Data Mining & CRM

Fosca Giannotti and Dino Pedreschi
Pisa KDD Lab, ISTI-CNR & Univ. Pisa

MAINS – Master in Management dell'Innovazione
Scuola Superiore S. Anna

Association rules and market basket analysis

Association rules - module outline

- What are association rules (AR) and what are they used for:
  - The paradigmatic application: Market Basket Analysis
  - The single dimensional AR (intra-attribute)
- How to compute AR
  - Basic Apriori Algorithm and its optimizations
  - Multi-Dimension AR (inter-attribute)
  - Quantitative AR
- How to reason on AR and how to evaluate their quality
  - Interestingness
  - Correlation vs. Association

Market Basket Analysis: the context

Customer buying habits by finding associations and correlations between the different items that customers place in their “shopping basket”

Milk, eggs, sugar, bread

Customer1

Milk, eggs, cereal, bread

Customer2

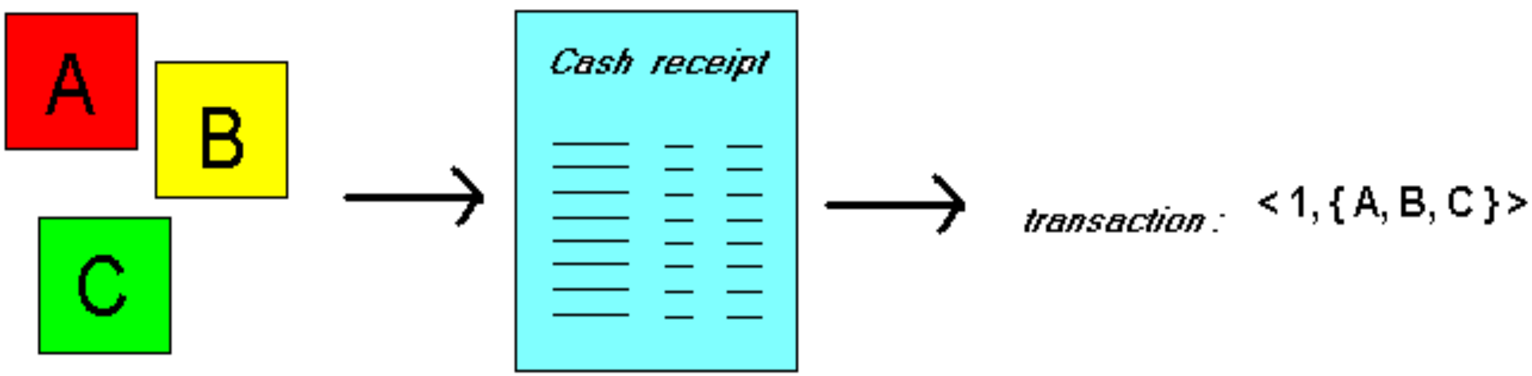
Eggs, sugar

Customer3

Market Basket Analysis: the context

Given: a database of customer transactions, where each transaction is a set of items

- Find groups of items which are frequently purchased together



Goal of MBA

- Extract information on purchasing behavior
- Actionable information: can suggest
  - new store layouts
  - new product assortments
  - which products to put on promotion
- MBA applicable whenever a customer purchases multiple things in proximity
  - credit cards
  - services of telecommunication companies
  - banking services
  - medical treatments

MBA: applicable to many other contexts

Telecommunication:

Each customer is a transaction containing the set of customer's phone calls

Atmospheric phenomena:

Each time interval (e.g. a day) is a transaction containing the set of observed event (rains, wind, etc.)

Etc.

Association Rules

- Express how product/services relate to each other, and tend to group together
  - "if *a customer purchases three-way calling*, then will also purchase call-waiting"
- Actionable information:
  - bundle three-way calling and call-waiting in a single package
- Rule form: "Body → Head [support, confidence]".
- Examples.
  - buys(x, "diapers") → buys(x, "beers") [0.5%, 60%]
  - major(x, "CS") ^ takes(x, "DB") → grade(x, "A") [1%, 75%]

Useful, trivial, unexplicable

- Useful: "On Thursdays, grocery store consumers often purchase diapers and beer together".
- Trivial: "Customers who purchase maintenance agreements are very likely to purchase large appliances".
- Unexplicable: "When a new hardaware store opens, one of the most sold items is toilet rings."

Basic Concepts

Transaction:

Relational format

<Tid,item>

<1, item1>

<1, item2>

<2, item3>

Compact format

<Tid,itemset>

<1, {item1,item2}>

<2, {item3}>

Item: single element, **Itemset**: set of items

Support of an itemset I: # of transactions containing I

Minimum Support σ : threshold for support

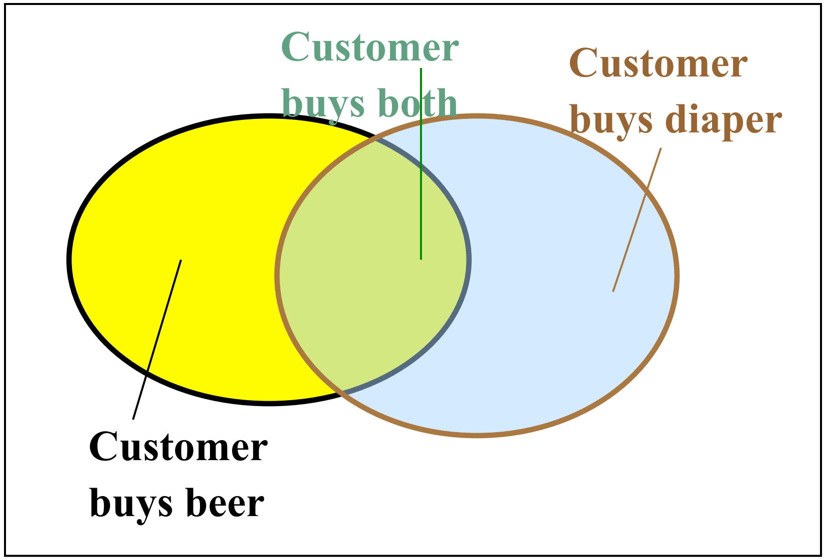
Frequent Itemset : with support $\geq \sigma$.

Basic Concepts: Frequent Patterns and Association Rules

- Itemset $X = \{x_1, \ldots, x_k\}$
- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - **support**, *s*, probability that a transaction contains $X \cup Y$
  - **confidence**, *c*, conditional probability that a transaction having X also contains Y

Basic Concepts: Frequent Patterns and Association Rules

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F



Let $sup_{min} = 50\%$, $conf_{min} = 50\%$

Freq. Pat.: {A:3, B:3, D:4, E:3, AD:3}

Association rules:

$A \rightarrow D$ (60%, 100%)

$D \rightarrow A$ (60%, 75%)

Association Rules: Measures

- Let A and B be a partition of an itemset I :

$$A \Rightarrow B\ [s, c]$$

A and B are itemsets

s = **support of A $\Rightarrow$ B** = support(A$\cup$B)

c = **confidence of A $\Rightarrow$ B** = support(A$\cup$B)/support(A)

- Measure for rules:
  - ✓ minimum support σ
  - ✓ minimum confidence γ
- The rules holds if : $s \geq \sigma$ and $c \geq \gamma$

Association Rules: Meaning

$$A \Rightarrow B \; [\, s, c \,]$$

Support: denotes the frequency of the rule within transactions. A high value means that the rule involve a great part of database.

$$\text{support}(A \Rightarrow B) = p(A \cup B)$$

Confidence: denotes the percentage of transactions containing A which contain also B. It is an estimation of conditioned probability .

$$\text{confidence}(A \Rightarrow B) = p(B|A) = p(A \;\&\; B)/p(A).$$

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} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {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}) = 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

Definition: Association Rule

- Association Rule
  - **An implication expression of the form X → Y, where X and Y are itemsets**
  - **Example:**
    {Milk, Diaper} → {Beer}

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**

- 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

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

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

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

Frequent Itemsets

Transaction ID	Items Bought
1	dairy,fruit
2	dairy,fruit, vegetable
3	dairy
4	fruit, cereals

Support({dairy}) = 3 (75%)
Support({fruit}) = 3 (75%)
Support({dairy, fruit}) = 2 (50%)

If σ = 60%, then

{dairy} and {fruit} are frequent while {dairy, fruit} is not.

