UNIT II

Association Rules Mining: Basics - The Task and a Naïve Algorithm - The Apriori Algorithm - Improving the Efficiency of the Apriori Algorithm - Apriori-TID – Direct Hashing and Pruning (DHP) - Dynamic Item set Counting (DIC) - Mining Frequent Patterns without Candidate Generation (FP–Growth) - Performance Evaluation of Algorithms – Case Study: Mining Customer Value: From Association Rules to Direct Marketing.

UNIT II – ASSOCIATION RULE MINING

Bloom's Taxonomy – Level 1

1. Introduction to Association Rule Mining

Association Rule Mining (ARM) is a data mining technique used to discover interesting relationships, patterns, and associations among items in large datasets.

Example: In a supermarket, customers who purchase bread may also purchase butter.

Association rule:

\[Bread \rightarrow Butter \]

This means customers who purchase bread may also purchase butter.

Applications:

- Market basket analysis.
- Customer purchasing behaviour.
- Product recommendation systems.
- Online shopping.
- Banking and insurance.
-



2. Basics of Association Rules

An association rule consists of two parts:

- Antecedent (X): If-part or left-hand side.
- Consequent (Y): Then-part or right-hand side.

General form:

General form:

$$X \rightarrow Y$$

Where X and Y are itemsets and $X \cap Y = \emptyset$.

Important measures

1. Support

Support indicates how frequently an itemset occurs in the transaction database.

$$Support(X) = \frac{\text{Transactions containing X}}{\text{Total transactions}}$$

2. Confidence

Confidence measures how often Y occurs when X occurs.

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

3. Lift

Lift measures the strength of association between two items.

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Support(Y)}$$

Lift value	Interpretation
Lift = 1	Independent
Lift greater than 1	Positive association
Lift less than 1	Negative association

3. The Task and a Naïve Algorithm

Task: Identify frequent itemsets and generate association rules from transaction databases.

Example transaction database:

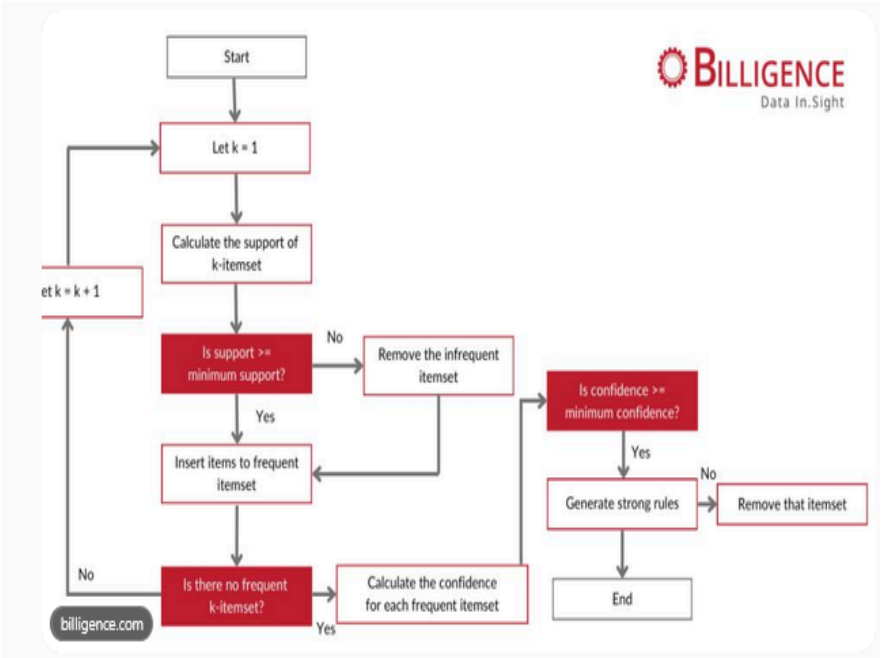
Transaction	Items
T1	Bread, Milk
T2	Bread, Butter
T3	Bread, Milk, Butter
T4	Milk, Butter
T5	Bread, Milk

Naïve Algorithm:

The naïve algorithm generates all possible item combinations and checks their support by scanning the transaction database.

Steps:

1. Identify all individual items.
2. Generate possible item combinations.
3. Calculate support for each combination.
4. Compare support with minimum support.
5. Identify frequent itemsets.
6. Generate association rules.
7. Calculate confidence.



Limitation: It generates a large number of candidate item sets and requires repeated database scans, making it computationally expensive.

4. Apriori Algorithm

The Apriori algorithm is a popular association rule mining algorithm used to find frequent itemsets.

It follows the Apriori Principle:

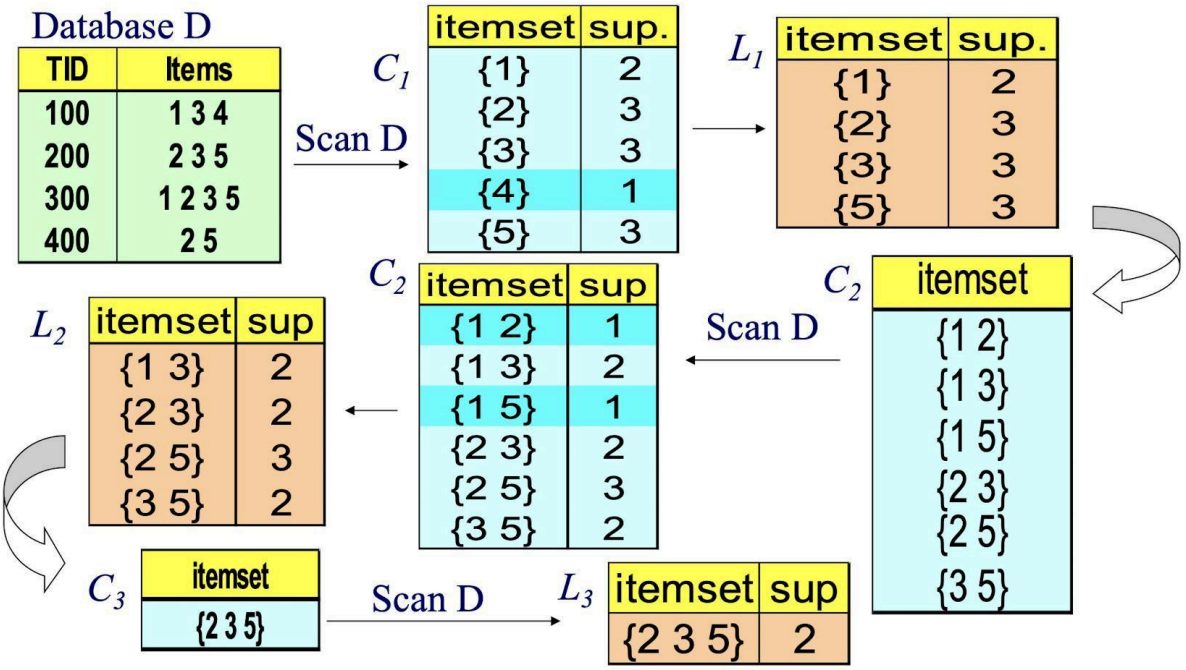
"If an itemset is frequent, all its non-empty subsets must also be frequent."

