Chapter 4:

Pattern mining: Basic Concepts and Methods

Data Mining

Instructor:

Debesh Jha, Ph.D.,

Visiting Assistant Professor,

University of South Dakota,

Vermillion, SD

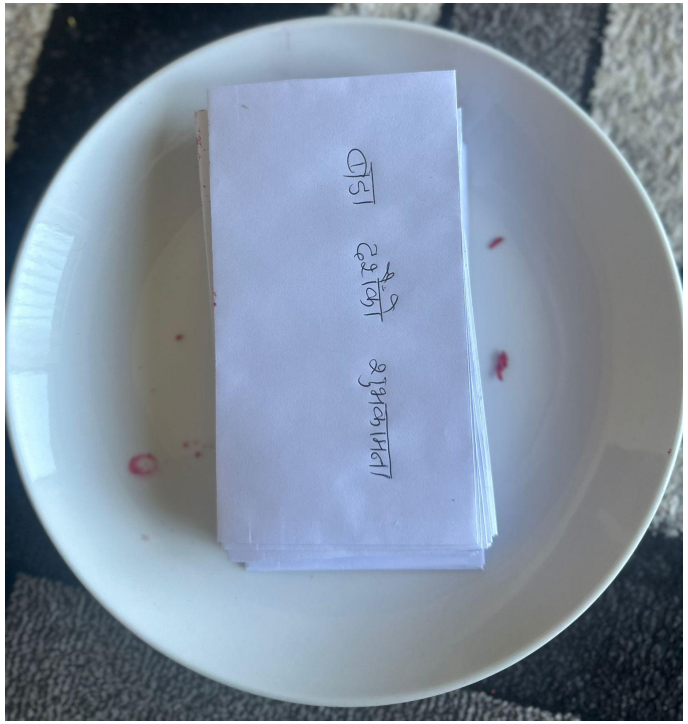
September October 19, 2024

Pattern Mining: Basic Concepts and Methods

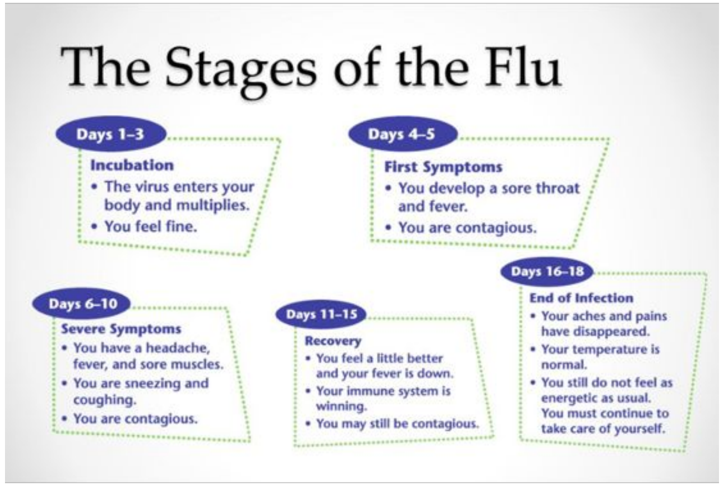
- Basic Concepts
- Frequent Itemset Mining Methods
- Which Patterns Are Interesting?—Pattern Evaluation Methods
- Summary

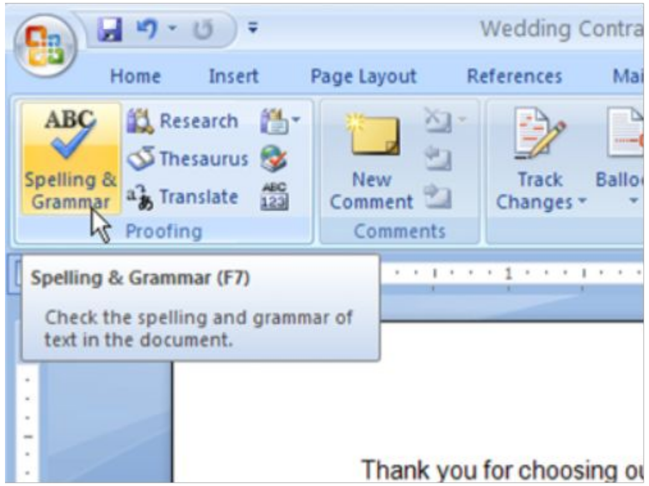
Patterns

- Patterns: A set of items, subsequences, or substructures that occur frequently together (or strongly correlated) in a data set.
- Patterns represent intrinsic and important properties of datasets.



Patterns

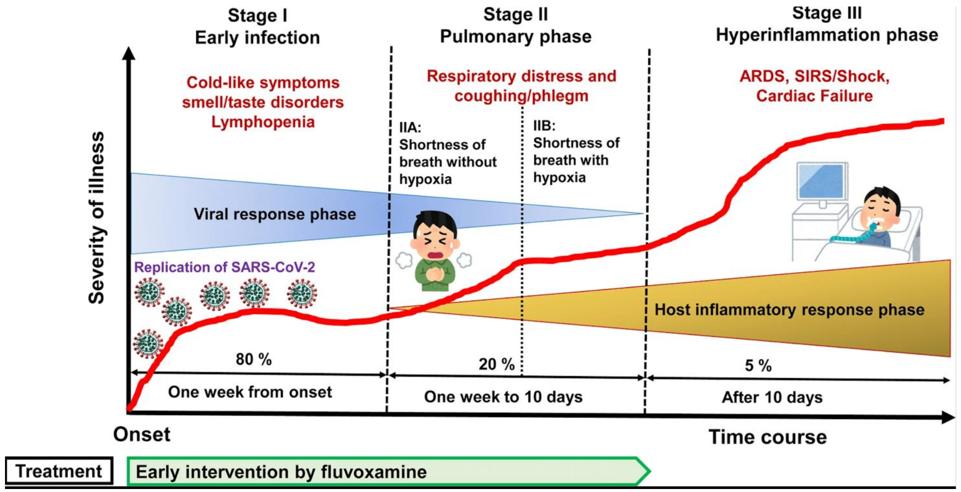




Frequent item set

Frequent sequences

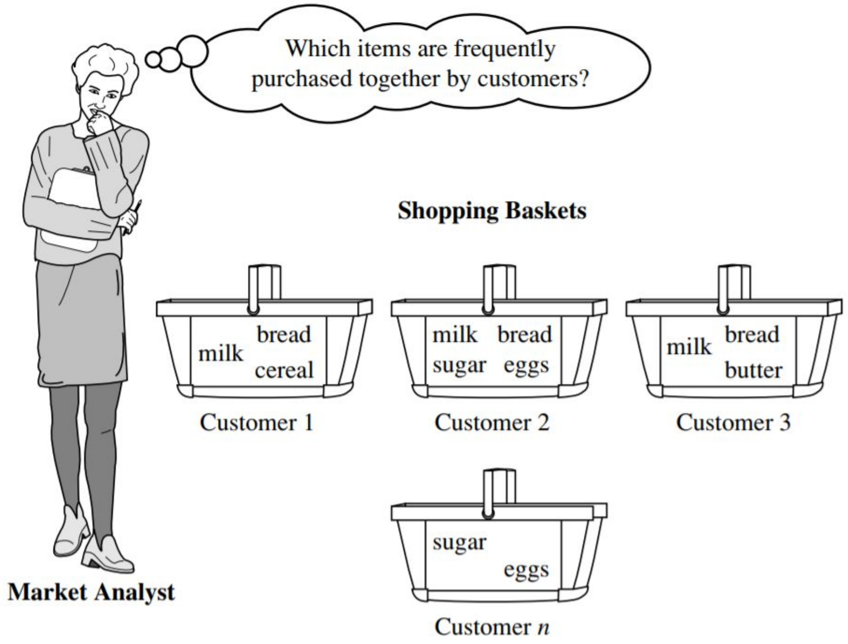
Frequent structures



What Is Pattern Discovery?

