Week 11

Mining Association Rules

Seokho Chi

Associate Professor | Ph.D.

SNU Construction Innovation Lab

Source: Tan, Kumar, Steinback (2006)

서울대학교 건설혁신연구실
Construction Innovation Lab SNU

Association Rule Mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Market-Basket transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Implication means co-occurrence, not causality!

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Mining

- Itemset: a collection of one or more items
  - 1 item set: {milk}, 3 item set: {milk, bread, diaper}
- Support count (θ): frequency of occurrence of an itemset
  - θ({milk, bread, diaper}) = 2
- Support (S): fraction of transactions that contain an itemset
  - S({milk, bread, diaper}) = 2/5
- Frequent itemset: an item set whose support is greater than or equal to a minimum support threshold(*minsup*)

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Mining

- Association rule: an implication expression of the form X → Y where X and Y are item sets
  - {milk, diaper} → {bread}, {milk} → {diaper, bread}

- Rule evaluation metrics
  - Support (S): fraction of transactions that contain both X and Y
  - Confidence (C): measure how often items in Y appear transactions that contain X

  - {Milk, Diaper} → {Beer}
  - S = 2/5 : milk, diaper & beer among total
  - C = 2/3: beer among milk, diaper

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Discovery: Application 2

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Association Rule Mining Task

- Given a set of transactions T, the goal of association rule mining is to find all rules having
  - support ≥ *minsup* threshold
  - confidence ≥ *minconf* threshold

- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the *minsup* and *minconf* thresholds

  ⇒ Computationally prohibitive!

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Mining Association Rules

TID	*Items*
1	**Bread, Milk**
2	**Bread, Diaper, Beer, Eggs**
3	**Milk, Diaper, Beer, Coke**
4	**Bread, Milk, Diaper, Beer**
5	**Bread, Milk, Diaper, Coke**

Example of Rules:

{Milk,Diaper} $\rightarrow$ {Beer} (s=0.4, c=0.67)
{Milk,Beer} $\rightarrow$ {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} $\rightarrow$ {Milk} (s=0.4, c=0.67)
{Beer} $\rightarrow$ {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} $\rightarrow$ {Milk,Beer} (s=0.4, c=0.5)
{Milk} $\rightarrow$ {Diaper,Beer} (s=0.4, c=0.5)

Observations:

- All the above rules are binary partitions of the same itemset:
  {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Mining Association Rules

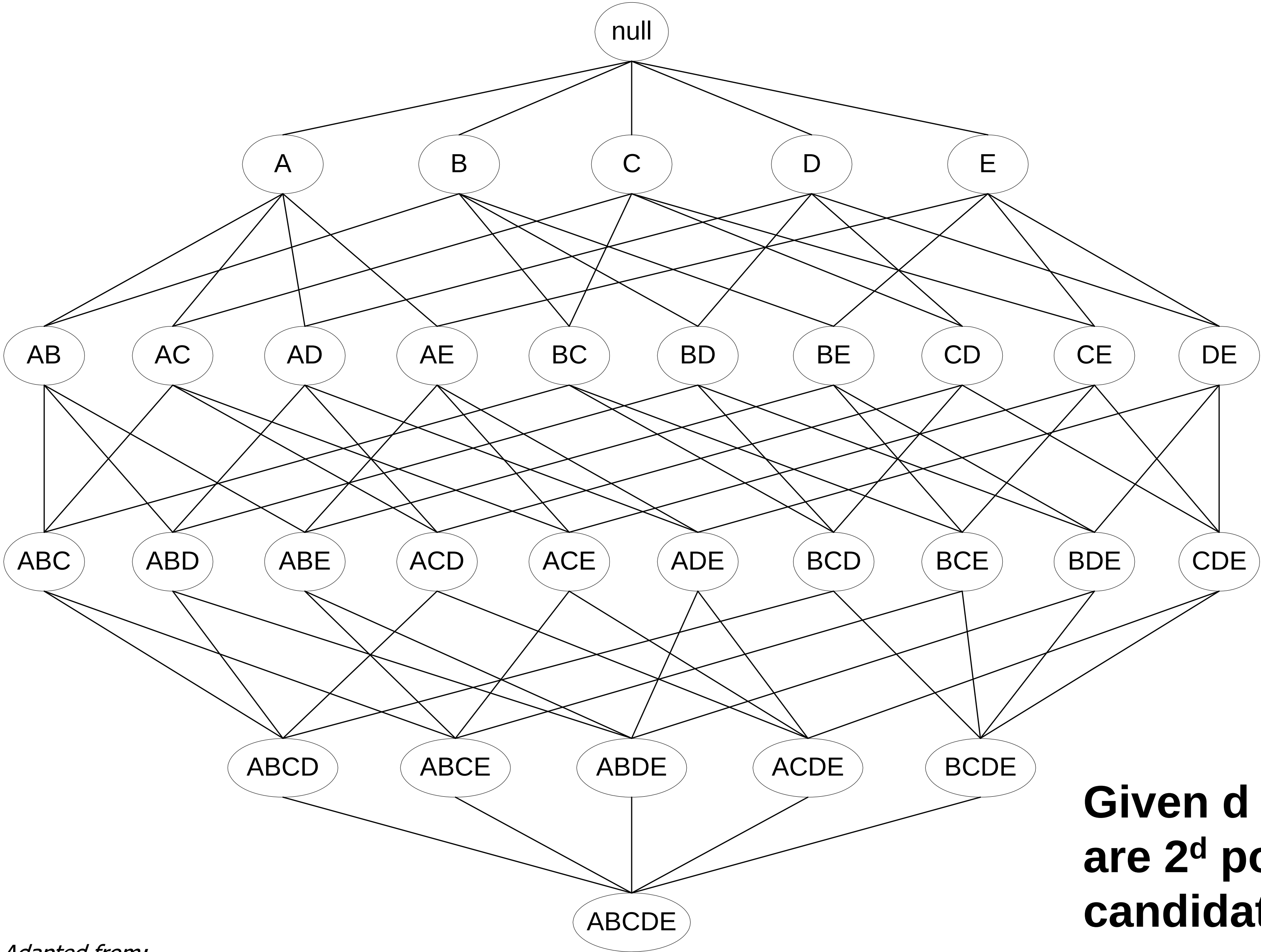
- Two-step approach:
  1. Frequent Itemset Generation
     - Generate all itemsets whose support ≥ minsup
  2. Rule Generation
     - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset
- Frequent itemset generation is still computationally expensive

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Frequent Itemset Generation



Given d items, there are 2^d possible candidate itemsets

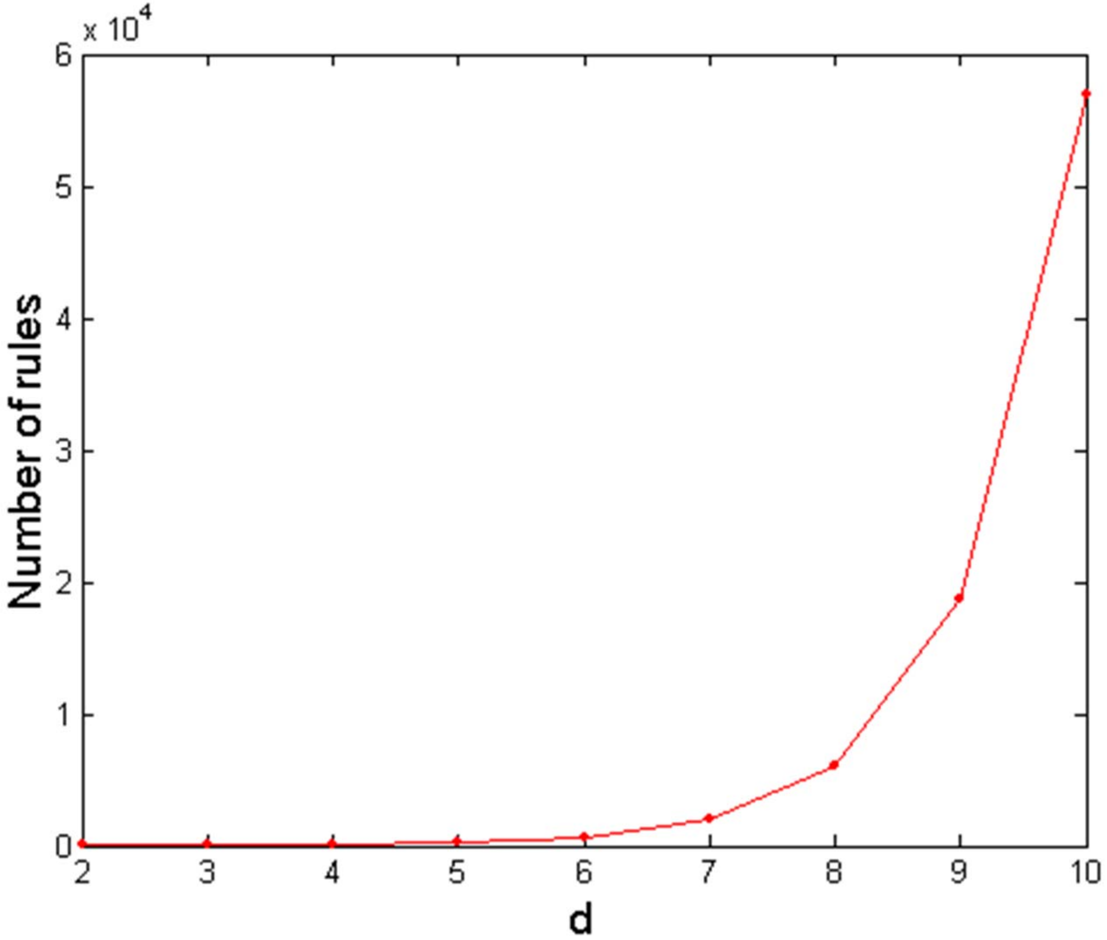
Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Computational Complexity

