie, song), and vector equations calculate the similarity. A collaborative filtering matrix can then suggest things to users, depending on their behaviors, actions, and intentions.
* **Diagram:**

|  |
| --- |
|  |

* **Advantages:**
* **Improved accuracy:** Hybrid filtering can provide more accurate recommendations by leveraging the strengths of multiple recommendation approaches and mitigating their weaknesses.
* **Increased coverage:** Hybrid filtering can recommend items to a wider range of users by combining multiple recommendation approaches that address different user needs and preferences.
* **Enhanced diversity:** Hybrid filtering can recommend diverse items by combining different recommendation approaches that focus on different aspects of the items.
* **Robustness:** Hybrid filtering can be more robust to changes in the data or the recommendation environment, as it can adapt to different situations by combining multiple recommendation approaches.
* **Disadvantages:**
* **Complexity:** Hybrid filtering can be more complex to implement and maintain than a single recommendation approach, as it requires integrating and coordinating multiple recommendation approaches.
* **Computational overhead:** Hybrid filtering can require more computational resources than a single recommendation approach, as it involves processing and integrating multiple data sources and algorithms.
* **Lack of transparency:** Hybrid filtering can be less transparent than a single recommendation approach, as it may be difficult to explain how the recommendations are generated or which recommendation approach contributed to each recommendation.
* **Higher data requirements:** Hybrid filtering can require more data than a single recommendation approach, as it involves multiple recommendation approaches that require their own data inputs.
* **Examples/Applications:**
* **Netflix uses a hybrid recommendation engine**: It makes recommendations by analyzing the user’s interests (collaborative) and recommending such shows/movies that share similar attributes with those rated highly by the user(content-based).

**(PPT Give you more understanding than PDF)** The material for the PDF has been compiled from various sources such as books, tutorials (offline and online), lecture notes, several resources available on Internet. The information contained in this PDF is for general information and education purpose only. While we endeavor to keep the information up to date and correct, **we make no representation of any kind about the completeness and accuracy of the material.** The information shared through this PDF material should be used for educational purpose only.

**Thank you…**