- Pattern discovery: Uncovering patterns from massive data sets
- It can answer questions such as:
  - What products were often purchased together?
  - What are the subsequent purchases after buying an iPad?

Pattern discovery



- **Market analyst:** to improve marketing strategies, store layouts, or cross-selling.
- **Shopping basket:** Displayed the items that are purchased
- **Purpose:** The goal is to find out which product are bought together so business can make informed decision.

Pattern Discovery: Why Is It Important?

- Foundation for many essential data mining tasks
  - Association, correlation, and causality analysis (e.g bread & butter)
  - Mining **sequential**, structural (e.g., sub-graph) patterns (e.g, DNA seq.)
  - **Classification**: Discriminative pattern-based analysis (e.g, email filtering)
  - **Cluster** analysis: Pattern-based subspace clustering (customer segmentation)
- Broad applications
  - Market basket analysis, cross-marketing, catalog design, sale campaign analysis, Web log analysis, biological sequence analysis
  - Many types of data: spatiotemporal, multimedia, time-series, and stream data

Pattern discovery practical examples

Some practical examples of Pattern discovery are:

- Shopping habits
- Music and movie recommendations
- Weather forecasting
- Health and fitness
- Social media

Basic Concepts: Transactional Database

- Transactional Database (TDB)
  - Each transaction is associated with an identifier, called a Transaction ID (TID).
  - May also have counts associated (quantity sold) with each item sold

Tid	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

- We can use association rule mining to identify frequent item sets.
- Real-world application: Placing frequently bought item together.

Basic Concepts: k-Itemsets and Their Supports

- Itemset: A set of ***one or more items***

$$I = \{I_1, I_2, \cdots, I_m\}$$

- k-itemset: An itemset containing ***k items***:

$$X = \{x_1, \ldots, x_k\}$$

  - Ex. {Beer, Nuts, Diaper} is a 3-itemset

- Absolute support (count)
  - sup{X} = occurrences of an itemset X
    - Ex. sup{Beer} = 3
    - Ex. sup{Diaper} = 4
    - Ex. sup{Beer, Diaper} = 3
    - Ex. sup{Beer, Eggs} = 1

Tid	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

- Relative support
  - *s*{*X*} = The fraction of transactions that contains X (i.e., the probability that a transaction contains X)
  - Ex. s{Beer} = 3/5 = 60%
  - Ex. s{Diaper} = 4/5 = 80%
  - Ex. s{Beer, Eggs} = 1/5 = 20%

Basic Concepts: Frequent Itemsets (Patterns)

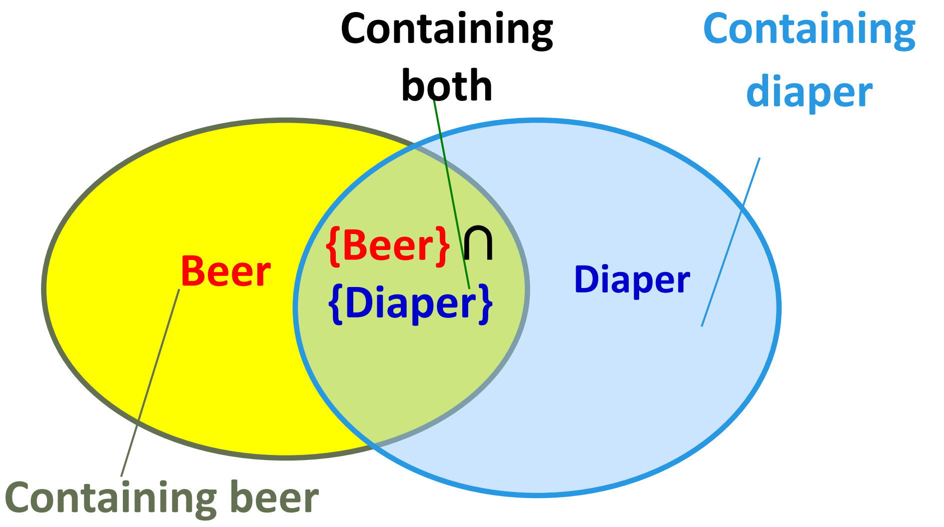
- An itemset (or a pattern) X is *frequent* if the support of X is no less than a *minsup* threshold σ
- Let σ = *50%* (σ: *minsup* threshold) for the given 5-transaction dataset
  - All the frequent 1-itemsets:
    - Beer: 3/5 (60%); Nuts: 3/5 (60%); Diaper: 4/5 (80%); Eggs: 3/5 (60%)
  - All the frequent 2-itemsets:
    - {Beer, Diaper}: 3/5 (60%)
  - All the frequent 3-itemsets?
    - None (no 3-itemset meets threshold)

Tid	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

- Why do these itemsets (shown on the left) form the complete set of frequent *k*-itemsets (patterns) for any *k*?
- **Observation**: We may need an efficient method to mine a complete set of frequent patterns

From Frequent Itemsets to Association Rules

- Compared with itemsets, association rules can be more telling
  - Ex. *Diaper* $\longrightarrow$ *Beer*
    - *Buying diapers may likely lead to buying beers*
    - *This rule helps make predictions or provide insights for decision making, such as marketing strategies.*
    - *The overlap shows the support for the itemset {{Beer} ∪ {Diaper}}, which is key to forming association rules like Diaper ⟹ Beer*



{Beer} ∪ {Diaper} = {Beer, Diaper}

Association Rule

- An association rule is a statement of the form X⟶Y, which suggests that when itemset X is present in a transaction, itemset Y is likely to also be present.
- For example, Diaper ⟶ Beer means that if someone buys a diaper, they are likely to also buy beer.
- **Support (s):** Measures how often both X and Y appear together in transactions.
- **Confidence (c):** Measures how often Y appears in transactions that contain X.

Association Rules

- How do we compute the strength of an association rule $X \rightarrow Y$ (Both X and Y are itemsets)?
- We first compute the following two metrics, s and c.
  - Support of X ∪ Y
    - Ex. s{Diaper, Beer} = 3/5 = 0.6 (i.e., 60%)
  - Confidence of $X \rightarrow Y$
    - The *conditional probability* that a transaction containing X also contains *Y:*

      c = sup(X, Y) / sup(X)
    - Ex. c = sup{Diaper, Beer}/sup{Diaper} = ¾ = 0.75

Tid	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

- This means that 75% of the transactions where Diaper is purchased, beer is also purchased.
- Real World example: If the support for Diaper → Beer is high, a store could create bundle promotions.

