Outline

- Basic concepts of association rule learning
- Apriori algorithm
- FP-Growth Algorithm
- Finding interesting rules

64

Two Problems with Association Rule Mining

- Quantity problem
  - Too many rules can be generated
    - Given a dataset, the number of rules generated depends on the support and confidence thresholds.
      - If the support threshold is high, a small number of rules are generated. But some interesting rules are missed.
      - If the support threshold is low, a huge number of rules are generated.
- Quality problem
  - Not all the generated rules are interesting

65

Number of Generated Patterns versus Support Threshold (An Example)

Support threshold	0.02	0.01	0.008	0.005	0.003	0.0028	0.0025	0.002	0.001
Num. of rules (conf. thres.=0.5)	2	14	39	88	723	4,556	74,565	4,800,070	$>10^9$
Num. of rules (conf. thres.=0.8)	1	7	17	38	591	4,172	65,615	3,584,339	$>10^9$

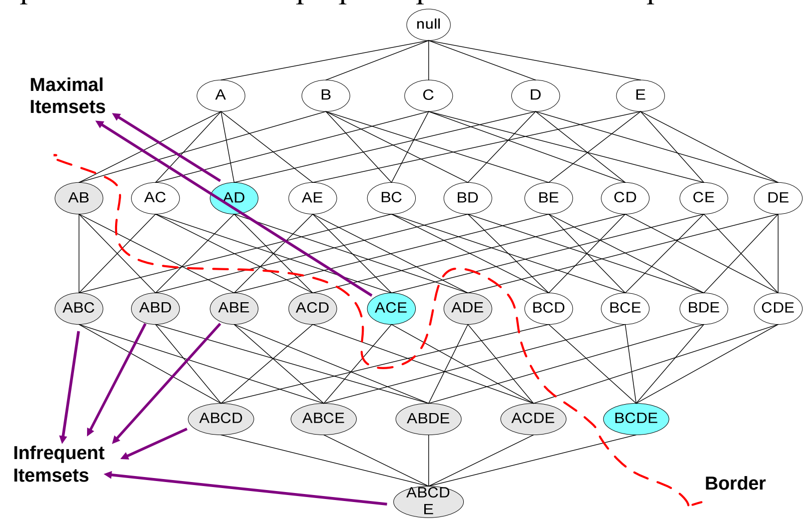
Number of sessions (transactions): 30586
Number of objects (items): 38679

Solutions to the Problems

- Finding only *maximum* or *closed* frequent patterns
  - Other frequent patterns can be generated from them
- Constraint-based data mining
  - Applying constraints in the mining process so the search can be more focused.
- Using interestingness measures to remove or rank rules
  - Remove misleading associations and find correlation rules
  - Prune patterns using other interestingness measures
- Using rule structures
  - Eliminate structurally and semantically redundant rules.
  - Group or summarize related rules

Maximal Frequent Itemset

An itemset *X* is a ***maximal frequent itemset*** in a data set *D* if *X* is frequent and none of the proper super-set of *X* is frequent in *D*.



Maximal Frequent Patterns

- Reducing the # of patterns returned to the user
- Maximal frequent patterns are a *lossy* compression of frequent patterns
  - Given the set of all maximal frequent patterns and their supports in a data set *D*, we can generate all the frequent patterns, *but not their supports*.
- Algorithm for mining maximal frequent itemsets: MaxMiner
  - R. Bayardo. Efficiently mining long patterns from databases. *SIGMOD'98*

69

Closed Patterns

- Problem with maximal frequent itemsets:
  - Supports of their subsets are not known – additional DB scans are needed (to get the supports)
- An itemset is ***closed*** if none of its proper supersets has the same support as the itemset