Association Rules - Example

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Min. support 50%
Min. confidence 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

For rule $A \Rightarrow C$:

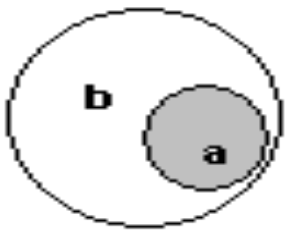
support = support({A, C}) = 50%

confidence = support({A, C})/support({A}) = 66.6%

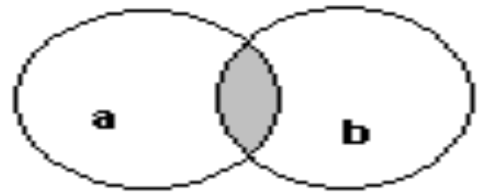
Frequent Itemsets vs. Logic Rules

Frequent itemset I = {a, b} does not distinguish between (1) and (2)

(1)

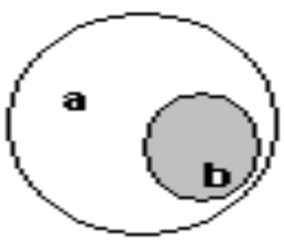


a => b

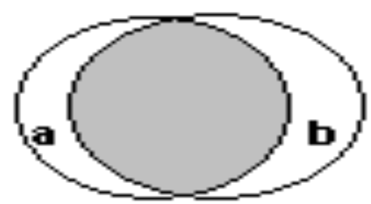


Almost no relation

(2)



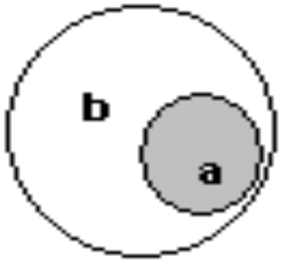
b => a



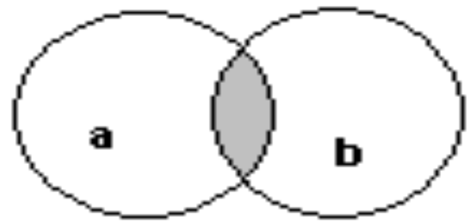
Almost a=>b e b=>a

Logic does: $x \Rightarrow y$ *iff* when x holds, y holds too

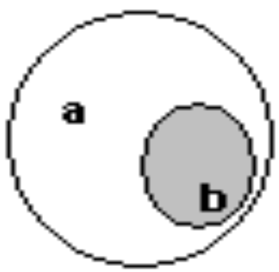
Association Rules – the effect



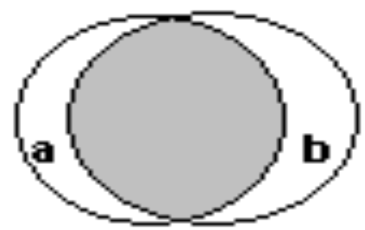
conf(a => b) = 100%
conf(b => a) = ~ 0%



conf(a => b) = ~ 0%
conf(b => a) = ~ 0%



conf(a => b) = ~ 0%
conf(b => a) = 100%



conf(a => b) = ~100%
conf(b => a) = ~100%

Association Rules – the parameters σ and γ

Minimum Support σ :

High ⇒ few frequent itemsets

⇒ few valid rules which occur very often

Low ⇒ many valid rules which occur rarely

Minimum Confidence γ :

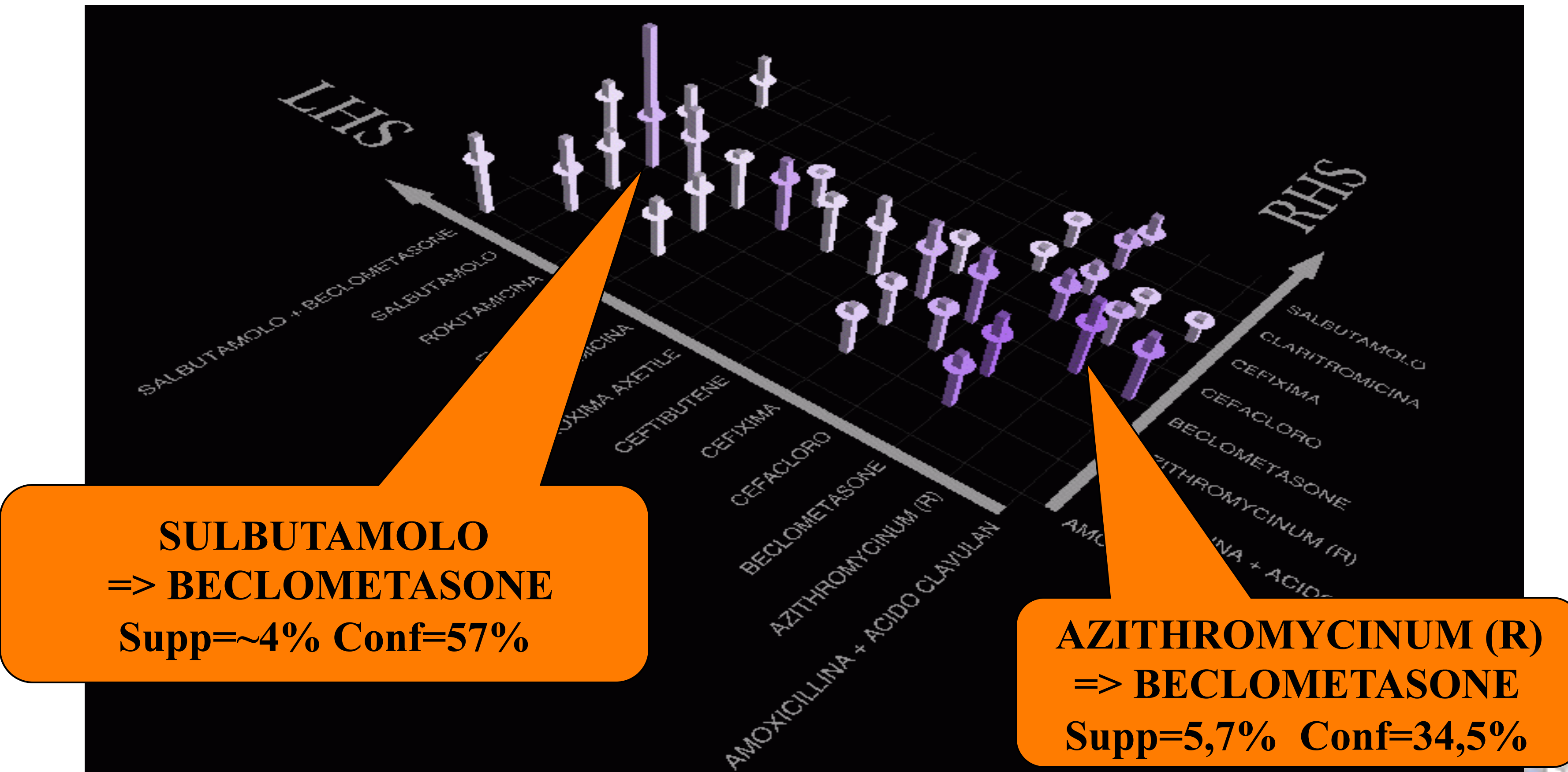
High ⇒ few rules, but all "almost logically true"

Low ⇒ many rules, but many of them very "uncertain"

