Association Analysis
Basic Concepts and Algorithms

INFO 523 - Lecture 8

Dr. Greg Chism

Topics

- Definition
- Mining Frequent Itemsets (APRIORI)
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- Association Rule Generation
- Support Distribution
- Pattern Evaluation

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

$\{Diaper\} \rightarrow \{Beer\},$
$\{Milk, Bread\} \rightarrow \{Eggs, Coke\},$
$\{Beer, Bread\} \rightarrow \{Milk\},$

Implication means co-occurrence, not causality!

Definition: Frequent Itemset

- **Itemset**
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- **Support count (σ)**
  - Frequency of occurrence of an itemset
  - E.g. σ({Milk, Bread, Diaper}) = 2
- **Support**
  - Fraction of transactions that contain an itemset
  - E.g. s({Milk, Bread, Diaper})
    = σ({Milk, Bread, Diaper}) / |T| = 2/5
- **Frequent Itemset**
  - An itemset whose support is greater than or equal to a *minsup* threshold

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

$$s(X) = \frac{\sigma(X)}{|T|}$$

Definition: Association Rule

- **Association Rule**
  - An implication expression of the form $X \to Y$, where X and Y are itemsets
  - Example:
    $\{Milk, Bread\} \to \{Diaper\}$

- **Rule Evaluation Metrics**
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X

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:

$\{Milk, Bread\} \to \{Diaper\}$

$$s = \frac{\sigma(\{Milk, Bread, Diaper\})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\{Milk, Bread, Diaper\})}{\sigma(\{Milk, Diaper\})} = \frac{2}{3} = 0.67$$

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{s(X \cup Y)}{s(X)}$$

Topics

- Definition
- **Mining Frequent Itemsets (APRIORI)**
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- Association Rule Generation
- Support Distribution
- Pattern Evaluation

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!

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

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

Frequent Itemset Generation
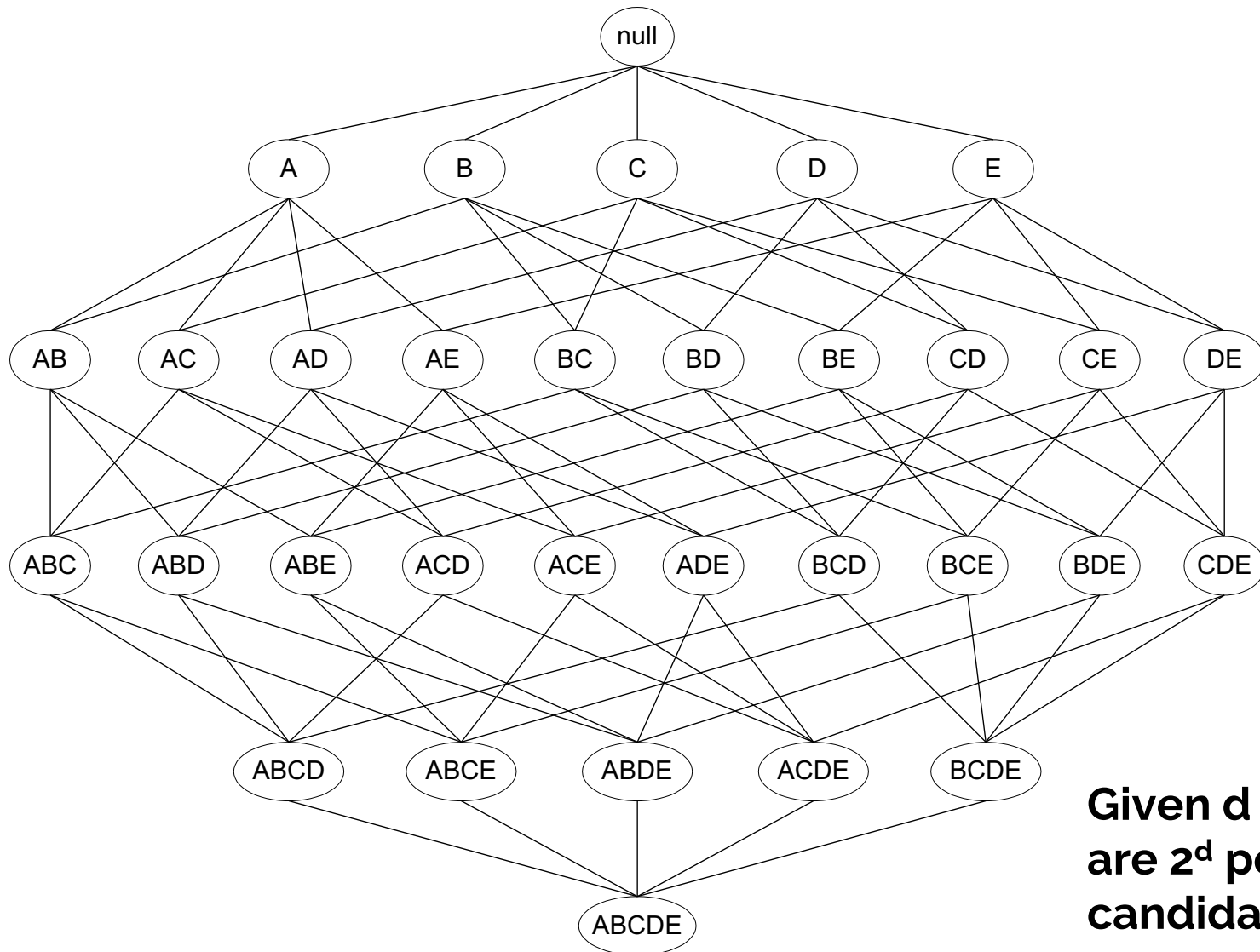
**Given d items, there are 2^d possible candidate itemsets**