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

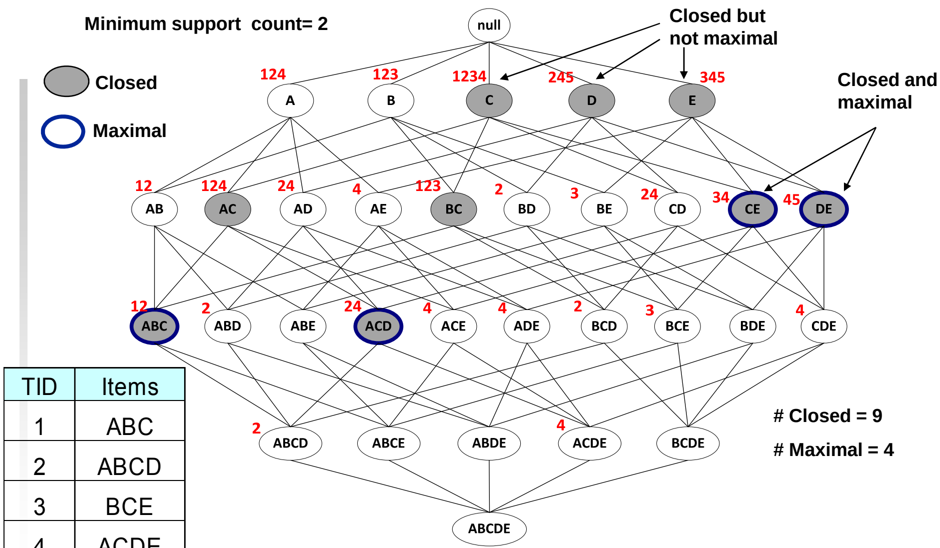
Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	2
{A,B,C,D}	2

Closed Frequent Patterns

- An itemset *X* is a ***closed frequent itemset*** in a data set *D* if *X* is both *closed* and *frequent* in *D* with respect to a support threshold.
- Closed frequent itemsets are a *lossless* compression of frequent patterns
  - Reducing the # of patterns returned to the user
  - Given the set of all closed frequent patterns and their supports in a data set *D*, the user can generate all the frequent patterns and their supports.
- Algorithm for finding closed frequent patterns: CLOSET
  - J. Pei, J. Han & R. Mao. “CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets", DMKD'00.

71

Maximal vs Closed Frequent Itemsets



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

Closed Patterns and Max-Patterns

- Exercise. DB = $\{\{a_1, \ldots, a_{100}\}, \{a_1, \ldots, a_{50}\}\}$
  - Min_sup_count = 1.
- What is the set of closed frequent itemsets?
  - $\{a_1, \ldots, a_{100}\}$: 1
  - $\{a_1, \ldots, a_{50}\}$: 2
- What is the set of maximal frequent itemsets?
  - $\{a_1, \ldots, a_{100}\}$: 1
- What is the set of all frequent itemsets?
  - !!

73

Solutions to the Problems

- Finding only *maximum* or *closed* frequent patterns
  - Other frequent patterns can be generated from them
- *Constraint-based data mining*
  - *Applying constraints in the mining process so the search can be more focused.*
- Using interestingness measures to remove or rank rules
  - Remove misleading associations and find correlation rules
  - Prune patterns using other interestingness measures
- Using rule structures
  - Eliminate structurally and semantically redundant rules.
  - Group or summarize related rules

74

Constrain-based Frequent Pattern Mining

- Mining frequent patterns with constraint C
  - find all patterns satisfying not only min_sup, but also constraint C
- Examples of Constraints
  - *?* → *a particular product*
  - *a particular product* → *?*
  - small sales (price < \$10) triggers big sales (sum > \$200)

75

Constrain-based Frequent Pattern Mining (Cont'd)

- A naïve solution
  - Testing frequent patterns on C as a post-processing process
- Some constraints can be incorporated into the mining process to improve the efficiency
- More efficient approaches
  - Analyze the properties of constraints comprehensively
  - Push the constraint as deeply as possible inside the frequent pattern mining
  - Example: find all frequent itemsets containing item "b"