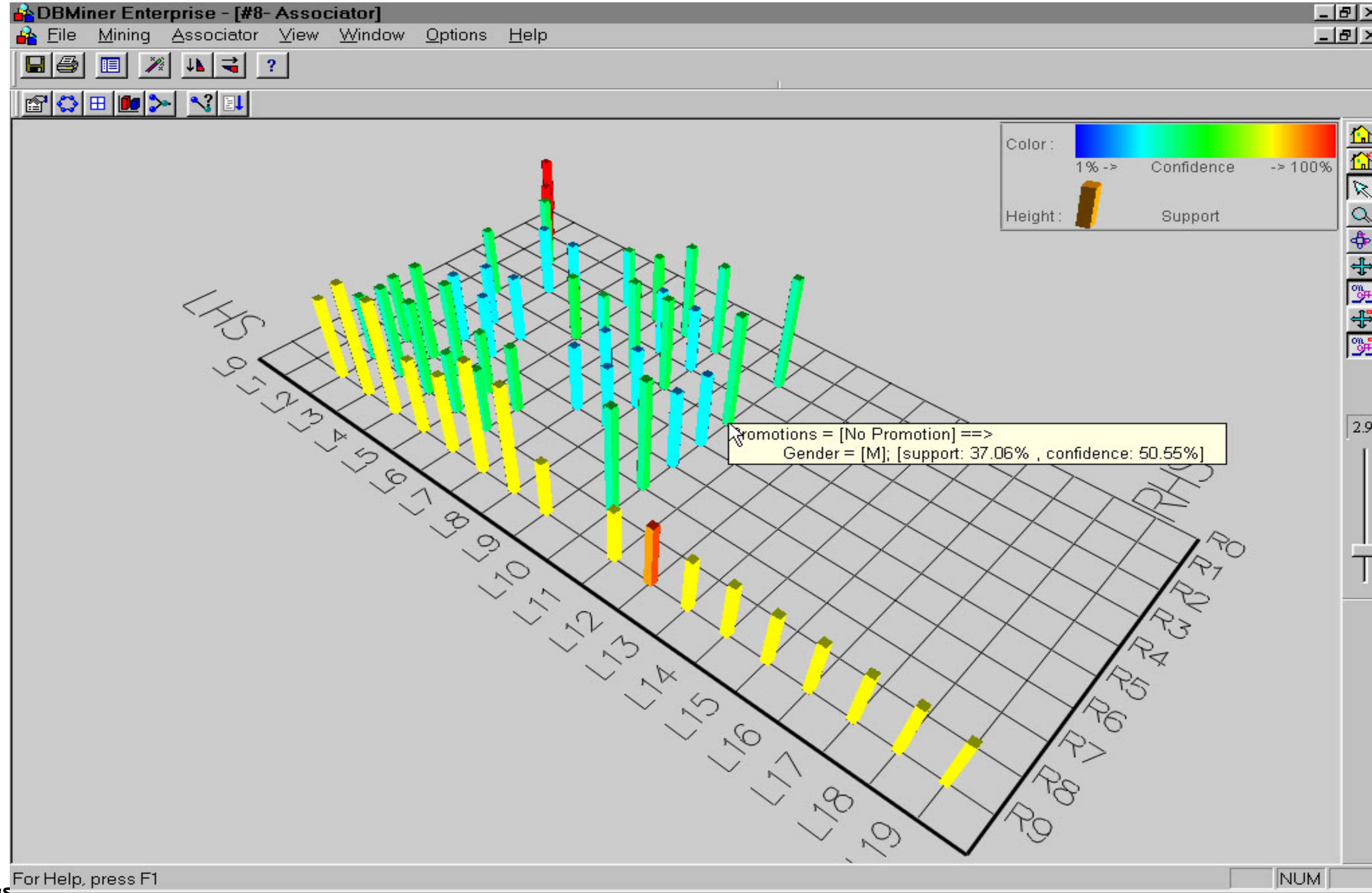
Typical Values: $\sigma = 2 \div 10\ \%$ $\gamma = 70 \div 90\ \%$

Association Rules – visualization

(Patients <15 old for USL 19 (a unit of Sanitary service), January-September 1997)



Visualization of Association Rules: Plane Graph



Association Rules – bank transactions

Step 1: Create groups of customers (cluster) on the base of demographical data.

Step 2: Describe customers of each cluster by mining association rules.

Example:

Rules on cluster 6 (23,7% of dataset):

File Edit Search Help

Group	Support	Confidence		Body	--> Head
1	0.277	91.4	.	1.3	[TERM DEPOSITS] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.164	86.4	.	1.3	[TERM DEPOSITS] AND [ATM CARD] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.104	85.7	.	1.9	[SAVINGS] AND [INTERNET BANKING] AND [LEASES] ==> [TELEBANKING]
1	0.138	84.2	.	1.2	[PERSONAL BANKING] AND [TERM DEPOSITS] AND [BUSINESS CREDIT CARD] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.251	82.9	.	1.2	[TERM DEPOSITS] AND [ATM CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.328	82.6	.	1.2	[ATM CARD] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.242	82.4	.	1.2	[PERSONAL BANKING] AND [TERM DEPOSITS] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.631	81.1	.	1.2	[BUSINESS CREDIT CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.138	80.0	.	1.2	[ATM CARD] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [INTERNET BANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.138	80.0	.	1.2	[TERM DEPOSITS] AND [TEL --> [SAVINGS]
1	0.458	79.1	.	1.2	[TERM DEPOSITS] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.130	78.9	.	1.2	[PERSONAL BANKING] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.346	78.4	.	1.2	[PERSONAL BANKING] AND [BUSINESS CREDIT CARD] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	1.037	77.9	.	1.1	[TERM DEPOSITS] AND [ATM CARD] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [INTERNET BANKING] ==> [SAVINGS]
1	0.102	77.8	.	1.7	[TERM DEPOSITS] AND [ATM CARD] AND [INTERNET BANKING] AND [BUSINESS SAVINGS] ==> [BUSINESS CREDIT CARD]

Cluster 6 (23.7% of customers)

File Edit Search Help

Group	Support	Confidence		Body	--> Head
1	0.277	91.4	.	1.3	[TERM DEPOSITS] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.164	86.4	.	1.3	[TERM DEPOSITS] AND [ATM CARD] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] --> [SAVINGS]
1	0.104	85.7	.	1.9	[SAVINGS] AND [INTERNET BANKING] AND [LEASES] --> [TELEBANKING]
1	0.138	84.2	.	1.2	[PERSONAL BANKING] AND [TERM DEPOSITS] AND [BUSINESS CREDIT CARD] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.251	82.9	.	1.2	[TERM DEPOSITS] AND [ATM CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] --> [SAVINGS]
1	0.328	82.6	.	1.2	[ATM CARD] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.242	82.4	.	1.2	[PERSONAL BANKING] AND [TERM DEPOSITS] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.631	81.1	.	1.2	[BUSINESS CREDIT CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.138	80.0	.	1.2	[ATM CARD] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [INTERNET BANKING] AND [BUSINESS SAVINGS] --> [SAVINGS]
1	0.138	80.0	.	1.2	[TERM DEPOSITS] AND [TEL --> [SAVINGS]
1	0.458	79.1	.	1.2	[TERM DEPOSITS] AND [TELEBANKING] AND [BUSINESS SAVINGS] --> [SAVINGS]
1	0.130	78.9	.	1.2	[PERSONAL BANKING] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [BUSINESS SAVINGS] ==> [SAVINGS]
1	0.346	78.4	.	1.2	[PERSONAL BANKING] AND [BUSINESS CREDIT CARD] AND [BUSINESS SAVINGS] --> [SAVINGS]
1	1.037	77.9	.	1.1	[TERM DEPOSITS] AND [ATM CARD] AND [BUSINESS CREDIT CARD] AND [TELEBANKING] AND [INTERNET BANKING] ==> [SAVINGS]
1	0.102	77.0	.	1.7	[TERM DEPOSITS] AND [ATM CARD] AND [INTERNET BANKING] AND [BUSINESS SAVINGS] --> [BUSINESS CREDIT CARD]

Esercizio 1

Table 6.1. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

Support?: e, (b,d), (b,d,e), quali regole? Quale supporto?