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

Illustrating Apriori Principle
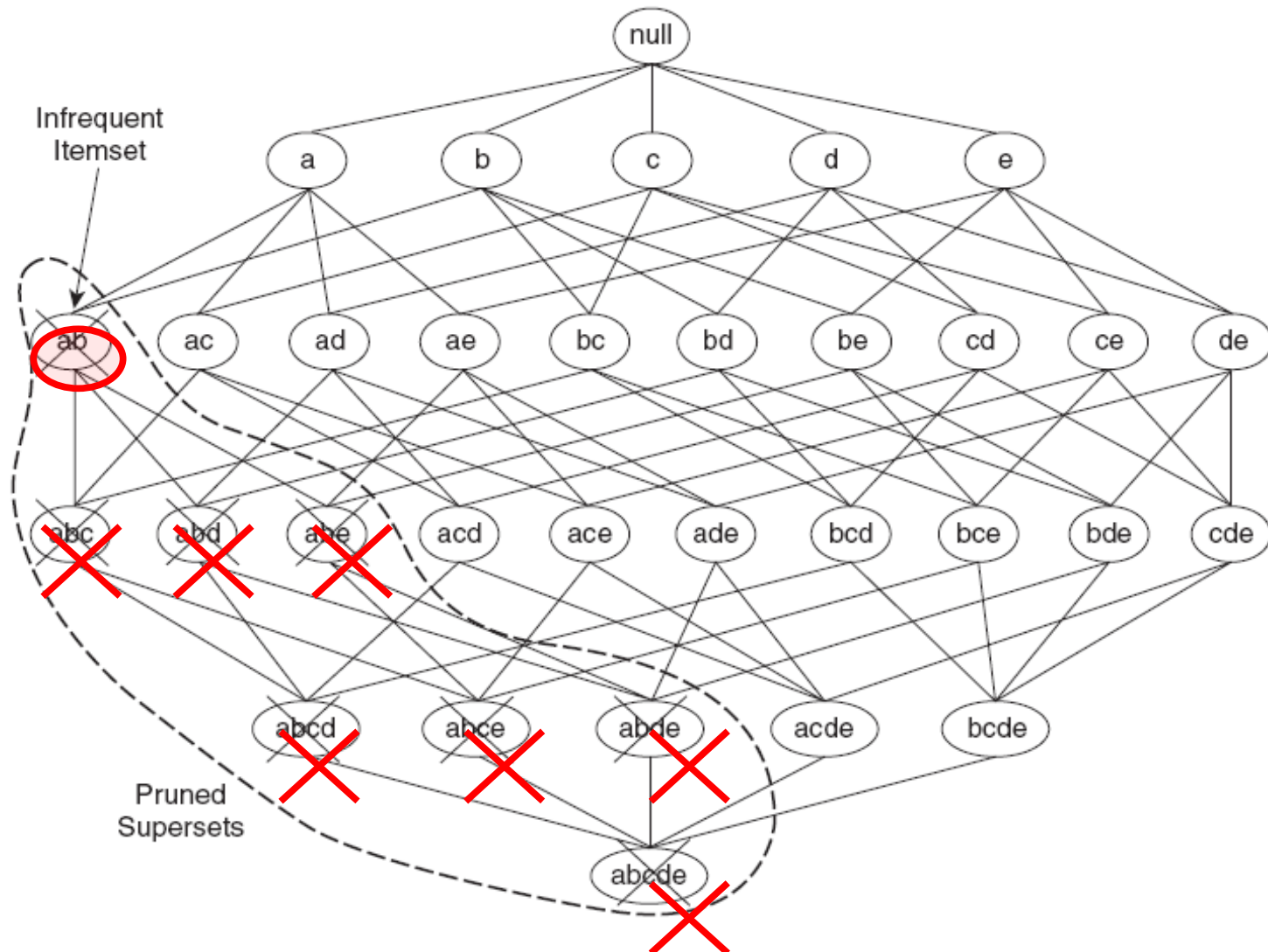
Figure 6.4. An illustration of support-based pruning. If $\{a, b\}$ is infrequent, then all supersets of $\{a, b\}$ are infrequent.

Illustrating Apriori Principle

Items (1-itemsets)

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

Pairs (2-itemsets)

(No need to generate candidates involving Coke or Eggs)