Similarly, if an itemset is infrequent, all its supersets must be infrequent.

Working procedure:

1. Find frequent 1-itemsets (L1).
2. Generate candidate 2-itemsets (C2).
3. Calculate support.
4. Remove infrequent itemsets.
5. Generate candidate 3-itemsets (C3).

6. Repeat until no more frequent itemsets are found.



Advantages:

- Simple and easy to understand.
- Uses the Apriori property.
- Eliminates infrequent itemsets.
- Suitable for discovering purchasing patterns.

Disadvantage: Requires multiple database scans and candidate generation.

5. Improving the Efficiency of Apriori Algorithm

The efficiency of Apriori can be improved by reducing candidate generation and database scanning.

Important techniques:

Technique	Description
Hash-based itemset counting	Uses hash tables to reduce candidates
Transaction reduction	Removes transactions that cannot contain further frequent itemsets
Partitioning	Divides database into smaller partitions

Sampling	Uses a sample of the database
Dynamic itemset counting	Introduces candidates during database scanning

6. Apriori-TID Algorithm

Apriori-TID is an improved version of the Apriori algorithm.

Main idea: Instead of scanning the original database in every iteration, it uses a candidate itemset database containing transaction identifiers.

Working:

1. Generate candidate 1-itemsets.
2. Construct candidate itemset database.
3. Store transaction identifiers associated with candidates.
4. Generate the next candidate itemsets.
5. Count their occurrences using the candidate database.
6. Continue until frequent itemsets are identified.

Advantages:

- Reduces repeated scanning of the original database.
- Can improve performance in later iterations.

Limitation: Candidate database may become large.

7. Direct Hashing and Pruning (DHP)

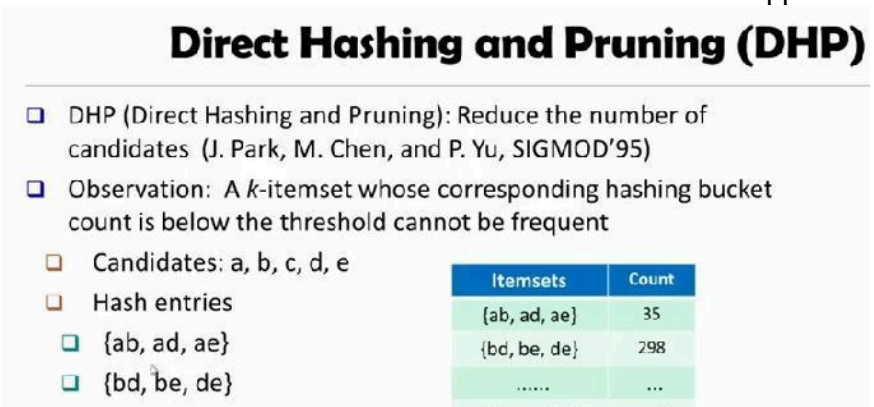
DHP stands for Direct Hashing and Pruning.

It is an improved technique for reducing the number of candidate itemsets in association rule mining.

Main concepts:

- Uses hash tables to reduce candidate itemsets.
- Uses pruning to eliminate unnecessary candidates.
- Reduces database size by removing transactions that cannot contribute to future frequent itemsets.

Working steps:

1. Scan the database.
2. Generate candidate item sets.
3. Hash item sets into buckets.
4. Count bucket frequencies.
5. Prune candidates whose bucket count is below minimum support.
6. 

8. Dynamic Itemset Counting (DIC)

Dynamic Itemset Counting is an association rule mining technique that introduces new candidate itemsets during database scanning instead of waiting until the complete scan finishes.

Main idea: Divide database scanning into intervals and dynamically introduce new candidates.

Four itemset states:

State	Meaning
Dashed circle	Suspected infrequent
Dashed square	Suspected frequent
Solid circle	Confirmed infrequent
Solid square	Confirmed frequent

Working steps:

1. Divide the database into intervals.
2. Count candidate itemsets.
3. Introduce new candidates at interval boundaries.
4. Update their support counts.
5. Identify frequent itemsets.

Advantage: Can reduce the number of complete database scans.

9. FP-Growth Algorithm

FP-Growth stands for Frequent Pattern Growth.

It discovers frequent itemsets without generating candidate itemsets.

Main concept: Uses a compact data structure called an FP-Tree (Frequent Pattern Tree).

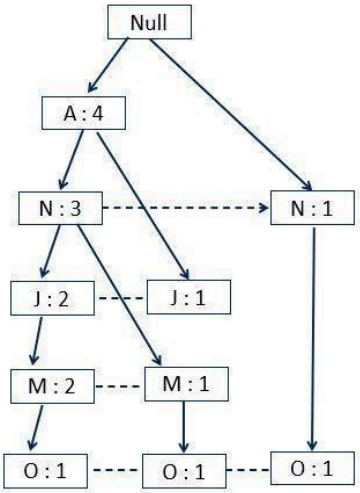
Working steps:

1. Scan the database and calculate item frequencies.
2. Remove items below minimum support.
3. Sort frequent items by descending support.
4. Construct the FP-Tree.
5. Generate conditional pattern bases.
6. Construct conditional FP-Trees.
7. Extract frequent patterns.

