Mining Frequent Patterns

Frequent patterns are patterns (e.g., itemsets, subsequences, or substructures) that appear frequently in a data set. For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent itemset.
A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it
occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as subgraphs, subtrees, or sublattices,
which may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data.

Associations and Corelations:-

What is Association?

The statistical term association is defined as a relationship between two random variables which makes them statistically dependent. It refers to rather a general relationship without specifics of the relationship being mentioned, and it is not necessary to be a causal relationship.

Many statistical methods are used to establish the association between two variables. Pearson's correlation coefficient, odds ratio, distance correlation, Goodman's and Kruskal's Lambda and Spearman's rho (ρ) are a few examples.

What is Correlation?

Correlation is a measure of the strength of the relationship between two variables. The correlation coefficient quantifies the degree of change of one variable based on the change of the other variable. In statistics, correlation is connected to the concept of dependence, which is the statistical relationship between two variables

The Pearson's correlation coefficient or just the correlation coefficient r is a value between -1 and 1 ($-1 \leq r \leq +1$). It is the most commonly used correlation coefficient and valid only for a linear relationship between the variables. If $r=0$, no relationship exist, and if $r \geq 0$, the relation is directly proportional; the value of one variable increases with the increase in the other. If $r \leq 0$, the relationship is inversely proportional; one variable decreases as the other increases.

What is the difference between Association and Correlation?

• Association refers to the general relationship between two random variables while the correlation refers to a more or less a linear relationship between the random variables.

• Association is a concept, but correlation is a measure of association and mathematical tools are provided to measure the magnitude of the correlation.

• Pearson’s product moment correlation coefficient establishes the presence of a linear relationship and determines the nature of the relationship (whether they are proportional or inversely proportional).

• Rank correlation coefficients are used to determine the nature of the relationship only, excluding the linearity of the relation (it may or may not be linear, but it will tell whether the variables increase together, decrease together or one increases while the other decreases or vice versa).

 # Basic concepts :-

Scatter plot
A scatter plot shows the association between two variables. A scatter plot matrix shows all pairwise scatter plots for many variables.

Covariance
Covariance is a measure of how much two variables change together. A covariance matrix measures the covariance between many pairs of variables.

Correlation coefficient
A correlation coefficient measures the association between two variables. A correlation matrix measures the correlation between many pairs of variables.

Inferences about association
Inferences about the strength of association between variables are made using a random bivariate sample of data drawn from the population of interest.

 # Efficient and scalable frequent item set mining methods :-

Frequent Item set in Data set (Association Rule Mining):

Frequent item sets, also known as association rules, are a fundamental concept in association rule mining, which is a technique used in data mining to discover relationships between items in a dataset. The goal of association rule mining is to identify relationships between items in a dataset that occur frequently together.
A frequent item set is a set of items that occur together frequently in a dataset. The frequency of an item set is measured by the support count, which is the number of transactions or records in the dataset that contain the item set. For example, if a dataset contains 100 transactions and the item set {milk, bread} appears in 20 of those transactions, the support count for {milk, bread} is 20.

Important Definitions :

- **Support :** It is one of the measures of interestingness. This tells about the usefulness and certainty of rules. **5% Support** means total 5% of transactions in the database follow the rule.

```
Support(A -> B) = Support_count(A ∪ B)
```

- **Confidence:** A confidence of 60% means that 60% of the customers who purchased a milk and bread also bought butter.

```
Confidence(A -> B) = Support_count(A ∪ B) / Supp
```

If a rule satisfies both minimum support and minimum confidence, it is a strong rule.

- **Support_count(X)**: Number of transactions in which X appears. If X is A **union** B then it is the number of transactions in which A and B both are present.

- **Maximal Itemset:** An itemset is maximal frequent if none of its supersets are frequent.
- **Closed Itemset:** An itemset is closed if none of its immediate supersets have same support count same as Itemset.
- **K- Itemset:** Itemset which contains K items is a K-itemset. So it can be said that an itemset is frequent if the corresponding support count is greater than the minimum support count.

Example On finding Frequent Itemsets – Consider the given dataset with given transactions.