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

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,
$6 + 6 + 1 = 13$

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

Factors Affecting Complexity

- Choice of minimum support threshold
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets
- Dimensionality (number of items) of the data set
  - more space is needed to store support count of each item
  - if number of frequent items also increases, both computation and I/O costs may also increase
- Size of database
  - since Apriori makes multiple passes, run time of algorithm may increase with number of transactions
- Average transaction width
  - transaction width increases with denser data sets
  - This may increase max length of frequent itemsets and traversals of hash tree (number of subsets in a transaction increases with its width)
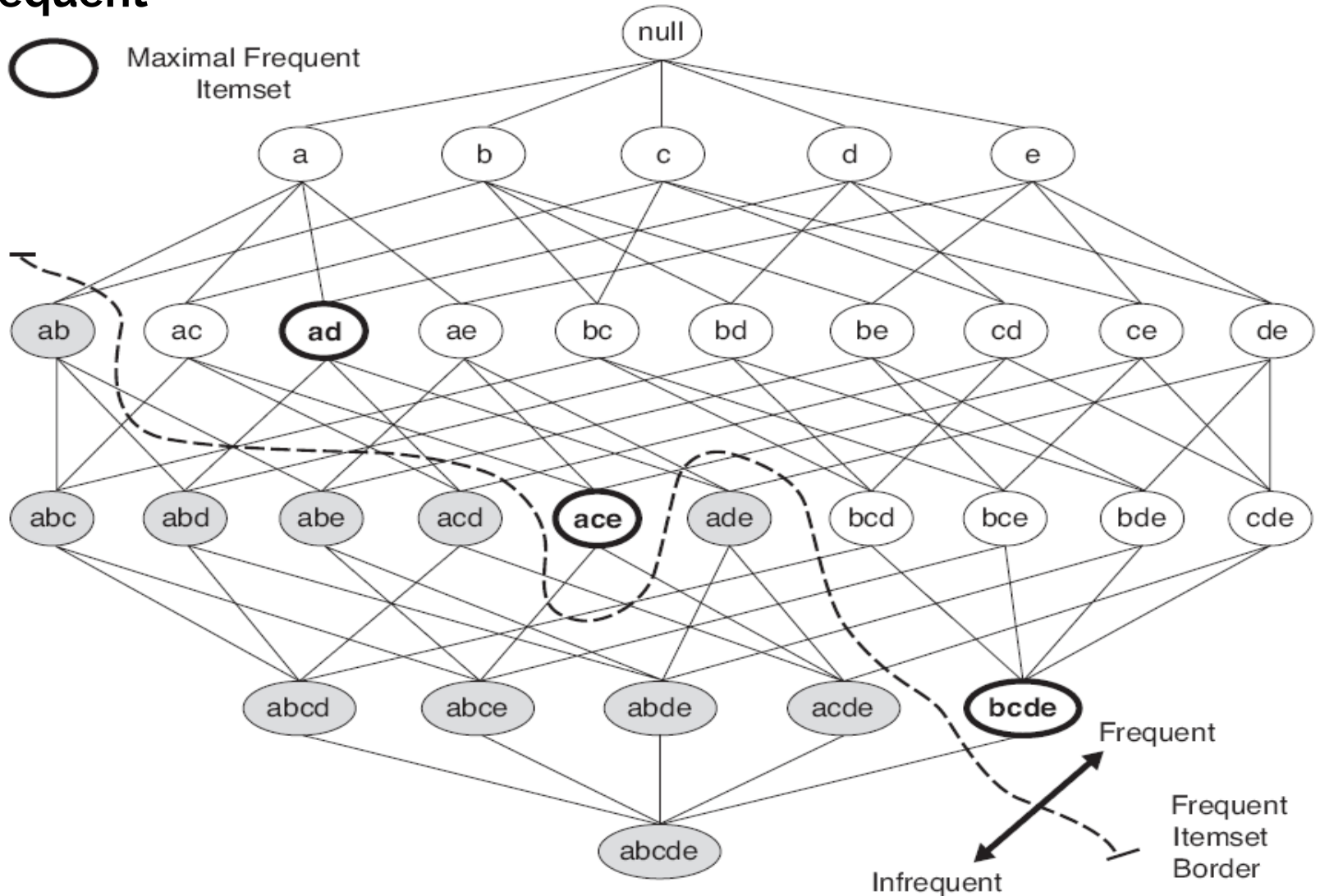
Topics

- Definition
- Mining Frequent Itemsets (APRIORI)
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- Association Rule Generation
- Support Distribution
- Pattern Evaluation

Maximal Frequent Itemset

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

Closed Itemset

- An itemset is closed if none of its immediate supersets has the same support as the itemset (can only have smaller support -> see APRIORI principle)
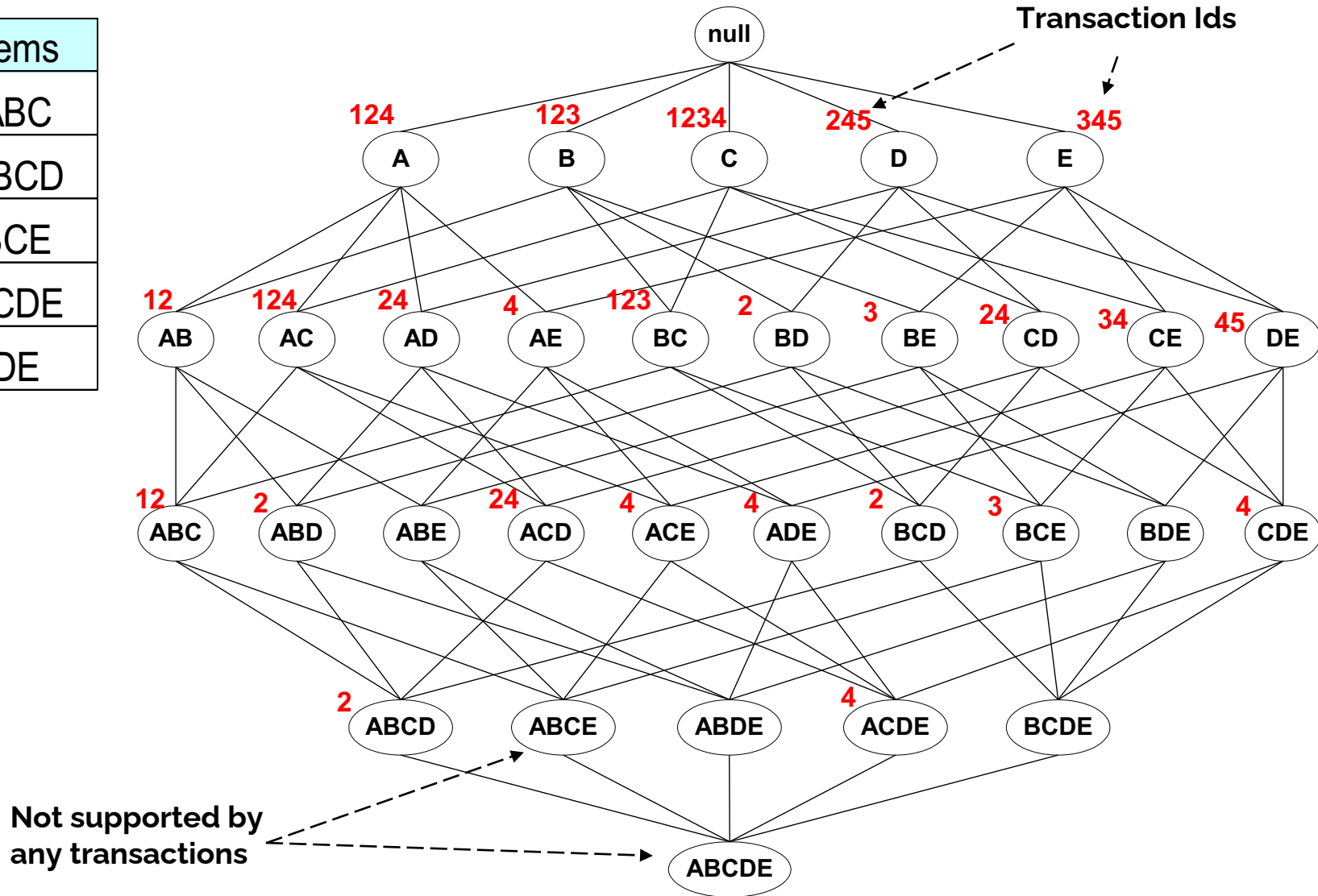
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