Mining frequent patterns using FP-tree

Creating conditional (sub) pattern-bases for item A

Item	Conditional Pattern Base
O	{(A,N,J,M :1), (A,N,M :1), (N :1)}
M	{(A,N,J :2), (A,N :1)}
J	{(A,N :2), (A :1)}
N	{(A :3)}
A	ϕ



Advantages:

- Avoids candidate generation.
- Reduces database scans.
- Efficient for large datasets.
- Uses a compact tree structure.

Limitation: FP-Tree construction and maintenance can be complex, especially for datasets with limited shared prefixes.

10. Performance Evaluation of Algorithms

Performance evaluation is used to compare different association rule mining algorithms.

Important parameters:

1. Execution time – time required to complete mining.
2. Memory usage – amount of memory consumed.
3. Database scans – number of database passes.
4. Candidate generation – number of candidates generated.
5. Scalability – ability to handle larger datasets.
6. Accuracy – correctness of the discovered patterns.

Algorithm	Candidate generation	Main feature
Naïve	Yes	Brute-force combinations
Apriori	Yes	Apriori property
Apriori-TID	Yes	Candidate transaction database
DHP	Yes	Hashing and pruning
DIC	Yes	Dynamic candidate counting
FP-Growth	No	FP-Tree structure

Important observation: FP-Growth avoids candidate generation, whereas Apriori relies on candidate generation and pruning.

11. Case Study – Mining Customer Value: From Association Rules to Direct Marketing

Introduction:

Customer Value Mining is a technique used to identify valuable customers and understand their purchasing behaviour.

Association rules help businesses discover products frequently purchased together and use these patterns for targeted marketing.



Example:

A supermarket discovers the following association:

\[Mobile ->Mobile\ Accessories \]

The supermarket can use this information to:

1. Identify customers purchasing mobile phones.
2. Recommend mobile accessories.
3. Offer personalized discounts.
4. Improve cross-selling.
5. Increase customer satisfaction.
6. Improve customer lifetime value.

Association Rules to Direct Marketing:

Association rule	Marketing strategy
Bread → Butter	Product bundle offer
Laptop → Mouse	Cross-selling
Mobile → Headphones	Product recommendation
Baby food → Diapers	Targeted promotion

Important parameters from the case study:

Parameter	Value	Meaning
Minimum Support	0.0045	0.45%
Minimum Confidence	0.2	20%
Minimum Lift	3	Strong positive association
Minimum Length	2	At least 2 items in a rule

12. Bloom's Taxonomy – Level 1

Level 1 – Remembering

Level 1 tests: Recall, recognize, identify, define, list, and remember basic concepts.

Section A – Basics of Association Rules

1. What is Association Rule Mining?

A) Data encryption
B) Discovering relationships among items
C) Data compression
D) Data deletion

Answer: B

2. Which technique is used in market basket analysis?

A) Classification
B) Regression
C) Association Rule Mining
D) Data encryption

Answer: C

3. What is the general form of an association rule?

A) X + Y
B) X → Y
C) X × Y
D) X − Y

Answer: B

4. What does support measure?

A) Execution time
B) Frequency of an itemset
C) Memory usage
D) Database size

Answer: B

5. Which measure determines the reliability of an association rule?

A) Support
B) Confidence
C) Distance
D) Variance

Answer: B

6. Which measure identifies the strength of association between items?

A) Mean
B) Median
C) Lift
D) Mode

Answer: C

7. What does a lift value equal to 1 indicate?

A) Positive association
B) Negative association
C) Independence
D) Strong association

Answer: C

8. Which part of an association rule represents the IF condition?

A) Consequent
B) Antecedent
C) Support
D) Lift

Answer: B

Section B – Apriori Algorithm

9. Who developed the Apriori algorithm?

A) Rakesh Agrawal and Ramakrishnan Srikant
B) Alan Turing
C) Charles Babbage
D) John McCarthy

Answer: A

10. Apriori is mainly used for which task?

A) Image processing
B) Frequent itemset mining
C) Sorting
D) Encryption

Answer: B

11. What is the main principle of the Apriori algorithm?

A) Regression principle
B) Apriori property
C) Clustering property
D) Encryption principle

Answer: B

12. Which itemsets are generated first in Apriori?

A) 2-itemsets
B) 3-itemsets
C) 1-itemsets
D) 4-itemsets

Answer: C

13. What does L1 represent in Apriori?

A) Candidate 1-itemsets
B) Frequent 1-itemsets
C) Frequent 2-itemsets
D) Infrequent itemsets

Answer: B

14. What does C2 represent?

A) Candidate 2-itemsets
B) Frequent 2-itemsets
C) Candidate 1-itemsets
D) Closed itemsets

Answer: A

15. What happens to an itemset whose support is below minimum support?

A) It is selected
B) It is pruned
C) It is duplicated
D) It is encrypted

Answer: B

16. Which is a major limitation of Apriori?

A) No candidate generation
B) Multiple database scans
C) No support calculation
D) No itemset generation

Answer: B

Section C – Apriori-TID and DHP

17. What does Apriori-TID use?

A) Image database
B) Candidate itemset database
C) Neural network
D) Decision tree

Answer: B

18. What is the main purpose of Apriori-TID?

A) Increase encryption
B) Reduce repeated scanning of the original database
C) Delete all transactions
D) Perform classification

Answer: B

19. What does DHP stand for?

A) Direct Hashing and Pruning
B) Data Handling Process
C) Dynamic Hash Processing
D) Direct Hierarchical Processing

Answer: A

20. Which technique is used in DHP?

A) Hashing
B) Regression
C) Classification
D) Normalization

Answer: A

21. What is the main objective of pruning in DHP?

A) Increase candidate item sets
B) Reduce unnecessary candidates
C) Increase database size
D) Duplicate transactions

Answer: B

22. What data structure is used in DHP for reducing candidates?

A) Stack
B) Queue
C) Hash table
D) Linked list

Answer: C

Section D – Dynamic Item set Counting

23. What does DIC stand for?

A) Data Item Classification
B) Dynamic Itemset Counting
C) Direct Item Calculation
D) Database Information Counting

Answer: B

24. What is the main feature of DIC?

A) Static candidate generation
B) Dynamic introduction of candidates
C) Data encryption
D) Image classification

Answer: B

25. DIC divides the database into:

A) Intervals
B) Images
C) Tables only
D) Clusters only

Answer: A

26. Which itemset state represents a confirmed frequent itemset in DIC?

A) Dashed circle
B) Dashed square
C) Solid square
D) Solid circle

Answer: C

Section E – FP-Growth Algorithm

27. What does FP-Growth stand for?

A) Frequent Pattern Growth
B) Fast Processing Growth
C) Frequent Processing Group
D) First Pattern Generation