- Given d unique items:
  - Total number of itemsets = 2^d
  - Total number of possible association rules:



$$R = \sum_{k=1}^{d-1} \left[\binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j} \right]$$

$$= 3^d - 2^{d+1} + 1$$

If d=6, R = 602 rules

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Frequent Itemset Generation

- Brute-force approach:
  - Each itemset in the lattice is a candidate frequent itemset
  - Count the support of each candidate by scanning the database

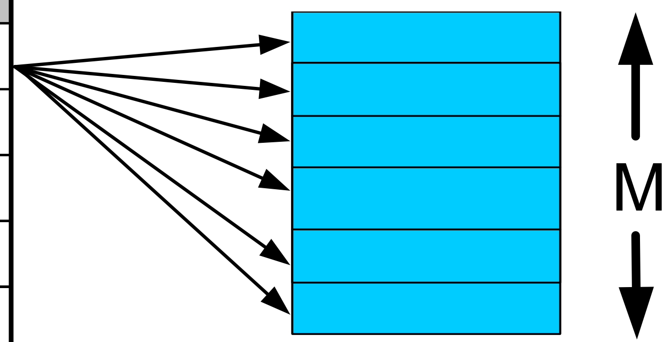
Transactions

N

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

w

List of Candidates



  - Match each transaction against every candidate

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Frequent Itemset Generation Strategies

- Reduce the number of candidates (M)
  - Complete search: $M=2^d$
  - Use pruning techniques to reduce M
- Reduce the number of transactions (N)
  - Reduce size of N as the size of itemset increases
  - Used by DHP and vertical-based mining algorithms
- Reduce the number of comparisons (NM)
  - Use efficient data structures to store the candidates or transactions
  - No need to match every candidate against every transaction

Reducing Number of Candidates

- Apriori principle:
  – If an itemset is frequent, then all of its subsets must also be frequent

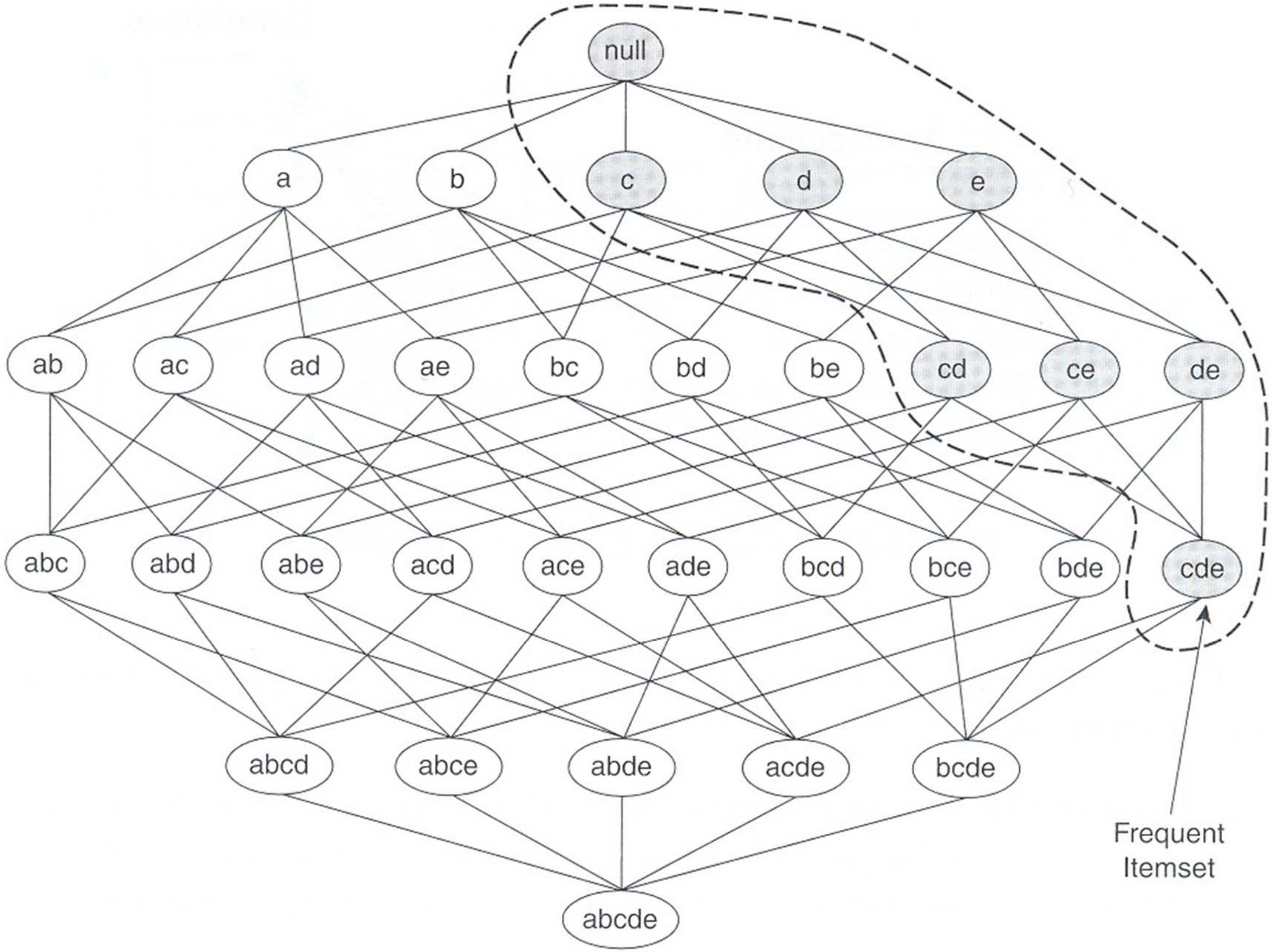
- Apriori principle holds due to the following property of the support measure:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

  – Support of an itemset never exceeds the support of its subsets
  – This is known as the anti-monotone property of support

