Recap – 2nd Lecture

Data Base and Data Mining Group of Politecnico di Torino

Danilo Giordano

Politecnico di Torino

Association rules

- Objective
  - extraction of frequent correlations or pattern from a transactional database

Tickets at a supermarket counter

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Coke, Diapers, Milk
…	…

- Association rule

  diapers $\Rightarrow$ beer
  - 2% of transactions contains both items
  - 30% of transactions containing diapers also contains beer

Association rule mining

- A collection of transactions is given
  - a transaction is a set of items
  - items in a transaction are *not ordered*
- Association rule

$$A, B \Rightarrow C$$

  - A, B = items in the rule body
  - C = item in the rule head
- The $\Rightarrow$ means co-occurrence
  - *not* causality
- Example
  - coke, diapers $\Rightarrow$ milk

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Coke, Diapers, Milk
…	…

Rule quality metrics

- Given the association rule

$$A \Rightarrow B$$

  - A, B are itemsets
- *Support* is the fraction of transactions containing both A and B

$$\frac{\#\{A,B\}}{|T|}$$

  - |T| is the cardinality of the transactional database
  - a priori probability of itemset AB
  - rule frequency in the database
- *Confidence* is the frequency of B in transactions containing A

$$\frac{sup(A,B)}{sup(A)}$$

  - conditional probability of finding B having found A
  - "strength" of the "$\Rightarrow$"

Association rule extraction

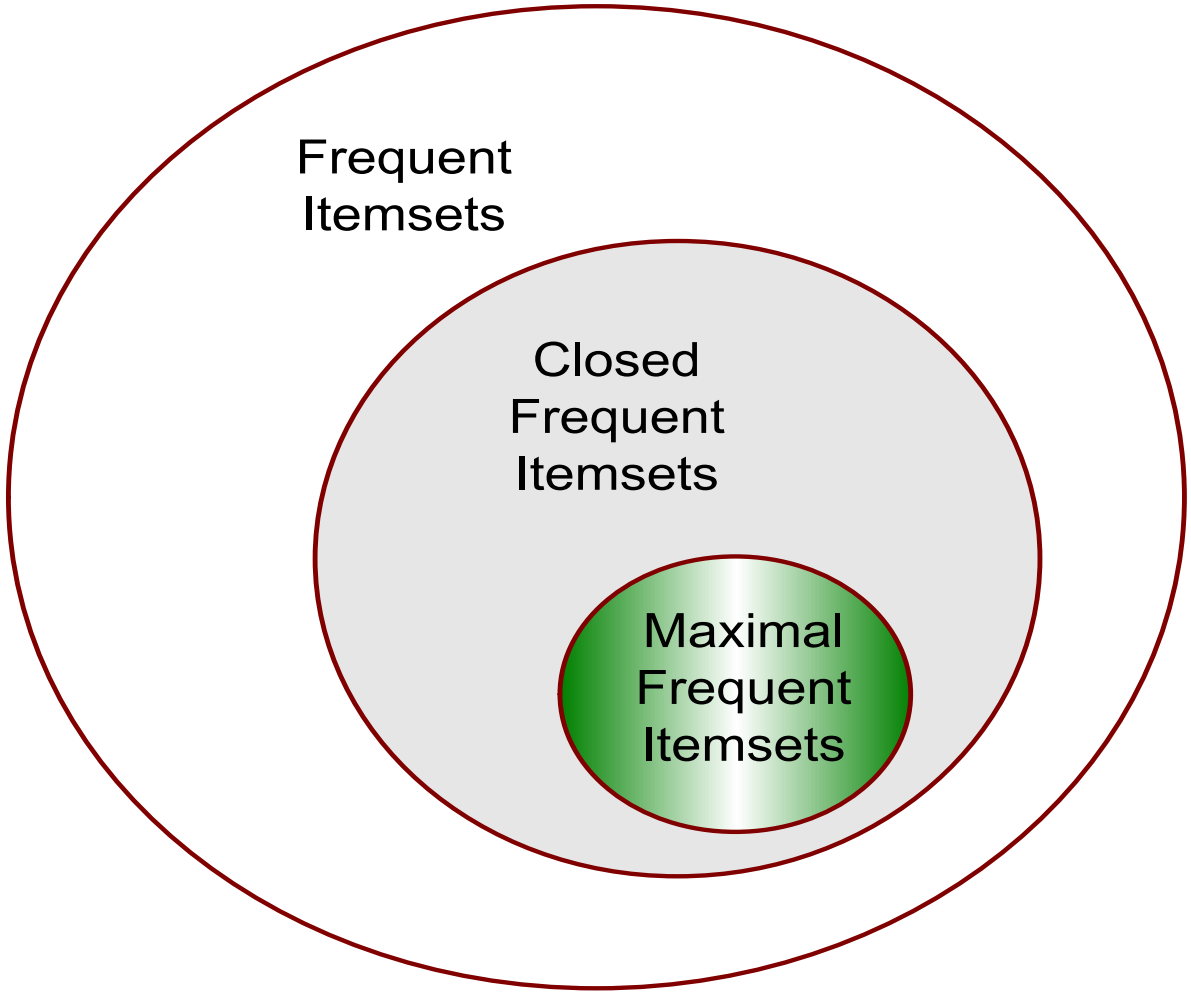
(1) Extraction of frequent itemsets

- many different techniques
  - level-wise approaches (Apriori, ...)
  - approaches without candidate generation (FP-growth, ...)
  - other approaches
- most computationally expensive step
  - limit extraction time by means of support threshold

(2) Extraction of association rules

- generation of all possible binary partitioning of each frequent itemset
  - possibly enforcing a confidence threshold

Maximal vs Closed Itemsets



From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Confidence measure: always reliable?

- 5000 high school students are given
  - 3750 eat cereals
  - 3000 play basket
  - 2000 eat cereals and play basket
- Rule

$$\text{play basket} \Rightarrow \text{eat cereals}$$
$$\text{sup} = 40\%, \text{conf} = 66{,}7\%$$

is misleading because eat cereals has sup 75% (>66,7%)

  - Problem caused by high frequency of rule head
    - negative correlation

	basket	not basket	total
cereals	2000	1750	3750
not cereals	1000	250	1250
total	3000	2000	5000

Correlation or lift

r: $A \Rightarrow B$

$$\text{Correlation} = \frac{P(A,B)}{P(A)P(B)} = \frac{\text{conf(r)}}{\text{sup(B)}}$$

- Statistical independence
  - Correlation = 1
- Positive correlation
  - Correlation > 1
- Negative correlation
  - Correlation < 1