Answer: A

28. Which data structure is used in FP-Growth?

A) B-Tree
B) FP-Tree
C) Binary Search Tree
D) Queue

Answer: B

29. Does FP-Growth generate candidate itemsets?

A) Yes
B) No
C) Sometimes always
D) None of the above

Answer: B

30. What is the full form of FP-Tree?

A) Frequent Pattern Tree
B) Fast Processing Tree
C) First Pattern Table
D) Frequent Processing Table

Answer: A

31. Which algorithm avoids candidate generation?

A) Apriori
B) Naïve
C) FP-Growth
D) DHP

Answer: C

32. What is a conditional pattern base used for?

A) Constructing conditional FP-Trees
B) Sorting transactions
C) Data encryption
D) Data deletion

Answer: A

Section F – Performance Evaluation and Case Study

33. Which parameter measures the time taken by an algorithm?

A) Support
B) Execution time
C) Confidence
D) Lift

Answer: B

34. Which parameter measures memory consumption?

A) Memory usage
B) Lift
C) Confidence
D) Support Answer: A

35. What does scalability refer to?

A) Ability to handle increasing data sizes
B) Removing data
C) Decreasing confidence
D) Deleting itemsets

Answer: A

36. Which technique is useful for product recommendations?

A) Association Rule Mining
B) Data encryption
C) File compression
D) Data deletion

Answer: A

37. What is cross-selling?

A) Selling unrelated products only
B) Offering complementary products
C) Deleting products
D) Reducing database size

Answer: B

38. What does customer lifetime value represent?

A) Customer's expected total value over the relationship
B) Product manufacturing cost
C) Database execution time
D) Number of transactions only

Answer: A

39. What is the minimum support specified in the case study?

A) 0.0045
B) 0.45
C) 0.20
D) 3

Answer: A

40. What is the minimum confidence specified in the case study?

A) 0.0045
B) 0.20
C) 3
D) 2

Answer: B

13. Quick Revision – Important One-Mark Questions

Question	Answer
What is ARM?	Association Rule Mining
What is FP?	Frequent Pattern
What is DHP?	Direct Hashing and Pruning
What is DIC?	Dynamic Itemset Counting
What is FP-Tree?	Frequent Pattern Tree
What is L1?	Frequent 1-itemsets
What is C2?	Candidate 2-itemsets
What is support?	Itemset frequency
What is confidence?	Rule reliability
What is lift?	Association strength
Which algorithm avoids candidates?	FP-Growth
Which algorithm uses hashing?	DHP
Which algorithm uses dynamic counting?	DIC
Which algorithm uses a candidate database?	Apriori-TID

14. Self-Assessment Test – Level 1

Association Rule Mining Quiz - 8 Marks

Select one correct answer for each question.

1. Which algorithm uses an FP-Tree?

A. Apriori
B. FP-Growth
C. DHP
D. DIC

2. What does confidence measure?

A. Rule reliability
B. Memory usage
C. Execution time
D. Database size

3. Which algorithm uses Direct Hashing and Pruning?

A. DIC
B. FP-Growth
C. DHP
D. Apriori-TID

4. Which itemsets are generated first in Apriori?

A. 3-itemsets
B. 2-itemsets
C. 1-itemsets
D. 4-itemsets

5. Which algorithm avoids candidate generation?

A. Apriori
B. DHP
C. Naïve
D. FP-Growth

6. What does lift equal to 1 indicate?

A. Positive association
B. Independence
C. Negative association
D. Strong association

7. What does DIC stand for?

A. Dynamic Itemset Counting
B. Data Item Classification
C. Direct Item Counting
D. Database Index Creation

8. What is the minimum lift in the case study?

A. 0.2
B. 0.0045
C. 3
D. 2

UNIT II – ASSOCIATION RULE MINING

Complete University Question Bank

Bloom's Taxonomy Levels 1–6

Subject: Data Mining and Data Warehousing
Department: Computer Science / Data Science
Coverage: Complete Unit II syllabus

This question bank is designed for undergraduate examinations, university internal assessments, semester examinations, and classroom teaching.

Question bank structure

Section	Question type	Bloom's level
A	MCQs with answers	L1–L6
B	Fill in the blanks	L1–L2
C	2-mark questions and answers	L1–L2
D	5-mark questions and answers	L2–L4
E	10-mark questions and answers	L3–L6
F	Important algorithm comparisons	L2–L4
G	Case-study questions	L3–L6

SECTION A – MCQs WITH ANSWERS

Part 1: Level 1 – Remembering

1. Association Rule Mining is used to identify:

- A) Data encryption
- B) Relationships among items
- C) Data deletion
- D) Sorting

Answer: B

2. The general form of an association rule is:

- A) $X + Y$
- B) $X \rightarrow Y$
- C) $X \times Y$
- D) $X - Y$

Answer: B

3. Which measure indicates the frequency of an itemset?

- A) Confidence
- B) Lift
- C) Support
- D) Variance

Answer: C

4. Which algorithm uses the Apriori property?

- A) K-means
- B) Apriori
- C) KNN
- D) Decision Tree

Answer: B

5. FP-Growth uses:

- A) Binary tree
- B) FP-Tree
- C) Queue
- D) Stack

Answer: B

6. DHP stands for:

- A) Direct Hashing and Pruning
- B) Data Handling Process
- C) Dynamic Hash Processing
- D) Direct Hierarchical Processing

Answer: A

7. DIC stands for:

- A) Data Item Classification
- B) Dynamic Itemset Counting
- C) Direct Item Calculation
- D) Database Information Counting

Answer: B

8. Which algorithm avoids candidate generation?

- A) Apriori
- B) DHP
- C) FP-Growth
- D) Naïve

Answer: C

9. A lift value of 1 indicates:

- A) Positive association
- B) Independence
- C) Negative association
- D) Strong association

Answer: B

10. Which measure determines the reliability of a rule?

- A) Confidence
- B) Mean
- C) Median
- D) Variance

Answer: A

Part 2: Level 2 – Understanding

11. The Apriori principle states that:

- A) All itemsets are frequent
- B) All non-empty subsets of a frequent itemset must be frequent
- C) All transactions are identical
- D) All itemsets have equal support