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

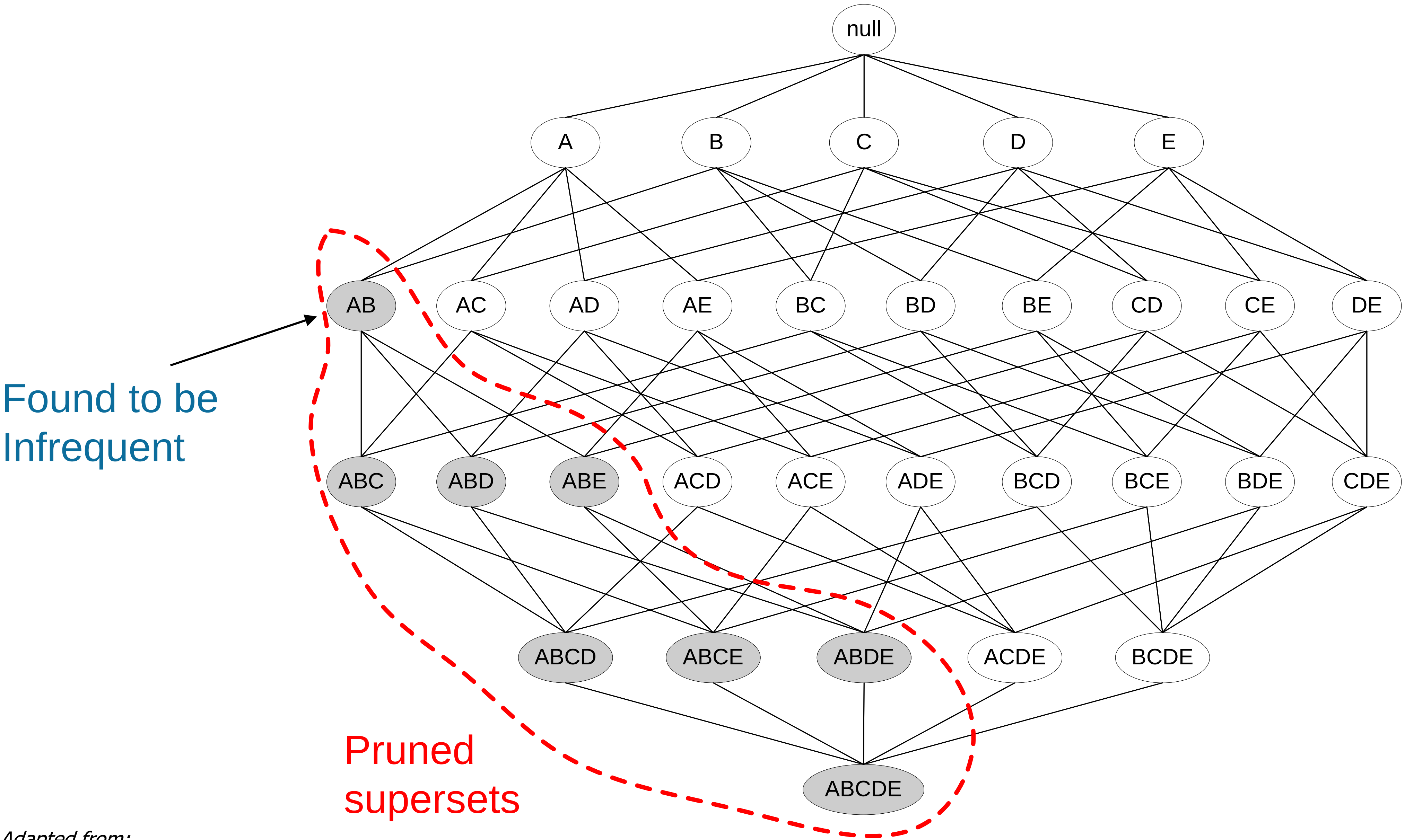
Illustrating Apriori Principle



If "cde" is frequent, all things are also frequent!

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Illustrating Apriori Principle



Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Illustrating Apriori Principle

Item	Count
Bread	**4**
Coke	**2**
Milk	**4**
Beer	**3**
Diaper	**4**
Eggs	**1**

Items (1-itemsets)

Itemset	Count
{Bread,Milk}	**3**
{Bread,Beer}	**2**
{Bread,Diaper}	**3**
{Milk,Beer}	**2**
{Milk,Diaper}	**3**
{Beer,Diaper}	**3**

Pairs (2-itemsets)

(No need to generate candidates involving Coke or Eggs)

Minimum Support = 3

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	**3**

If every subset is considered,

$$^6C_1 + {}^6C_2 + {}^6C_3 = 41$$

With support-based pruning,

$$6 + 6 + 1 = 13$$

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Apriori Algorithm

- Method:
  - Let k=1
  - Generate frequent itemsets of length 1
  - Repeat until no new frequent itemsets are identified
    - Generate length (k+1) candidate itemsets from length k frequent itemsets
    - Prune candidate itemsets containing subsets of length k that are infrequent
    - Count the support of each candidate by scanning the DB
    - Eliminate candidates that are infrequent, leaving only those that are frequent

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Candidate Generation

Items

Item
Beer
Bread
Cola
Diapers
Milk
Eggs

Candidate Generation

Itemset
{Beer, Bread, Cola}
{Beer, Bread, Diapers}
{Beer, Bread, Milk}
{Beer, Bread, Eggs}
{Beer, Cola, Diapers}
{Beer, Cola, Milk}
{Beer, Cola, Eggs}
{Beer, Diapers, Milk}
{Beer, Diapers, Eggs}
{Beer, Milk, Eggs}
{Bread, Cola, Diapers}
{Bread, Cola, Milk}
{Bread, Cola, Eggs}
{Bread, Diapers, Milk}
{Bread, Diapers, Eggs}
{Bread, Milk, Eggs}
{Cola, Diapers, Milk}
{Cola, Diapers, Eggs}
{Cola, Milk, Eggs}
{Diapers, Milk, Eggs}

Candidate Pruning

Itemset
{Bread, Diapers, Milk}

Figure 6.6. A brute-force method for generating candidate 3-itemsets.

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Candidate Generation (2)

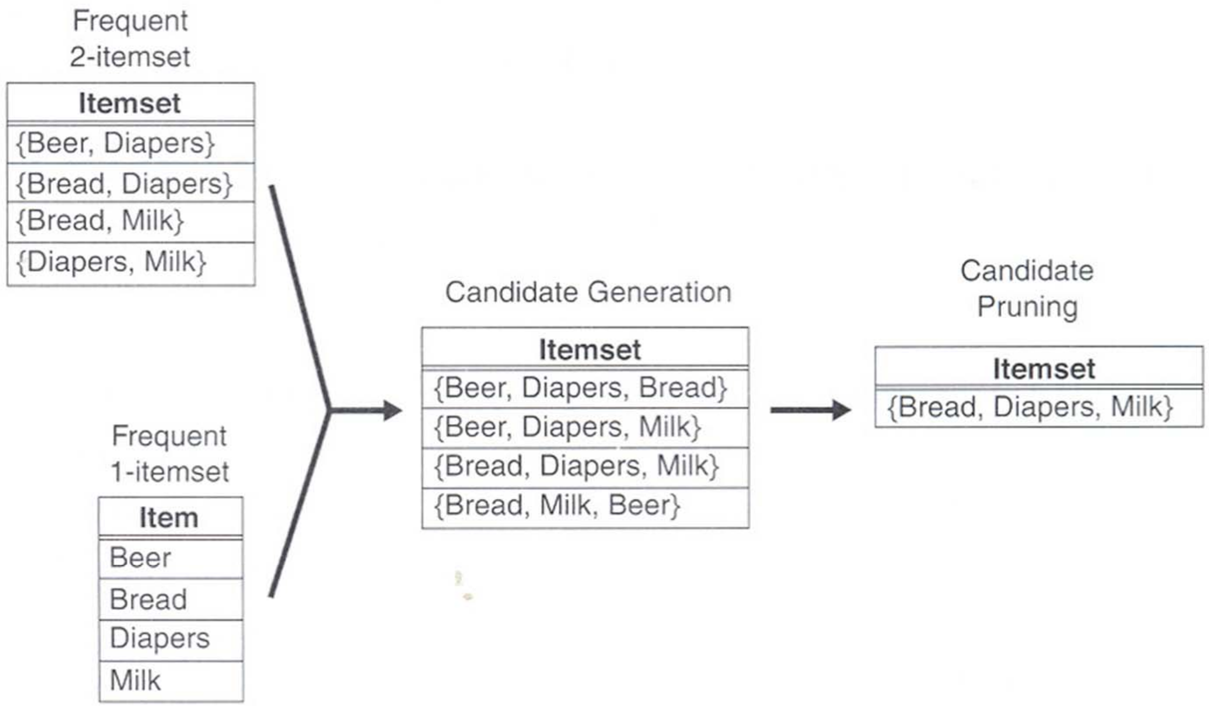


Figure 6.7. Generating and pruning candidate k-itemsets by merging a frequent $(k-1)$-itemset with a frequent item. Note that some of the candidates are unnecessary because their subsets are infrequent.

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Candidate Generation (3)