Types of Constraints

- Anti-monotonic constraints
  - An itemset S satisfies the constraint, so does any of its subset (That is, S violates the constraint, so does any of its superset).
- Monotonic constraints
  - An itemset S satisfies the constraint, so does any of its superset
- Examples
  - Sum of the prices of items in S ≤ 100 is anti-monotone
  - Maximum price in S ≤ 15 is anti-monotone
  - Sum of the prices of items in S ≥ 100 is monotone
  - Minimum price in S ≤ 15 is monotone

How to Use Antimonotonic or Monotonic Constraints in Mining

- Antimonotonic constraints
  - In Apriori:
    - Use it to prune candidates in each iteration
  - In FP-growth
    - Use it to stop growing a pattern
- Monotonic constraints
  - If an itemset satisfies a monotonic constraint, no need to check its supersets on the constraint
    - Only checks their support

78

Types of Constraints (Cont'd)

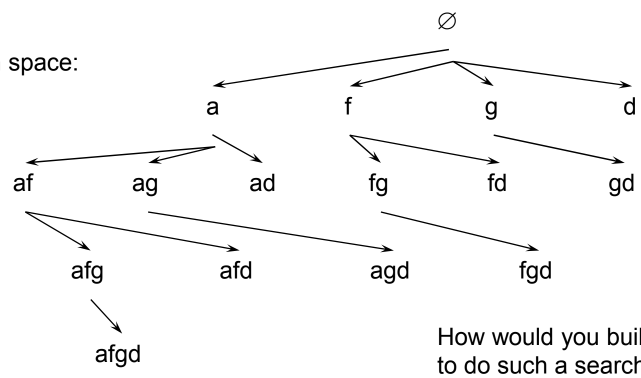
- Convertible constraints
  - Some constraints are not anti-monotonic or monotonic
  - But can be converted to anti-monotonic or monotonic by properly ordering items
- Example of convertible constraint:
  - Average price of the items in $S \geq 25$
  - Order items in price-descending order
    - <a, f, g, d, b, h, c, e>
  - If an itemset afb violates C
    - So does afbh, afb*
    - It becomes anti-monotone!

79

Example of Convertible Constraints

- Convertible constraint:
  - Average price of the items in S $\geq$ 25
- Price-descending order of items: a, f, g, d

Search space:



How would you build an FP-tree to do such a search?

Solutions to the Problems

- Finding only *maximum* or *closed* frequent patterns
  - Other frequent patterns can be generated from them
- Constraint-based data mining
  - Applying constraints in the mining process so the search can be more focused.
- *Using interestingness measures to remove or rank rules*
  - *Remove misleading associations and find correlation rules*
  - Prune patterns using other interestingness measures
- Using rule structures
  - Eliminate structurally and semantically redundant rules.
  - Group or summarize related rules

Misleading Association Rules

	Basketball	Not basketball	Sum (row)
Cereal	2000	1750	3750
Not cereal	1000	250	1250
Sum(col.)	3000	2000	5000

- *play basketball* $\Rightarrow$ *eat cereal* [40%, 66.7%] is misleading
  - The overall percentage of students eating cereal is 75% which is higher than 66.7%
- *play basketball* $\Rightarrow$ *not eat cereal* [20%, 33.3%] is more accurate, although with lower support and confidence

Association ≠ Correlation

82

Interestingness Measure: Correlation

- Correlation
  - If $P(A|B)>P(A)$, A and B are *positively correlated.*

    Note: $P(A \mid B) > P(A) \Leftrightarrow P(B \mid A) > P(B) \Leftrightarrow P(A\ B) > P(A)P(B)$
  - If $P(A|B)<P(A)$, A and B are *negatively correlated.*

    Note: $P(A \mid B) < P(A) \Leftrightarrow P(B \mid A) < P(B) \Leftrightarrow P(A\ B) < P(A)P(B)$
  - If $P(A|B)=P(A)$, A and B are *independent.*