Mining Frequent Itemsets and Association Rules

- **Association rule mining**
  - Given two thresholds: *minsup, minconf*
  - Find all of the rules, $X \rightarrow Y$ (s, c) such that s ≥ *minsup* and *c* ≥ *minconf*
- **Setting up the threshold**
- Let *minsup = 50%*
  - **Freq. 1-itemsets:** Beer: 3, Nuts: 3, Diaper: 4, Eggs: 3
  - **Freq. 2-itemsets:** {Beer, Diaper}: 3
- *Let minconf = 50% (Rules satisfy condition?)*
  - *Beer* → *Diaper* (60%, 100%)
  - *Diaper* → *Beer* (60%, 75%)
- Marketing campaign targeting customer with babies.

Tid	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

- **Observations:**
  - Mining association rules and mining frequent patterns are very close problems
  - Scalable methods are needed for mining large datasets
  - Efficient algorithms like Apriori or FP-growth are necessary.

Pattern Mining: Basic Concepts and Methods

- Basic Concepts

- Frequent Itemset Mining Methods

- Which Patterns Are Interesting?—Pattern Evaluation Methods

- Summary

Efficient Pattern Mining Methods

- The Downward Closure Property of Frequent Patterns
  - The Apriori Algorithm
  - Extensions or Improvements of Apriori
- Mining Frequent Patterns by Exploring Vertical Data Format
- FPGrowth: A Frequent Pattern-Growth Approach
- Mining Closed Patterns

The Downward Closure Property of Frequent Patterns

- The downward closure property (also called the Apriori property) states that if an itemset is frequent, then all of its subsets must also be frequent.
- Observation: From TDB_1: T_1: $\{a_1, \ldots, a_{50}\}$; T_2: $\{a_1, \ldots, a_{100}\}$
  - We get a frequent itemset: $\{a_1, \ldots, a_{50}\}$
  - Also, its subsets are all frequent: $\{a_1\}, \{a_2\}, \ldots, \{a_{50}\}, \{a_1, a_2\}, \ldots, \{a_1, \ldots, a_{49}\}, \ldots$
  - There are some hidden relationships among frequent patterns!
- The downward closure (also called "Apriori") property of frequent patterns
  - If **{beer, diaper, nuts}** is frequent, so is **{beer, diaper}**
  - Every transaction containing {beer, diaper, nuts} also contains {beer, diaper}
  - Apriori: Any subset of a frequent itemset must be frequent
- Efficient mining methodology
  - If any subset of an itemset S is infrequent, then there is no chance for S to be frequent—why do we even have to consider S!?
  - This **property is critical** because it allows us to prune itemsets that have infrequent subsets, saving time and computation.

Apriori Pruning and Scalable Mining Methods

- **Apriori pruning principle:** If there is any itemset which is infrequent, its superset should **not even be generated/tested**!
- For example, If we find that the itemset {bread, butter} is infrequent, we don't need to check whether {bread, butter, milk} or any other larger sets containing {bread, butter} are frequent because they can't be.
- **Scalable mining Methods: Three major approaches**
  - **Level-wise, join-based approach: Apriori**
  - **Vertical data format approach: Eclat**
  - **Frequent pattern projection and growth: FPgrowth**

Scalable Mining Methods

- **Level-wise, join based approach: Apriori**
  - This is the original Apriori algorithm, proposed by Agrawal & Srikant in 1994.
  - It works in **levels**, first *finding frequent 1-itemsets*, then *frequent 2-itemsets* by joining 1-itemsets, and so on.
- **Vertical Data Format Approach: Eclat**
  - Eclat uses a vertical format where each item is stored with the list of transaction IDs (TIDs)
  - Instead of scanning whole database repeatedly, it intersects these TID to find frequent patterns.
- **Frequent Pattern Projection and Growth: FP-Growth**
  - It is depth-first search approach that avoids generating candidate sets like Apriori.
  - It uses a data structure called an FP-tree to compactly store the data and grow patterns recursively.

Apriori: A Candidate Generation & Test Approach

- Outline of Apriori (level-wise, candidate generation and test)
  - Scan DB once to get frequent 1-itemset
    - Bread: 4 times
    - Butter: 4 times
    - Milk: 4 times
  - Set a minimum threshold of 3.
  - Frequent itemsets: {Bread}, {Butter}, {Milk}

Tid	Items bought
1	Bread, butter, milk
2	Bread, milk
3	Bread, Butter
4	Milk, Butter
5	Bread, milk, Butter

- Repeat
  - Generate length-(k+1) candidate itemsets from length-k frequent itemsets.
  - Possible 2-itemsets: {Bread, Butter}, {Bread, Milk}, {Butter, Milk}
  - Test the candidates against DB to find frequent (k+1)-itemsets
  - {Bread, Butter}: 3 times, {Bread, Milk} = 3 times, {Butter, Milk} = 3 times
  - Since, 2-itemsets have support ≧ 3, they are considered as **frequent 2-itemsets**.
  - Set k := k +1 (for example, now K =3)
- Repeat until no frequent or candidate set can be generated
- Return all the frequent itemsets derived
  - Frequent 1-itemsets: {Bread}, {Butter}, {Milk}
  - Frequent 2-itemsets: {Bread, Butter}, {Bread, Milk}, {Butter, Milk}

Apriori Algorithm

Step 1: Frequent 1-itemsets (items that appear in at least 3 out of 5 transactions):

- Bread: 4/5 transactions = 80% (frequent)
- Butter: 4/5 transactions = 80% (frequent)
- Milk: 4/5 transactions = 80% (frequent)

Tid	Items bought
1	Bread, butter, milk
2	Bread, milk
3	Bread, Butter
4	Milk, Butter
5	Bread, milk, Butter

Step 2: Frequent 2-itemsets (combinations of 2 items that appear in at least 3 out of 5 transactions):

- {Bread, Butter}: 3/5 transactions = 60% (frequent)
- {Bread, Milk}: 3/5 transactions = 60% (frequent)
- {Milk, Butter}: 3/5 transactions = 60% (frequent)

Step 3: Frequent 3-itemsets (combinations of 3 items that appear in at least 3 out of 5 transactions):

- {Bread, Butter, Milk}: 2/5 transactions = 40% (not frequent)

Step 4: Association Rules:

- From the frequent itemsets, we can generate rules like:
    - If someone buys **Bread**, they are likely to buy **Butter** (with 60% confidence).
    - If someone buys **Milk**, they are likely to buy **Butter** (with 60% confidence).