TransactionId	Items
1	{A,C,D}
2	{B,C,D}
3	{A,B,C,D}
4	{B,D}
5	{A,B,C,D}

Advantages and Disadvantages

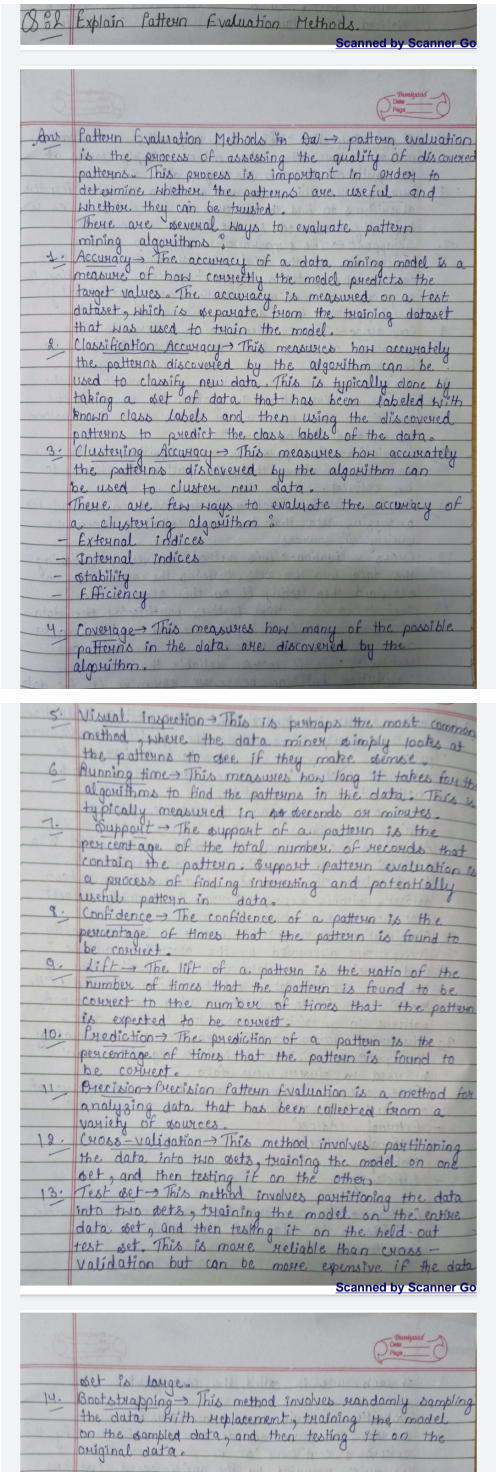
Advantages of using frequent item sets and association rule mining include:

1. Efficient discovery of patterns: Association rule mining algorithms are efficient at discovering patterns in large datasets, making them useful for tasks such as market basket analysis and recommendation systems.
2. Easy to interpret: The results of association rule mining are easy to understand and interpret, making it possible to explain the patterns found in the data.
3. Can be used in a wide range of applications: Association rule mining can be used in a wide range of applications such as retail, finance, and healthcare, which can help to improve decision-making and increase revenue.
4. Handling large datasets: These algorithms can handle large datasets with many items and transactions, which makes them suitable for big-data scenarios.

Disadvantages of using frequent item sets and association rule mining include:

1. Large number of generated rules: Association rule mining can generate a large number of rules, many of which may be irrelevant or uninteresting, which can make it difficult to identify the most important patterns.
2. Limited in detecting complex relationships: Association rule mining is limited in its ability to detect complex relationships between items, and it only considers the co-occurrence of items in the same transaction.
3. Can be computationally expensive: As the number of items and transactions increases, the number of candidate item sets also increases, which can make the algorithm computationally expensive.
4. Need to define the minimum support and confidence threshold: The minimum support and confidence threshold must be set before the association rule mining process, which can be difficult and requires a good understanding of the data.

Pattern Evaluation Methods :-



Q.02 Explain Pattern Evaluation Methods.

Scanned by Scanner Go

Ans Pattern Evaluation Methods in DM → pattern evaluation is the process of assessing the quality of discovered patterns. This process is important in order to determine whether the patterns are useful and whether they can be trusted.

There are several ways to evaluate pattern mining algorithms:

1. Accuracy → The accuracy of a data mining model is a measure of how correctly the model predicts the target values. The accuracy is measured on a test dataset, which is separate from the training dataset that was used to train the model.
2. Classification Accuracy → This measures how accurately the patterns discovered by the algorithm can be used to classify new data. This is typically done by taking a set of data that has been labeled with known class labels and then using the discovered patterns to predict the class labels of the data.
3. Clustering Accuracy → This measures how accurately the patterns discovered by the algorithm can be used to cluster new data.