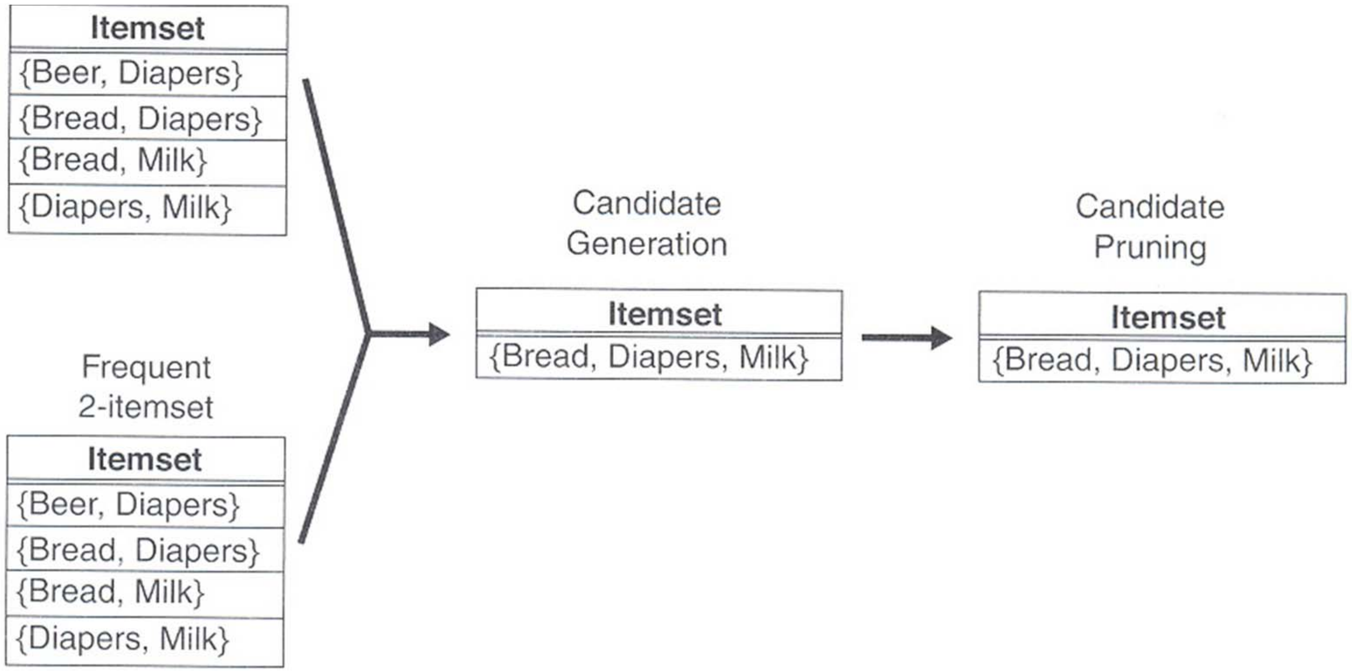


Figure 6.8. Generating and pruning candidate k-itemsets by merging pairs of frequent $(k-1)$-itemsets.

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Rule Generation

Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

- Given a frequent itemset *L*, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement
  - If {A,B,C,D} is a frequent itemset, candidate rules:

ABC $\rightarrow$D,	ABD $\rightarrow$C,	ACD $\rightarrow$B,	BCD $\rightarrow$A,
A $\rightarrow$BCD,	B $\rightarrow$ACD,	C $\rightarrow$ABD,	D $\rightarrow$ABC
AB $\rightarrow$CD,	AC $\rightarrow$ BD,	AD $\rightarrow$ BC,	BC $\rightarrow$AD,
BD $\rightarrow$AC,	CD $\rightarrow$AB,		

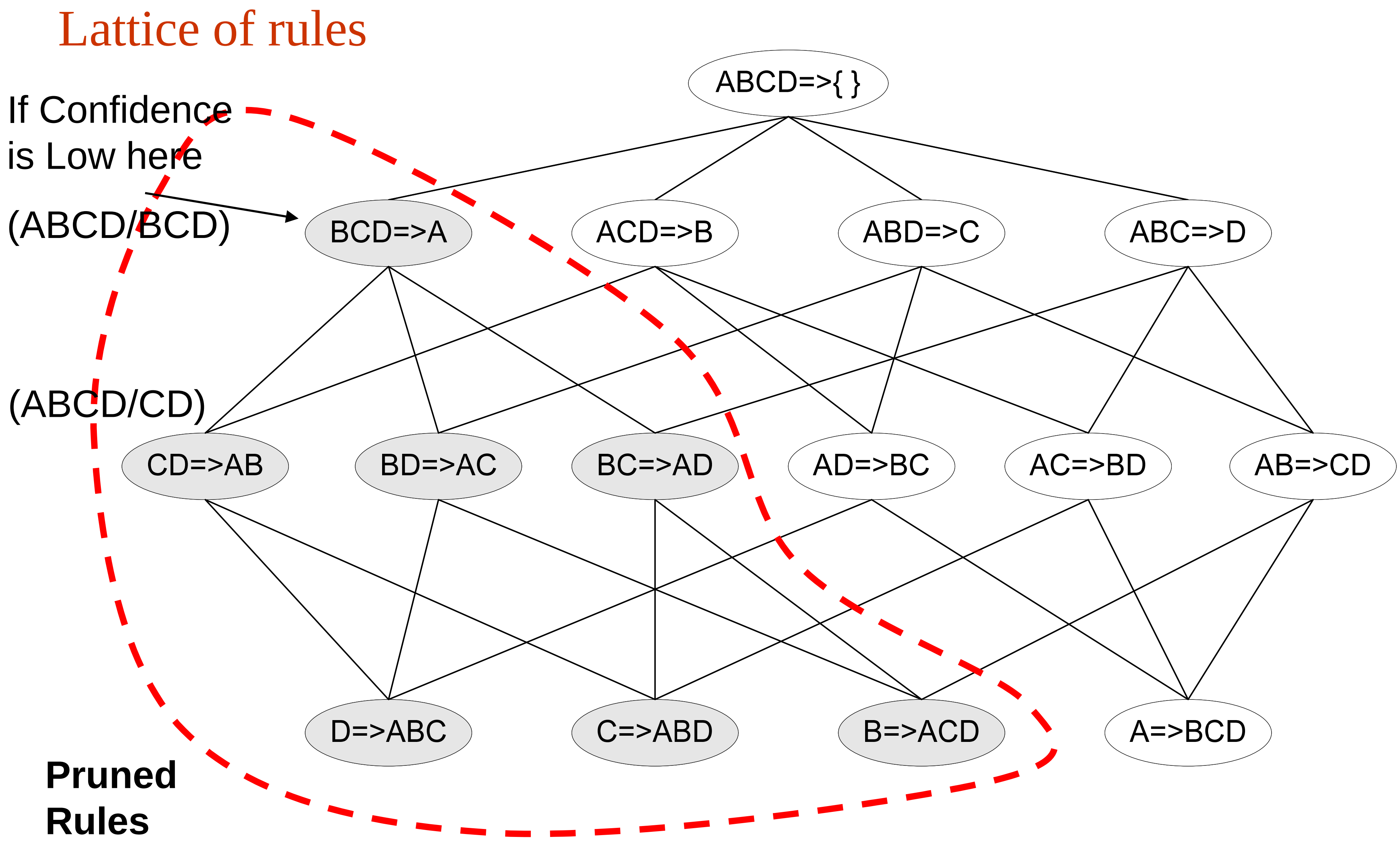
- If |L| = k, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \emptyset$ and $\emptyset \rightarrow L$)

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Rule Generation for Apriori Algorithm



Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Compact Representation of Frequent Itemsets

- Some itemsets are redundant because they have identical support as their supersets

- Need a compact representation

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Closed Itemset

- An itemset is closed if none of its immediate supersets has the same support as the itemset (5-4, 5-3)
- An itemset is not closed if at least one of its immediate supersets has the same support (5-5)

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

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

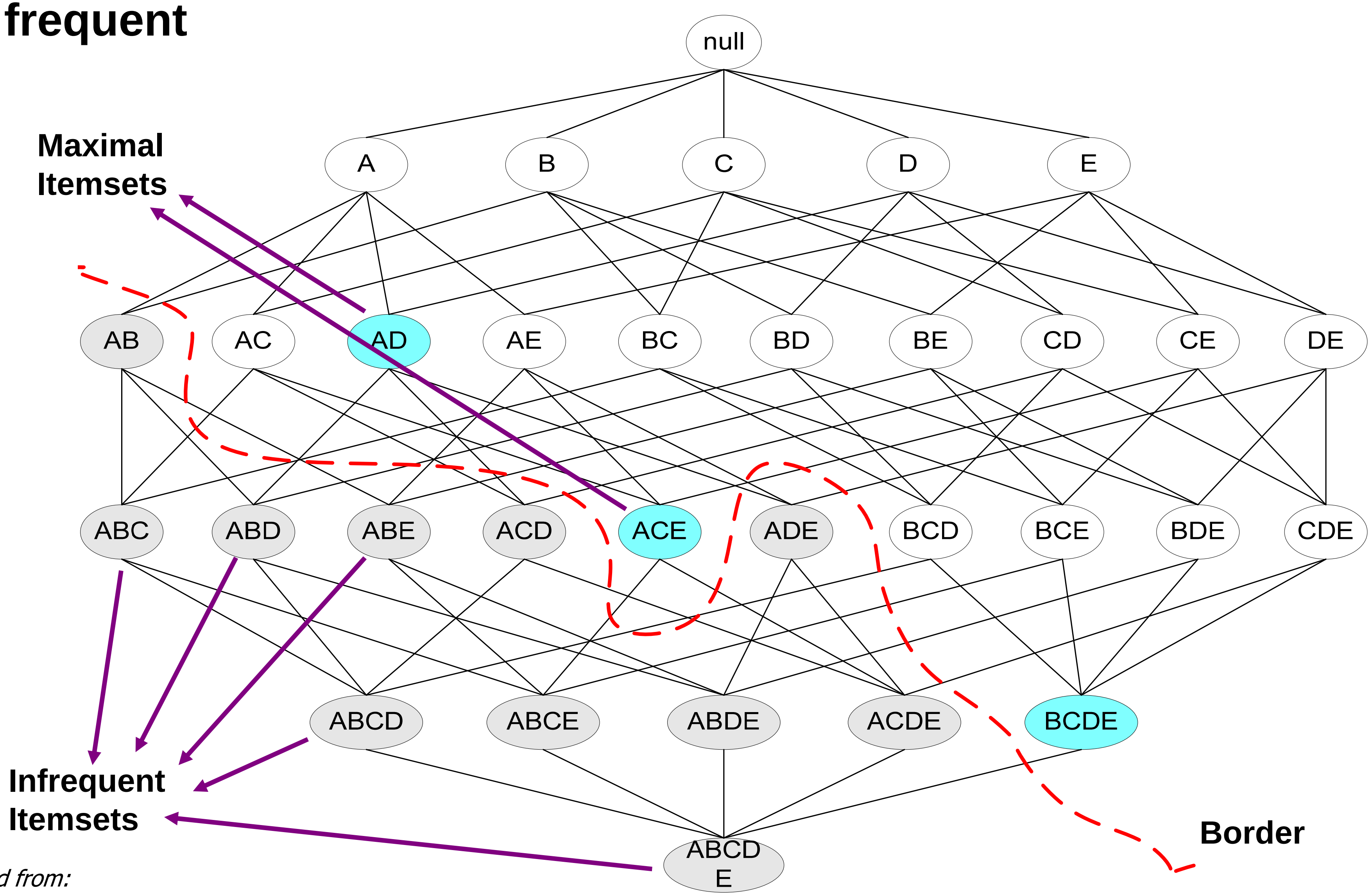
Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Maximal Frequent Itemset