The Apriori Algorithm (Pseudo-Code)

C_k: Candidate itemset of size k

F_k : Frequent itemset of size k

1. Initialization

K := 1; // (looking for frequent 1-itemset)

F_k := {frequent items}; // find frequent 1-itemset

2. Main Loop

While (F_k != ∅) **do {** // when F_k is non-empty

C_{k+1} := candidates generated from F_k; // **candidate generation**

Derive F_{k+1} by counting candidates in C_{k+1} with respect to *TDB* at minsup; // **Pruning and count**

k := k + 1 // **increment K**

}

return $\cup_k F_k$ // return F_k generated at each level

Apriori algorithm with example

1. **Start with 1-itemsets** (Bread, butter, milk)

2. **Move to 2-itemsets** (e.g, {Bread, Milk}, {Milk, Butter})
3. **Repeat for larger itemsets** (after finding 2-itemset, it will combine to find 3 itemset)
4. **Return to all frequent itemsets** (1-itemsets, 2-itemsets, etc.),

- This results will be used for further analysis (finding which items are often bought together)

Tid	Items bought
1	Bread, butter, milk
2	Bread, milk
3	Bread, Butter
4	Milk, Butter
5	Bread, milk, Butter

The Apriori Algorithm—An Example

minsup = 2

Database TDB

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

1^{st} scan

C_1

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

F_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}

2^{nd} scan

C_2

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

F_2

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

C_3

Itemset
{B, C, E}

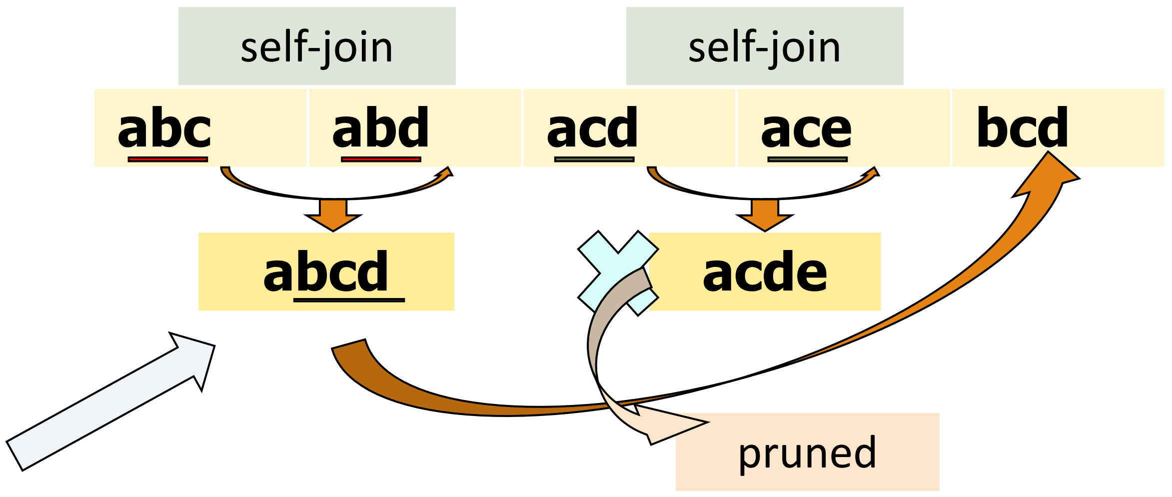
3^{rd} scan

F_3

Itemset	sup
{B, C, E}	2

Apriori: Implementation Tricks

- How to generate candidates?
  - Step 1: self-joining F_k
  - Step 2: pruning
- Example of candidate-generation
  - F_3 = {*abc, abd, acd, ace, bcd*}
  - Self-joining: F_3*F_3
    - *abcd* from *abc* and *abd*
    - *acde* from *acd* and *ace*
  - Pruning:
    - *acde* is removed because *ade* is not in F_3
  - C_4 = {*abcd*}



Exploring Vertical Data Format: ECLAT

- ECLAT (Equivalence Class Transformation): A **depth-first search** algorithm using set intersection [Zaki et al. @KDD'97]
- **Vertical format**
- **Properties of Tid-Lists**
  - t(X) = t(Y): X and Y always happen together (e.g., t(ac} = t(d})
  - t(X) ⊂ t(Y): transaction having X always has Y (e.g., t(ac) ⊂ t(ce))
- **Frequent patterns by vertical intersections**
- **Using diffset to accelerate mining**
  - Only keep track of differences of tids
  - t(e) = {T_{10}, T_{20}, T_{30}}, t(ce) = {T_{10}, T_{30}} → Diffset (ce, e) = {T_{20}}
  - This method *reduces the amount of data stored* and *speeds up* the mining process

A transaction DB in Horizontal Data Format

Tid	Itemset
10	a, c, d, e
20	a, b, e
30	b, c, e

The transaction DB in Vertical Data Format

Item	TidList
a	10, 20
b	20, 30
c	10, 30
d	10
e	10, 20, 30

t(e) = {T_{10}, T_{20}, T_{30}};
t(a) = {T_{10}, T_{20}};
t(ae) = {T_{10}, T_{20}}

ECLAT algorithm use case

- Transactional DB
- Convert it into vertical format
- Step 1: Intersect TidLists to generate frequent item
  a. Bread: {1, 2, 3, 5} → frequent
  b. Butter: {1, 3, 4, 5} → frequent
  c. Milk: {1, 2, 4, 5} → frequent
- Step 2: Generate frequent 2-itemsets
  a. Bread ∩ Butter = {1, 3, 5} → appears in 3 transactions → frequent
  b. Bread ∩ Milk = {1, 2, 5} → appears in 3 transactions → frequent
  c. Butter ∩ Milk = {1, 4, 5} → appears in 3 transactions → frequent
- Step 3: Generate frequent 3-itemsets:
  a. **Bread** ∩ **Butter** ∩ **Milk** = {1, 5} → appears in 2 transactions → frequent
- Outputs
- Frequent 1-itemsets: {Bread}, {Butter}, {Milk}
- Frequent 2-itemsets: {Bread, Butter}, {Bread, Milk}, {Butter, Milk}
- Frequent 3-itemsets: {Bread, Butter, Milk}
- useful for marketing and promotion, product placement, cross-selling opportunities.
  a.

TID	Item bought
1	{Bread, Butter, Milk}
2	{Bread, Milk}
3	{Bread, Butter}
4	{Milk, Butter}
5	{Bread, Milk, Butter}

Item	TID
Bread	{1, 2, 3, 5}
Butter	{1, 3, 4, 5}
Milk	{1, 2, 4, 5}

FP-Growth algorithm

- **FP-Growth (Frequent Pattern Growth)** is an algorithm designed to find **frequent itemsets** without candidate generation, making it more efficient than algorithms like Apriori.
- **Retail & Ecommerce**: Market basket analysis, recommendation systems.
- **Healthcare**: Medical diagnosis, disease prediction.
- **Finance**: Fraud detection, risk management.
- **Web & Social Media**: Website navigation optimization, content recommendation, social network analysis.
- **Bioinformatics**: Gene expression analysis, DNA sequencing.
- **Manufacturing**: Quality control, defect analysis.

Example: From Transactional DB to Ordered Frequent Itemlist

- Example: A Sample Transactional Database
  - **Let min_support = 3**

TID	Items in the Transaction
100	{*f, a, c, d, g, i, m, p*}
200	{*a, b, c, f, l, m, o*}
300	{*b, f, h, j, o, w*}
400	{*b, c, k, s, p*}
500	{*a, f, c, e, l, p, m, n*}

- Scan DB once, find single item frequent pattern: f:4, a:3, c:4, b:3, m:3, p:3
- Sort frequent items in frequency descending order, f-list F-list = f-c-a-b-m-p
- Scan DB again, use the ordered frequent itemlist for each transaction to construct an FP-tree