Answer: B

12. Why is candidate pruning performed?

- A) To increase database size
- B) To remove unnecessary candidates
- C) To duplicate data
- D) To increase memory consumption

Answer: B

13. What is the main purpose of an FP-Tree?

- A) Store compressed frequent-pattern information
- B) Encrypt transactions
- C) Delete itemsets
- D) Sort customer names

Answer: A

14. Apriori-TID differs from Apriori because it:

- A) Uses a candidate itemset database
- B) Uses only neural networks
- C) Avoids support calculation
- D) Uses image data

Answer: A

15. The purpose of hashing in DHP is to:

- A) Reduce candidate itemsets
- B) Increase duplicate data
- C) Encrypt customer records
- D) Increase transaction count

Answer: A

16. Confidence is calculated using:

- A) Support(X) / Support(Y)
- B) Support(X ∪ Y) / Support(X)
- C) Support(X) × Support(Y)
- D) Support(Y) / Support(X ∪ Y)

Answer: B

17. FP-Growth is particularly useful because:

- A) It generates all candidates
- B) It avoids candidate generation
- C) It does not need transactions
- D) It ignores support

Answer: B

18. Direct marketing uses association rules to:

- A) Identify customer buying patterns
- B) Delete customers
- C) Compress databases
- D) Encrypt products

Answer: A

Part 3: Level 3 – Applying

19. If 40 out of 100 transactions contain bread, support is:

- A) 20%
- B) 30%
- C) 40%

- D) 50%

Answer: C

20. If support(X ∪ Y) = 0.3 and support(X) = 0.5, confidence is:

- A) 0.2
- B) 0.4
- C) 0.6
- D) 0.8

Answer: C

21. If confidence is 0.8 and support(Y) is 0.4, lift is:

- A) 1
- B) 2
- C) 3
- D) 4

Answer: B

22. If minimum support is 30%, which itemset with 20% support is removed?

- A) Frequent itemset
- B) Infrequent itemset
- C) Closed itemset
- D) Maximal itemset

Answer: B

23. A customer buys a laptop and receives a mouse recommendation. This is:

- A) Cross-selling
- B) Encryption
- C) Clustering
- D) Classification

Answer: A

Part 4: Level 4 – Analyzing

24. Which algorithm is generally preferred when candidate generation becomes expensive?

- A) Apriori
- B) FP-Growth
- C) Naïve
- D) Sequential search

Answer: B

25. Which factor can make an FP-Tree less compact?

- A) Highly shared prefixes
- B) Few shared prefixes

- C) Repeated common patterns
- D) Identical transactions

Answer: B

26. Which algorithm dynamically introduces candidate itemsets?

- A) DIC
- B) FP-Growth
- C) K-means
- D) Naïve Bayes

Answer: A

27. Which algorithm uses hash buckets for pruning?

- A) DHP
- B) DIC
- C) FP-Growth
- D) KNN

Answer: A

Part 5: Level 5 – Evaluating

28. Which is a major advantage of FP-Growth over Apriori?

- A) More candidate generation
- B) No candidate generation
- C) More database scans
- D) No frequent itemsets

Answer: B

29. Which metric is appropriate for evaluating algorithm efficiency?

- A) Execution time
- B) Customer name
- C) Product colour
- D) Item description

Answer: A

30. A rule with lift greater than 1 indicates:

- A) Positive association
- B) Independence
- C) No relationship
- D) Negative association

Answer: A

Part 6: Level 6 – Creating

31. Which algorithm would you choose to build a frequent-pattern mining system without candidate generation?

- A) Apriori
- B) FP-Growth
- C) Naïve
- D) DIC

Answer: B

32. Which approach can be designed to reduce candidate itemsets using hashing?

- A) DHP
- B) KNN
- C) Linear regression
- D) Decision tree

Answer: A

33. A retailer wants to create personalized product recommendations. Which approach is most appropriate?

- A) Association Rule Mining
- B) Data encryption
- C) File compression
- D) Data deletion

Answer: A

SECTION B – FILL IN THE BLANKS

20 Questions – L1 & L2

No.	Question	Answer
1	Association Rule Mining discovers _____ among items.	Relationships
2	The general form of an association rule is _____.	X → Y
3	_____ measures the frequency of an itemset.	Support
4	_____ measures the reliability of a rule.	Confidence
5	_____ measures the strength of association.	Lift
6	Apriori follows the _____ principle.	Apriori
7	L1 represents frequent _____.	1-itemsets
8	C2 represents candidate _____.	2-itemsets
9	FP-Growth uses a data structure called _____.	FP-Tree
10	DHP stands for _____.	Direct Hashing and Pruning
11	DIC stands for _____.	Dynamic Itemset Counting

12	Apriori-TID uses a _____ itemset database.	Candidate
13	FP-Growth avoids _____ generation.	Candidate
14	The left-hand side of a rule is called _____.	Antecedent
15	The right-hand side of a rule is called _____.	Consequent
16	A lift value of 1 indicates _____.	Independence
17	DHP uses _____ tables.	Hash
18	The process of removing unnecessary candidates is called _____.	Pruning
19	Offering complementary products is called _____.	Cross-selling
20	The minimum support in the case study is _____.	0.0045

SECTION C – TWO-MARK QUESTIONS WITH ANSWERS

Important Short Answers

1. Define Association Rule Mining.

Association Rule Mining is a data mining technique used to discover relationships and frequent patterns among items in large databases.

2. What is an association rule?

An association rule represents a relationship between two itemsets.

$$X \rightarrow Y$$

Here, X is the antecedent and Y is the consequent.

3. Define support.

Support measures how frequently an itemset occurs in the transaction database.

$$Support(X)=\frac{Count(X)}{Total\ Transactions}$$

4. Define confidence.

Confidence measures the probability of finding Y in transactions that contain X.

$$Conf(X\rightarrow Y)=\frac{Support(X\cup Y)}{Support(X)}$$

5. What is lift?

Lift measures the strength of association between two itemsets.

$$Lift=\frac{Confidence(X\rightarrow Y)}{Support(Y)}$$

6. State the Apriori principle.

If an itemset is frequent, all its non-empty subsets must also be frequent. If an itemset is infrequent, all its supersets must be infrequent.