An itemset is maximal frequent if none of its immediate supersets is frequent



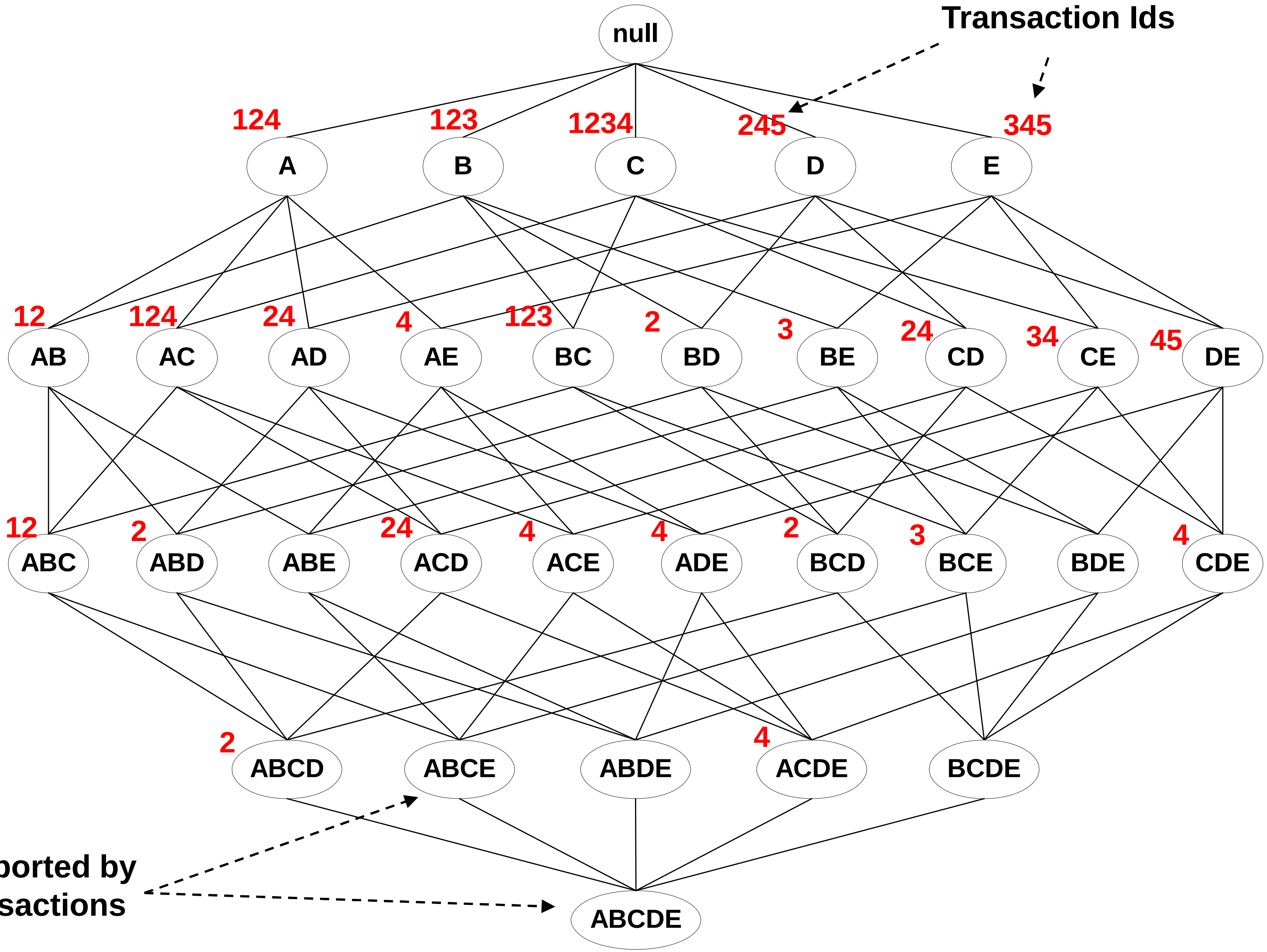
Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Maximal vs Closed Itemsets

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

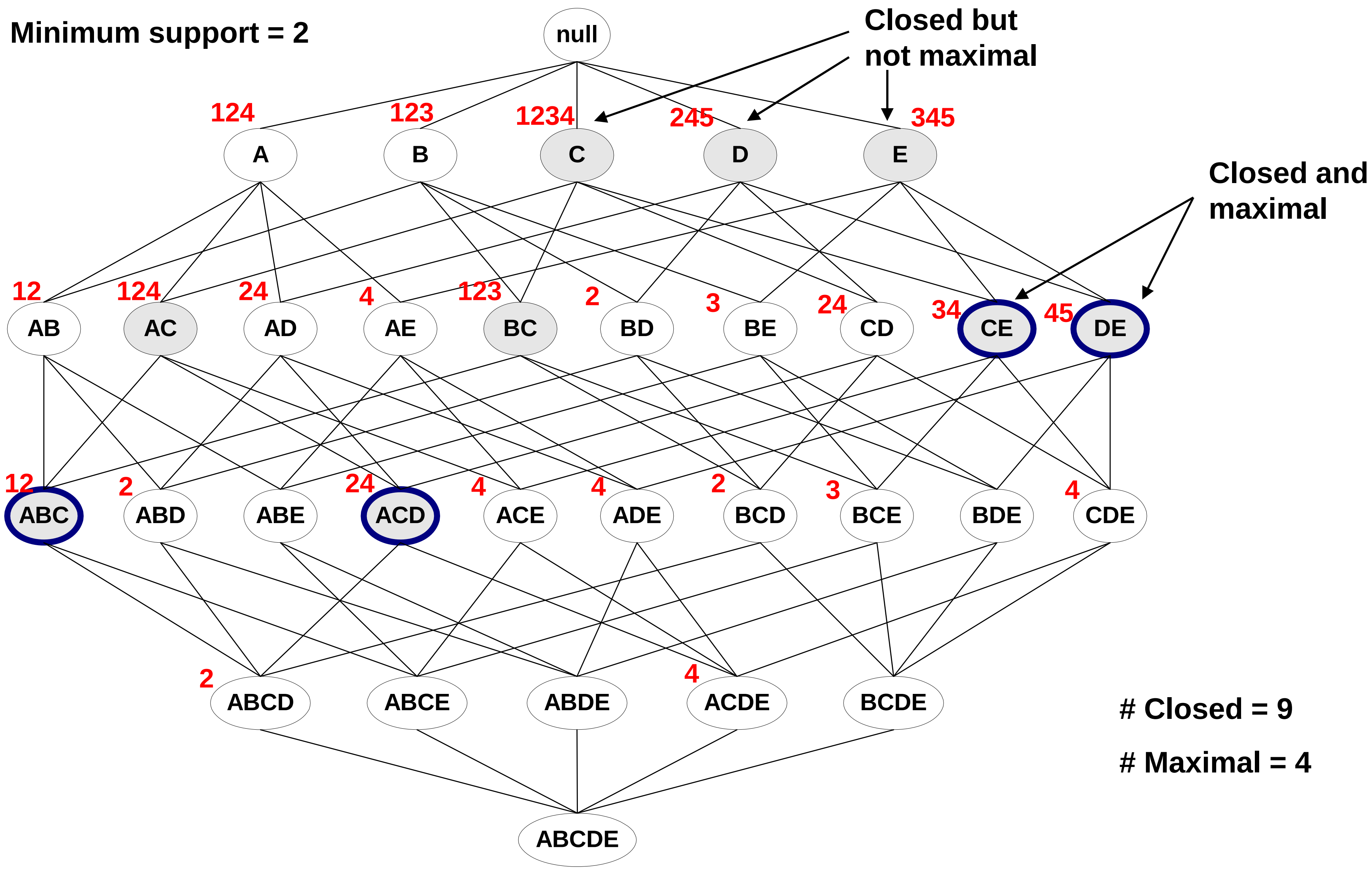


Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Maximal vs Closed Frequent Itemsets