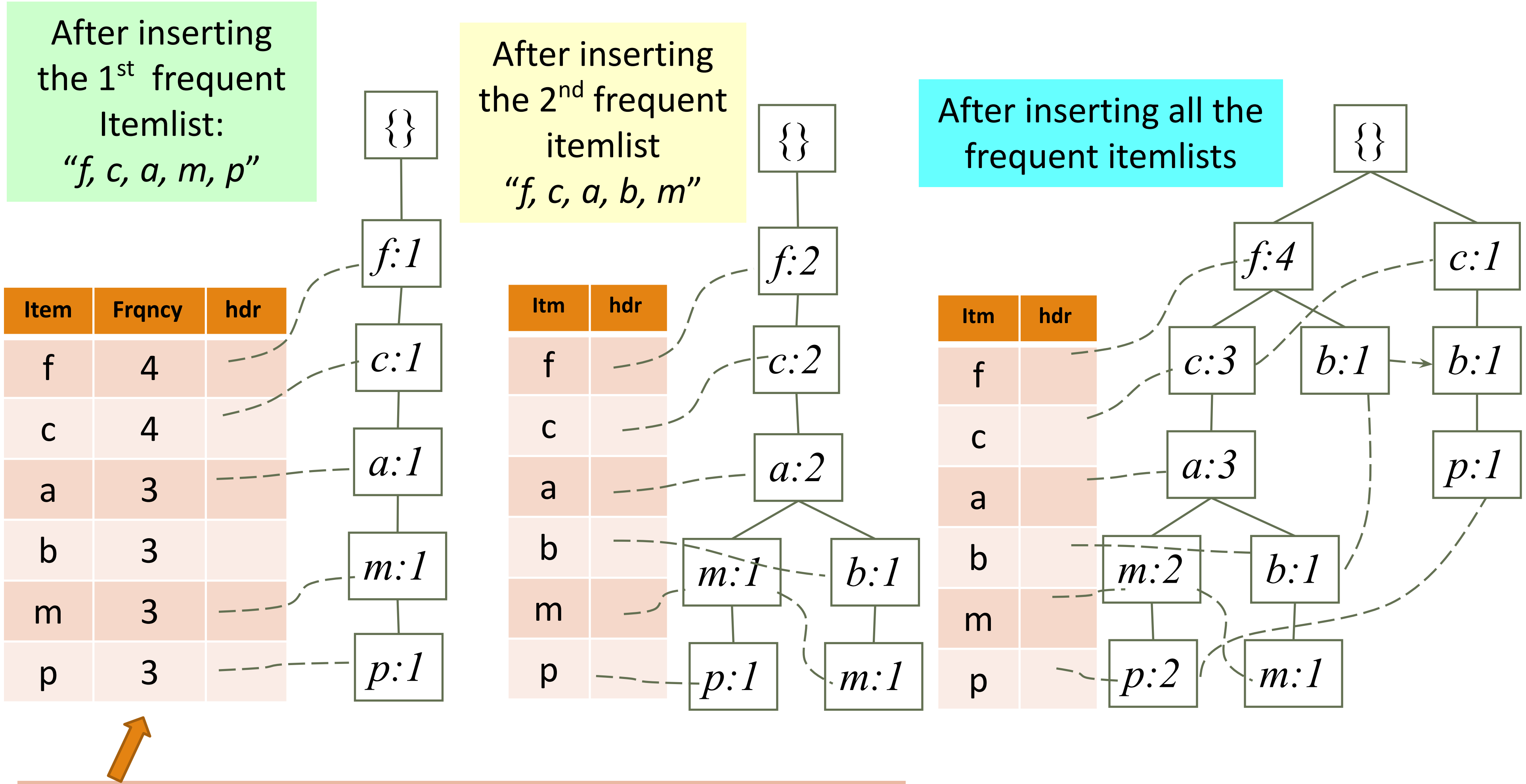
TID	Items in the Transaction	Ordered, frequent itemlist
100	{*f, a, c, d, g, i, m, p*}	*f, c, a, m, p*
200	{*a, b, c, f, l, m, o*}	*f, c, a, b, m*
300	{*b, f, h, j, o, w*}	*f, b*
400	{*b, c, k, s, p*}	*c, b, p*
500	{*a, f, c, e, l, p, m, n*}	*f, c, a, m, p*

Example: Construct FP-tree from Transaction DB

TID	Ordered, frequent itemlist
100	*f, c, a, m, p*
200	*f, c, a, b, m*
300	*f, b*
400	*c, b, p*
500	*f, c, a, m, p*

FP-Tree Construction:

For each transaction, insert the ordered frequent itemlist into an FP-tree, with shared sub-branches merged, counts accumulated



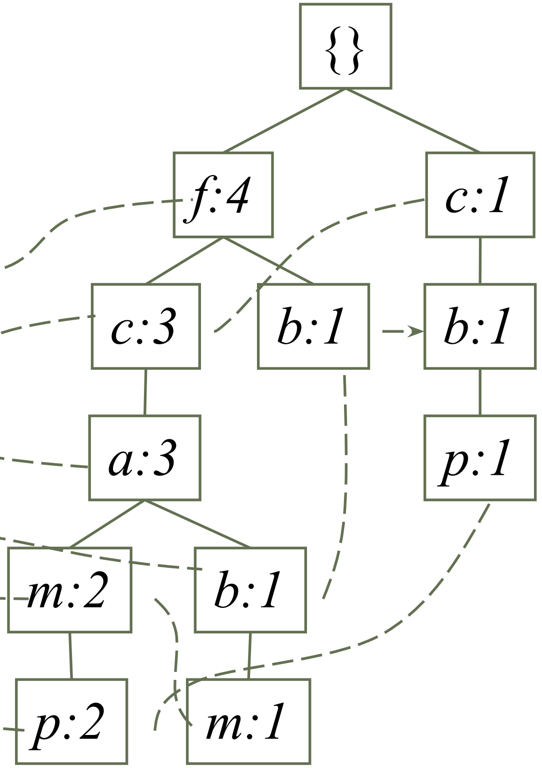
Header Table: keeps track of all the frequent items and links them to the nodes in the FP-tree.

Mining FP-Tree: Divide and Conquer Based on Patterns and Data

- Pattern mining can be partitioned according to current patterns
  - Patterns containing *p*: *p*'s conditional database: *fcam:2, cb:1*
    - *p*'s conditional database (i.e., the database under the condition that *p* exists):
      - *transformed prefix paths* of item *p*
  - Patterns having m but no p: m's conditional database: *fca:2, fcab:1*
  - …… ……

min_support = 3

Item	Frequency	Header
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	



Conditional database of each pattern

Item	***Conditional database***
c	***f:3***
a	***fc:3***
b	***fca:1, f:1, c:1***
m	***fca:2, fcab:1***
p	***fcam:2, cb:1***

Mine Each Conditional Database Recursively

min_support = 3

1. Conditional Data Bases

item	*cond. data base*
c	*f:3*
a	*fc:3*
b	*fca:1, f:1, c:1*
m	*fca:2, fcab:1*
p	*fcam:2, cb:1*

2. Recursive mining process
Then, mining m's FP-tree: fca:3

- For each conditional database
  - Mine single-item patterns
  - Construct its FP-tree & mine it

p's conditional DB: ***fcam:2, cb:1*** → ***c: 3***

m's conditional DB: ***fca:2, fcab:1*** → ***fca: 3***

b's conditional DB: ***fca:1, f:1, c:1*** → ***ϕ***

Actually, for single branch FP-tree, all the frequent patterns can be generated in one shot

m: 3
fm: 3, cm: 3, am: 3
fcm: 3, fam:3, cam: 3
fcam: 3

Pattern Mining: Basic Concepts and Methods

- **Basic Concepts**
- **Frequent Itemset Mining Methods**
- **Which Patterns Are Interesting?—Pattern Evaluation Methods**
- **Summary**

How to Judge if a Rule/Pattern Is Interesting?