7. What is a frequent itemset?

An itemset whose support is greater than or equal to the minimum support threshold is called a frequent itemset.

8. What is candidate generation?

Candidate generation is the process of producing possible itemsets that may satisfy the minimum support requirement.

9. What is pruning?

Pruning is the process of eliminating itemsets that cannot become frequent, thereby reducing computation.

10. What is Apriori-TID?

Apriori-TID is an improved version of Apriori that uses a candidate itemset database containing transaction-related information rather than repeatedly scanning the original database.

11. What is DHP?

DHP stands for Direct Hashing and Pruning. It uses hash-based techniques to reduce candidate itemsets.

12. What is DIC?

Dynamic Itemset Counting is an algorithm that introduces new candidate itemsets dynamically during database scanning.

13. What is an FP-Tree?

An FP-Tree is a compact tree-based data structure used by FP-Growth to store frequent-pattern information.

14. What is FP-Growth?

FP-Growth is an association rule mining algorithm that finds frequent patterns without generating candidate itemsets.

15. Mention two applications of Association Rule Mining.

1. Market basket analysis.
2. Product recommendation systems.

16. What is direct marketing?

Direct marketing is a marketing strategy in which businesses communicate targeted offers and promotions directly to selected customers.

17. What is customer lifetime value?

Customer lifetime value is the estimated total value a customer contributes to a business throughout their relationship with it.

18. Mention any two performance evaluation measures.

1. Execution time.
2. Memory usage.

19. What is a conditional pattern base?

A conditional pattern base is the collection of prefix paths associated with a particular item in an FP-Tree.

20. What is cross-selling?

Cross-selling is a marketing technique in which complementary products are recommended to customers based on their purchasing behaviour.

SECTION D – FIVE-MARK QUESTIONS WITH ANSWERS

1. Explain the basics of Association Rule Mining.

Association Rule Mining identifies relationships between items in a database.

Important concepts:

1. Item: An individual product or object.
2. Itemset: A collection of items.
3. Transaction: A set of items purchased together.
4. Support: Frequency of an itemset.
5. Confidence: Reliability of a rule.
6. Lift: Strength of association.

Example:

\[Bread\rightarrow Butter \]

This rule indicates that customers purchasing bread may also purchase butter.

Applications include retail marketing, e-commerce, healthcare and customer recommendations.

2. Explain the Naïve Algorithm.

The Naïve Algorithm discovers frequent itemsets by generating possible combinations.

Steps:

1. Read the transaction database.
2. Identify individual items.
3. Generate all possible item combinations.
4. Calculate the support of each itemset.
5. Compare support with minimum support.
6. Select frequent itemsets.
7. Generate association rules.
8. Calculate confidence and lift.

Advantages:

- Easy to implement.
- Simple to understand.

Disadvantages:

- Generates many candidates.
- Requires considerable computation.
- Not efficient for large databases.

3. Explain the Apriori Algorithm with an example.

The Apriori algorithm identifies frequent itemsets using the Apriori property.

Consider the following transactions:

Transaction	Items
T1	A, B, C
T2	A, B
T3	A, C
T4	B, C
T5	A, B, C

Assume minimum support count = 2.

Step 1: Generate L1

Item	Support count
A	4
B	4
C	4

Therefore:

$$L1=\{A,B,C\}$$

Step 2: Generate C2

Candidate	Support count
AB	3
AC	3
BC	3

All candidates satisfy minimum support.

$$L2=\{AB,AC,BC\}$$

Step 3: Generate C3

Candidate	Support count
ABC	2

Therefore:

$$L3=\{ABC\}$$

Conclusion: The frequent itemsets are A, B, C, AB, AC, BC and ABC.

4. Explain the techniques used to improve Apriori efficiency.

The following methods improve the performance of Apriori:

1. Hash-based itemset counting – reduces unnecessary candidate itemsets.

2. Transaction reduction – removes transactions that cannot contribute to future frequent itemsets.
3. Partitioning – divides the database into smaller partitions.
4. Sampling – uses a representative sample of the database.
5. Dynamic itemset counting – introduces candidates during scanning.

These techniques reduce computation time and improve mining efficiency.

5. Explain Apriori-TID.

Apriori-TID is an extension of the Apriori algorithm.

Working:

1. Generate candidate itemsets.
2. Create a candidate itemset database.
3. Store transaction-related candidate information.
4. Count candidate itemsets.
5. Generate subsequent candidate itemsets.
6. Repeat until no more frequent itemsets exist.

Advantages:

- Reduces repeated scanning of the original database.
- Can improve performance in later iterations.

Limitation: The candidate database may consume more memory.

6. Explain DHP with its working procedure.

DHP means Direct Hashing and Pruning.

Working:

1. Scan the transaction database.
2. Generate candidate itemsets.
3. Hash itemsets into buckets.
4. Calculate bucket counts.
5. Remove candidates whose bucket count is below minimum support.
6. Generate the next candidate itemsets.
7. Identify frequent itemsets.

Advantages:

- Reduces candidate generation.
- Uses hash-based pruning.
- Improves mining efficiency.

7. Explain Dynamic Itemset Counting.

DIC introduces candidate itemsets dynamically during database scanning.

Working:

1. Divide the database into intervals.
2. Begin counting candidate itemsets.
3. Introduce new candidates at interval boundaries when their subsets qualify.
4. Continue counting candidates.
5. Identify frequent itemsets after sufficient support counts are obtained.

Four states:

- Dashed circle – suspected infrequent.
- Dashed square – suspected frequent.
- Solid circle – confirmed infrequent.
- Solid square – confirmed frequent.

8. Explain FP-Growth with its advantages.

FP-Growth is a frequent-pattern mining algorithm that does not generate candidate itemsets.

Working:

1. Scan the database.
2. Find the frequency of items.
3. Remove infrequent items.
4. Sort frequent items.
5. Construct an FP-Tree.
6. Generate conditional pattern bases.
7. Construct conditional FP-Trees.
8. Extract frequent patterns.

Advantages:

- Avoids candidate generation.
- Compresses transaction information.
- Reduces database scans.
- Efficient for large transaction datasets.