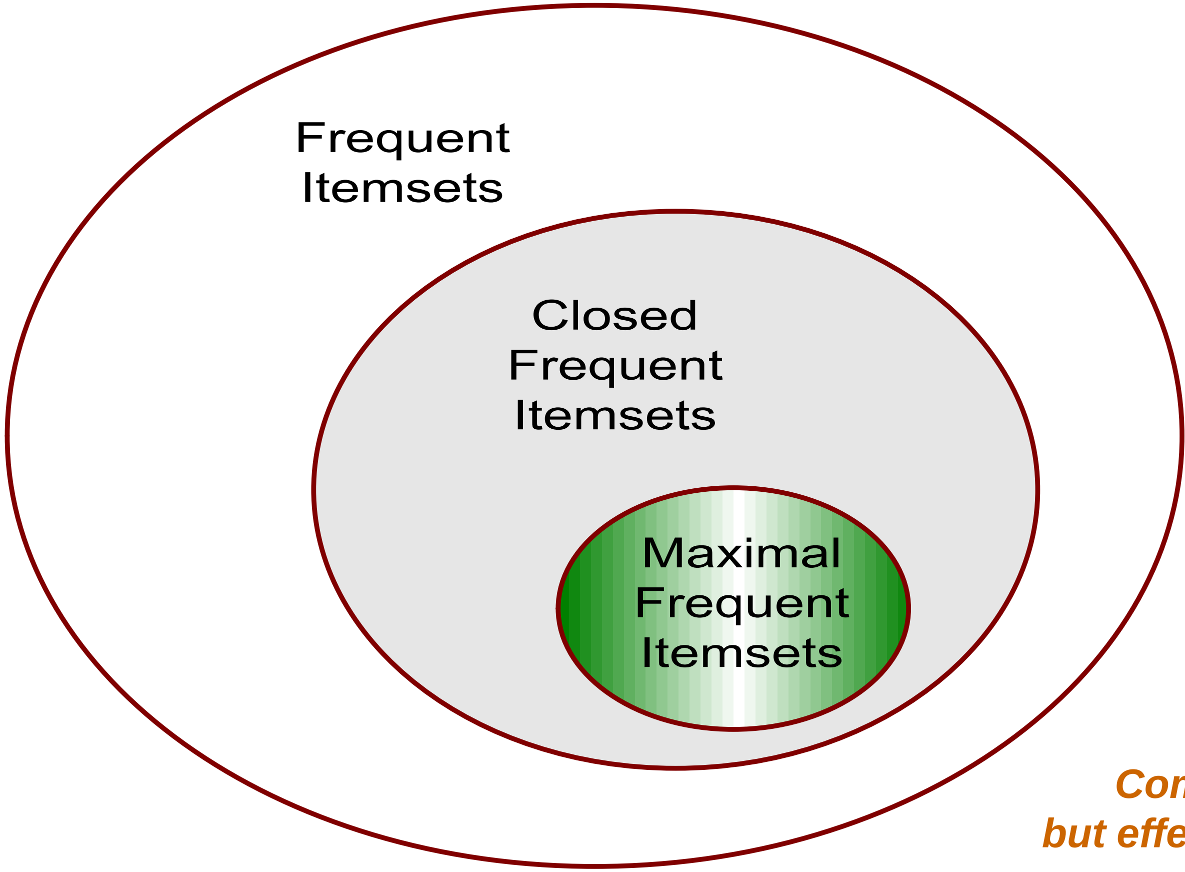


Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Maximal vs Closed Itemsets



Computationally efficient
but effective rule generation!

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

FP-growth Algorithm

- Use a compressed representation of the database using an FP-tree (Frequent-Pattern Tree)

- Once an FP-tree has been constructed, it uses a recursive divide-and-conquer approach to mine the frequent itemsets

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

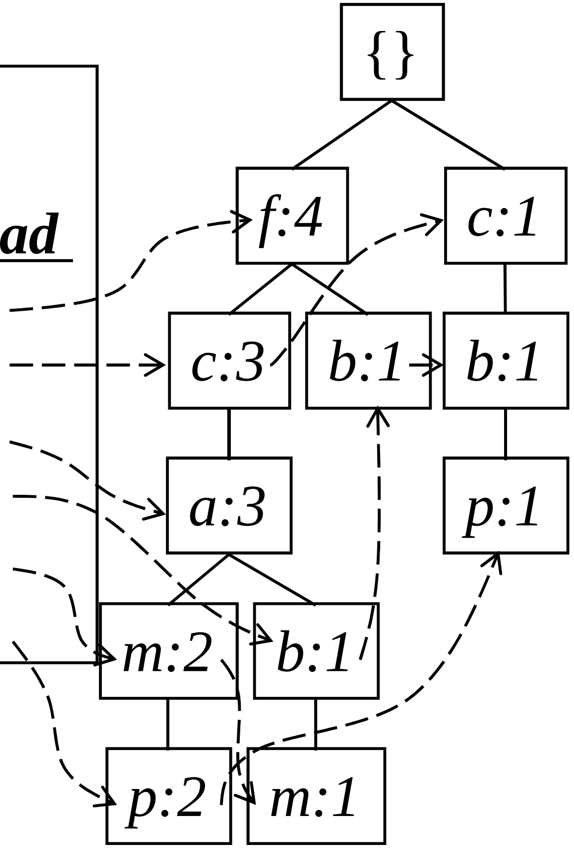
Construct FP-tree from a Transaction Database

TID	***Items bought***	***(ordered) frequent items***
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}***

min_support = 3

Header Table

Item	***frequency***	***head***
f	*4*	
c	*4*	
a	*3*	
b	*3*	
m	*3*	
p	*3*	



1. Scan DB once, find frequent 1-itemset (single item pattern)
2. Sort frequent items in frequency descending order, f-list
3. Scan DB again, construct FP-tree

F-list=f-c-a-b-m-p

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Effect of Support Distribution

- How to set the appropriate *minsup* threshold?
  – If *minsup* is set too high, we could miss itemsets involving interesting rare items (e.g., expensive products)

  – If *minsup* is set too low, it is computationally expensive and the number of itemsets is very large

- Using a single minimum support threshold may not be effective

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Pattern Evaluation

- Association rule algorithms tend to produce too many rules
  - many of them are uninteresting or redundant
  - Redundant if $\{A,B,C\} \rightarrow \{D\}$ and $\{A,B\} \rightarrow \{D\}$ have same support & confidence

- In the original formulation of association rules, support & confidence are the only measures used

- Interestingness measures can be used to prune/rank the derived patterns

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Computing Interestingness Measure

- Given a rule X $\rightarrow$ Y, information needed to compute rule interestingness can be obtained from a contingency table

Contingency table for X $\rightarrow$ Y

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	|T|

f_{11}: support of X and Y
f_{10}: support of X and $\overline{Y}$
f_{01}: support of $\overline{X}$ and Y
f_{00}: support of $\overline{X}$ and $\overline{Y}$

Used to define various measures

- support, confidence, lift, Gini, J-measure, etc.

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Drawback of Confidence

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea $\rightarrow$ Coffee

Support (Tea $\rightarrow$ Coffee) = 15 / 100 = 15%

Confidence (Tea $\rightarrow$ Coffee) = 15 / 20 = 75%

but the fraction of people who drink coffee, regardless of whether they drink tea is 90% while the fraction of tea drinkers who drink coffee is only 75%

$\Rightarrow$ Although confidence is high, rule is misleading

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Statistical-based Measures

- Measures that take into account statistical dependence

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

$$Interest = \frac{P(X,Y)}{P(X)P(Y)}$$

$$PS = P(X,Y) - P(X)P(Y)$$

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1 - P(X)]P(Y)[1 - P(Y)]}}$$

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Example: Lift

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea $\rightarrow$ Coffee

Confidence= P(Coffee|Tea) = 0.75

but P(Coffee) = 0.9

$\Rightarrow$ Lift = 0.75/0.9= 0.8333 (< 1, therefore is negatively associated)

**Measure of the performance of a targeting model with respect to the population as a whole*

Good if the response within the target is much better than the average for the population

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

There are lots of measures proposed in the literature

Some measures are good for certain applications, but not for others

What criteria should we use to determine whether a measure is good or bad?