- Pattern-mining will generate a large set of patterns/rules
  - Not all the generated patterns/rules are interesting
- Interestingness measures: Objective vs. subjective
  - Objective interestingness measures
    - Support, confidence, correlation, …
  - Subjective interestingness measures:
        - Different users may judge interestingness differently
    - Let a user specify
      - **Query-based:** Relevant to a user's particular request
    - Judge against **one's knowledge-base**
      - unexpected, freshness, timeliness

Limitation of the Support-Confidence Framework

- Are *s* and *c* interesting in association rules: "A ⇒ B" [*s*, *c*]? Be careful!
- Example: Suppose one school may have the following statistics on # of students who may play basketball and/or eat cereal:

	play-basketball	not play-basketball	sum (row)
eat-cereal	400	350	750
not eat-cereal	200	50	250
sum(col.)	600	400	1000

2-way contingency table

- Association rule mining may generate the following:
  - ***play-basketball* ⇒ *eat-cereal* [40%, 66.7%]** (higher s & c)
- But this strong association rule is misleading: The overall % of students eating cereal is 75% > 66.7%, a more telling rule:
  - **¬ *play-basketball* ⇒ *eat-cereal* [35%, 87.5%]** (high s & c)

Lift and X^2

- Lift and x^2 *are used to evaluate* ***how strongly two events are correlated*** *in the dataset.*

Interestingness Measure: Lift

- Measure of dependent/correlated events: **lift**

$$lift(B,C) = \frac{c(B \rightarrow C)}{s(C)} = \frac{s(B \cup C)}{s(B) \times s(C)}$$

Lift is more telling than s & c

	B	¬B	Σ_{row}
C	400	350	750
¬C	200	50	250
$\Sigma_{col.}$	600	400	1000

- Lift(B, C) may tell how B and C are correlated
  - Lift(B, C) = 1: B and C are independent
  - > 1: positively correlated
  - < 1: negatively correlated
- For our example,

$$lift(B,C) = \frac{400/1000}{600/1000 \times 750/1000} = 0.89$$

$$lift(B,\neg C) = \frac{200/1000}{600/1000 \times 250/1000} = 1.33$$

- Thus, B and C are negatively correlated since lift(B, C) < 1;
  - B and ¬C are positively correlated since lift(B, ¬C) > 1

Interestingness Measure: χ2

- Another measure to test correlated events: χ^2

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

	B	¬B	Σ_{row}
C	400 (450)	350 (300)	750
¬C	200 (150)	50 (100)	250
Σ_{col}	600	400	1000

Expected value

Observed value

- For the table on the right,

$$\chi^2 = \frac{(400-450)^2}{450} + \frac{(350-300)^2}{300} + \frac{(200-150)^2}{150} + \frac{(50-100)^2}{100} = 55.56$$

Some general rules:

- **χ^2 = 0 : Independent**
- **χ^2 > 0 : Correlated either positive or negative.**

- Lookup χ^2 distribution table → B, C are correlated
- χ^2-test shows B and C are negatively correlated since the expected value is 450 but the observed is only 400
- Thus, χ^2 is also more telling than the support-confidence framework

Lift and χ2 : Are They Always Good Measures?

- **Null transactions:** Transactions that contain neither B nor C
- Let's examine the new dataset D
  - BC (100) is much rarer than B¬C (1000) and ¬BC (1000), but there are many ¬B¬C (100000)
  - Unlikely B & C will happen together!

	B	¬B	Σ_{row}
C	100	1000	1100
¬C	1000	100000	101000
$\Sigma_{col.}$	1100	101000	102100

null transactions

Let's take new dataset,

- B and C occur together 100 times.
- B and C appear together **only 100 times**.
- B occurs without C **1000 times**.
- Neither B nor C appears in **100,000 transactions**.
- From here it is obvious that B and C happening together is rare.

Contingency table with expected values added

	B	¬B	Σ_{row}
C	100 (11.85)	1000	1100
¬C	1000 (988.15)	100000	101000
$\Sigma_{col.}$	1100	101000	102100

- however, Lift(B, C) = 8.44 >> 1 (Lift shows B and C are strongly positively correlated!)
- This value of Lift suggests a strong positive correlation between B and C and it is misleading.

Lift and χ2

- χ^2 = 670: Observed(BC) >> expected value (11.85)
- *Too many null transactions may "spoil the soup"!*

	B	¬B	Σ_{row}
C	100 (11.85)	1000	1100
¬C	1000 (988.15)	100000	101000
$\Sigma_{col.}$	1100	101000	102100

Conclusion: ***Although Lift and χ^2 suggest a strong positive correlation, this is mainly because of the overwhelming number of transactions that contain neither B nor C, which skews the results.***

Quiz

1. **What is pattern discovery?**
- A) Cleaning a dataset
- B) Discovering frequent patterns
- C) Sorting data
- D) Creating new data

2. **What is an itemset?**
- A) A group of unrelated items
- B) A collection of one or more items
- C) A pattern found in a data stream
- D) A subset of random transactions

3. **What does the Apriori property state?**
- A) Only large datasets can be used
- B) If an itemset is frequent, so are all of its subsets
- C) Subsets do not affect frequent itemsets
- D) Confidence is higher than support

4. **Which is an example of a 3-itemset?**
- A) {Bread, Butter, Eggs}
- B) {Apple}
- C) {Milk}
- D) {Diaper, Beer}

5. **How is support calculated?**
- A) Number of occurrences of an itemset
- B) The correlation between items
- C) Transaction IDs
- D) Sum of itemset frequencies

Quiz

6. Which of the following is a frequent itemset mining algorithm?

- A) K-Means
- B) Apriori
- C) SVM
- D) Decision Tree

7. Which property allows Apriori to prune itemsets?

- A) Minimum support threshold
- B) Confidence
- C) Downward closure
- D) Random sampling

8. What does FP-Growth do?

- A) Generate candidate itemsets
- B) Construct an FP-tree for frequent pattern mining
- C) Clean noisy data
- D) Create decision trees

9. What is the primary use of the Eclat algorithm?

- A) To find frequent itemsets
- B) To create association rules
- C) To mine sequential patterns
- D) To prune infrequent itemsets

Quiz

10. Which of the following is a limitation of the support-confidence framework?

- A) It is time-efficient
- B) It handles large datasets effectively
- C) It cannot detect correlations between items
- D) It can lead to misleading rules

11. What is the purpose of lift in pattern mining?

- A) To measure support
- B) To measure correlation between two items
- C) To increase the confidence of an association rule
- D) To reduce dataset size

12. What happens when the lift of two items is greater than 1?

- A) The items are independent
- B) The items are negatively correlated
- C) The items are positively correlated
- D) The items cannot occur together

13. What is a frequent pattern tree (FP-tree)?

- A) A binary tree
- B) A compact structure used for frequent itemset mining
- C) A decision tree
- D) A clustering algorithm

Quiz

14. What is the goal of association rule mining?

- A) To discover patterns in unrelated items
- B) To identify the strongest item correlations
- C) To clean large datasets
- D) To create sequential patterns

15. What is the minimum confidence threshold used for?

- A) To prune infrequent itemsets
- B) To determine the smallest size for an association rule
- C) To measure the strength of association rules
- D) To reduce dataset size