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

Maximal vs Closed Itemsets

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

Transaction Ids

Not supported by any transactions

Maximal vs Closed Frequent Itemsets

Minimum support = 2

Closed but not maximal

Closed and maximal

null

124 A   **123** B   **1234** C   **245** D   **345** E

12 AB   **124** AC   **24** AD   **4** AE   **123** BC   **2** BD   **3** BE   **24** CD   **34** CE   **45** DE

12 ABC   **2** ABD   ABE   **24** ACD   **4** ACE   **4** ADE   **2** BCD   **3** BCE   BDE   **4** CDE

2 ABCD   ABCE   ABDE   **4** ACDE   BCDE

ABCDE
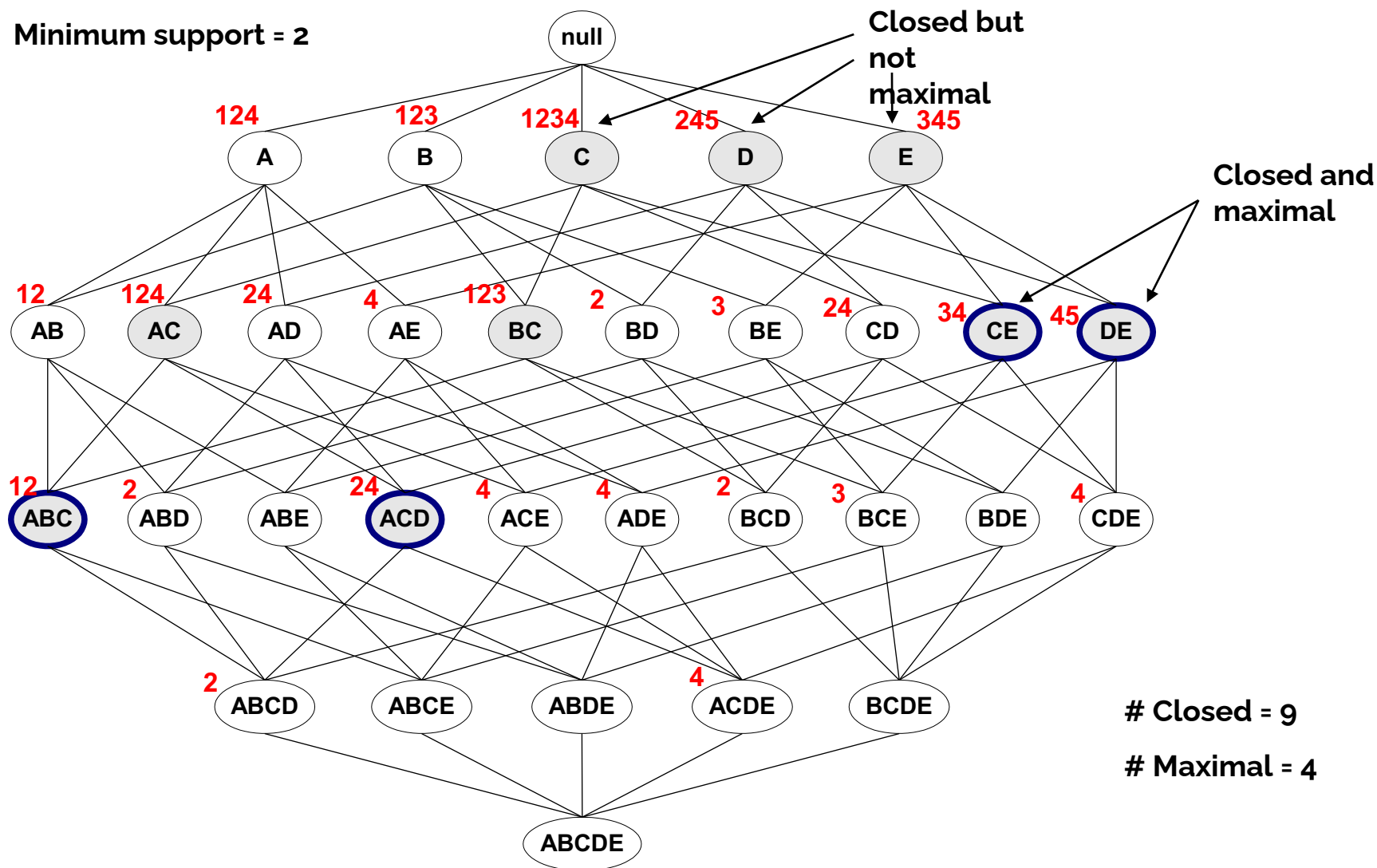
Closed = 9

# Maximal = 4

Maximal vs Closed Itemsets
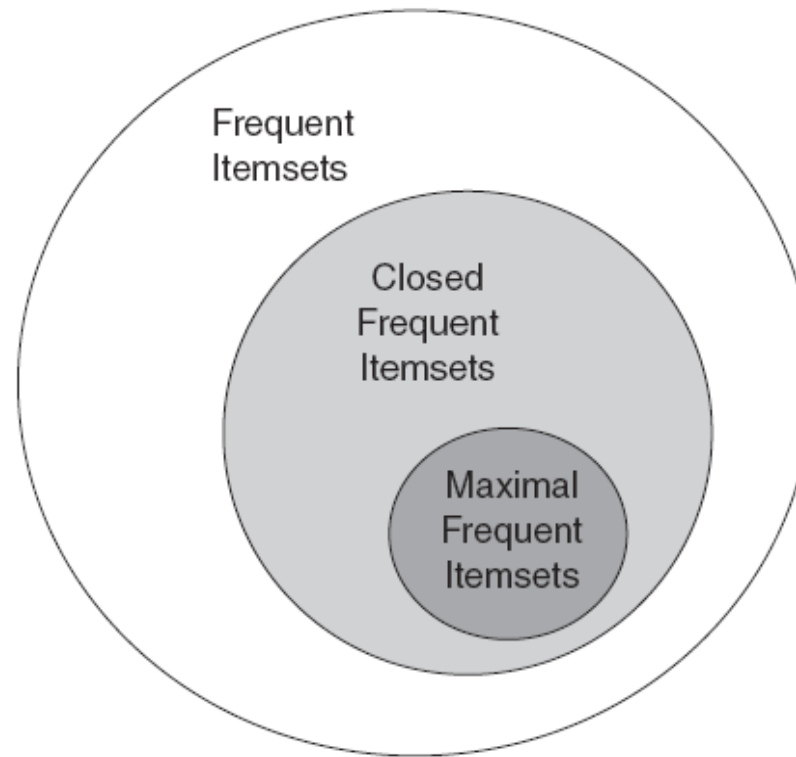
Figure 6.18. Relationships among frequent, maximal frequent, and closed frequent itemsets.
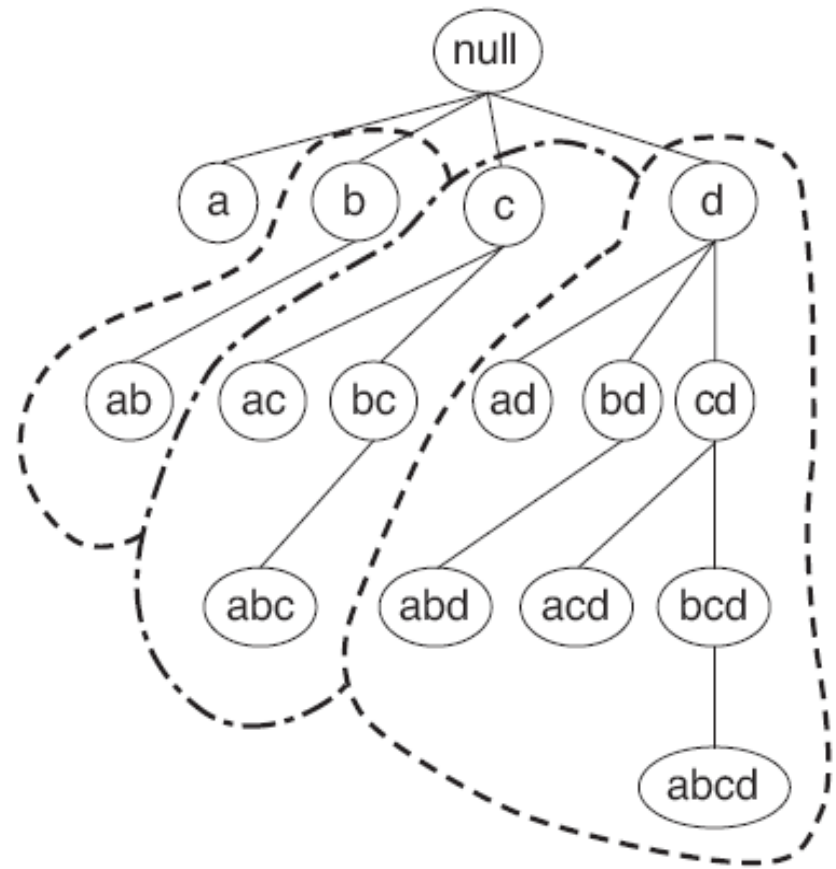
Topics

- Definition
- Mining Frequent Itemsets (APRIORI)
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- Association Rule Generation
- Support Distribution
- Pattern Evaluation

Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice
  - Equivalent Classes
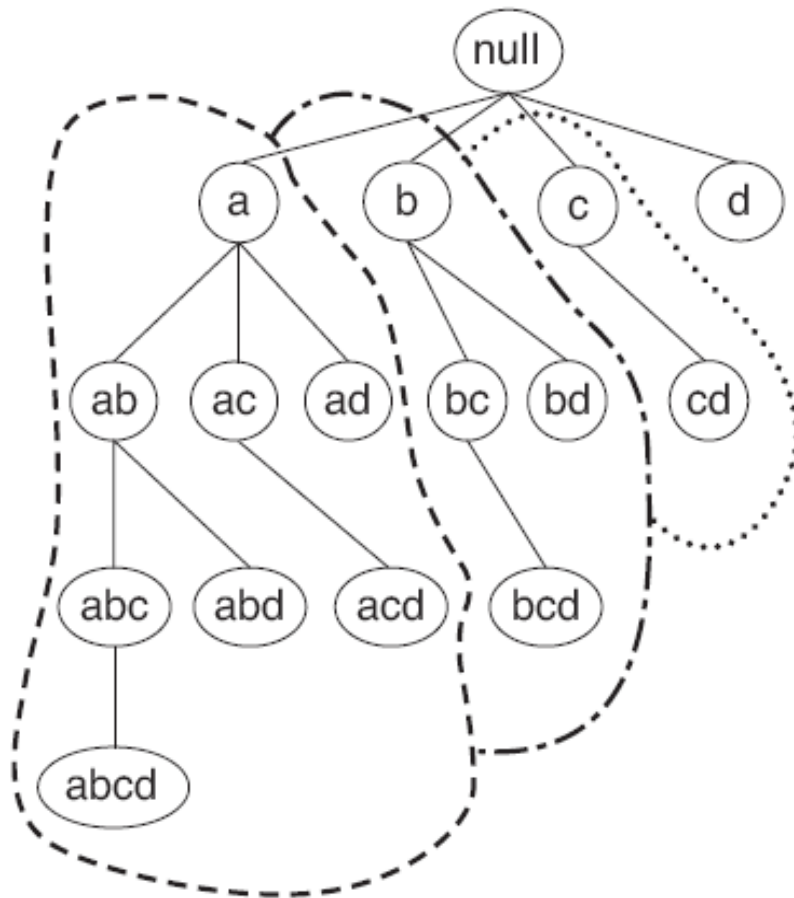
(a) Prefix tree.