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A),P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)},\frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B),P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

Continuous and Categorical Attributes

How to apply association analysis formulation to non-asymmetric binary variables?

Session Id	Country	Session Length (sec)	Number of Web Pages viewed	Gender	Browser Type	Buy
1	USA	982	8	Male	IE	No
2	China	811	10	Female	Netscape	No
3	USA	2125	45	Female	Mozilla	Yes
4	Germany	596	4	Male	IE	Yes
5	Australia	123	9	Male	Mozilla	No
…	…	…	…	…	…	…

Example of Association Rule:

$$\{\text{Number of Pages} \in [5,10) \wedge (\text{Browser=Mozilla})\} \rightarrow \{\text{Buy = No}\}$$

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Handling Categorical Attributes

Too many categories will create less important rules with low support and confidence!

- Potential Issues
  - What if attribute has many possible values
    - Example: attribute country has more than 200 possible values
    - Many of the attribute values may have very low support
      - » Potential solution: Aggregate the low-support attribute values
  - What if distribution of attribute values is highly skewed
    - Example: 95% of the visitors have Buy = No
    - Most of the items will be associated with (Buy=No) item
      - » Potential solution: drop the highly frequent items

Handling Categorical Attributes

- Transform categorical attribute into asymmetric binary variables
- Introduce a new "item" for each distinct attribute-value pair
  - Example: replace Browser Type attribute with
    - Browser Type = Internet Explorer
    - Browser Type = Mozilla
    - Browser Type = Netscape
    - Then, YES/NO

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Handling Continuous Attributes

- Different kinds of rules:
  - Age $\in$[21,35) $\wedge$ Salary $\in$[70k,120k) $\rightarrow$ Buy
  - Salary $\in$[70k,120k) $\wedge$ Buy $\rightarrow$ Age: μ=28, σ=4

- Different methods:
  - Discretization-based
  - Statistics-based
  - Non-discretization based

Discretization Issues

- Size of the discretized intervals affect support & confidence

$$\{Refund = No, (Income = \$51,250)\} \rightarrow \{Cheat = No\}$$

$$\{Refund = No, (60K \leq Income \leq 80K)\} \rightarrow \{Cheat = No\}$$

$$\{Refund = No, (0K \leq Income \leq 1B)\} \rightarrow \{Cheat = No\}$$

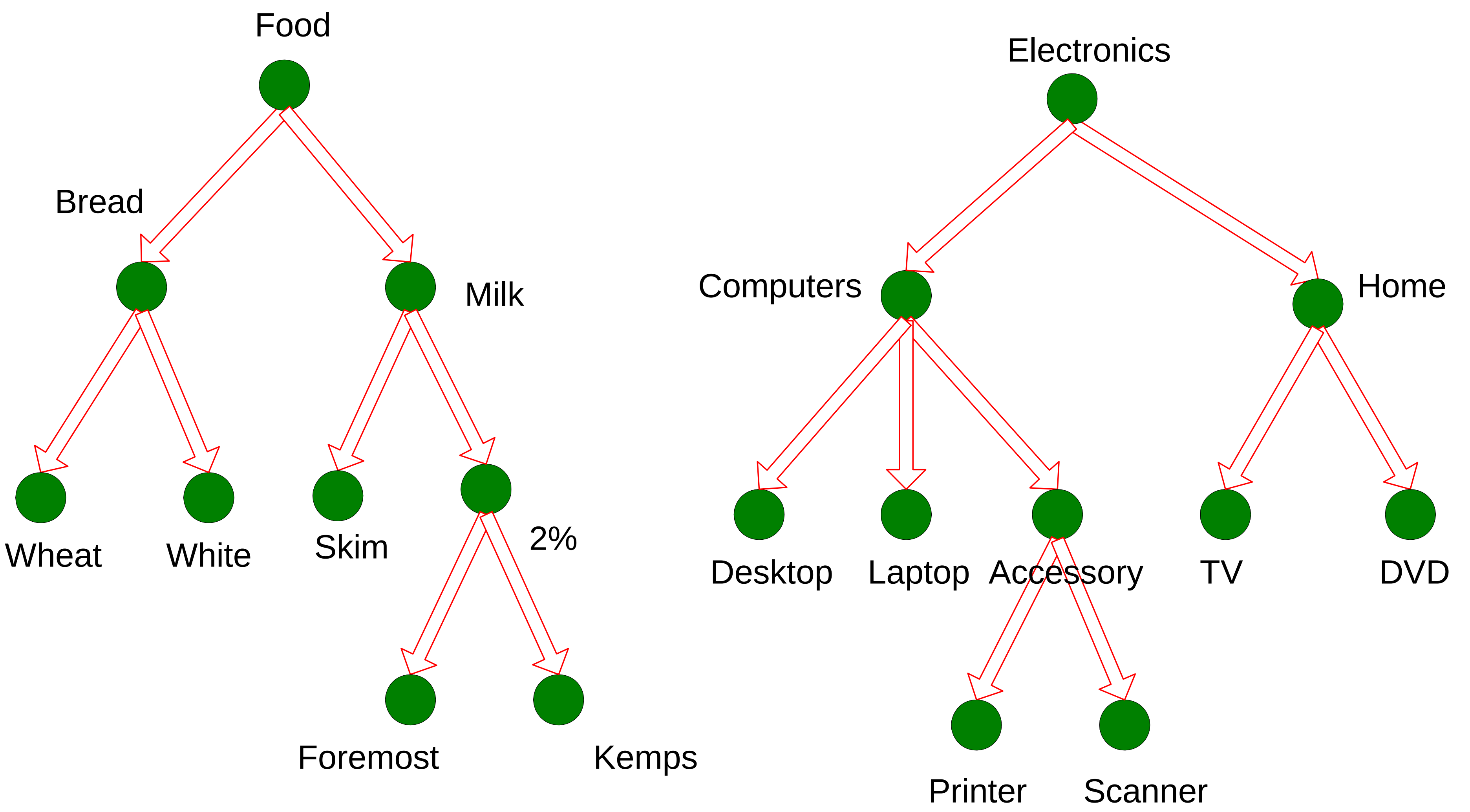
  - If interval is too small
    - may not have enough support
  - If interval is too large
    - may not have enough confidence
- Potential solution: try all possible intervals

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Handling a Concept Hierarchy



Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Where to stop??