9. Explain the performance evaluation of Association Rule Mining algorithms.

Performance evaluation compares mining algorithms based on:

Measure	Explanation
Execution time	Time taken to complete mining
Memory usage	Memory required
Database scans	Number of database passes
Candidate generation	Number of candidate itemsets
Scalability	Ability to process larger datasets

Pattern quality	Usefulness and validity of patterns

The choice of algorithm depends on the size, density and characteristics of the database.

10. Explain the Customer Value Mining case study.

Customer Value Mining identifies customer behaviour and helps businesses develop effective marketing strategies.

Working:

1. Collect customer transaction data.
2. Identify frequent product combinations.
3. Generate association rules.
4. Calculate support, confidence and lift.
5. Identify customer purchasing patterns.
6. Design personalized offers.
7. Improve cross-selling and customer retention.

Case-study parameters:

- Minimum support = 0.0045
- Minimum confidence = 0.2
- Minimum lift = 3
- Minimum length = 2

Benefits:

- Targeted marketing.
- Increased sales.
- Customer segmentation.
- Better product recommendations.
- Improved customer value.

SECTION E – TEN-MARK QUESTIONS WITH ANSWERS

Important Long-Answer Questions

1. Explain Association Rule Mining, its basic concepts and applications.

Introduction:

Association Rule Mining is an important data mining technique used to discover hidden relationships among items in large transactional databases.

1. Basic concepts

- Item: Individual product.
- Itemset: Group of items.

- Transaction: Collection of purchased items.
- Frequent itemset: Itemset satisfying minimum support.
- Association rule: Relationship between two itemsets.

2. Important measures

Support:

$$Support(X)=\frac{Count(X)}{Total\ Transactions}$$

Confidence:

$$Confidence(X\rightarrow Y)= \frac{Support(X\cup Y)}{Support(X)}$$

Lift:

$$Lift(X\rightarrow Y)= \frac{Confidence(X\rightarrow Y)}{Support(Y)}$$

3. Working process

1. Collect transaction data.
2. Preprocess the data.
3. Generate candidate itemsets.
4. Identify frequent itemsets.
5. Generate association rules.
6. Evaluate rules using support, confidence and lift.
7. Apply discovered rules.

4. Applications

- Retail business.
- E-commerce.
- Banking.
- Healthcare.
- Telecommunications.
- Recommendation systems.

Conclusion: Association Rule Mining helps organizations discover useful purchasing patterns and improve business decisions.

2. Explain the Apriori Algorithm with its working procedure and example.

Introduction: Apriori is a frequent itemset mining algorithm that uses the Apriori property to eliminate unnecessary candidates.

Apriori property: Every non-empty subset of a frequent itemset must also be frequent.

Algorithm:

```
Input:
    Transaction Database D
    Minimum Support minsup

Output:
    Frequent Itemsets L
```

```
1. Find frequent 1-itemsets L1
2. Set k = 2
3. Repeat
4.      Generate candidate itemsets Ck from Lk-1
5.      Prune candidates using Apriori property
6.      Scan database D
7.      Count support of each candidate
8.      Select candidates satisfying minsup
9.      Lk = frequent candidate itemsets
10.     k = k + 1
11. Until Lk becomes empty
12. Return all frequent itemsets
```

Example:

Transaction	Items
T1	A, B, C
T2	A, B
T3	A, C
T4	B, C
T5	A, B, C

Minimum support count = 2.

Itemset	Support count	Result
A	4	Frequent
B	4	Frequent
C	4	Frequent
AB	3	Frequent
AC	3	Frequent
BC	3	Frequent
ABC	2	Frequent

Advantages:

1. Simple implementation.
2. Uses pruning.
3. Finds frequent itemsets.
4. Easy to understand.

Limitations:

1. Repeated database scans.
2. Large candidate generation.
3. High computation cost for dense datasets.

Conclusion: Apriori is a fundamental algorithm for discovering frequent itemsets and association rules.

3. Compare Apriori, Apriori-TID, DHP and DIC algorithms.

Feature	Apriori	Apriori-TID	DHP	DIC
Main technique	Candidate generation	Candidate database	Hashing and pruning	Dynamic counting
Candidate generation	Yes	Yes	Yes	Yes
Database processing	Repeated scans	Uses candidate database in later passes	Hash-based pruning	Interval-based counting
Main advantage	Simple	Reduces original DB scanning in later passes	Reduces candidates	Dynamic introduction
Limitation	Many scans	Candidate database memory	Hash bucket overhead	Complex counting

Explanation:

- Apriori uses frequent itemset generation and pruning.
- Apriori-TID uses transaction-related candidate information.
- DHP uses hashing and pruning.
- DIC introduces candidates dynamically during scanning.

Conclusion: All four algorithms aim to improve frequent itemset discovery, but they use different optimization strategies.

4. Explain the FP-Growth algorithm with its tree construction.

Introduction: FP-Growth discovers frequent patterns without candidate generation.

Working procedure:

Step 1 – Scan database: Calculate support counts.

Step 2 – Remove infrequent items: Eliminate items below minimum support.

Step 3 – Sort items: Arrange remaining items in descending support order.

Step 4 – Construct FP-Tree: Insert transactions into a shared-prefix tree.

Step 5 – Generate conditional pattern bases: Identify prefix paths for each item.

Step 6 – Construct conditional FP-Trees: Build smaller trees for frequent patterns.

Step 7 – Extract frequent patterns: Generate all frequent itemsets.

Advantages:

- No candidate generation.
- Compact data representation.
- Fewer database scans.
- Efficient pattern mining.

Limitations:

- Tree construction is complex.
- Conditional pattern generation requires additional processing.

- Performance depends on data structure and transaction patterns.

Conclusion: FP-Growth is a useful alternative to Apriori, especially when candidate generation becomes expensive.

5. Explain the techniques used to improve Apriori algorithm efficiency.

Introduction: Apriori performance can be improved by reducing the number of candidates, database scans and computation.

Techniques:

1. Hash-based itemset counting: Uses hash buckets to eliminate unlikely candidates.
2. Transaction reduction: Removes transactions that cannot contain future frequent itemsets.
3. Partitioning: Divides the database into independent partitions.
4. Sampling: Processes a representative sample.
5. Dynamic itemset counting: Introduces candidates during scanning.

Benefits:

- Less computation.
- Reduced candidate generation.
- Improved scalability.
- Better memory utilization in suitable datasets.
- Faster frequent-pattern discovery.

Conclusion: Optimization techniques improve the performance of frequent itemset mining.

