DATA WAREHOUSING & MINING

Unit-3: Data Mining

Parameshwar R Hegde, PhD

What is Frequent Pattern Analysis?

- **Frequent patterns:** A pattern (set of items, subsequences, substructures, etc) that occurs frequently in a dataset
- First proposed by Agrawal, Imielinski and Swami in the context of frequent itemsets and association rule mining
- **Motivation:** Finding inherent regularities in data
  - What products were often purchased? – Milk and Bread?
  - What are the subsequent purchases after buying a PC?
  - What kind of DNAs are sensitive to new drug
  - Can we automatically classify web documents?
- **Applications**
  - Basket data analysis, cross-marketing, catalog design, web log analysis , DNA sequence analysis etc.

Why Frequent Pattern Analysis?

- **Discloses an intrinsic and important property of data sets**

- **Forms the foundation of many essential data mining tasks**
  - Association, correlation and causality analysis
  - Sequential and structured (sub-graphs) patterns
  - Pattern analysis in spatiotemporal, multimedia, time-series and stream data
  - Classification: associative classification
  - Cluster analysis: frequent pattern based cluster
  - Data warehousing: iceburg cube and cube gradient
  - Semantic data compression: fascicles
  - Broad applications

Basic Concepts – Frequent Patterns and Association

- "support" and "confidence" they can provide a measure if the Association rule is qualified or not for a particular data set

- **Support**: how often a given rule appears in the database being mined

  Support = Frequency (A, C) / N

- **Confidence**: the number of times a given rule turns out to be true in practice

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

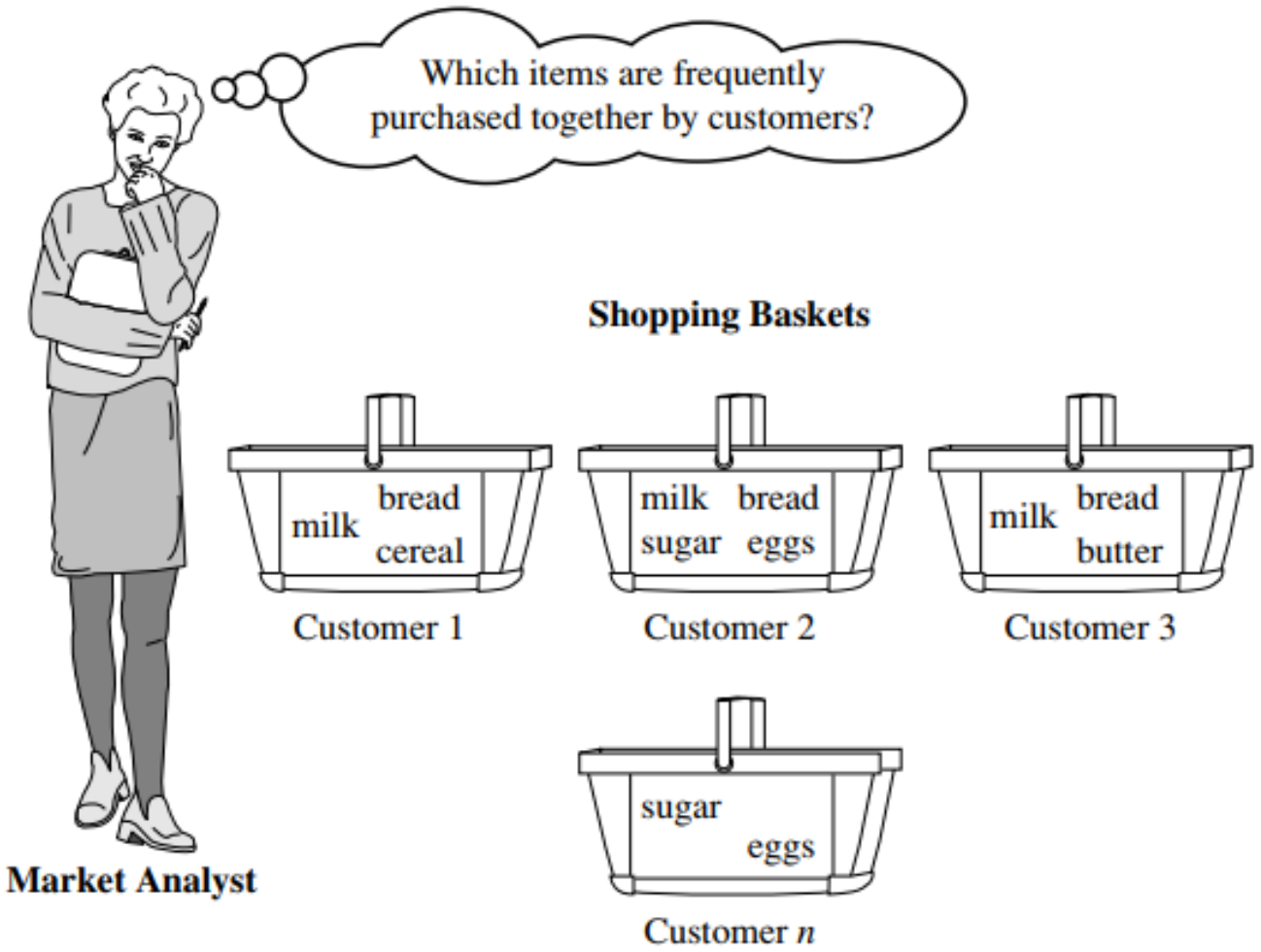
Transactions	Occurrences
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, B, C, E