There are few ways to evaluate the accuracy of a clustering algorithm:

- External Indices
- Internal Indices
- Stability
- Efficiency

4. Coverage → This measures how many of the possible patterns in the data are discovered by the algorithm.
5. Visual Inspection → This is perhaps the most common method, where the data miner simply looks at the patterns to see if they make sense.
6. Running time → This measures how long it takes for the algorithms to find the patterns in the data. This is typically measured in ~~s~~ seconds or minutes.
7. Support → The support of a pattern is the percentage of the total number of records that contain the pattern. Support pattern evaluation is a process of finding interesting and potentially useful pattern in data.
8. Confidence → The confidence of a pattern is the percentage of times that the pattern is found to be correct.
9. Lift → The lift of a pattern is the ratio of the number of times that the pattern is found to be correct to the number of times that the pattern is expected to be correct.
10. Prediction → The prediction of a pattern is the percentage of times that the pattern is found to be correct.
11. Precision → Precision Pattern Evaluation is a method for analyzing data that has been collected from a variety of sources.
12. Cross-validation → This method involves partitioning the data into two sets, training the model on one set, and then testing it on the other.
13. Test set → This method involves partitioning the data into two sets, training the model on the entire data set, and then testing it on the held-out test set. This is more reliable than cross-validation but can be more expensive if the data set is large.
14. Bootstrapping → This method involves randomly sampling the data with replacement, training the model on the sampled data, and then testing it on the original data.

Scanned by Scanner Go

 # Application of frequent pattern and association:-

Frequent pattern mining has several applications in different areas, including:

- **Market Basket Analysis:** This is the process of analyzing customer purchasing patterns in order to identify items that are frequently bought together. This information can be used to optimize product placement, create targeted marketing campaigns, and make other business decisions.

- **Recommender Systems:** Frequent pattern mining can be used to identify patterns in user behavior and preferences in order to make personalized recommendations.
- **Fraud Detection:** Frequent pattern mining can be used to identify abnormal patterns of behavior that may indicate fraudulent activity.
- **Network Intrusion Detection:** Network administrators can use frequent pattern mining to detect patterns of network activity that may indicate a security threat.
- **Medical Analysis:** Frequent pattern mining can be used to identify patterns in medical data that may indicate a particular disease or condition.
- **Text Mining:** Frequent pattern mining can be used to identify patterns in text data, such as keywords or phrases that appear frequently together in a document.
- **Web usage mining:** Frequent pattern mining can be used to analyze patterns of user behavior on a website, such as which pages are visited most frequently or which links are clicked on most often.
- **Gene Expression:** Frequent pattern mining can be used to analyze patterns of gene expression in order to identify potential biomarkers for different diseases.

Issues of frequent pattern mining

- flexibility and reusability for creating frequent patterns
- most of the algorithms used for mining frequent item sets do not offer flexibility for reusing
- much research is needed to reduce the size of the derived patterns.

Frequent Patterns and Association Mining :-
Frequent Pattern Mining in Data Mining

Frequent pattern mining in data mining is the process of identifying patterns or associations within a dataset that occur frequently. This is typically done by analyzing large datasets to find items or sets of items that appear together frequently.

Frequent pattern mining is a major concern it plays a major role in associations and correlations and disclose an intrinsic and important property of dataset.

There are several different algorithms used for frequent pattern mining, including:

Apriori algorithm: This is one of the most commonly used algorithms for frequent pattern mining. It uses a “bottom-up” approach to identify frequent itemsets and then generates association rules from those itemsets.

ECLAT algorithm: This algorithm uses a “depth-first search” approach to identify frequent itemsets. It is particularly efficient for datasets with a large number of items.

FP-growth algorithm: This algorithm uses a “compression” technique to find frequent patterns efficiently. It is particularly efficient for datasets with a large number of transactions.

–Frequent pattern mining has many applications, such as Market Basket Analysis, Recommender Systems, Fraud Detection, and many more.

Advantages:

- It can find useful information which is not visible in simple data browsing
- It can find interesting association and correlation among data items

Disadvantages:

- It can generate a large number of patterns
- With high dimensionality, the number of patterns can be very large, making it difficult to interpret the results.

Frequent data mining can be done by using association rules with particular algorithms eclat and apriori algorithms. Frequent pattern mining searches for recurring relationships in a data set. It also helps to find the inheritance regularities. to make fast processing software with a user interface and used for a long time without any error.

Association Rule Mining:

It is easy to find associations in frequent patterns:

- for each frequent pattern x for each subset y c x.
- calculate the support of y-> x – y.

if it is greater than the threshold, keep the rule. There are two algorithms that support this lattice

1. Apriori algorithm
2. eclat algorithm

Apriori	Eclat
It performs “perfect” pruning of infrequent item sets.	It reduces memory requirements and is faster.
It requires a lot of memory(all frequent item sets are represented) and support counting takes very long for large transactions. But this is not efficient in practice.	Its storage of transaction list.