6. Explain performance evaluation of Association Rule Mining algorithms.

Introduction: Performance evaluation determines how efficiently algorithms discover frequent itemsets and association rules.

Evaluation parameters:

1. Execution time.
2. Memory consumption.
3. Number of database scans.
4. Candidate itemsets generated.
5. Scalability.
6. Quality of discovered patterns.

Comparison:

Algorithm	Candidate generation	Important feature
Naïve	Yes	Exhaustive combinations
Apriori	Yes	Apriori pruning

Apriori-TID	Yes	Candidate database
DHP	Yes	Hash-based pruning
DIC	Yes	Dynamic counting
FP-Growth	No	Tree-based mining

Factors affecting performance:

- Database size.
- Number of distinct items.
- Transaction length.
- Minimum support.
- Data density.
- Available memory.

Conclusion: FP-Growth is often advantageous when candidate generation is expensive, while Apriori remains useful for understanding and implementing basic association mining.

7. Explain the case study: Mining Customer Value – From Association Rules to Direct Marketing.

Introduction:

Customer Value Mining uses transaction information to identify valuable customers, product relationships and purchasing behaviour.

Objectives:

1. Identify valuable customer groups.
2. Discover purchasing patterns.
3. Develop targeted promotions.
4. Improve cross-selling.
5. Increase customer retention.

Step 1 – Data collection

Collect transaction records, customer purchase history and product information.

Step 2 – Data preprocessing

Remove inconsistent records, handle missing values and prepare transactional data.

Step 3 – Frequent itemset discovery

Use Apriori or FP-Growth to identify frequent product combinations.

Step 4 – Rule generation

Generate rules such as:

\[Laptop\rightarrow Mouse \]

Step 5 – Rule evaluation

Evaluate rules using support, confidence and lift.

Step 6 – Direct marketing

Use discovered patterns to prepare personalized offers.

Case-study parameters:

Parameter	Value
Minimum support	0.0045
Minimum confidence	0.2
Minimum lift	3
Minimum length	2

Benefits:

- Better customer understanding.
- Improved promotional campaigns.
- Increased cross-selling opportunities.
- Better marketing decisions.
- Improved customer lifetime value.

Conclusion: Association Rule Mining transforms transaction data into useful marketing knowledge and supports customer-oriented business strategies.

8. Design an Association Rule Mining system for supermarket transactions.

Problem: A supermarket wants to identify frequently purchased product combinations and recommend complementary products.

Proposed system:

Modules:

1. Transaction data collection.
2. Data preprocessing.
3. Minimum support selection.
4. Frequent itemset mining.
5. Association rule generation.
6. Confidence and lift calculation.
7. Recommendation generation.

Example rules:

Rule	Business action
Bread → Butter	Bundle promotion
Laptop → Mouse	Cross-selling
Mobile → Headphones	Accessory recommendation
Milk → Cereal	Combo offer

Expected results:

- Better product recommendations.

- Increased sales opportunities.
- Better customer satisfaction.
- More effective marketing.

Conclusion: The proposed system converts shopping transactions into actionable business recommendations.

SECTION F – IMPORTANT DIFFERENCES

1. Support vs Confidence vs Lift

Support	Confidence	Lift
Measures frequency	Measures rule reliability	Measures association strength
Based on itemset occurrence	Based on conditional occurrence	Compares confidence with consequent support
Indicates popularity	Indicates predictive reliability	Indicates positive, negative or independent association

2. Apriori vs FP-Growth

Apriori	FP-Growth
Candidate generation	No candidate generation
Multiple database scans	Typically two database scans to build the FP-Tree
Uses frequent itemset tables	Uses FP-Tree
Simple implementation	More complex implementation
Can generate many candidates	Avoids candidate explosion

3. DHP vs DIC

DHP	DIC
Direct Hashing and Pruning	Dynamic Itemset Counting
Uses hash buckets	Uses database intervals
Reduces candidates	Dynamically introduces candidates
Hash-based optimization	Counting-based optimization

SECTION G – UNIVERSITY EXAMINATION MODEL QUESTIONS

Part A – One-mark MCQs

- 10–20 questions from Association Rule Mining basics.
- 10 questions from Apriori and its improvements.
- 10 questions from FP-Growth and DIC.

- 5 questions from performance evaluation.
- 5 questions from the customer value case study.

Part B – Two-mark questions

1. Define Association Rule Mining.
2. Define support and confidence.
3. What is lift?
4. State the Apriori principle.
5. What is an FP-Tree?
6. Define DHP.
7. Define DIC.
8. What is a conditional pattern base?
9. What is customer lifetime value?
10. Define direct marketing.

Part C – Five-mark questions

1. Explain the Naïve Algorithm.
2. Explain the Apriori algorithm with an example.
3. Explain Apriori-TID.
4. Explain DHP.
5. Explain DIC.
6. Explain FP-Growth.
7. Explain the efficiency improvements of Apriori.
8. Discuss the performance evaluation parameters.

Part D – Ten-mark questions

1. Explain Association Rule Mining and its basic concepts.
2. Describe the Apriori algorithm with a suitable example.
3. Explain the different techniques for improving Apriori efficiency.
4. Explain Apriori-TID, DHP and DIC.
5. Explain FP-Growth with FP-Tree construction.
6. Compare the performance of Association Rule Mining algorithms.
7. Discuss the customer value mining case study.
8. Design an Association Rule Mining system for supermarket transactions.

SECTION H – MOST IMPORTANT EXAM TOPICS

High-priority revision

Priority	Topic	Important question
1	Apriori	Algorithm with example
2	FP-Growth	FP-Tree construction
3	Association Rule Basics	Support, confidence, lift
4	DHP	Hashing and pruning
5	DIC	Dynamic itemset counting
6	Apriori-TID	Working procedure
7	Algorithm comparison	Apriori vs FP-Growth
8	Case study	Customer value and direct marketing
9	Performance evaluation	Time and memory analysis
10	Naïve Algorithm	Working and limitations

Teaching recommendation: Give special attention to Apriori numerical problems, FP-Tree diagrams, algorithm comparisons and the customer value case study. These cover the main conceptual and procedural areas of the unit.

The question bank above covers the complete syllabus at Bloom's Levels 1–6 and can be used to prepare classroom notes, internal assessment tests and semester examination materials.