16. What does downward closure allow in Apriori?

- A) Generate new candidate sets
- B) Prune infrequent itemsets and their supersets
- C) Create association rules
- D) Increase the confidence of an itemset

17. Which of the following is an example of an association rule?

- A) {Bread} → {Butter}
- B) {Eggs} ∩ {Milk}
- C) {Diaper} + {Beer}
- D) {Apples} - {Bananas}

Quiz

18. What is the relationship between support and confidence in pattern mining?

- A) Support is always larger than confidence
- B) Confidence depends on the support of the itemset
- C) Confidence is independent of support
- D) Support is equal to confidence

19. What is an objective interestingness measure in pattern mining?

- A) User-defined constraints
- B) Support
- C) Random sampling
- D) Interestingness defined by domain knowledge

20. What is the χ^2 statistic used for in pattern mining?

- A) Measure dataset size
- B) Test correlations between items
- C) Prune infrequent itemsets
- D) Create association rules

21. How does FP-growth differ from Apriori?

- A) It avoids generating candidate sets
- B) It uses downward closure to prune
- C) It generates more itemsets
- D) It requires more memory

Quiz

22. What is the main advantage of vertical data formats in pattern mining?

- A) It simplifies dataset cleaning
- B) It reduces the number of transactions to scan
- C) It increases the size of the database
- D) It improves confidence values

23. Which is a subjective interestingness measure in pattern mining?

- A) Support
- B) Confidence
- C) User-defined relevance
- D) Correlation

24. What does the Eclat algorithm primarily use to find frequent patterns?

- A) Candidate generation
- B) Set intersections
- C) Pruning
- D) Association rules

25. Which patterns are considered closed patterns in pattern mining?

- A) Patterns that are maximal in size
- B) Patterns that are frequent and have no superset with the same support
- C) Patterns that overlap with other itemsets
- D) Patterns that are pruned after each iteration

Quiz

26. Which metric in association rules measures how often Y appears in transactions that contain X?

- A) Support
- B) Confidence
- C) Lift
- D) χ^2

27. What is the purpose of a conditional FP-tree in FP-growth?

- A) To generate association rules
- B) To store transaction IDs
- C) To mine patterns under specific conditions
- D) To remove null transactions

28. What are the conditions for a frequent pattern to be valid in Apriori?

- A) The pattern must contain more than three items
- B) The pattern must have high confidence
- C) The pattern must have frequent subsets
- D) The pattern must be maximal

29. What type of database format does the Eclat algorithm use?

- A) Vertical format
- B) Horizontal format
- C) Relational format
- D) Graph format

Quiz

30. Which of the following describes the purpose of support in association rule mining?

- A) To measure how frequent an itemset occurs in a database
- B) To identify the strongest rules
- C) To prune large datasets

Answers

1. **What is papattern discovery?**
    - **Answer**: B) Discovering frequent patterns
2. **What is an itemset?**
    - **Answer**: B) A collection of one or more items
3. **What does the Apriori property state?**
    - **Answer**: B) If an itemset is frequent, so are all of its subsets
4. **Which is an example of a 3-itemset?**
    - **Answer**: A) {Bread, Butter, Eggs}
5. **How is support calculated?**
    - **Answer**: A) Number of occurrences of an itemset
6. **Which of the following is a frequent itemset mining algorithm?**
    - **Answer**: B) Apriori
7. **Which property allows Apriori to prune itemsets?**
    - **Answer**: C) Downward closure
8. **What does FP-Growth do?**
    - **Answer**: B) Construct an FP-tree for frequent pattern mining
9. **What is the primary use of the Eclat algorithm?**
    - **Answer**: A) To find frequent itemsets
10. **Which of the following is a limitation of the support-confidence framework?**
    - **Answer**: C) It cannot detect correlations between item

Answers

11. What is the purpose of lift in pattern mining?

- **Answer**: B) To measure correlation between two items

12. What happens when the lift of two items is greater than 1?

- **Answer**: C) The items are positively correlated

13. What is a frequent pattern tree (FP-tree)?

- **Answer**: B) A compact structure used for frequent itemset mining

14. What is the goal of association rule mining?

- **Answer**: B) To identify the strongest item correlations

15. What is the minimum confidence threshold used for?

- **Answer**: C) To measure the strength of association rules

16. What does downward closure allow in Apriori?

- **Answer**: B) Prune infrequent itemsets and their supersets

17. Which of the following is an example of an association rule?

- **Answer**: A) {Bread} → {Butter}

18. What is the relationship between support and confidence in pattern mining?

- **Answer**: B) Confidence depends on the support of the itemset

19. What is an objective interestingness measure in pattern mining?

- **Answer**: B) Support

Answers

20. What is the χ^2 statistic used for in pattern mining?

- **Answer**: B) Test correlations between items

21. How does FP-growth differ from Apriori?

- **Answer**: A) It avoids generating candidate sets

22. What is the main advantage of vertical data formats in pattern mining?

- **Answer**: B) It reduces the number of transactions to scan

23. Which is a subjective interestingness measure in pattern mining?

- **Answer**: C) User-defined relevance

24. What does the Eclat algorithm primarily use to find frequent patterns?

- **Answer**: B) Set intersections

25. Which patterns are considered closed patterns in pattern mining?

- **Answer**: B) Patterns that are frequent and have no superset with the same support

26. Which metric in association rules measures how often Y appears in transactions that contain X?

- **Answer**: B) Confidence

27. What is the purpose of a conditional FP-tree in FP-growth?

- **Answer**: C) To mine patterns under specific conditions

Answers

28. What are the conditions for a frequent pattern to be valid in Apriori?

- **Answer**: C) The pattern must have frequent subsets

29. What type of database format does the Eclat algorithm use?

- **Answer**: A) Vertical format

30. Which of the following describes the purpose of support in association rule mining?

- **Answer**: A) To measure how frequent an itemset occurs in a database

Quiz

What is pattern mining primarily concerned with?

- A) Reducing dataset size
- B) Identifying patterns in large datasets
- C) Clustering unrelated data
- D) Labeling datasets

What does a high confidence value in an association rule suggest?

- A) Strong likelihood of co-occurrence
- B) Weak association between items
- C) Random data occurrence
- D) Strong correlation with other rules

Which of the following does support represent in association rule mining?

- A) The total number of transactions
- B) The correlation between two items
- C) The frequency of an itemset appearing in transactions
- D) The overall data quality

What is a common application of pattern mining?

- A) Real-time video processing
- B) Market basket analysis
- C) Linear regression modeling
- D) Cloud storage

Quiz

What does the term 'minsup' refer to in Apriori?

- A) Minimum confidence level
- B) Maximum support threshold
- C) Minimum support threshold
- D) Number of itemsets in a transaction

What does the Apriori algorithm require for efficient itemset generation?

- A) Hierarchical clustering
- B) Support values only
- C) A minimum support threshold
- D) A classification model

Which of the following is the first step in the Apriori algorithm?

- A) Generate k-itemsets
- B) Create frequent 1-itemsets
- C) Create an association rule
- D) Calculate confidence

How does the FP-growth algorithm improve efficiency?

- A) By generating candidate sets faster
- B) By pruning irrelevant data
- C) By using an FP-tree to avoid candidate generation
- D) By storing frequent transactions only
-

Quiz

What is an advantage of the Eclat algorithm?