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

- Support = 3 / 5
- Confidence = 3 / 3

Basic Concepts – Frequent Patterns and Association

- **Market Basket analysis**



Basic Concepts – Frequent Patterns and Association

- **Frequent itemset:** An itemset whose support is greater than some user specified minimum support
  - **Apriori principle**
    1. If an itemset is infrequent then all its supersets are infrequent
    2. If an itemset is frequent then all its subsets are frequent

Itemsets	Occurrences	F/I
A	3/5	F
B	4/5	F
C	4/5	F
D	1/5	I
E	4/5	F
…	…	…

2^n-1 itemsets

Basic Concepts – Frequent Patterns and Association

- **Closed frequent itemset:** If none of its immediate supersets has the same support as that of the itemset

Itemsets	Occurrences	F/I	C/NC
A	3/5	F	NC
B	4/5	F	NC
C	4/5	F	C
D	1/5	I	NC
E	4/5	F	NC
…	…	…	…

Itemset C is closed because AC, BC and CD frequencies are less than C's frequency

Basic Concepts – Frequent Patterns and Association

- **Maximal frequent itemset:** If none of its immediate supersets has the same support as that and tat is frequent in the itemset (No immediate superset is frequent)

Itemsets	Occurrences	F/I	M/NM
A	3/5	F	NM
B	4/5	F	NM
C	4/5	F	NM
D	1/5	I	NM
E	4/5	F	NM
…	…	…	…
AB	2/5	F	NM
…	…	…	…
ABC	2/5	F	NM
…	…	…	…
ABCE	2/5	F	M