Table 6.2. Market basket transactions.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

Max size of itemset, 2-itemsets with larger support

Association rules - module outline

- **What are association rules (AR) and what are they used for:**
  - The paradigmatic application: Market Basket Analysis
  - The single dimensional AR (intra-attribute)
- **How to compute AR**
  - Basic Apriori Algorithm and its optimizations
  - Multi-Dimension AR (inter-attribute)
  - Quantitative AR
  - Constrained AR
- **How to reason on AR and how to evaluate their quality**
  - Multiple-level AR
  - Interestingness
  - Correlation vs. Association

Association rules - module outline

- **What are association rules (AR) and what are they used for:**
  - The paradigmatic application: Market Basket Analysis
  - The single dimensional AR (intra-attribute)
- **How to compute AR**
  - Basic Apriori Algorithm and its optimizations
  - Multi-Dimension AR (inter-attribute)
  - Quantitative AR
  - Constrained AR
- **How to reason on AR and how to evaluate their quality**
  - Multiple-level AR
  - Interestingness
  - Correlation vs. Association

Basic Apriori Algorithm

Problem Decomposition

① Find the *frequent itemsets*: the sets of items that satisfy the support constraint

- *A subset of a frequent itemset is also a frequent itemset*, i.e., if {*A*,*B*} is a frequent itemset, both {*A*} and {*B*} should be a frequent itemset
- Iteratively find frequent itemsets with cardinality from 1 to *k* (*k*-itemset)

② Use the frequent itemsets to generate association rules.

Problem Decomposition

Transaction ID	Purchased Items
1	{1, 2, 3}
2	{1, 4}
3	{1, 3}
4	{2, 5, 6}

- **For minimum support = 50% = 2 transactions and minimum confidence = 50%**

Frequent Itemsets	Support
{1}	75%
{2}	50%
{3}	50%
{1,3}	50%

For the rule $1 \Rightarrow 3$:

- Support = Support({1, 3}) = 50%
- Confidence = Support({1,3})/Support({1}) = 66%

The Apriori property

- **If B is frequent and A ⊆ B then A is also frequent**
  - Each transaction which contains B contains also A, which implies supp.(A) ≥ supp.(B))

- **Consequence:** if A is not frequent, then it is not necessary to generate the itemsets which include A.

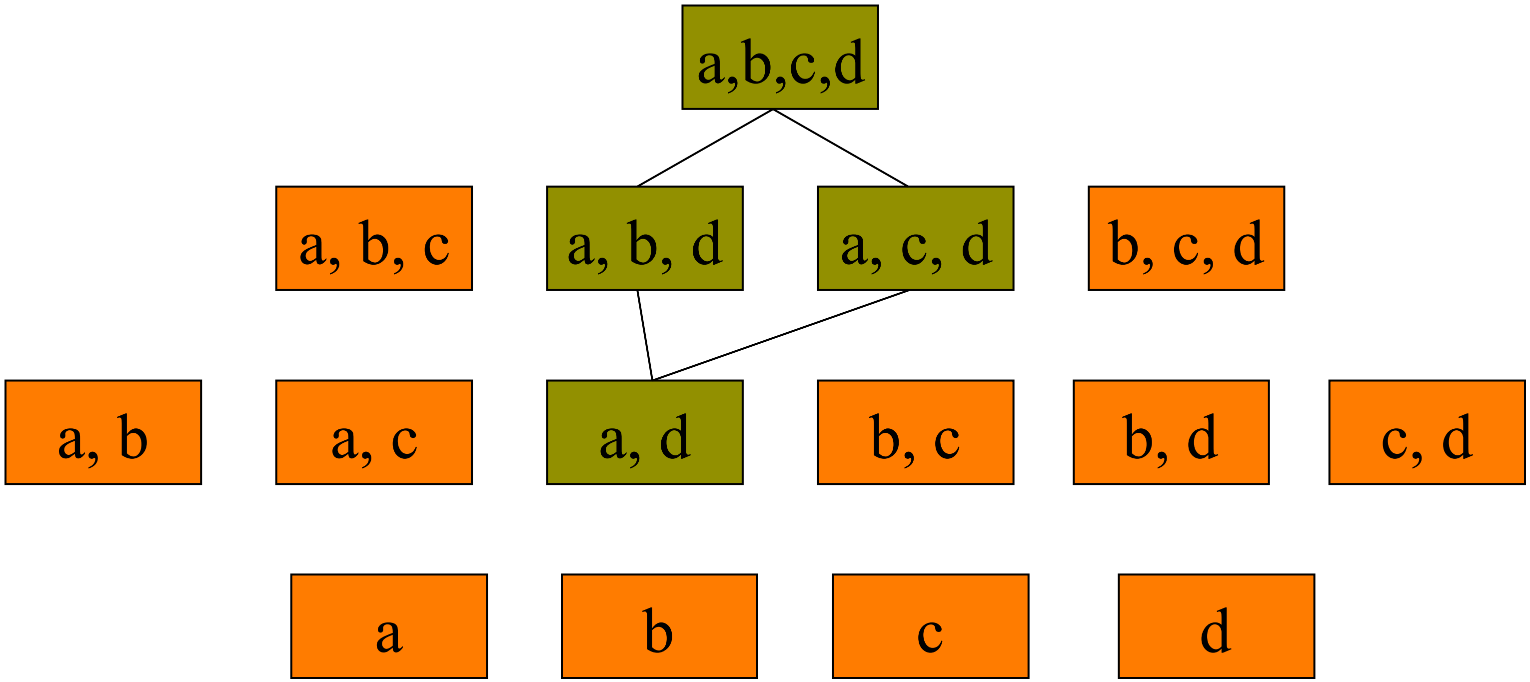
- **Example:**
  - <1, {a, b}> <2, {a} >
  - <3, {a, b, c}> <4, {a, b, d}>

  with minimum support = 30%.

  The itemset {c} is not frequent so is not necessary to check for:

  {c, a}, {c, b}, {c, d}, {c, a, b}, {c, a, d}, {c, b, d}

Apriori - Example



{a,d} is not frequent, so the 3-itemsets {a,b,d}, {a,c,d} and the 4-itemset {a,b,c,d}, are not generated.

Apriori Execution Example (min_sup = 2)

Database TDB

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan TDB

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan TDB

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

Scan TDB

L_3

itemset	sup
{2 3 5}	2

The Apriori Algorithm

- **Join Step:** C_k is generated by joining L_{k-1} with itself
- **Prune Step:** Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- **Pseudo-code:**

C_k: Candidate itemset of size k
L_k : frequent itemset of size k

L_1 = {frequent items};
for (k = 1; L_k !=∅; k++) **do begin**
    C_{k+1} = candidates generated from L_k;
    for each transaction t in database do
        increment the count of all candidates in C_{k+1}
        that are contained in t
    L_{k+1} = candidates in C_{k+1} with min_support
    end
return $\cup_k L_k$;

How to Generate Candidates?

- **Suppose the items in L_{k-1} are listed in an order**
- **Step 1: self-joining L_{k-1}**

  insert into C_k

  select $p.item_1, p.item_2, \ldots, p.item_{k-1}, q.item_{k-1}$

  from $L_{k-1}\ p, L_{k-1}\ q$

  where $p.item_1 = q.item_1, \ldots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$
- **Step 2: pruning**

  forall *itemsets c in* C_k **do**

      forall ***(k-1)-subsets s of c*** do

          if *(s is not in* L_{k-1}*)* **then delete** *c* **from** C_k

Example of Generating Candidates

- L_3={*abc, abd, acd, ace, bcd*}
- Self-joining: L_3*L_3
  - *abcd* from *abc* and *abd*
  - *acde* from *acd* and *ace*
- Pruning:
  - *acde* is removed because *ade* is not in L_3
- C_4={*abcd*}

Generating Association Rules from Frequent Itemsets

- **Only strong association rules are generated**
- **Frequent itemsets satisfy minimum support threshold**
- **Strong rules are those that satisfy minimum confidence threshold**

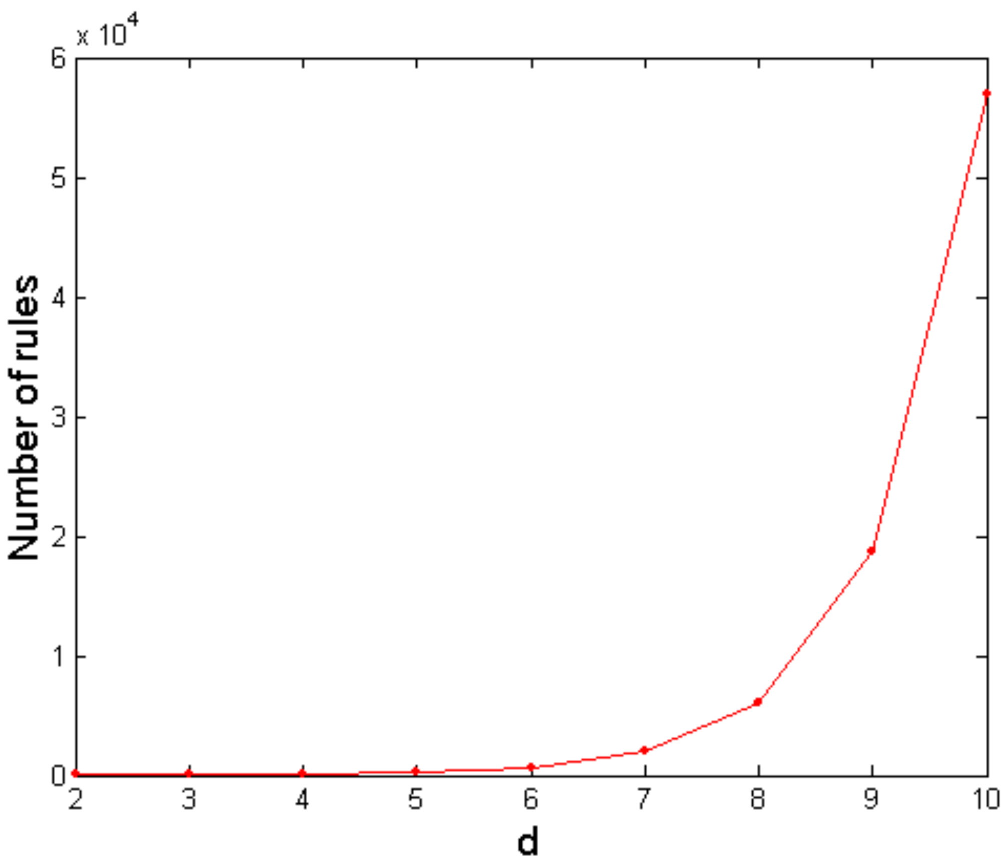
$$\frac{support(A \cup B)}{support(A)}$$

- *c*

For each frequent itemset, **f**, generate all non-empty subsets of **f**
For every non-empty subset **s** of **f** **do**
    if support(**f**)/support(**s**) ≥ min_confidence **then**
        output rule **s** ==> **(f-s)**
end

Computational Complexity

- Given d unique items:
  - Total number of itemsets = 2^d
  - Total number of possible association rules:



$$R = \sum_{k=1}^{d-1} \left[\binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j} \right]$$
$$= 3^d - 2^{d+1} + 1$$

If d=6, R = 602 rules

Rule Generation

- Given a frequent itemset L, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement
  - If {A,B,C,D} is a frequent itemset, candidate rules:

    $ABC \rightarrow D$, $ABD \rightarrow C$, $ACD \rightarrow B$, $BCD \rightarrow A$,
    $A \rightarrow BCD$, $B \rightarrow ACD$, $C \rightarrow ABD$, $D \rightarrow ABC$
    $AB \rightarrow CD$, $AC \rightarrow BD$, $AD \rightarrow BC$, $BC \rightarrow AD$,
    $BD \rightarrow AC$, $CD \rightarrow AB$,

- If $|L| = k$, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \varnothing$ and $\varnothing \rightarrow L$)

Esercizio 1

Table 6.1. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

Support?: e, (b,d), (b,d,e),

ID Transazione Items

1 {f,a,d,b}

2 {d,a,c,e,b}

3 {c,a,b,e}

4 {b,a,d}

Fissati il supporto minimo σ = 60% e la confidenza minima γ = 80%

a) Indicare quali tra questi itemset sono frequenti.

- 1) {a}
- 2) {c}
- 3) {b,c}
- 4) {b,d}
- 5) {a,b,d}
- 6) {a,b,e}

b) Indicare quali tra queste regole sono valide

- 1) {a}=>{b}
- 2) {a}=>{d}
- 3) {d}=>{a}
- 4) {d}=>{a,b}
- 5) {a,b}=>{d}
- 6) {a,d}=>{b}

Association rules - module outline

- **What are association rules (AR) and what are they used for:**
  - The paradigmatic application: Market Basket Analysis
  - The single dimensional AR (intra-attribute)
- **How to compute AR**
  - Basic Apriori Algorithm and its optimizations
  - Multi-Dimension AR (inter-attribute)
  - Quantitative AR
  - Constrained AR
- **How to reason on AR and how to evaluate their quality**
  - Multiple-level AR
  - Interestingness
  - Correlation vs. Association

Multidimensional AR

Associations between values of different attributes :

CID	**nationality**	**age**	**income**
1	Italian	50	low
2	French	40	high
3	French	30	high
4	Italian	50	medium
5	Italian	45	high
6	French	35	high

RULES:

nationality = French ⇒ **income** = high [50%, 100%]

income = high ⇒ **nationality** = French [50%, 75%]

age = 50 ⇒ **nationality** = Italian [33%, 100%]

Single-dimensional vs Multi-dimensional AR

Multi-dimensional

<1, Italian, 50, low>

<2, French, 45, high>

Single-dimensional

<1, {nat/Ita, age/50, inc/low}>

<2, {nat/Fre, age/45, inc/high}>

Schema: <ID, a?, b?, c?, d?>

<1, yes, yes, no, no>

<2, yes, no, yes, no>

<1, {a, b}>

<2, {a, c}>

Quantitative Attributes

- Quantitative attributes (e.g. age, income)
- Categorical attributes (e.g. color of car)

CID	height	weight	income
1	168	75,4	30,5
2	175	80,0	20,3
3	174	70,3	25,8
4	170	65,2	27,0

Problem: too many distinct values

Solution: transform quantitative attributes in categorical ones via **discretization**.

Quantitative Association Rules

CID	Age	Married	NumCars
1	23	No	1
2	25	Yes	1
3	29	No	0
4	34	Yes	2
5	38	Yes	2

[Age: 30..39] and [Married: Yes] ⇒ [NumCars:2]

support = 40%
confidence = 100%

Discretization of quantitative attributes

Solution: each value is replaced by the interval to which it belongs.

height: 0-150cm, 151-170cm, 171-180cm, >180cm

weight: 0-40kg, 41-60kg, 60-80kg, >80kg

income: 0-10ML, 11-20ML, 20-25ML, 25-30ML, >30ML

CID	**height**	**weight**	**income**
1	151-171	60-80	>30
2	171-180	60-80	20-25
3	171-180	60-80	25-30
4	151-170	60-80	25-30

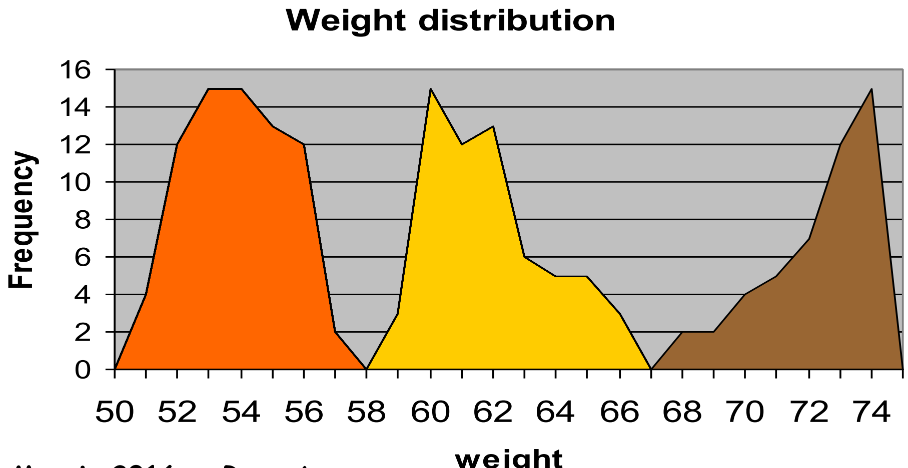
Problem: the discretization may be useless (see **weight**).