The words support and confidence support the association rule.

- **Support:** how often a given rule in a database is mined? support the transaction contains x U y
- **Confidence:** the number of times the given rule in a practice is true. The conditional probability is a transaction having x as well as y.

Let's practice it through a sample data set;

Transaction	Item Occurrence
T1	A, B
T2	A, C, D
T3	A, B, C, D
T4	A, D, E
T5	B, C

Image by Author

Support = Frequency (A, B) / N

support calculation — Image by Author

Confidence = Frequency (A, B) / Frequency (A)

confidence calculation — Image by Author

Example: One of possible Association Rule is
A => D

Total no of Transactions(N) = 5

Frequency(A, D) = > Total no of instances together A with D is 3

Frequency(A) => Total no of occurrence in A

Support = 3 / 5

Confidence = 3 / 4

 # Mining various kinds of Association Rules :-

Association rule learning is a machine learning technique used for discovering interesting relationships between variables in large databases. It is designed to detect strong rules in the database based on some interesting metrics. For any given multi-item transaction, association rules aim to obtain rules that determine how or why certain items are linked.

Types of Association Rules:
There are various types of association rules in data mining:-

- Multi-relational association rules
- Generalized association rules
- Quantitative association rules
- Interval information association rules

1. Multi-relational association rules: Multi-Relation Association Rules (MRAR) is a new class of association rules, different from original, simple, and even multi-relational association rules (usually extracted from multi-relational databases), each rule element consists of one entity but many a relationship. These relationships represent indirect relationships between entities.

2. Generalized association rules: Generalized association rule extraction is a powerful tool for getting a rough idea of interesting patterns hidden in data. However, since patterns are extracted at each level of abstraction, the mined rule sets may be too large to be used effectively for decision-making. Therefore, in order to discover valuable and interesting knowledge, post-processing steps are often required. Generalized association rules should have categorical (nominal or discrete) properties on both the left and right sides of the rule.

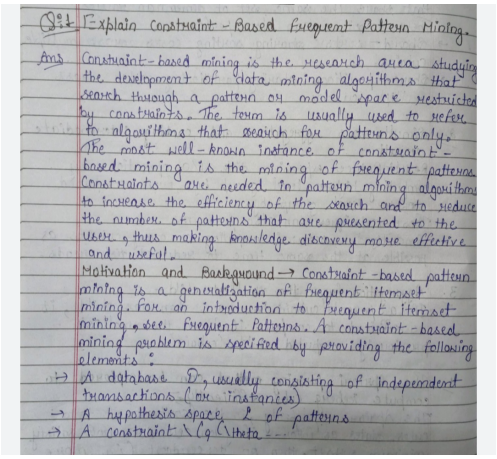
3. Quantitative association rules: Quantitative association rules is a special type of association rule. Unlike general association rules, where both left and right sides of the rule should be categorical (nominal or discrete) attributes, at least one attribute (left or right) of quantitative association rules must contain numeric attributes.

4. Interval information association rules: Next, interval association rules are generated which involved data partitioning via clustering before the rules are generated using an Apriori algorithm. Finally, these rules are used to identify data values that fall outside the expected intervals.

Uses of Association Rules
Some of the uses of association rules in different fields are given below:

- **Medical Diagnosis:** Association rules in medical diagnosis can be used to help doctors cure patients. As all of us know that diagnosis is not an easy thing, and there are many errors that can lead to unreliable end results. Using the multi-relational association rule, we can determine the probability of disease occurrence associated with various factors and symptoms.
- **Market Basket Analysis:** It is one of the most popular examples and uses of association rule mining. Big retailers typically use this technique to determine the association between items.

 # Constraint-Based Frequent Pattern Mining :-


Q.11 Explain Constraint-Based Frequent Pattern Mining.

Ans Constraint-based mining is the research area studying the development of data mining algorithms that search through a pattern or model space restricted by constraints. The term is usually used to refer to algorithms that search for patterns only. The most well-known instance of constraint-based mining is the mining of frequent patterns. Constraints are needed to pattern mining algorithms to increase the efficiency of the search and to reduce the number of patterns that are presented to the user, thus making knowledge discovery more effective and useful.

Motivation and Background → Constraint-based pattern mining is a generalization of frequent itemset mining. For an introduction to frequent itemset mining, see Frequent Patterns. A constraint-based mining problem is specified by providing the following elements:

→ A database D, usually consisting of independent transactions (or instances)
→ A hypothesis space, L of patterns
→ A constraint C(theta ...