(b) Suffix tree.

Alternative Methods for Frequent Itemset Generation

- Representation of Database: horizontal vs vertical data layout

Horizontal Data Layout

TID	Items
1	a,b,e
2	b,c,d
3	c,e
4	a,c,d
5	a,b,c,d
6	a,e
7	a,b
8	a,b,c
9	a,c,d
10	b

Vertical Data Layout

a	b	c	d	e
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

Alternative Algorithms

- FP-growth
  - Use a compressed representation of the database using an FP-tree
  - Once an FP-tree has been constructed, it uses a recursive divide-and-conquer approach to mine the frequent itemsets
- ECLAT
  - Store transaction id-lists (vertical data layout).
  - Performs fast tid-list intersection (bit-wise XOR) to count itemset frequencies

Topics

- Definition
- Mining Frequent Itemsets (APRIORI)
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- ~~Association Rule Generation~~
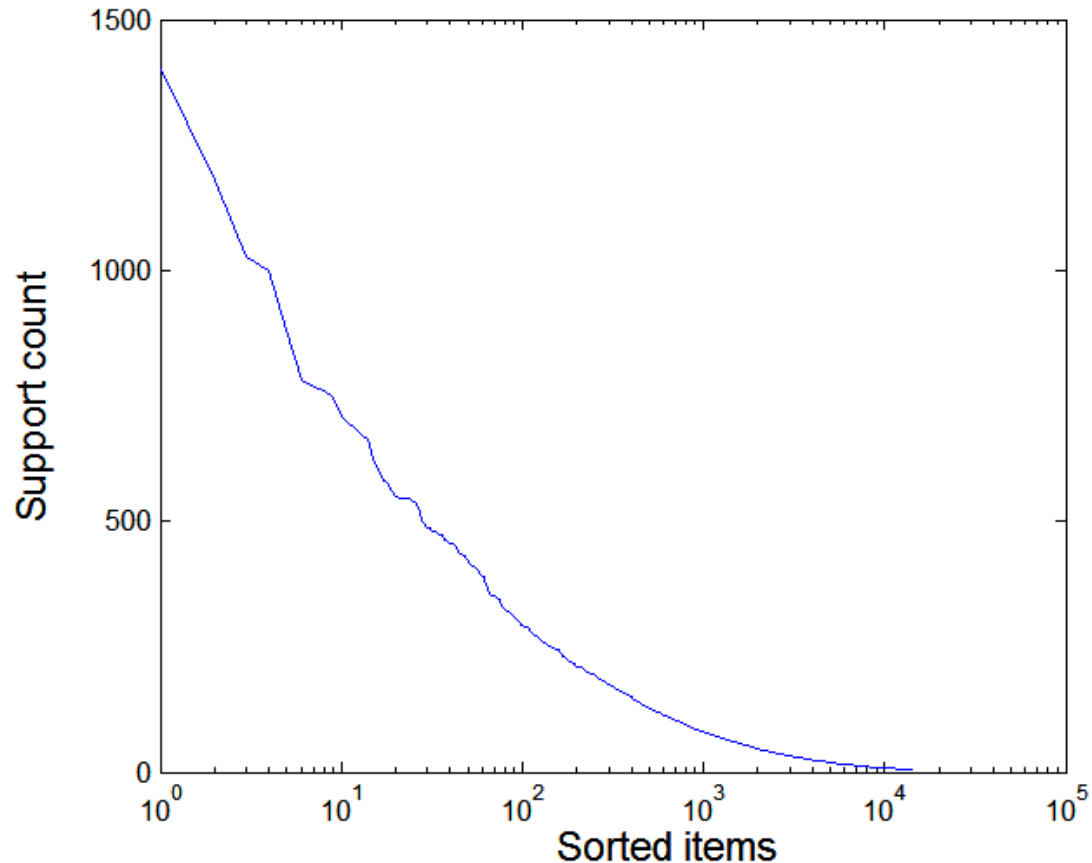- Support Distribution
- Pattern Evaluation

Topics

- Definition
- Mining Frequent Itemsets (APRIORI)
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- Association Rule Generation
- Support Distribution
- Pattern Evaluation