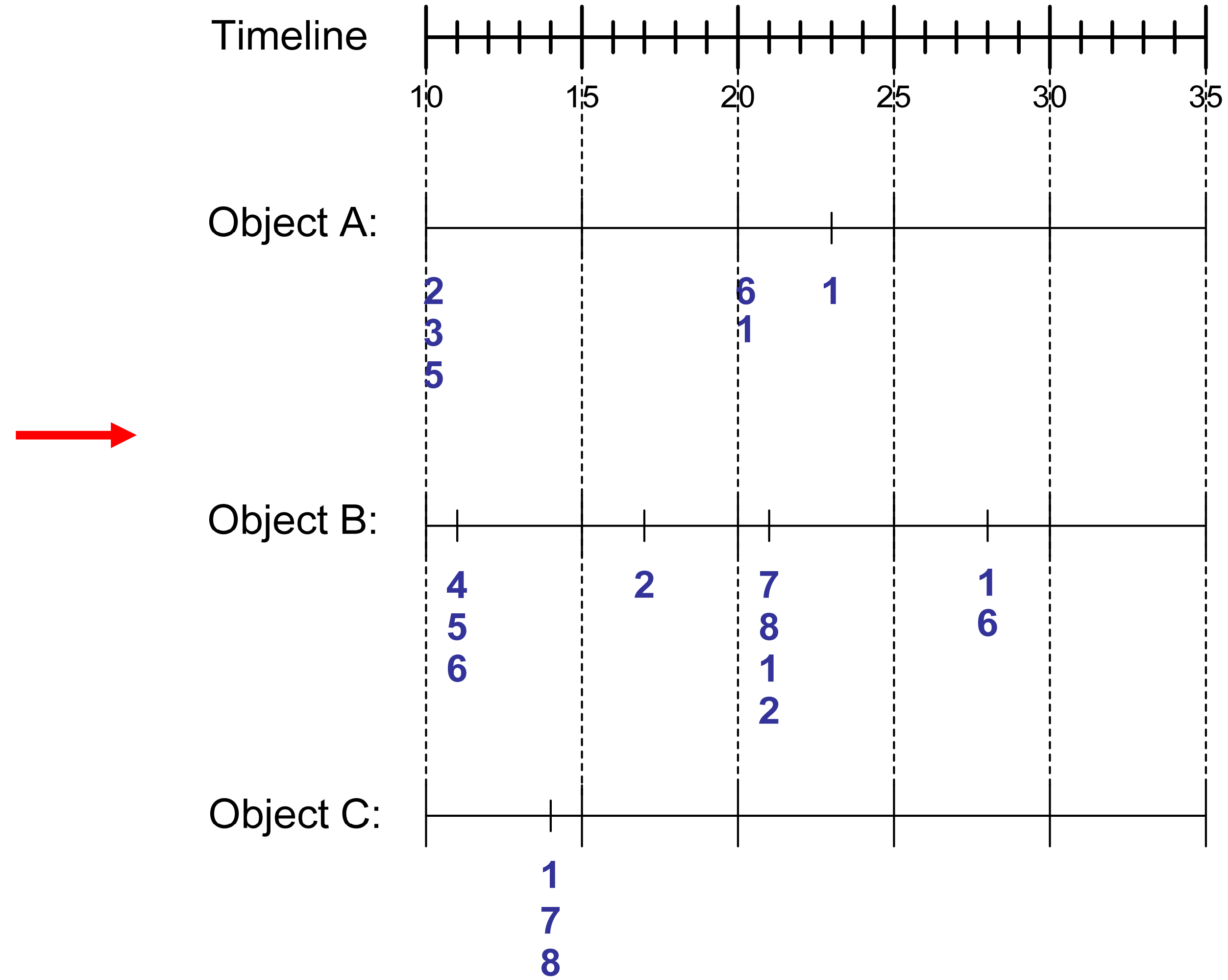
Multi-level Association Rules

- Issues with concept hierarchy:
  - Rules at lower levels may not have enough support to appear in any frequent itemsets
  - Rules at lower levels of the hierarchy are overly specific (upper level rules can cover lower level rules)
    - e.g., skim milk $\rightarrow$ white bread, 2% milk $\rightarrow$ wheat bread, skim milk $\rightarrow$ wheat bread, etc. are indicative of association between milk and bread

Sequence Data

Sequence Database:

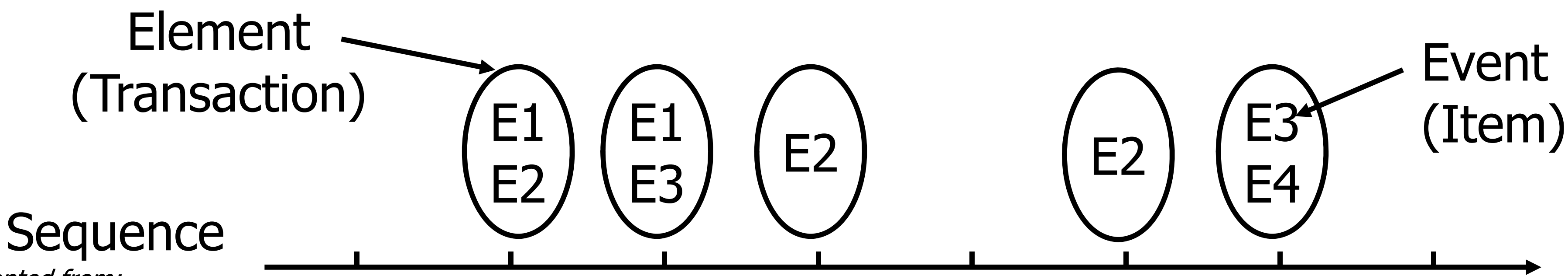
Object	Timestamp	Events
A	10	2, 3, 5
A	20	6, 1
A	23	1
B	11	4, 5, 6
B	17	2
B	21	7, 8, 1, 2
B	28	1, 6
C	14	1, 8, 7



Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Examples of Sequence Data

Sequence Database	Sequence	Element (Transaction)	Event (Item)
Customer	Purchase history of a given customer	A set of items bought by a customer at time t	Books, diary products, CDs, etc
Web Data	Browsing activity of a particular Web visitor	A collection of files viewed by a Web visitor after a single mouse click	Home page, index page, contact info, etc
Event data	History of events generated by a given sensor	Events triggered by a sensor at time t	Types of alarms generated by sensors
Genome sequences	DNA sequence of a particular species	An element of the DNA sequence	Bases A,T,G,C



Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Formal Definition of a Sequence

- A sequence is an ordered list of elements (transactions)

$$s = < e_1 \, e_2 \, e_3 \dots >$$

  - Each element contains a collection of events (items)

$$e_i = \{i_1, i_2, \dots, i_k\}$$

  - Each element is attributed to a specific time or location
- Length of a sequence, $|s|$, is given by the number of elements of the sequence
- A k-sequence is a sequence that contains k events (items)

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Examples of Sequence

- Web sequence:

  < {Homepage} {Electronics} {Digital Cameras} {Canon Digital Camera} {Shopping Cart} {Order Confirmation} {Return to Shopping} >

- Sequence of initiating events causing the nuclear accident at 3-mile Island:

  (http://stellar-one.com/nuclear/staff_reports/summary_SOE_the_initiating_event.htm)

  < {clogged resin} {outlet valve closure} {loss of feedwater} {condenser polisher outlet valve shut} {booster pumps trip} {main waterpump trips} {main turbine trips} {reactor pressure increases}>

- Sequence of books checked out at a library:

  <{Fellowship of the Ring} {The Two Towers} {Return of the King}>

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Formal Definition of a Subsequence

- A sequence $<a_1\ a_2 \ldots a_n>$ is contained in another sequence $<b_1\ b_2 \ldots b_m>$ ($m \geq n$) if there exist integers $i_1 < i_2 < \ldots < i_n$ such that $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i1}, \ldots, a_n \subseteq b_{in}$

Data sequence	Subsequence	Contain?
< {2,4} {3,5,6} {8} >	< {2} {3,5} >	Yes
< {1,2} {3,4} >	< {1} {2} >	No
< {2,4} {2,4} {2,5} >	< {2} {4} >	Yes

- The support of a subsequence w is defined as the fraction of data sequences that contain w
- A *sequential pattern* is a frequent subsequence (i.e., a subsequence whose support is ≥ *minsup*)

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Sequential Pattern Mining: Definition

- Given:
  - a database of sequences
  - a user-specified minimum support threshold, *minsup*

- Task:
  - Find all subsequences with support ≥ *minsup*

Adapted from:
Tan,Steinbach, Kumar - Introduction to Data Mining
Han, Kamber - Data Mining: Concepts and Techniques

Sequential Pattern Mining: Example

Object	Timestamp	Events
A	1	1,2,4
A	2	2,3
A	3	5
B	1	1,2
B	2	2,3,4
C	1	1, 2
C	2	2,3,4
C	3	2,4,5
D	1	2
D	2	3, 4
D	3	4, 5
E	1	1, 3
E	2	2, 4, 5

Minsup = 50%

Examples of Frequent Subsequences:

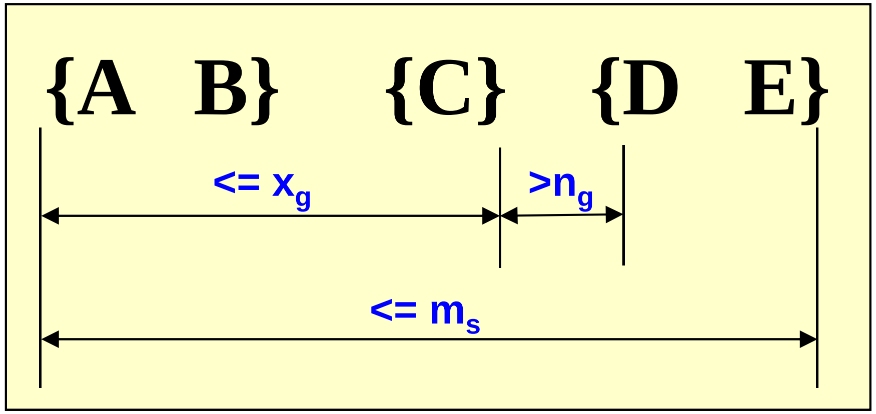
< {1,2} >	s=60%
< {2,3} >	s=60%
< {2,4}>	s=80%
< {3} {5}>	s=80%
< {1} {2} >	s=80%
< {2} {2} >	s=60%
< {1} {2,3} >	s=60%
< {2} {2,3} >	s=60%
< {1,2} {2,3} >	s=60%

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

Timing Constraints



x_g: max-gap

n_g: min-gap

m_s: maximum span

$x_g = 2$, $n_g = 0$, $m_s = 4$

Data sequence	Subsequence	Contain?
< {2,4} {3,5,6} {4,7} {4,5} {8} >	< {6} {5} >	Yes
< {1} {2} {3} {4} {5}>	< {1} {4} >	No
< {1} {2,3} {3,4} {4,5}>	< {2} {3} {5} >	Yes
< {1,2} {3} {2,3} {3,4} {2,4} {4,5}>	< {1,2} {5} >	No

Adapted from:

Tan,Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques