Mining Multi-Dimensional Association

- Single-dimensional rules:

  buys(X, "milk") $\Rightarrow$ buys(X, "bread")

- Multi-dimensional rules: $\geq$ 2 dimensions or predicates
  - Inter-dimension assoc. rules (*no repeated predicates*)

    age(X,"19-25") $\wedge$ occupation(X,"student") $\Rightarrow$ buys(X, "coke")
  - hybrid-dimension assoc. rules (*repeated predicates*)

    age(X,"19-25") $\wedge$ buys(X, "popcorn") $\Rightarrow$ buys(X, "coke")

Constraint-based (Query-Directed) Mining

- Finding all the patterns in a database autonomously? — unrealistic!
  - The patterns could be too many but not focused!
- Data mining should be an interactive process
  - User directs what to be mined using a data mining query language (or a graphical user interface)
- Constraint-based mining
  - User flexibility: provides constraints on what to be mined

Constraints in Data Mining

- Knowledge type constraint:
  - classification, association, etc.
- Data constraint
  - find product pairs sold together in stores in Chicago this year
- Dimension/level constraint
  - in relevance to region, price, brand, customer category
- Rule (or pattern) constraint
  - small sales (price < $10) triggers big sales (sum > $200)
- Interestingness constraint
  - strong rules: min_support ≥ 3%, min_confidence ≥ 60%

Meta-Rule Guided Mining

- Meta-rule can be in the rule form with partially instantiated predicates and constants

  $P_1(X, Y) \wedge P_2(X, W) \Rightarrow buys(X, \text{“iPad”})$

- The resulting rule derived can be

  age(X, “15-25”) ^ profession(X, “student”) => buys(X, “iPad”)

- In general, it can be in the form of

  $P_1 \wedge P_2 \wedge \ldots \wedge P_l \Rightarrow Q_1 \wedge Q_2 \wedge \ldots \wedge Q_r$

Challenges

- A major challenge in mining frequent itemsets from a large data set is the fact that such mining often generates a huge number of itemsets satisfying the minimum support threshold, especially when *minsup is* set low.

- This is because if an itemset is frequent, each of its subsets is frequent as well.

- A long itemset will contain a combinatorial number of shorter, frequent sub-itemsets.

Closed Patterns and Max-Patterns

- Example : A long pattern contains a combinatorial number of sub-patterns, e.g., $\{a_1, \ldots, a_{100}\}$ contains $\binom{100}{1} + \binom{100}{2} + \ldots + \binom{100}{100} = 2^{100} - 1 = 1.27*10^{30}$ sub-patterns!
- Solution: *Mine closed patterns and max-patterns instead*
- An itemset X is closed if X is *frequent* and there exists *no super-pattern* Y ɔ X, *with the same support* as X
- An itemset X is a max-pattern if X is frequent and there exists no frequent super-pattern Y ɔ X
- Closed pattern is a lossless compression of freq. patterns

Closed Patterns and Max-Patterns

- Exercise. DB = $\{<a_1, \ldots, a_{100}>, < a_1, \ldots, a_{50}>\}$
  - Min_sup = 1.
- What is the set of closed itemset?
  - $<a_1, \ldots, a_{100}>$: 1
  - $< a_1, \ldots, a_{50}>$: 2
- What is the set of max-pattern?
  - $<a_1, \ldots, a_{100}>$: 1

Colossal itemset

- The result of frequent closed itemset mining algorithms includes small and mid-sized itemsets, which does not enclose valuable and complete information in many applications.
- In application dealing with high dimensional datasets such as bioinformatics (Micro array analysis, biological sequence analysis) , association rule mining gives greater importance to the large sized itemsets called as colossal Itemsets
- An itemset X is called frequent colossal closed itemset if and only if it is frequent closed and card(X) ≥ mincard, where mincard is user specified least cardinality threshold

Table 1

Tid	features
1	$f_1, f_2, f_4, f_6, f_{10}$
2	f_1, f_2, f_4, f_7, f_8
3	f_2, f_4, f_7, f_8
4	$f_1, f_2, f_6, f_8, f_9, f_{10}$
5	$f_1, f_3, f_4, f_7, f_8, f_{10}$
6	f_2, f_4, f_9
7	f_5, f_7
8	f_5, f_{11}

- In Table 1, the itemset $X = \{f_2, f_4, f_7, f_8\}$, is frequent colossal closed itemset with minimum support threshold set to 2 and minimum cardinality threshold set to 4, $sup(X) \geq 2$ and $card(X) \geq 4$.

Colossal Patterns: A Motivating Example

Let's make a set of 40 transactions

T1 = 1 2 3 4 ….. 39 40

T2 = 1 2 3 4 ….. 39 40

: .

: .

: .

: .

T40=1 2 3 4 ….. 39 40

Then delete the items on the diagonal

T_1 = 2 3 4 ….. 39 40

T_2 = 1 3 4 ….. 39 40

: .

: .

: .

: .

T_{40}=1 2 3 4 …… 39

A Show of Colossal Pattern Mining!

T_1 = 2 3 4 ..... 39 40
T_2 = 1 3 4 ..... 39 40
: .
: .
: .
: .
T_{40}=1 2 3 4 39
T_{41}= 41 42 43 79
T_{42}= 41 42 43 79
: .
: .
T_{60}= 41 42 43 ... 79

Let the min-support threshold σ= 20

Then there are $\binom{40}{20}$ closed/maximal frequent patterns of size 20

However, there is only one colossal pattern with size greater than 20,

α= {41,42,…,79} of size 39