    Note: $P(A \mid B) = P(A) \Leftrightarrow P(B \mid A) = P(B) \Leftrightarrow P(A\ B) = P(A)P(B)$
- A measure of correlation (called lift)

$$corr_{A,B} = \frac{P(AB)}{P(A)P(B)}$$

83

Pruning Misleading Rules (Keep Correlation Rules)

- A measure of correlation (lift) for rule A→ B

$$lift(A \rightarrow B) = \frac{P(AB)}{P(A)P(B)}$$

- Rules whose lift ≤ 1 is ***misleading***, which should be removed
  - E.g. the following rule:

    play basketball ⇒ *eat cereal* [40%, 66.7%]

    should be removed because its lift is 0.89

84

Solutions to the Problems

- Finding only *maximum* or *closed* frequent patterns
  - Other frequent patterns can be generated from them
- Constraint-based data mining
  - Applying constraints in the mining process so the search can be more focused.
- *Using interestingness measures to remove or rank rules*
  - Remove misleading associations and find correlation rules
  - *Prune patterns using other interestingness measures*
- Using rule structures
  - Eliminate structurally and semantically redundant rules.
  - Group or summarize related rules

85

Many interestingness measures for A→ B

symbol	measure	range	formula
ϕ	ϕ-coefficient	-1…1	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
Q	Yule's Q	-1 …1	$\frac{P(A,B)P(\overline{A},\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A},\overline{B})+P(A,\overline{B})P(\overline{A},B)}$
Y	Yule's Y	-1 …1	$\frac{\sqrt{P(A,B)P(\overline{A},\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A},\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}$
k	Cohen's	-1 …1	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
PS	Piatetsky-Shapiro's	-0.25 …0.25	$P(A,B)-P(A)P(B)$
F	Certainty factor	-1 …1	$\max(\frac{P(B\mid A)-P(B)}{1-P(B)}, \frac{P(A\mid B)-P(A)}{1-P(A)})$
AV	added value	-0.5 …1	$\max(P(B\mid A)-P(B), P(A\mid B)-P(A))$
K	Klosgen's Q	-0.33 …0.38	$\sqrt{P(A,B)}\max(P(B\mid A)-P(B), P(A\mid B)-P(A))$
g	Goodman-kruskal's	0 …1	$\frac{\Sigma_j \max_k P(A_j,B_k)+\Sigma_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-max_k P(B_k)}$
M	Mutual Information	0 …1	$\frac{\Sigma_i \Sigma_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\Sigma_i P(A_i)\log P(A_i)\log P(A_i), -\Sigma_i P(B_i)\log P(B_i)\log P(B_i))}$
J	J-Measure	0 …1	$\max(P(A,B)\log(\frac{P(B\mid A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\mid A)}{P(\overline{B})}))$ $P(A,B)\log(\frac{P(A\mid B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\mid B)}{P(\overline{A})})$
G	Gini index	0 …1	$\max(P(A)[P(B\mid A)^2+P(\overline{B}\mid A)^2]+P(\overline{A}[P(B\mid\overline{A})^2+P(\overline{B}\mid\overline{A})^2]-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\mid B)^2+P(\overline{A}\mid B)^2]+P(\overline{B}[P(A\mid\overline{B})^2+P(\overline{A}\mid\overline{B})^2]-P(A)^2-P(\overline{A})^2)$
s	support	0 …1	$P(A,B)$
c	confidence	0 …1	$max(P(B\mid A), P(A\mid B))$
L	Laplace	0 …1	$\max(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2})$
IS	Cosine	0 …1	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
γ	coherence(Jaccard)	0 …1	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
α	all_confidence	0 …1	$\frac{P(A,B)}{\max(P(A),P(B))}$
o	odds ratio	0 …∞	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(\overline{A},B)P(A,\overline{B})}$
V	Conviction	0.5 …∞	$\max(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})})$
λ	lift	0 …∞	$\frac{P(A,B)}{P(A)P(B)}$
S	Collective strength	0 …∞	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
χ^2	χ^2	0 …∞	$\Sigma_i\frac{(P(A_i)-E_i)^2}{E_i}$

Pruning rules with interestingness measure

- Choose a measure in your belief to assess the significance of a rule A→ B.
- Rank the rules according to their interestingness value.
- Remove rules with small interestingness values

Solutions to the Problems

- Finding only *maximum* or *closed* frequent patterns
  - Other frequent patterns can be generated from them
- Constraint-based data mining
  - Applying constraints in the mining process so the search can be more focused.
- Using interestingness measures to remove or rank rules
  - Remove misleading associations and find correlation rules
  - Prune patterns using other interestingness measures
- *Using rule structures to prune rules*
  - Eliminate structurally and semantically redundant rules.
  - Group or summarize related rules

88

Pruning Redundant Rules

- **Pruning Rule 1:** If there are two rules of the form $A \rightarrow C$ and $A \wedge B \rightarrow C$, and the interestingness value of rule $A \wedge B \rightarrow C$ is not significantly better than rule $A \rightarrow C$, then rule $A \wedge B \rightarrow C$ is redundant and should be pruned.

- **Pruning Rule 2:** If there are two rules of the form $A \rightarrow C_1$ and $A \rightarrow C_1 \wedge C_2$, and the interestingness value of rule $A \rightarrow C_1$ is not significantly better than rule $A \rightarrow C_1 \wedge C_2$, then rule $A \rightarrow C_1$ is redundant and should be pruned.

89

Summarizing and Grouping Association Rules

- Toivonen et al. (KDD'95)
  - Compute a subset of rules, called a structural rule cover, to reduce the number of rules and further grouped the rules in the cover using clustering
- Cristofor and Simovici (2002)
  - Define another type of rule cover, called informative cover, to group and summarize related rules.
- Khan, An and Huang (ICDM'03)
  - Proposed two algorithms
    - Objective grouping of rules according to the rule structure
    - Subjective grouping of rules according to the semantic relationship among items.

90

Related Topics

- Mining high utility patterns
  - Consider
    - the quantity $q(i, T_j)$ of an item i in a transaction T_j
    - the value (e.g., price $p(i)$) of an item i
  - Utility of an item i in a transaction T_j:

$$u(i, T_j) = q(i, T_j) \times p(i)$$

  - Utility of an itemset X in a transaction T_j:

$$(\quad)$$

  - Utility of an itemset X in a dataset D:

$$(\quad)$$

  - *High utility pattern*: itemsets whose utility in the dataset is no less than a minimum utility threshold

91

Related Topics

- Mining high utility patterns (*cont'd*)
  - Challenge: utility does not have the downward closure property. That is,

    The utility of a subset /superset of a set S may be smaller or larger than the utility of S

  - This means we cannot use Apriori or FP-growth to find high utility patterns directly since
    - the two algorithms use the downward closure property of support to cut down the search space
  - Solution: use an upper bound of utility with downward closure property to generate candidates first, and then scan DB to find high utility patterns from the set of candidates

92

Related Topics

- Mining frequent patterns over data streams
  - A continuous flow of data generated often at high-speed in a dynamic, time-changing environment
  - Memory is limited to hold all the data
  - Processing time may be limited by the rate of arrival of instances
  - One scan of data set is required for online mining
  - Pattern changes over time
    - Incremental learning
    - Change detection
    - etc

93

Related Topics

- Contrast pattern mining
  - Finding patterns and models contrasting two or more classes or conditions
  - Contrasting groups:
    - Objects at different time periods
    - Objects at different spatial locations
    - Objects across different classes.
  - Measures for measuring the difference
    - Frequent/infrequent
    - Frequency ratio
    - Odds ratio, etc.
  - A challenge: need to find infrequent itemsets in a group.

94

Next Class

- Sequential pattern mining (papers on the supplementary reading list)

95