Itemset ABCE is maximal because it has no superset

Frequent Itemset Mining Methods - Apriori

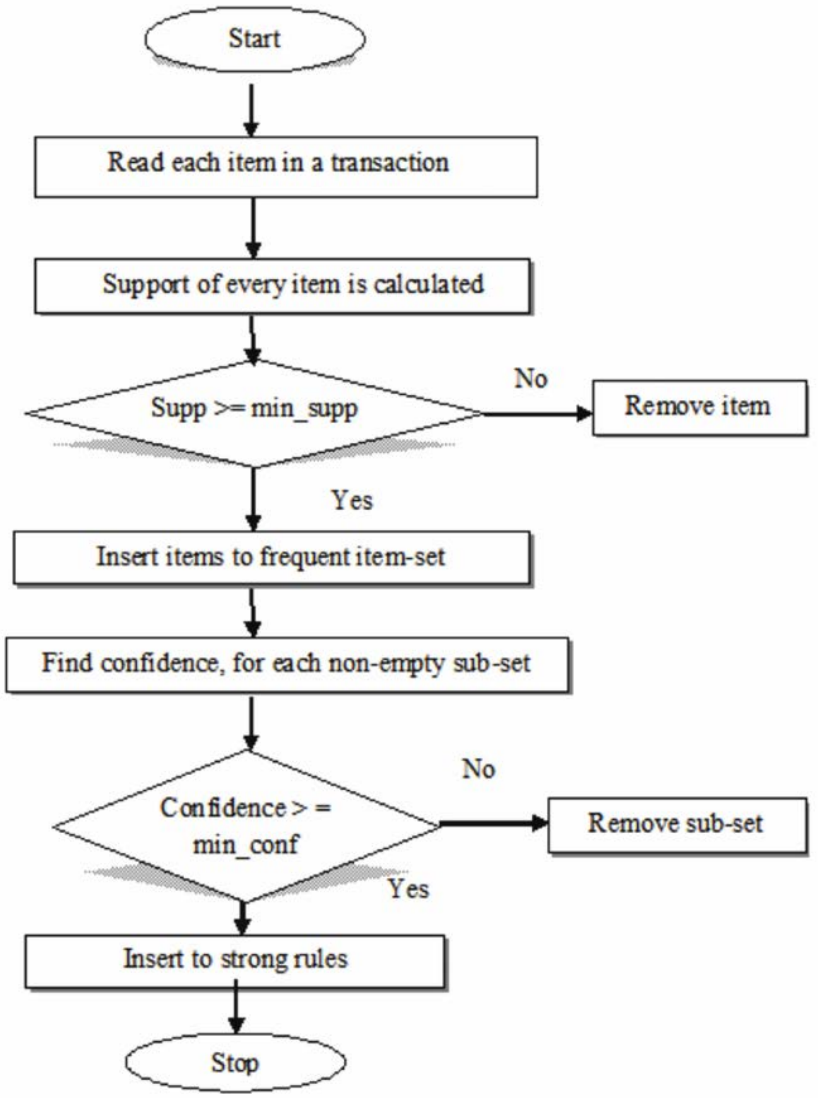
- **We have different approaches; lets begin by Apriori, the basic algorithm for finding frequent itemsets**

- **Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules**

- **The name of the algorithm is based on the fact that the algorithm uses *prior knowledge* of frequent itemset properties**

- **3 steps**
  1. Count
  2. Pruning
  3. Joining

Frequent Itemset Mining Methods - Apriori



https://www.researchgate.net/figure/Flow-chart-of-Apriori-algorithm_fig1_265051526

Frequent Itemset Mining Methods - Apriori

- **Example**

Transactions	Occurrences
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E

Min_sup = 2

The problem

Step 1

Itemsets	Occurrences
A	2
B	3
C	3
D	1
E	3

Frequent Itemset Mining Methods - Apriori

- **Example**

Itemsets	Occurrences
A	2
B	3
C	3
E	3

Step 2

Step 3

Itemsets	Occurrences
A, B	1
A, C	2
A, E	1
B, C	2
B, E	3
C, E	2

Frequent Itemset Mining Methods - Apriori

- **Example**

Itemsets	Occurrences
A, C	2
B, C	2
B, E	3
C, E	2

Repeat Pruning

Repeat Step 3

Itemsets	Occurrences
A, B, C	1
B, C, E	2

Itemset {B, C, E} = 2

Frequent Itemset Mining Methods - Apriori

- **Limitations**
  - It is not an efficient approach for large number of datasets

- **Improving the efficiency of apriori**
  1. Hash-based technique (hashing itemsets into corresponding buckets): A hash-based technique can be used to reduce the size of the candidate *k*-itemsets
     - Used to reduce the size of the given itemset
     - E.g., when scanning each transaction in the database to generate the frequent 1-itemsets, L1, we can generate all the 2-itemsets for each transaction, hash (i.e., map) them into the different buckets of a hash table structure, and increase the corresponding bucket counts

Frequent Itemset Mining Methods - Apriori

- **Improving the efficiency of apriori**
  - **Hash table**

Create hash table H_2 using hash function $h(x, y) = ((order\ of\ x) \times 10 + (order\ of\ y))\ mod\ 7$