How to choose intervals?

1. Interval with a fixed "reasonable" granularity
   Ex. intervals of 10 cm for height.

2. Interval size is defined by some domain dependent criterion
   Ex.: 0-20ML, 21-22ML, 23-24ML, 25-26ML, >26ML

3. Interval size determined by analyzing data, studying the distribution or using clustering



50 - 58 kg
59-67 kg
> 68 kg

Discretization of quantitative attributes

1. Quantitative attributes are statically discretized by using predefined concept hierarchies:
   - elementary use of background knowledge

Loose interaction between Apriori and discretizer

2. Quantitative attributes are dynamically discretized
   - into “bins” based on the distribution of the data.
   - considering the distance between data points.

Tighter interaction between Apriori and discretizer

Constraints and AR

- **Preprocessing**: use constraints to focus on a subset of transactions
  - Example: find association rules where the prices of all items are at most 200 Euro
- **Optimizations:** use constraints to optimize Apriori algorithm
  - Anti-monotonicity: when a set violates the constraint, so does any of its supersets.
  - Apriori algorithm uses this property for pruning
- **Push constraints as deep as possible** inside the frequent set computation

Constraint-based AR

- **What kinds of constraints can be used in mining?**
  - Data constraints:
    - ✓ SQL-like queries
      - Find product pairs sold together in Vancouver in Dec.'98.
    - ✓ OLAP-like queries (Dimension/level)
      - in relevance to region, price, brand, customer category.
  - Rule constraints:
    - ✓ specify the form or property of rules to be mined.
    - ✓ Constraint-based AR

Rule Constraints

- Two kind of constraints:
  - Rule form constraints: meta-rule guided mining.
    - ✓ P(x, y) ^ Q(x, w) → takes(x, "database systems").
  - Rule content constraint: constraint-based query optimization (Ng, et al., SIGMOD'98).
    - ✓ sum(LHS) < 100 ^ min(LHS) > 20 ^ sum(RHS) > 1000
- 1-variable vs. 2-variable constraints (Lakshmanan, et al. SIGMOD'99):
  - 1-var: A constraint confining only one side (L/R) of the rule, e.g., as shown above.
  - 2-var: A constraint confining both sides (L and R).
    - ✓ sum(LHS) < min(RHS) ^ max(RHS) < 5* sum(LHS)

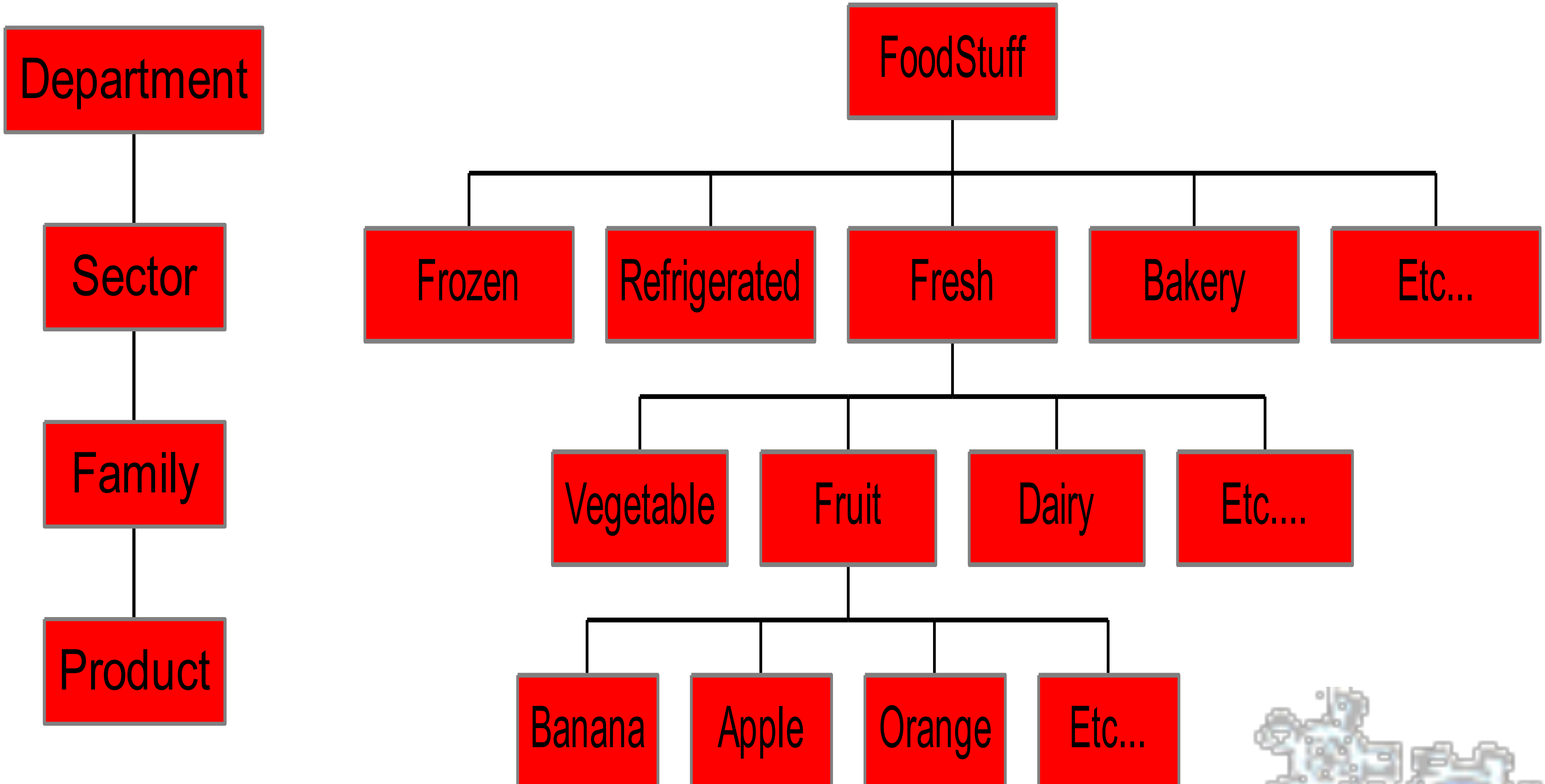
Association rules - module outline

- **What are association rules (AR) and what are they used for:**
  - The paradigmatic application: Market Basket Analysis
  - The single dimensional AR (intra-attribute)
- **How to compute AR**
  - Basic Apriori Algorithm and its optimizations
  - Multi-Dimension AR (inter-attribute)
  - Quantitative AR
  - Constrained AR
- **How to reason on AR and how to evaluate their quality**
  - Multiple-level AR
  - Interestingness
  - Correlation vs. Association