Effect of Support Distribution

- Many real data sets have skewed support distribution

Support distribution of a retail data set

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

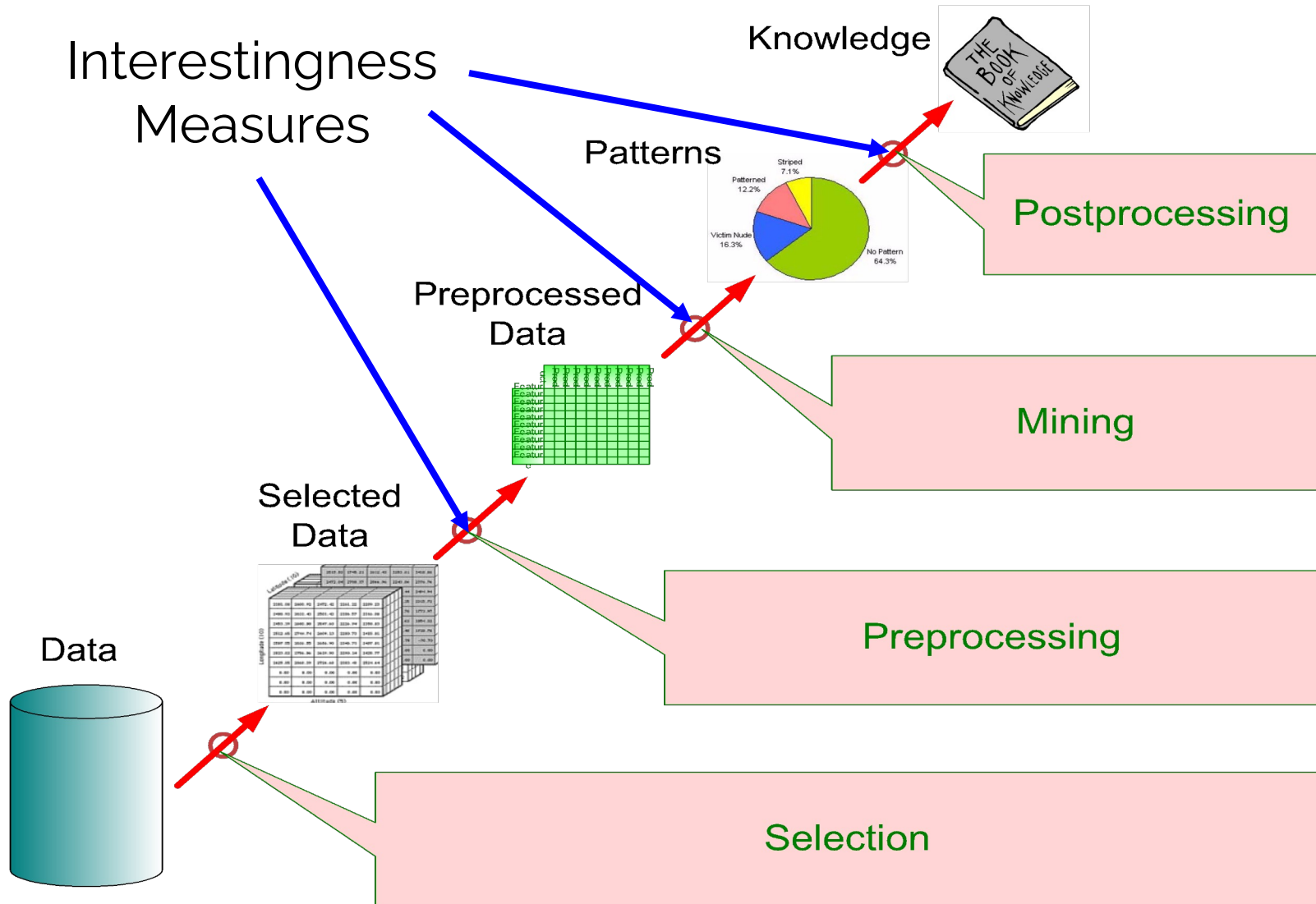- Using a single minimum support threshold may not be effective

Topics

- Definition
- Mining Frequent Itemsets (APRIORI)
- Concise Itemset Representation
- Alternative Methods to Find Frequent Itemsets
- Association Rule Generation
- Support Distribution
- Pattern Evaluation

Pattern Evaluation

- Association rule algorithms tend to produce **too many rules.** Many of them are
  - uninteresting or
  - redundant

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

- A rule {A,B,C} $\rightarrow$ {D} can be considered **redundant** if {A,B} $\rightarrow$ {D} has the same or higher confidence.

Application of Interestingness Measure

Interestingness Measures

Knowledge

Patterns

Preprocessed Data

Selected Data

Data

Postprocessing

Mining

Preprocessing

Selection

Statistical Independence

- Population of 1000 students
  - 600 students know how to swim (S)
  - 700 students know how to bike (B)
  - 450 students know how to swim and bike (S,B)

  - $P(S,B) = 450/1000 = 0.45$ (observed joint prob.)
  - $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$ (expected under indep.)

  - $P(S,B) = P(S) \times P(B) \Rightarrow$ Statistical independence
  - $P(S,B) > P(S) \times P(B) \Rightarrow$ Positively correlated
  - $P(S,B) < P(S) \times P(B) \Rightarrow$ Negatively correlated

Statistical-based Measures

- Measures that take statistical dependence into account for rule: $X \rightarrow Y$

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

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

Deviation from independence

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

Correlation

Example: Lift/Interest

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

Association Rule: Tea → Coffee

Conf(Tea → Coffee)= P(Coffee|Tea) = P(Coffee,Tea)/P(Tea)

= .15/.2 = 0.75

but P(Coffee) = 0.9

⇒ Lift(Tea → Coffee) = P(Coffee,Tee)/(P(Coffee)P(Tee))

= .15/(.9 x .2) = **0.8333**

Note: Lift < 1, therefore Coffee and Tea are negatively associated

Many measures have been 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?

What about Apriori-style support-based pruning? How does it affect these measures?

Source: The list is from Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. Information Systems, 29(4):293--313, 2004.