- A) Low memory requirement
- B) Fast set intersections in vertical format
- C) Reduced computational time for clustering
- D) High lift values for frequent itemsets

In Apriori, which of the following is true of frequent k-itemsets?

- A) They appear in all transactions
- B) They must meet the minsup threshold
- C) They depend only on lift
- D) They are pruned if confidence is low

Which of the following best describes a conditional FP-tree?

- A) An FP-tree constructed for all transactions
- B) A subset of an FP-tree for a specific item
- C) A horizontal list of itemsets
- D) A pruned tree for null transactions

What is the goal of pattern evaluation?

- A) To identify the largest itemset
- B) To test the interestingness of discovered patterns
- C) To calculate the sum of all supports
- D) To reduce dataset size

Quiz

Which interestingness measure considers user preferences?

- A) Support
- B) Confidence
- C) Subjective measures
- D) Frequency

What is the purpose of sequential pattern mining?

- A) To find clusters in a dataset
- B) To identify frequently occurring sequences over time
- C) To classify transactions
- D) To discover trends between unrelated items

How is the downward closure property useful in Apriori?

- A) It speeds up clustering
- B) It helps prune non-frequent itemsets
- C) It ensures maximum confidence
- D) It calculates minimum support

What are association rules most commonly used for?

- A) Classification
- B) Linear regression
- C) Understanding customer purchasing behavior
- D) Image segmentation

Quiz

What defines a closed pattern in pattern mining?

- A) No subsets are frequent
- B) It has no supersets with the same support
- C) It appears in all transactions
- D) It is pruned in the final stage

How does confidence differ from lift?

- A) Confidence measures support only, while lift measures correlation
- B) Confidence is about rule strength, lift is about correlation strength
- C) Lift measures frequency, confidence measures size
- D) Confidence only applies to Apriori, not FP-Growth

Which algorithm stores itemsets as transaction lists for easy intersection?

- A) Apriori
- B) FP-Growth
- C) Eclat
- D) K-Means

What is an itemset that includes a sequence of purchases called?

- A) A frequent transaction
- B) A sequential pattern
- C) A closed itemset
- D) A frequent k-itemset

Quiz

What is meant by the 'lift' of an association rule?

- A) The frequency of the rule
- B) The strength of the rule’s correlation
- C) The rule’s support value
- D) The confidence threshold for the rule

What does support represent in the context of a transactional database?

- A) The total frequency of items bought
- B) The likelihood of a transaction containing an itemset
- C) The confidence of all association rules
- D) The total number of transactions

How is lift calculated in an association rule?

- A) Confidence / Support
- B) Confidence of X and Y
- C) Support(X∩Y) / (Support(X) * Support(Y))
- D) Support(X) * Support(Y)

Which algorithm uses a depth-first search with vertical data format?

- A) Apriori
- B) Eclat
- C) FP-Growth
- D) K-Means

Quiz

What does the Eclat algorithm primarily depend on for efficiency?

- A) Candidate generation
- B) Set intersections
- C) Data clustering
- D) Neural networks

Which of the following describes a maximal frequent itemset?

- A) The most frequent itemset
- B) An itemset that has no frequent supersets
- C) An itemset that appears only in certain transactions
- D) The largest itemset found in clustering

Which of the following best describes association rules?

- A) Patterns that only use frequent 1-itemsets
- B) Rules that show the frequency of unrelated items
- C) Patterns that show a strong relationship between itemsets
- D) Rules used for data transformation

What is a frequent itemset with all its frequent subsets called?

- A) Closed itemset
- B) Maximal itemset
- C) Sequential pattern
- D) Frequent k-itemset

Quiz

How does FP-Growth differ from Apriori in terms of memory use?

- A) FP-Growth uses less memory
- B) FP-Growth requires frequent candidate generation
- C) Apriori has lower memory usage
- D) Both use the same memory

What is the purpose of the support-confidence framework?

- A) To reduce dataset size
- B) To identify and evaluate association rules
- C) To analyze clustering results
- D) To measure the quality of patterns found

Answers

What is pattern mining primarily concerned with?

- **Answer**: B) Identifying patterns in large datasets

What does a high confidence value in an association rule suggest?

- **Answer**: A) Strong likelihood of co-occurrence

Which of the following does support represent in association rule mining?

- **Answer**: C) The frequency of an itemset appearing in transactions

What is a common application of pattern mining?

- **Answer**: B) Market basket analysis

What does the term 'minsup' refer to in Apriori?

- **Answer**: C) Minimum support threshold

What does the Apriori algorithm require for efficient itemset generation?

- **Answer**: C) A minimum support threshold

Which of the following is the first step in the Apriori algorithm?

- **Answer**: B) Create frequent 1-itemsets

Answers

How does the FP-growth algorithm improve efficiency?

- **Answer**: C) By using an FP-tree to avoid candidate generation

What is an advantage of the Eclat algorithm?

- **Answer**: B) Fast set intersections in vertical format

In Apriori, which of the following is true of frequent k-itemsets?

- **Answer**: B) They must meet the minsup threshold

Which of the following best describes a conditional FP-tree?

- **Answer**: B) A subset of an FP-tree for a specific item

What is the goal of pattern evaluation?

- **Answer**: B) To test the interestingness of discovered patterns

Which interestingness measure considers user preferences?

- **Answer**: C) Subjective measures

What is the purpose of sequential pattern mining?

- **Answer**: B) To identify frequently occurring sequences over time

Answers

How is the downward closure property useful in Apriori?

- **Answer**: B) It helps prune non-frequent itemsets

What are association rules most commonly used for?

- **Answer**: C) Understanding customer purchasing behavior

What defines a closed pattern in pattern mining?

- **Answer**: B) It has no supersets with the same support

How does confidence differ from lift?

- **Answer**: B) Confidence is about rule strength, lift is about correlation strength

Which algorithm stores itemsets as transaction lists for easy intersection?

- **Answer**: C) Eclat

What is an itemset that includes a sequence of purchases called?

- **Answer**: B) A sequential pattern

What is meant by the 'lift' of an association rule?

- **Answer**: B) The strength of the rule’s correlation

Answers

What does support represent in the context of a transactional database?

- **Answer**: B) The likelihood of a transaction containing an itemset

How is lift calculated in an association rule?

- **Answer**: C) Support(X∩Y) / (Support(X) * Support(Y))

Which algorithm uses a depth-first search with vertical data format?

- **Answer**: B) Eclat

What does the Eclat algorithm primarily depend on for efficiency?

- **Answer**: B) Set intersections

Which of the following describes a maximal frequent itemset?

- **Answer**: B) An itemset that has no frequent supersets

Which of the following best describes association rules?

- **Answer**: C) Patterns that show a strong relationship between itemsets

What is a frequent itemset with all its frequent subsets called?

- **Answer**: A) Closed itemset

Answers

How does FP-Growth differ from Apriori in terms of memory use?

- **Answer**: A) FP-Growth uses less memory

What is the purpose of the support-confidence framework?

- **Answer**: B) To identify and evaluate association rules

- **Slide Credits:**

Some slides from this chapter is adapted from: Jiawei Han, Jian Pei, and Hanghang Tong, Data Mining: Concepts and Techniques, 4th ed. Morgan Kaufmann Publishers, 2023. ISBN 978-0-12-811760-6. Quiz are generated by using LLMs for academic purpose only.

Thank you!