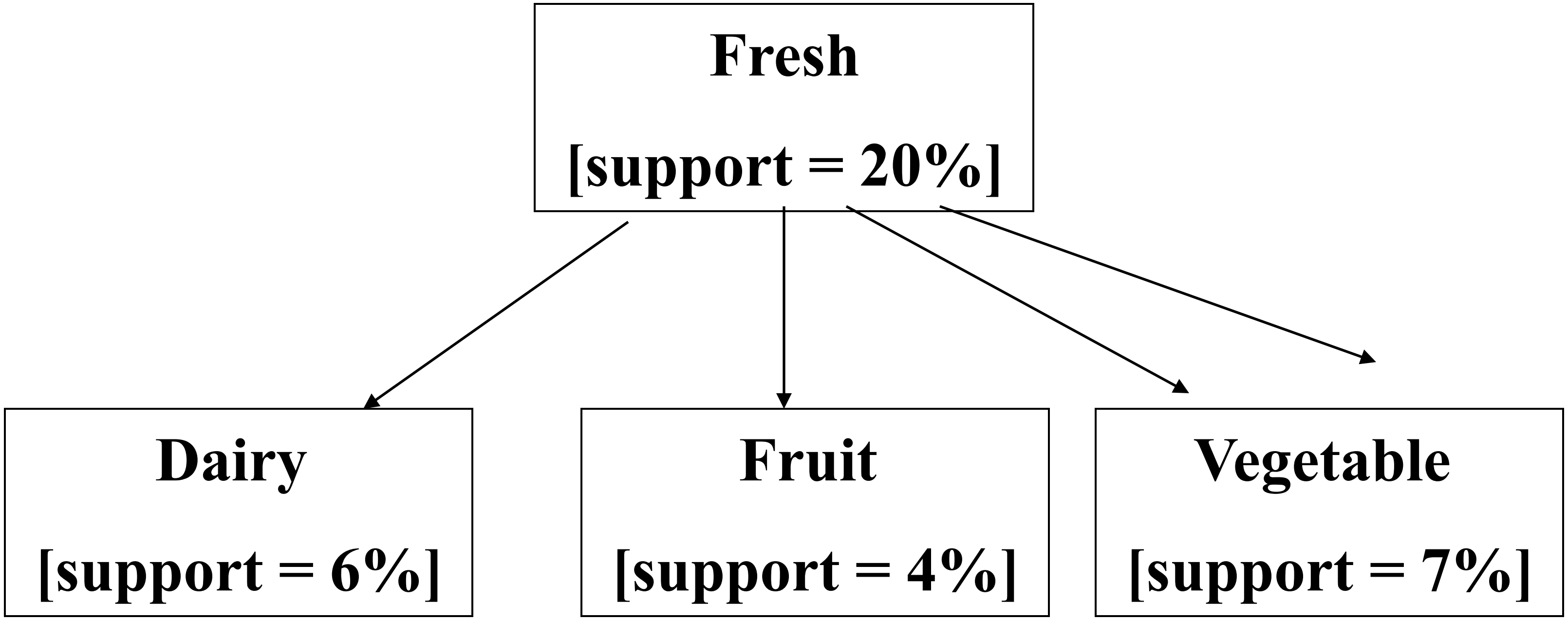
Multilevel AR

- Is difficult to find interesting patterns at a too primitive level
  - high support = too few rules
  - low support = too many rules, most uninteresting
- Approach: reason at suitable level of abstraction
- A common form of background knowledge is that an attribute may be generalized or specialized according to a hierarchy of concepts
- Dimensions and levels can be efficiently encoded in transactions
- Multilevel Association Rules : rules which combine associations with hierarchy of concepts

Hierarchy of concepts



Multilevel AR



Fresh ⇒ Bakery [20%, 60%]
Dairy ⇒ Bread [6%, 50%]
Fruit ⇒ Bread [1%, 50%] is not valid

Progetto "COOL PATTERNS"

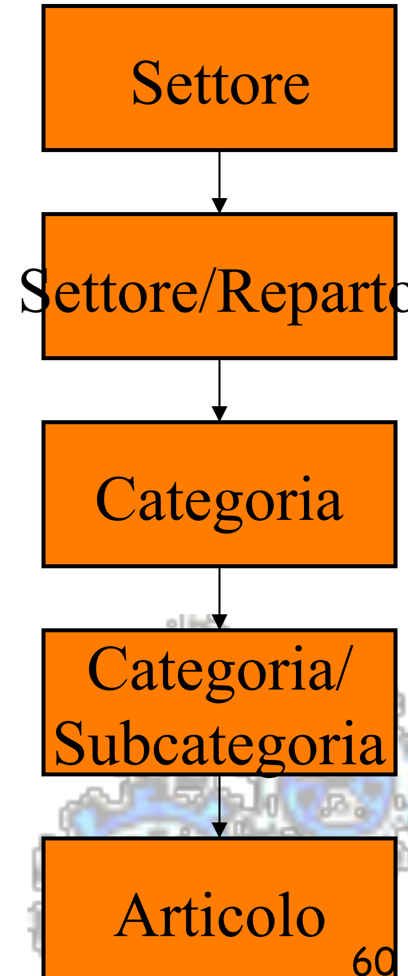
Analisi delle vendite nella grande distribuzione

Analisi dei Dati ed Estrazione di Conoscenza 2004/2005

Federico Colla

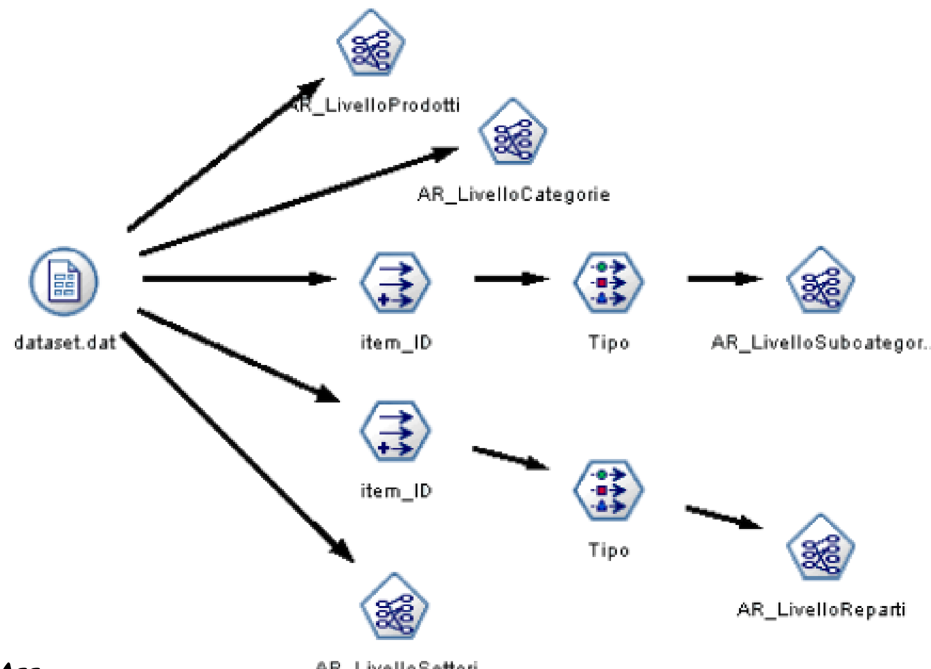
Data Understanding:
Data description – Gerarchia prodotti

- **La descrizione della gerarchia degli articoli è specificata nel file Excel `Classificazione Marketing.xls`.**
- **Si estraggono 4 tabelle che descrivono ciascuna un livello della gerarchia (chiave, descrizione)**
  - **Settori**
    - ✓ 9 record, 2 campi (chiave: *cod_settore)*
  - **Reparti**
    - ✓ 54 record, 3 campi (chiave: *cod_settore + cod_reparto)*
  - **Categorie**
    - ✓ 402 record, 4 campi (chiave: *cod_categ*)
  - **Subcategorie**
    - ✓ 1 516 record, 5 campi (chiave: *cod_categ + cod_subcateg*)



Modeling – Obj 1
Estrazione regole associative

- *Clementine* ha permesso di effettuare l'analisi usando i dati in formato *transazionale*.
- L'attributo *key* identifica ogni transazione.
- A seconda del livello di astrazione considerato, i codici di articolo, subcategoria, categoria, reparto e settore sono gli attributi di input/output.



