

Internal Assessment Test 2 – June 2022

Scheme of Evaluation

| | | | | | | | | | | |
|--|--|---|--|---|-----------|----------|---------------|-------|----|-----|
| Sub: | DATA MINING AND DATA WAREHOUSING | | | | Sub Code: | 18CS641 | Branch: | ISE | | |
| Date: | 09/06/2022 | Duration: | 90 min's | Max Marks: | 50 | Sem/Sec: | VI / A, B & C | OBE | | |
| <u>Answer any FIVE FULL Questions</u> | | | | | | | | MARKS | CO | RBT |
| 1 (a) | Describe the challenges that motivated the development of Data mining. Scheme: List and Explanation of each challenge carries 1 Mark each. Solution: Scalability, High Dimensionality, Heterogeneous and Complex Data, Data Ownership and Distribution, Non-traditional Analysis. | | | | | | [05] | CO2 | L2 | |
| (b) | Differentiate between ROLAP vs MOLAP vs HOLAP. | | | | | | [05] | CO2 | L1 | |
| | Comparison | MOLAP | ROLAP | HOLAP | | | | | | |
| | Meaning | Multi-Dimensional Online Analytical Processing | Relational Online Analytical Processing | Hybrid Online Analytical Processing | | | | | | |
| | Data Storage | It stores data in a multi-dimensional database. | It stores data in a relational database. | It stores data in a relational database | | | | | | |
| | Technique | It utilizes the Sparse Matrix technique. | It employs Structured Query Language (SQL). | It uses a combination of SQL and Sparse Matrix technique. | | | | | | |
| | Volume of data | It can process a limited volume of data. | It processes enormous data. | It can process huge volumes of data. | | | | | | |
| | Designed view | The multi-dimensional view is static. | The multi-dimensional view is dynamic. | The multi-dimensional view is dynamic. | | | | | | |
| | Data arrangement | It arranges data in data cubes. | It arranges data in rows and columns (tables). | There is a multi-dimensional arrangement of data | | | | | | |

| | | | | |
|---|---|------|-----|----|
| 2 | <p>Explain any 5 data pre-processing methods with examples.</p> <p>Scheme: Explanation of any 5 data pre-processing carries 2 Marks each.</p> <p>Solution:</p> <ul style="list-style-type: none"> • Aggregation • Sampling • Dimensionality Reduction • Feature subset selection • Feature creation • Discretization and Binarization • Attribute Transformation <p>Aggregation</p> <p><input type="checkbox"/> Combining two or more attributes (or objects) into a single attribute (or object)</p> <p>Purpose:</p> <ul style="list-style-type: none"> o Data reduction o Reduce the number of attributes or objects o Change of scale o Aggregated data tends to have less variability o More “stable” data <p>Sampling</p> <p><input type="checkbox"/> Sampling is the main technique employed for data selection.</p> <p><input type="checkbox"/> Selecting a subset of the data objects to be analyzed</p> <p><input type="checkbox"/> It is often used for both the preliminary investigation of the data and the final data analysis.</p> <p><input type="checkbox"/> Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.</p> <p>Dimensionality Reduction:</p> <ul style="list-style-type: none"> • A key benefit is that many data mining algorithms work better if the dimensionality the number of attributes in the data-is lower. • This is partly because dimensionality reduction can eliminate irrelevant features and reduce noise <p>Purpose:</p> <ul style="list-style-type: none"> o Avoid curse of dimensionality o Reduce amount of time and memory required by data mining algorithms o Allow data to be more easily visualized o May help to eliminate irrelevant features or reduce noise <p>Feature Subset Selection:</p> <ul style="list-style-type: none"> • Another way to reduce dimensionality of data • Use only a subset of the features • Redundant and irrelevant features can reduce classification accuracy and the quality of the clusters that are found. <p>Redundant features</p> <ul style="list-style-type: none"> – Duplicate much or all of the information contained in one or more other attributes – Example: purchase price of a product and the amount of sales tax paid almost same <p>Irrelevant features</p> <ul style="list-style-type: none"> – Contain no information that is useful for the data mining task at hand – Example: students' ID is often irrelevant to the task of predicting students' GPA <p>Feature Creation</p> <p><input type="checkbox"/> Create new attributes that can capture the important information in a data set much more efficiently than the original attributes</p> <p>Three general methodologies:</p> <p><input type="checkbox"/> Feature Extraction- Example: extracting edges from images domain-specific-</p> <p><input type="checkbox"/> Mapping Data to New Space Example: Fourier and wavelet analysis</p> | [10] | CO2 | L2 |
|---|---|------|-----|----|