A larger list of measures is available at: A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules

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

Support-based Pruning

- Most of the association rule mining algorithms use support measure to prune rules and itemsets

- Study effect of support pruning on correlation of itemsets
  - Generate 10,000 random contingency tables
  - Compute support and pairwise correlation for each table
  - Apply support-based pruning and examine the tables that are removed

Effect of Support-based Pruning
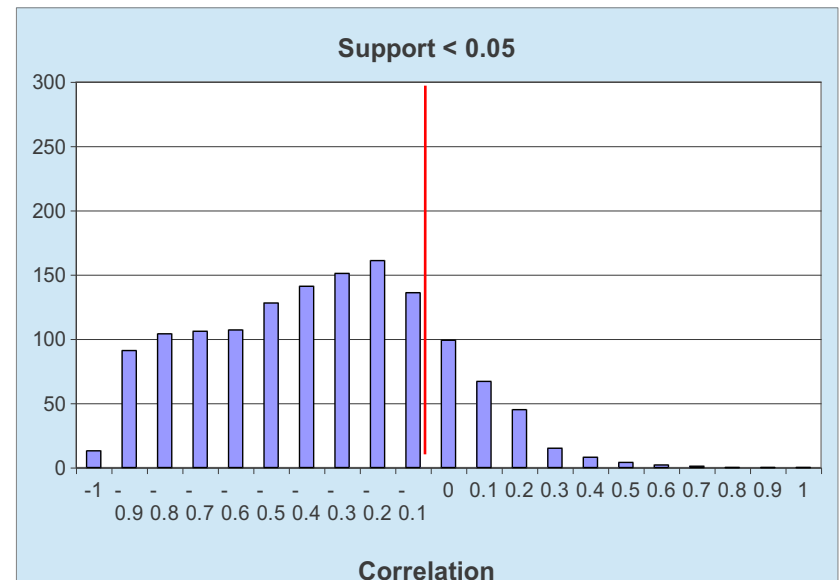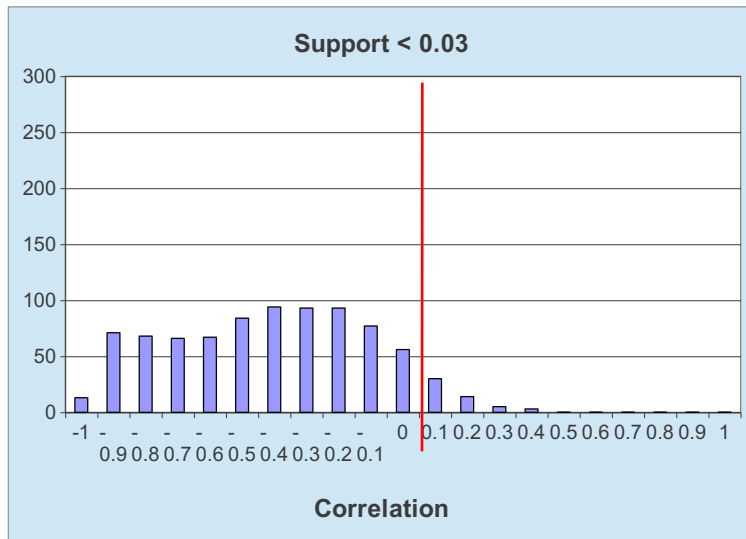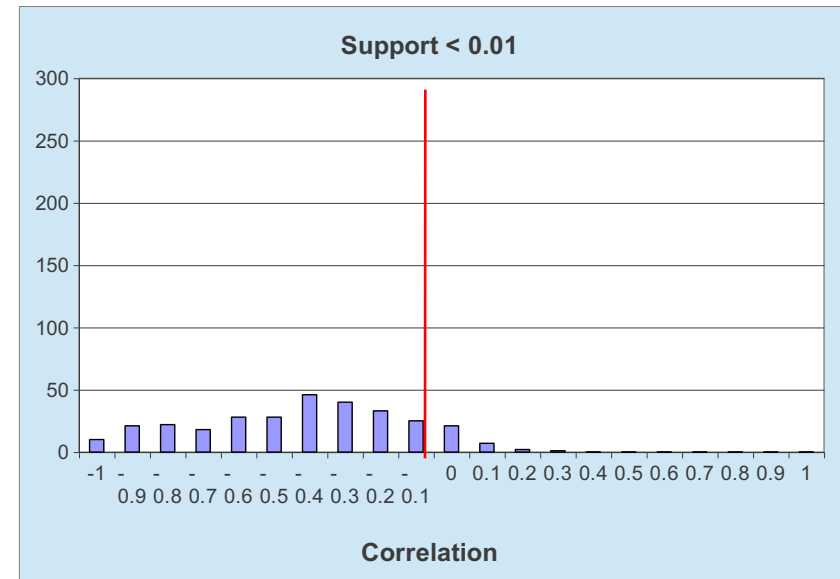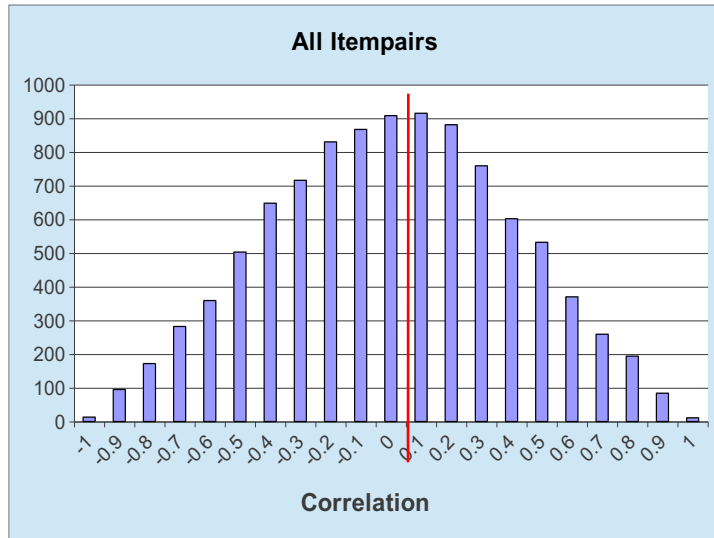
Support-based pruning eliminates mostly negatively correlated itemsets

Applications for Association Rules

- **Market Basket Analysis**
  Marketing & Retail. E.g., frequent itemsets give information about "other customer who bought this item also bought X"

- **Exploratory Data Analysis**
  Find correlation in very large (= many transactions), high-dimensional (= many items) data

- **Intrusion Detection**
  Rules with low support but very high lift

- **Build Rule-based Classifiers**
  Class association rules (CARs)