H_2

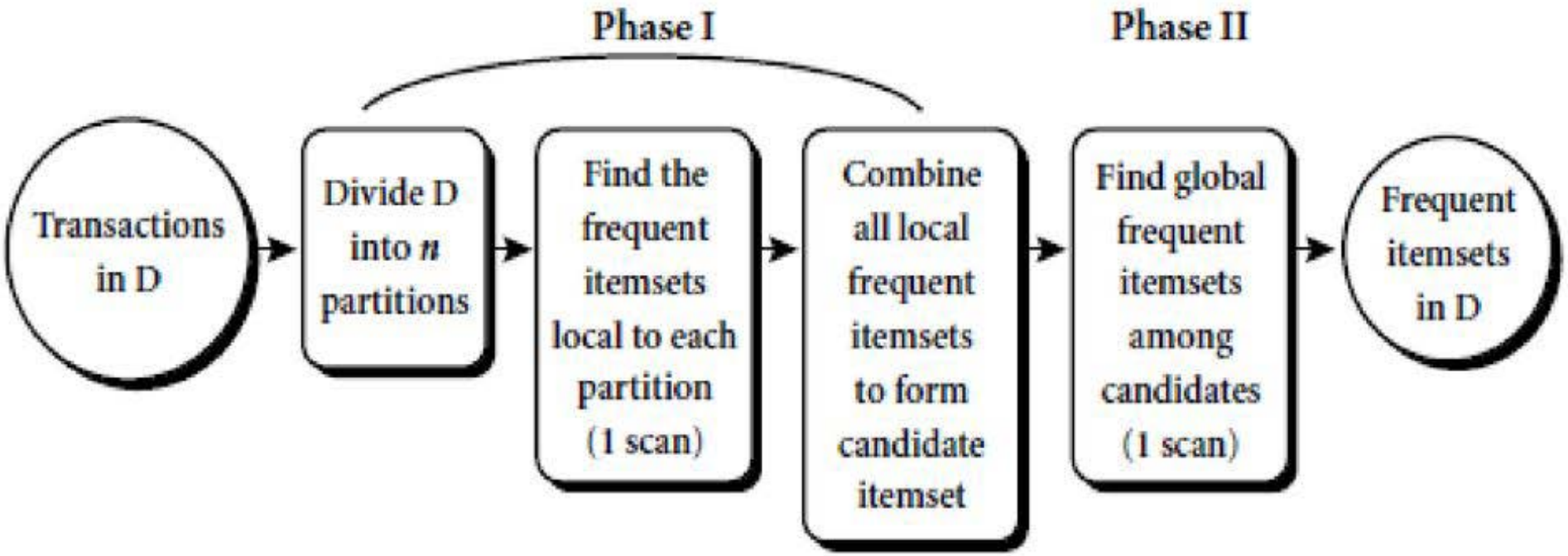
bucket address	0	1	2	3	4	5	6
bucket count	2	2	4	2	2	4	4
bucket contents	{I1, I4} {I3, I5}	{I1, I5} {I1, I5}	{I2, I3} {I2, I3} {I2, I3} {I2, I3}	{I2, I4} {I2, I4}	{I2, I5} {I2, I5}	{I1, I2} {I1, I2} {I1, I2} {I1, I2}	{I1, I3} {I1, I3} {I1, I3} {I1, I3}

Frequent Itemset Mining Methods - Apriori

- **Improving the efficiency of apriori**
  2. Transaction reduction (reducing the number of transactions scanned in future iterations): A transaction that does not contain any frequent *k*-itemsets cannot contain any frequent (*k*+1)-itemsets
     - Therefore, such a transaction can be marked or removed from further consideration because subsequent database scans for *j*-itemsets, where , $j > k$ will not need to consider such a transaction

Frequent Itemset Mining Methods - Apriori

- **Improving the efficiency of apriori**
  3. A partitioning technique can be used that requires just two database scans to mine the frequent itemsets