| | | | | |
|-------|---|------|-----|----|
| 3 | <p>For the following vectors, x and y, calculate the indicated similarity or distance measures.</p> <p>Scheme: Computation of Cosine, Correlation, Euclidean, Jaccard carries 4,3,4 Marks Each.</p> <p>Solution:</p> <p>a. $x = (0, 1, 0, 1), y = (1, 0, 1, 0)$ cosine, correlation, Euclidean, Jaccard $\cos(x, y) = 0, \text{corr}(x, y) = -1, \text{Euclidean}(x, y) = 2, \text{Jaccard}(x, y) = 0$</p> <p>b. $x = (0, -1, 0, 1), y = (1, 0, -1, 0)$ cosine, correlation, Euclidean $\cos(x, y) = 0, \text{corr}(x, y) = 0, \text{Euclidean}(x, y) = 2$</p> <p>c. $x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1)$ cosine, correlation, Jaccard $\cos(x, y) = 0.75, \text{corr}(x, y) = 0.25, \text{Jaccard}(x, y) = 0.6$</p> | [10] | CO2 | L3 |
| 4 (a) | <p>Discuss whether or not each of the following activities is a data mining task.</p> <p>Scheme: Defining each statement is a data mining Task or not with conclusion carries 1 Mark each.</p> <p>Solution:</p> <p>a. Dividing the customers of a company according to their gender. No. This is a simple database query.</p> <p>b. Dividing the customers of a company according to their profitability. No. This is an accounting calculation, followed by the application of a threshold. However, predicting the profitability of a new customer would be data mining.</p> <p>c. Computing the total sales of a company. No. Again, this is simple accounting.</p> <p>d. Sorting a student database based on student identification numbers. No. Again, this is a simple database query.</p> <p>e. Predicting the outcomes of tossing a (fair) pair of dice. No. Since the die is fair, this is a probability calculation.</p> <p>f. Predicting the future stock price of a company using historical records. Yes. We would attempt to create a model that can predict the continuous value of the stock price.</p> | [06] | CO2 | L2 |
| (b) | <p>Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio).</p> <p>Scheme: Defining each statement as a type of attributes with conclusion carries 1 Mark each.</p> <p>Solution:</p> <p>a. Brightness as measured by people's judgments. Discrete, qualitative, ordinal</p> <p>b. Angles as measured in degrees between 0^0 and 360^0. Continuous, quantitative, ratio</p> <p>c. Bronze, Silver, and Gold medals as awarded at the Olympics. Discrete, qualitative, ordinal</p> <p>d. Height above sea level. Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)</p> | [04] | CO2 | L2 |

5 **Explain Frequent Item Set Generation with example.**

Scheme: Explanation of Frequent Itemset with Brute-Force approach and No. of candidates with examples carries **5 Marks** each.

Solution:

- Generate all itemsets whose support \geq minsup

Brute-force approach:

- Each itemset in the lattice is a candidate frequent itemset
- Determine the support count of each candidate by scanning the database.

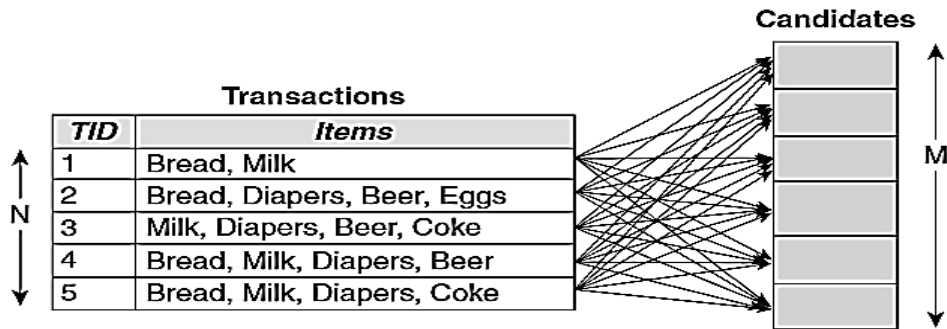


Figure 5.2. Counting the support of candidate itemsets.

Reduce the number of candidates (M)

The Apriori principle, is an effective way to eliminate some of the candidate itemsets without counting their support values.

[10]

CO3

L2

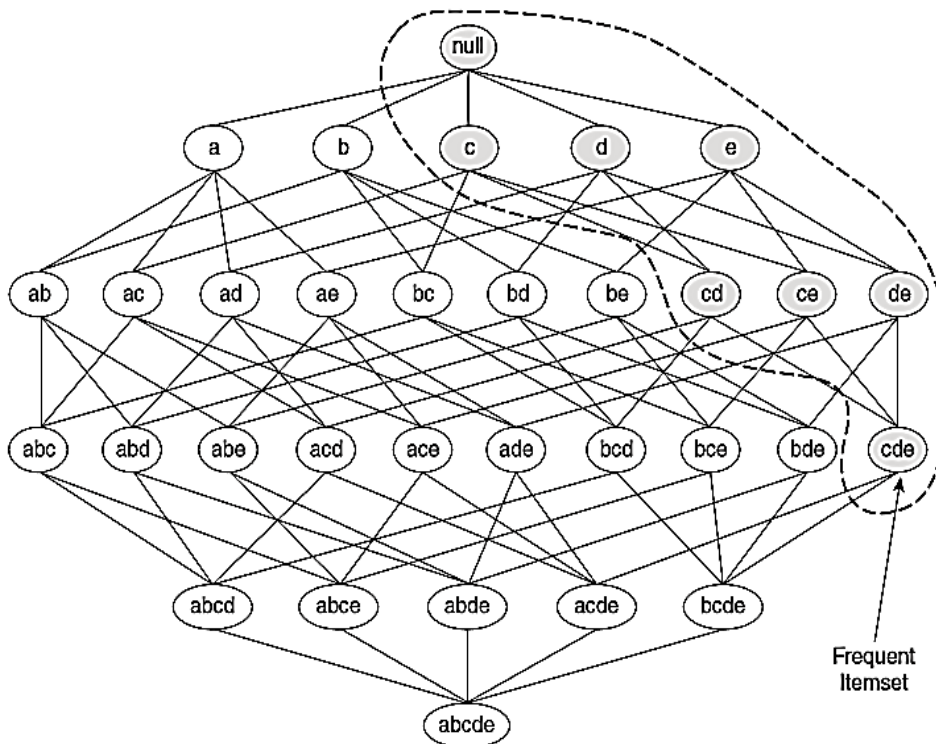


Figure 5.3. An illustration of the *Apriori* principle. If $\{c, d, e\}$ is frequent, then all subsets of this itemset are frequent.

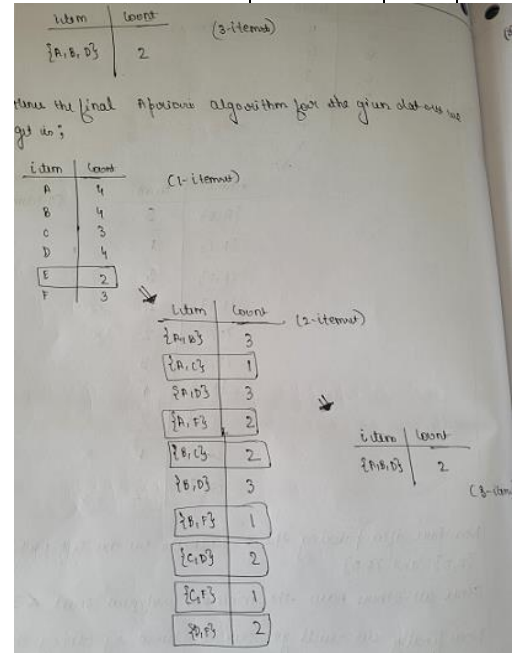
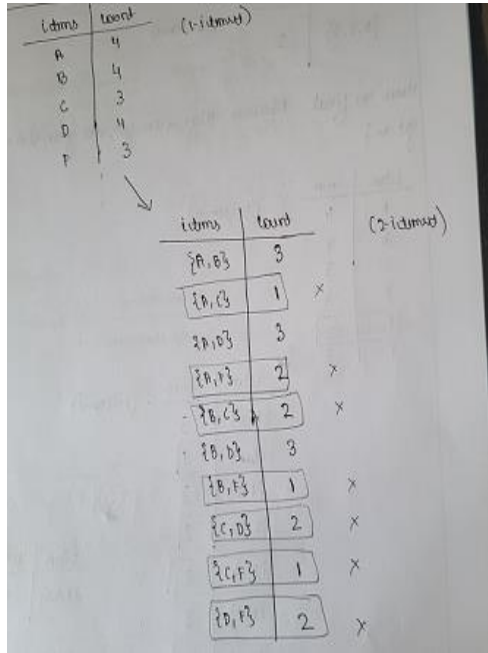
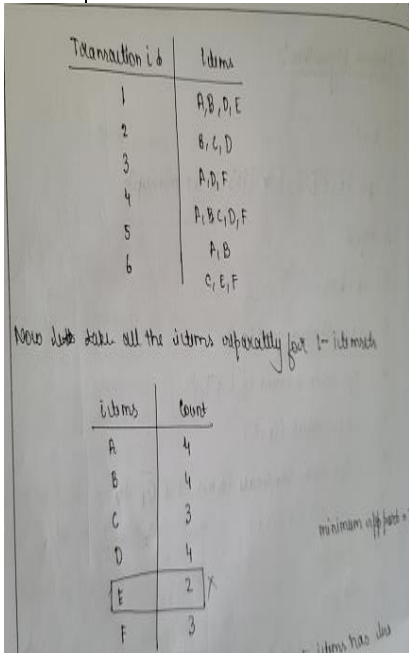
6 Apply Apriori Algorithm for the following dataset and Write the algorithm for the same.

[10] CO3 L3

| Transaction Id | Items Purchased |
|----------------|-----------------|
| 1 | A,B,D,E |
| 2 | B,C,D |
| 3 | A,D,F |
| 4 | A,B,C,D,F |
| 5 | A,B |
| 6 | C,E,F |

Scheme : Applying Apriori algorithm for the given dataset (6 Marks) and algorithm (4 Marks).

Solution:



Algorithm 6.1 Frequent itemset generation of the Apriori algorithm.

- 1: $k = 1$.
- 2: $F_k = \{ i \mid i \in I \wedge \sigma(\{i\}) \geq N \times \text{minsup} \}$. {Find all frequent 1-itemsets}
- 3: repeat
- 4: $k = k + 1$.
- 5: $C_k = \text{apriori-gen}(F_{k-1})$. {Generate candidate itemsets}
- 6: for each transaction $t \in T$ do
- 7: $C_t = \text{subset}(C_k, t)$. {Identify all candidates that belong to t }
- 8: for each candidate itemset $c \in C_t$ do
- 9: $\sigma(c) = \sigma(c) + 1$. {Increment support count}
- 10: end for
- 11: end for
- 12: $F_k = \{ c \mid c \in C_k \wedge \sigma(c) \geq N \times \text{minsup} \}$. {Extract the frequent k -itemsets}
- 13: until $F_k = \emptyset$
- 14: Result = $\bigcup F_k$.

Faculty Signature

CCI Signature

HOD Signature