Association Rules Fundamentals

Data Base and Data Mining Group of Politecnico di Torino

Elena Baralis, Tania Cerquitelli, Silvia Chiusano

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
…	…

Transactional formats

- Association rule extraction is an *exploratory technique* that can be applied to any data type
- A transaction can be any set of items
  - Market basket data
  - Textual data
  - Structured data
  - …

Transactional formats

- Textual data
  - A document is a transaction
  - Words in a document are items in the transaction

- Data example
  - Doc1: algorithm analysis customer data mining relationship
  - Doc2: customer data management relationship
  - Doc3: analysis customer data mining relationship social
- Rule example

  customer, relationship $\Rightarrow$ data, mining

Transactional formats

- Structured data
  - A table row is a transaction
  - Pairs (attribute, value) are items in the transaction

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

- Data example

Refund	Marital Status	Taxable Income	Cheat
No	Married	< 80K	**No**

  - Transaction

    Refund=no, MaritalStaus=married, TaxableIncome<80K, Cheat=No

- Rule example

  Refund=No, MaritalStatus=Married $\Rightarrow$ Cheat = No

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

Definitions

- *Itemset* is a set including one or more items
  - Example: {Beer, Diapers}
- *k-itemset* is an itemset that contains k items
- *Support count* (#) is the frequency of occurrence of an itemset
  - Example: #{Beer,Diapers} = 2
- *Support* is the fraction of transactions that contain an itemset
  - Example: sup({Beer, Diapers}) = 2/5
- *Frequent itemset* is an itemset whose support is greater than or equal to a *minsup* threshold

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 "⇒"

Rule quality metrics: example

- From itemset {Milk, Diapers} the following rules may be derived
- Rule: Milk $\Rightarrow$ Diapers
  - support
    sup=#{Milk,Diapers}/#trans. =3/5=60%
  - confidence
    conf=#{Milk,Diapers}/#{Milk}=3/4=75%
- Rule: Diapers $\Rightarrow$ Milk
  - same support
    s=60%
  - confidence
    conf=#{Milk,Diapers}/#{Diapers}=3/3 =100%

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 extraction

- Given a set of transactions T, association rule mining is the extraction of the rules satisfying the constraints
  - support ≥ *minsup* threshold
  - confidence ≥ *minconf* threshold
- The result is
  - complete (*all* rules satisfying both constraints)
  - correct (*only* the rules satisfying both constraints)
- May add other more complex constraints

Association rule extraction

- Brute-force approach
  - enumerate all possible permutations (i.e., association rules)
  - compute support and confidence for each rule
  - prune the rules that do not satisfy the *minsup* and *minconf* constraints
- Computationally *unfeasible*
- Given an itemset, the extraction process may be split
  - first generate frequent itemsets
  - next generate rules from each frequent itemset
- Example
  - Itemset

    {Milk, Diapers} sup=60%
  - Rules

    Milk $\Rightarrow$ Diapers (conf=75%)

    Diapers $\Rightarrow$ Milk (conf=100%)

Association rule extraction

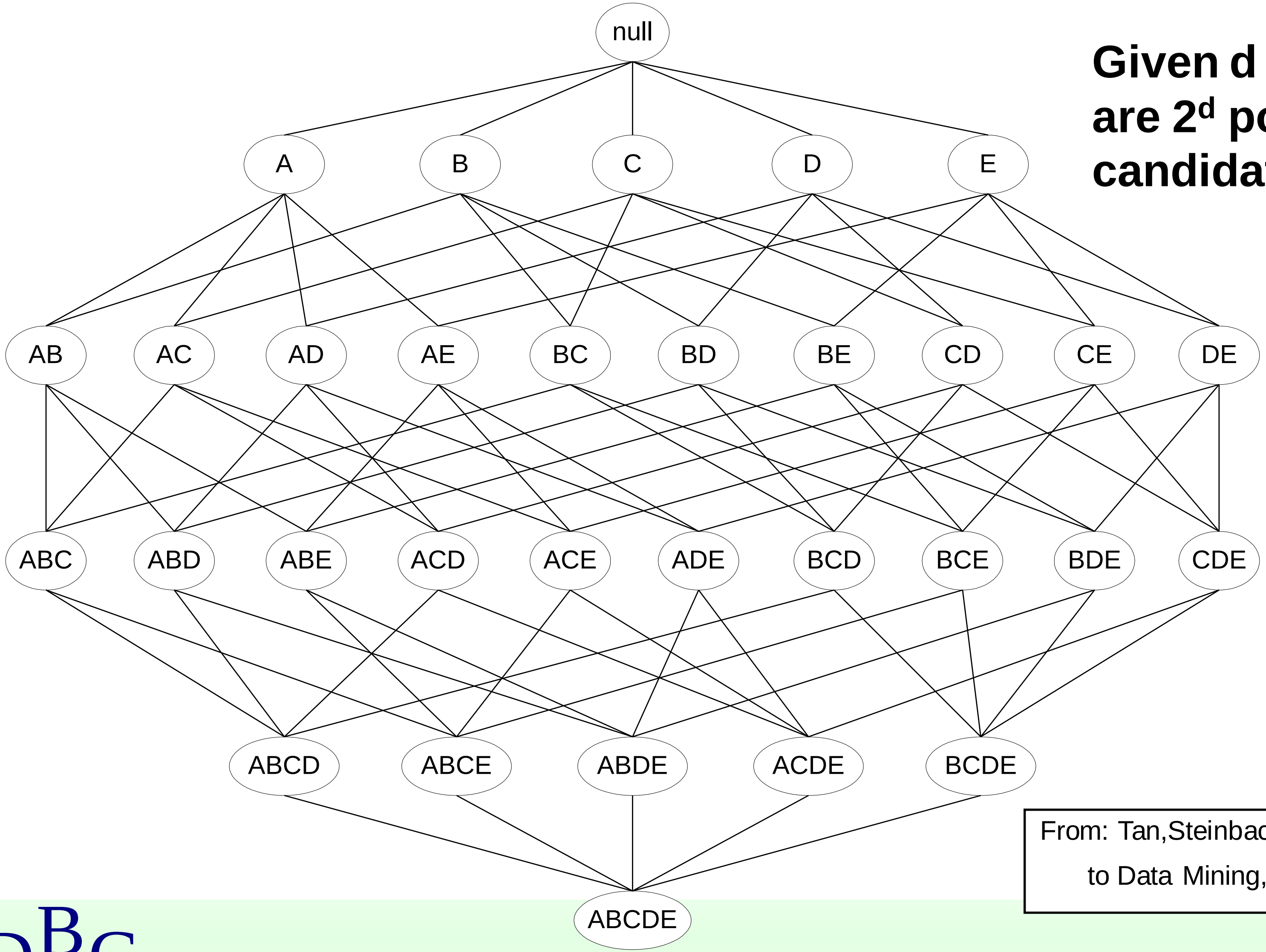
(1) Extraction of frequent itemsets

- 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

Frequent Itemset Generation



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

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

Frequent Itemset Generation

- Brute-force approach
  - each itemset in the lattice is a *candidate* frequent itemset
  - scan the database to count the support of each candidate
    - match each transaction against every candidate
  - Complexity ~ $O(|T|\ 2^d\ w)$
    - $|T|$ is number of transactions
    - d is number of items
    - w is transaction length

Improving Efficiency

- Reduce the number of candidates
  - Prune the search space
    - complete set of candidates is 2^d
- Reduce the number of transactions
  - Prune transactions as the size of itemsets increases
    - reduce $|T|$
- Reduce the number of comparisons
  - Equal to $|T| \, 2^d$
  - Use efficient data structures to store the candidates or transactions

The Apriori Principle

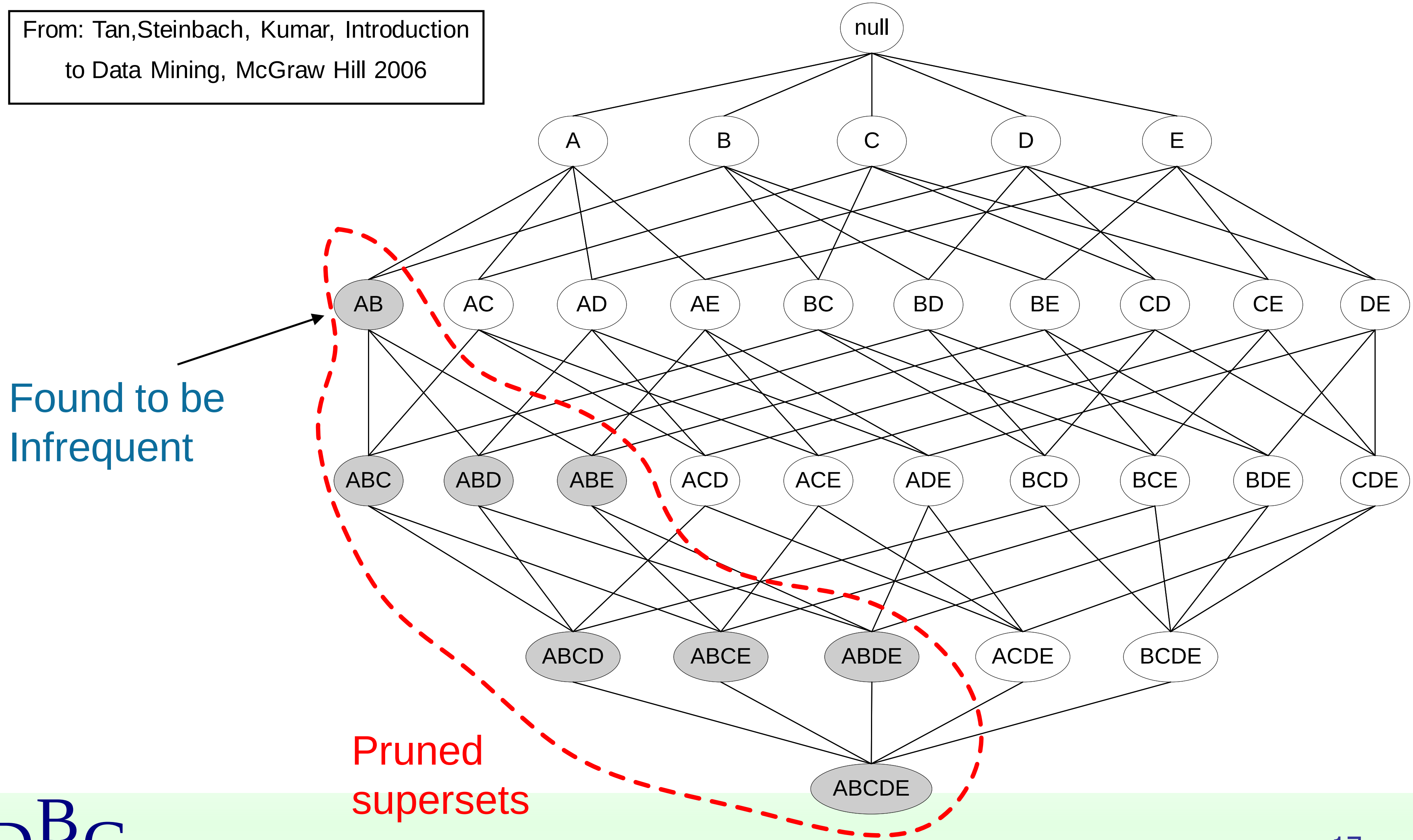
"If an itemset is frequent, then all of its subsets must also be frequent"

- The support of an itemset can never exceed the support of any of its subsets
- It holds due to the antimonotone property of the support measure
  - Given two arbitrary itemsets A and B

    if $A \subseteq B$ then $sup(A) \geq sup(B)$
- It reduces the number of candidates

The Apriori Principle

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



Factors Affecting Performance

- Minimum support threshold
  - lower support threshold increases number of frequent itemsets
    - larger number of candidates
    - larger (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 in dense data sets
  - may increase max length of frequent itemsets and traversals of hash tree
    - number of subsets in a transaction increases with its width

FP-growth Algorithm [Han00]

- Exploits a main memory compressed rappresentation of the database, the FP-tree
  - high compression for dense data distributions
    - less so for sparse data distributions
  - complete representation for frequent pattern mining
    - enforces support constraint
- Frequent pattern mining by means of FP-growth
  - recursive visit of FP-tree
  - applies divide-and-conquer approach
    - decomposes mining task into smaller subtasks
- Only two database scans
  - count item supports + build FP-tree

Other approaches

- Many other approaches to frequent itemset extraction
  - some covered later
- May exploit a different database representation
  - represent the tidset of each item [Zak00]

Horizontal Data Layout

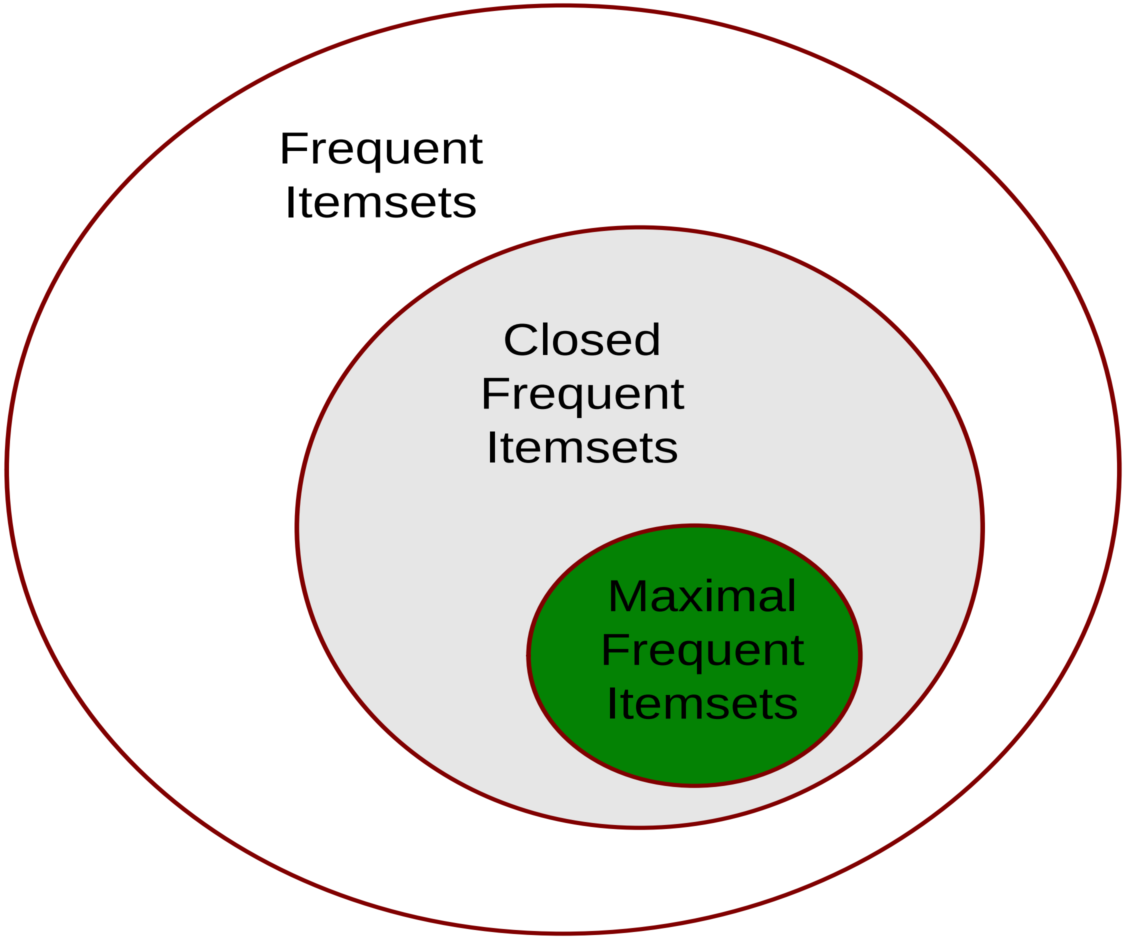
TID	Items
1	A,B,E
2	B,C,D
3	C,E
4	A,C,D
5	A,B,C,D
6	A,E
7	A,B
8	A,B,C
9	A,C,D
10	B

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				

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

Maximal vs Closed Itemsets



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

Effect of Support Threshold

- Selection of the appropriate *minsup* threshold is not obvious
  - If *minsup* is too high
    - itemsets including rare but interesting items may be lost
      - example: pieces of jewellery (or other expensive products)
  - If *minsup* is too low
    - it may become computationally *very expensive*
    - the number of frequent itemsets becomes *very large*

Interestingness Measures

- A large number of pattern may be extracted
  - rank patterns by their interestingness
- Objective measures
  - rank patterns based on statistics computed from data
  - initial framework [Agr94] only considered support and confidence
    - other statistical measures available
- Subjective measures
  - rank patterns according to user interpretation [Silb98]
    - interesting if it contradicts the expectation of a user
    - interesting if it is actionable

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

Example

- Association rule

  $$\text{play basket} \Rightarrow \text{eat cereals}$$

  has corr = 0.89
  - negative correlation
- but rule

  $$\text{play basket} \Rightarrow \text{not (eat cereals)}$$

  has corr = 1,34

#	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)	$\max\Big(P(A,B)\log(\frac{P(B\mid A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\mid A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\mid B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\mid B)}{P(A)})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\mid A)^2 + P(\overline{B}\mid A)^2] + P(\overline{A})[P(B\mid \overline{A})^2 + P(\overline{B}\mid\overline{A})^2]$ $-P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\mid B)^2 + P(\overline{A}\mid B)^2] + P(\overline{B})[P(A\mid\overline{B})^2 + P(\overline{A}\mid\overline{B})^2]$ $-P(A)^2 - P(\overline{A})^2\Big)$
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)	$\frac{P(A,B)}{P(A)P(B)}$
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)	$\max\left(\frac{P(B\mid A)-P(B)}{1-P(B)}, \frac{P(A\mid B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\mid A) - P(B), P(A\mid B) - P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\mid A) - P(B), P(A\mid B) - P(A))$