Frequent Itemset Mining Methods - Apriori

- **Improving the efficiency of apriori**
  5. Sampling (mining on a subset of the given data): The basic idea of the sampling approach is to pick a random sample *S* of the given data *D*, and then search for frequent itemsets in *S* instead of *D*
     - Trade off some degree of accuracy against efficiency

  6. Dynamic itemset counting (adding candidate itemsets at different points during a scan): A dynamic itemset counting technique was proposed in which the database is partitioned into blocks marked by start points
     - New candidate itemsets can be added at any start point, unlike in Apriori, which determines new candidate itemsets only immediately before each complete database scan

Frequent Itemset Mining Methods – FP Growth

- **FP – Frequent Pattern**

- **Efficient and scalable method for mining the complete set of FP using a tree structure for storing information about FP called FP tree**

- **Example**

Min_sup = 2

Transactions	Items
T1	E, A, D, B
T2	D, A, E, C, B
T3	C, A, B, E
T4	B, A, D
T5	D
T6	D, B
T7	A, D, E
T8	B, C

Frequent Itemset Mining Methods – FP Growth

- **Step 1: List out individual items**

Itemset	Occurrences	Priority
A	5	3
B	6	1
C	3	5
D	6	2
E	4	4

Write priorities

- *More frequency – more priority*
- *Same frequency – first come first serve*

Our priority is - *B, D, A, E, C*

Frequent Itemset Mining Methods – FP Growth

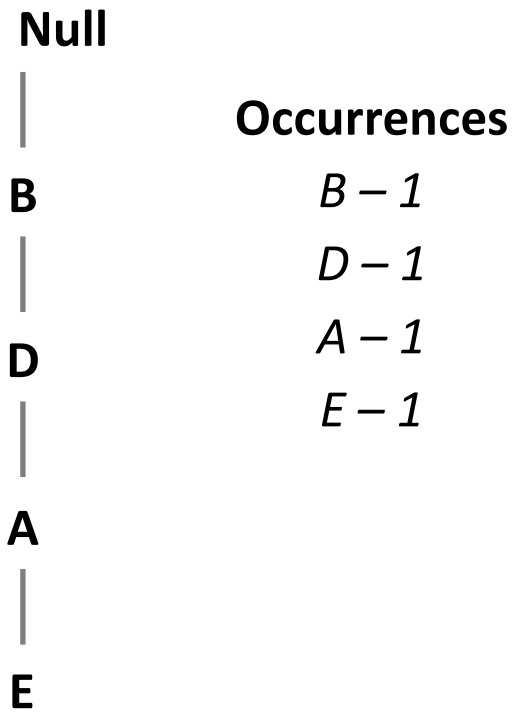
- **Step 2: Arrange itemset as per the priority**

Transactions	Items	Ordered Items
T1	E, A, D, B	B, D, A, E
T2	D, A, E, C, B	B, D, A, E, C
T3	C, A, B, E	B, A, E, C
T4	B, A, D	B, D, A
T5	D	D
T6	D, B	B, D
T7	A, D, E	D, A, E
T8	B, C	B, C

Frequent Itemset Mining Methods – FP Growth

- **Step 3: Construct the tree for ordered items**

T1:



Occurrences

B – 1

D – 1

A – 1

E – 1

T2:

Occurrences

B – 1, 2

D – 1, 2

A – 1, 2

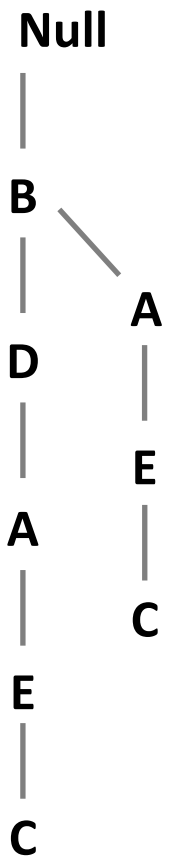
E – 1, 2

C – 1

Frequent Itemset Mining Methods – FP Growth

- **Step 3: Construct the tree for ordered items**

T3:



Occurrences

B – 1, 2, 3

D – 1, 2

A – 1, 2

E – 1, 2

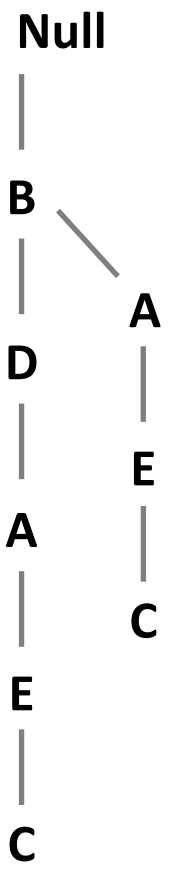
C – 1

A – 1

E – 1

C – 1

T4:



Occurrences

B – 1, 2, 3, 4

D – 1, 2, 3

A – 1, 2, 3

E – 1, 2

C – 1

A – 1

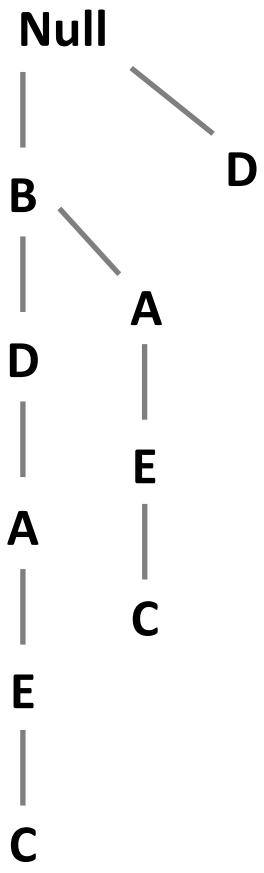
E – 1

C – 1

Frequent Itemset Mining Methods – FP Growth

- **Step 3: Construct the tree for ordered items**

T5:



Occurrences

B – 1, 2, 3, 4

D – 1, 2, 3

A – 1, 2, 3

E – 1, 2

C – 1

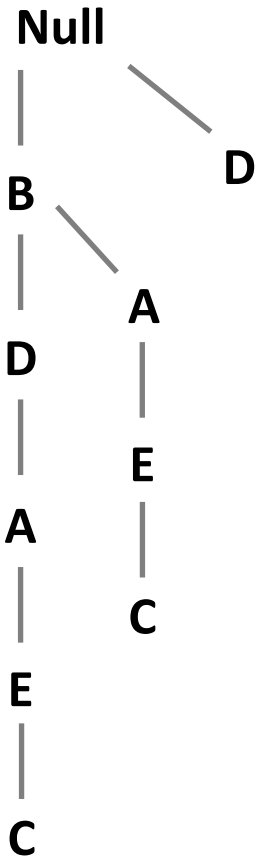
A – 1

E – 1

C – 1

D – 1

T6:



Occurrences

B – 1, 2, 3, 4, 5

D – 1, 2, 3, 4

A – 1, 2, 3

E – 1, 2

C – 1

A – 1

E – 1

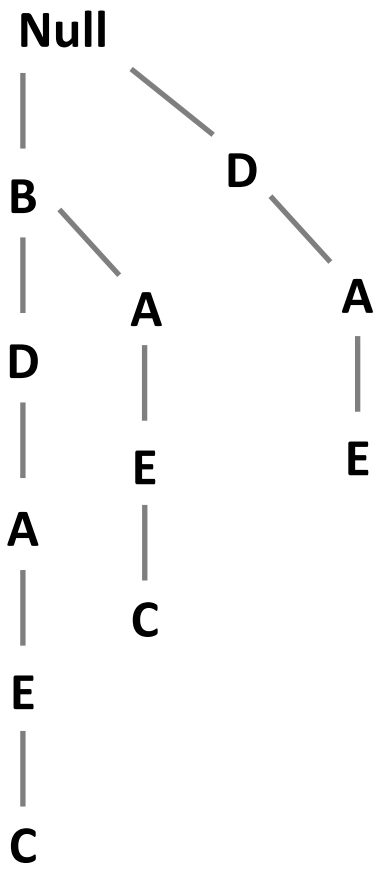
C – 1

D – 1

Frequent Itemset Mining Methods – FP Growth

- **Step 3: Construct the tree tree for ordered items**

T7:



Occurrences

B – 1, 2, 3, 4, 5

D – 1, 2, 3, 4

A – 1, 2, 3

E – 1, 2

C – 1

A – 1

E – 1

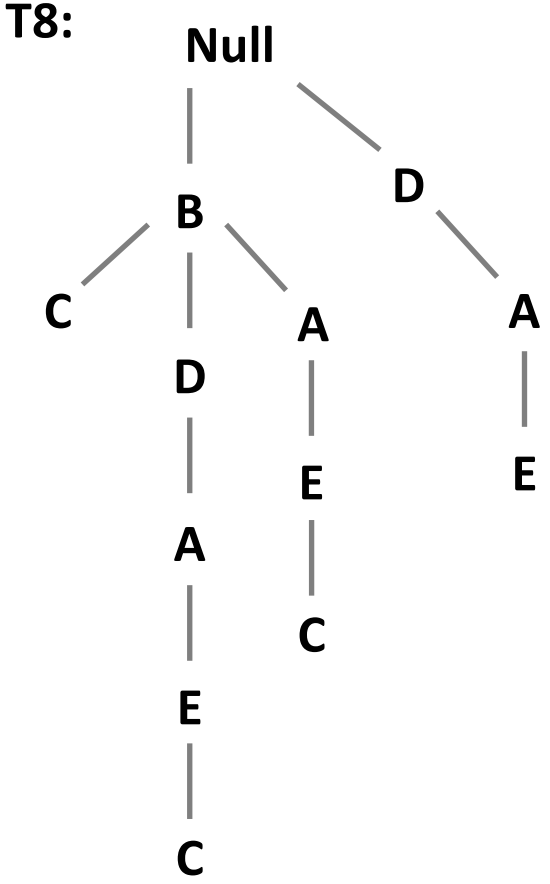
C – 1

D – 1, 2

A – 1

E – 1

T8:



Occurrences

B – 1, 2, 3, 4, 5, 6

D – 1, 2, 3, 4

A – 1, 2, 3

E – 1, 2

C – 1

A – 1

E – 1

C – 1

D – 1, 2

A – 1

E – 1

C – 1

Frequent Itemset Mining Methods – FP Growth

- **Frequent Pattern Generation**

Items	Conditional Pattern Base	Conditional FP Tree	Frequent Pattern Generated
C	{B: 1}, {B,D,A,E: 1}, {B, A, E: 1}	{B: 3}, {A: 2}, {E: 2}	{B, C: 3}, {A, C: 2}, {E, C: 2}, {B, A, C: 2}, {B, E, C: 2}, {B, A, E, C: 2}
E	{B,D,A: 2} , {B,A: 1}, {D,A: 1}	{B: 3}, {A: 4}, {D: 3}	{B, E: 3}, {A, E: 4}, {D, E: 3}, {B,A,E: 3}, {A,D,E: 3}, {B,D,E: 3}, {B,A,D,E: 3}
A	{B,D: 3}, {B: 1}, {D: 1}	{B: 4}, {D: 4}	{B,A: 4} , {D,A: 4}, {B,D,A: 4}
D	{B: 4}	{B: 4}	{B,D: 4}

Frequent Itemset Mining Methods

- **Class Activity – Apply both Apriory and FP Growth Algorithm**

Transactions	Items
T1	I1, I2, I5
T2	I2, I4
T3	I2 I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

Min_sup = 2

Pattern Evaluation Methods

- **Process of assessing the quality of discovered patterns**
- **This process is important in order to determine whether the patterns are useful and whether they can be trusted**
- **There are a number of different measures that can be used to evaluate patterns, and the choice of measure will depend on the application**

Pattern Evaluation Methods

- **Strong rules are not necessarily interesting**
  - Whether or not rule is interesting can be assessed either subjectively or objectively
  - Ultimately, only the user can judge if a given rule is interesting
  - Being subjective and may differ from one user to another
  - However objective interestingness measures based on the statistics “behind” the data
  - It can be used as one step towards the goal of weeding out uninteresting rules that would be otherwise presented to the user

Pattern Evaluation Methods

- **Strong rules are not necessarily interesting**
  - Whether or not rule is interesting can be assessed either subjectively or objectively
  - Ultimately, only the user can judge if a given rule is interesting
  - Being subjective and may differ from one user to another
  - However objective interestingness measures based on the statistics “behind” the data
  - It can be used as one step towards the goal of weeding out uninteresting rules that would be otherwise presented to the user

Pattern Evaluation Methods

- **A misleading "strong" association rule**
  - Of the 10,000 transactions analyzed, the data show that 6000 of the customer transactions included computer games, while 7500 included videos, and 4000 included both computer games and videos
  - Discovering association rules is run on the data, using a minimum support of, say, 30% and a minimum confidence of 60%

Pattern Evaluation Methods

- **From association analysis to correlation analysis**
  - The support and confidence measures are insufficient at filtering out uninteresting association rules
  - To tackle this weakness, a correlation measure can be used to augment the support–confidence framework for association rules
    - A correlation measure can be used to augment the support-confidence framework for association rules
    - This will lead to the correlation rules of the form,

      $A \rightarrow B$ [support, confidence, correlation]
  - A correlation rule is measured not only by its support and confidence but also by the correlation between itemsets A and B

Pattern Evaluation Methods

- **From association analysis to correlation analysis**
  - Lift is a simple correlation measure that is given as,
  - The occurrence of itemset
    - A is independent of the occurrence of itemset B

      If P(AUB) = P(A)P(B)
  - Otherwise,
    - Itemsets A and B are dependent and correlated as events
  - This definition can easily be extended to more than two itemsets
  - The lift between the occurrence of A and B can be measured by computing

$$\text{Lift}(A, B) = \frac{P(AUB)}{P(A)P(B)}$$

Pattern Evaluation Methods

- **From association analysis to correlation analysis**
  - If the resulting value is less than 1, then the occurrence of A is negatively correlated with the occurrence of B
    - i.e., the occurrence of one likely leads to the absence of the other one
  - If the resulting value is greater than 1, then A and B are positively correlated
    - i.e., the occurrence of one implies the occurrence of the other
  - If the resulting value is equal to 1, then A and B are independent and there is no correlation between them

Pattern Evaluation Methods

- **From association analysis to correlation analysis**
  - If we apply the lift rule to the previous purchasing of game and video example then,
    - Probability of purchasing a computer game P({game}) = 0.60
    - Probability of purchasing a video P({video}) = 0.75
    - Probability of purchasing both is P({game, video}) = 0.40
    - As per the lift rule,

      P({game, video}) / P({game}) P({video}) = 0.40 /(0.60*0.75) = 0.89
    - Because the value is less than 1, there is a negative correlation between the occurrence of {game} and {video}

Such a negative correlation cannot be identified by a support-confidence framework

Pattern Evaluation Methods

- **Correlation analysis using X^2**
  - To do this, we take the squared difference between the observed and expected value for a slot (A and B pair) in the contingency table divided by the expected value
  - We need the observed value and expected value for each slot of the contingency table as,

2 X 2 Contingency table summarizing the transactions with respect to Game and Video purchases

	Game	Not Game	Total
Video	4000 (4500)	3500 (3000)	7500
Not Video	2000 (1500)	500 (1000)	2500
Total	6000	4000	10000

Pattern Evaluation Methods

- **Correlation analysis using X^2**
  - From the table we can compute,

$$X2 = \sum \frac{(Observed - Expected)^2}{Expected} = \frac{(4000-4500)^2}{4500} + \frac{(3500-3000)^2}{3000} + \frac{(2000-1500)^2}{1500}$$

$$+ \frac{(500-1000)^2}{1000} = 555.6$$

  - Because the X2 value is greater than 1, the observed value of the slot (game, video) = 4000, which is than the expected value of 4500
  - Hence, buying game and video are negatively correlated
  - This is consistent with the conclusion derived from the analysis of the lift measure