Modeling – Obj 1
Estrazione regole associative

- La strategia utilizzata per l'estrazione delle regole è quella del *reduced support*.
- Ogni livello di astrazione ha la sua soglia di supporto minimo
  - più basso è il livello nella gerarchia, più piccola è la soglia di supporto minimo corrispondente.

Livello	Supporto minimo	Confidenza minima
Articoli	0,01%	80%
Subcategorie	0,2%	75%
Categorie	0,7%	75%
Reparti	4%	75%
Settori	8%	80%

Regole interessanti → Lift maggiore di 1

Evaluation – Obj 1
Regole associative interessanti

- L'insieme di regole ottenuto è stato esportato in un file di testo in cui esiste un record per ogni regola

Istanze	Supporto	Confidenza	Lift	Conseguente	Antecedente 1
53	0.01	92.5	4237.263	283917	283920

- Le regole ottenute non sono direttamente interpretabili.
- E' stato scritto il programma *PrettyPrinterApriori* che, data una regola "grezza", restituisce la corrispondente descrizione testuale.

[10 BICCH.CART.BIBO CIRC.200CC] → [PIATTI CART.BIBO CIRCUS D23X10]

Line: 70 Support: 0,01 Confidence: 92,5 Lift: 4237,263

Evaluation – Obj 1
Regole associative – Articoli

- [10 BICCH.CART.BIBO CIRC.200CC] → [PIATTI CART.BIBO CIRCUS D23X10]
  - Support: 0,01 Confidence: 92,5 Lift: 4237,263
- [TELO 100X150 460 GR/MQ TU] [OSPITE 40X60 460 GR/MQ TU] → [ASCIUGAMANO 60X110 460 GR TU]
  - Support: 0,01 Confidence: 91,4 Lift: 965,993
- [BOCC.CANI POLLO/TACCH.KG1.23] [BOC/NI GATTO VITELLO SIM.KG415] → [BOCC.GATTI CONIGLIO SIMBA G415]
  - Support: 0,01 Confidence: 91,4 Lift: 390,042
- [PIATTO FRUTTA MAZIME B.CO CM21] [PIATTO F.DO MAXIME B.CO CM.17] → [PIATTO P.NO MAXIME B.CO CM.25]
  - Support: 0,01 Confidence: 90 Lift: 3052,386
- [LENZUOLO PIANO 150X280 RIGHE] [LENZUOLO ANGOLI 90X200 TU] → [FEDERA 50X80 STAMPA RIGHE]
  - Support: 0,01 Confidence: 87,8 Lift: 809,222

Evaluation – Obj 1
Regole associative – Articoli

- [GOURM.GOLD DADINI GELLEE G85X8] [GOURMET PERLE FIL.C/MANZO G85] →[GOURMET PERLE FIL.CONIGLIO G85]
  - Support: 0,01 Confidence: 87,8 Lift: 492,757
- [CUCCHIAIONE ACCIAIO INOX] [PALA FRITTO ACCIAIO INOX] [FORCHETTONE ACCIAIO INOX] → [SCHIUMAROLA IN ACCIAIO INOX]
  - Support: 0,01 Confidence: 85,7 Lift: 1912,523
- [APER.CAMPARI MIXX PEACH ML275] [APERIT.CAMPARI MIXX LIME ML275] [APERITIVO CAMP.GRADI 6,5 ML275] → [CAMPARI MIXX ORANGE ML275]
  - Support: 0,01 Confidence: 83,3 Lift: 1314,55
- [GASSOSA S. BENEDETTO LT.1.5] [CEDRATA SAN BENEDETTO LT.1.5] [ARANCIATA S.BENEDETTO LT.1,5] → [SPUMA BIONDA LT1.5 S.BENEDETTO]
  - Support: 0,01 Confidence: 83 Lift: 172,76
- [BARAT.OVALE LT1,7 VTR COP.ACC.] [BARAT.OVALE LT0,84 VTR COP.ACC] → [BARAT.OVALE LT1,2 VTR COP.ACC.]
  - Support: 0,01 Confidence: 82,9 Lift: 1993,002
- [MOUSSE GAT.COOP MANZ/FEGAT.G85] [MOUSSE GAT.COOP PES/TROTAG85] → [MOUSSE GATTO COOP POL/TAC.G85]
  - Support: 0,1 Confidence: 81,7 Lift: 712,617

Evaluation – Obj 1
Regole associative – Subcategorie

- [BIBITE-ARANCIATE] [SNACK SALATI-PATATINE] [USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI] [USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA] → [BIBITE-COLE]
  - Support: 0,1 Confidence: 88,2 Lift: 11,084
- [USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA] [USA E GETTA TAVOLA-ACCESSORI USA E GETTA] [USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA] → [USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI]
  - Support: 0,1 Confidence: 84,7 Lift: 12,767
- [USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA] [USA E GETTA TAVOLA-STOV. PLAST. COLORATA DECORATA] → [USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI]
  - Support: 0,1 Confidence: 82,2 Lift: 12,391
- [SNACK SALATI-POP CORN/CEREALI] [SNACK SALATI-ESTRUSI] [BIBITE-ARANCIATE] → [BIBITE-COLE]
  - Support: 0,1 Confidence: 82,2 Lift: 10,34
- [CARAMELLE/PROD. BASE ZUCCH.-ALTRE CARAMELLE] [CARAMELLE/PROD. BASE ZUCCH.-CARAM.NORMALI] [CARAMELLE/PROD. BASE ZUCCH.-GOMME DA MASTICARE] → [PRODOTTI BASE CIOCCOLATO-SNACK]
  - Support: 0,1 Confidence: 81,2 Lift: 8,693

Evaluation – Obj 1
Regole associative – Categorie

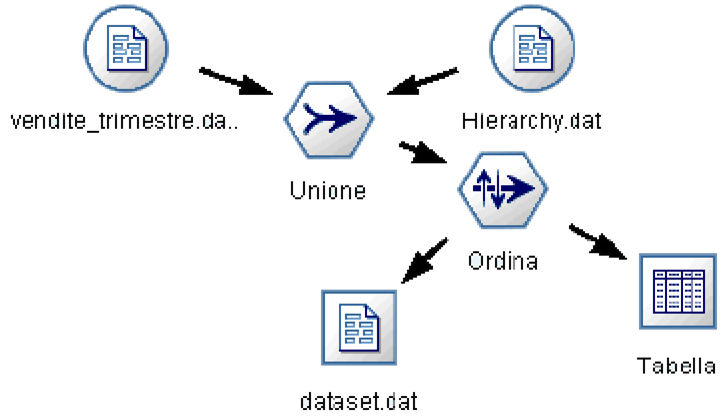
- [UOVA] [OF PREPARATA] [VERDURA FRESCA] [LATTE] [FRUTTA FRESCA] → [ORTAGGI]
  - Support: 0,8 Confidence: 85,2 Lift: 1,893
- [CAFFE] [UOVA] [VERDURA FRESCA] [FRUTTA FRESCA] → [ORTAGGI]
  - Support: 0,7 Confidence: 84,3 Lift: 1,871
- [UOVA] [GRASSI] [VERDURA FRESCA] [AVICUNICOLO] → [ORTAGGI]
  - Support: 0,9 Confidence: 83,5 Lift: 1,854
- [OLIO DI OLIVA] [UOVA] [SUINO] → [BOVINO]
  - Support: 0,7 Confidence: 78,9 Lift: 1,757
- [ZUCCHERO] [IGIENE CARTA] [DETERGENTI SUPERFICI] → [DETERGENZA TESSUTI]
  - Support: 0,7 Confidence: 76,6 Lift: 2,247

Evaluation – Obj 1
Regole associative – Reparti

- FRESCHI-CARNI BIANCHE] [FRESCHI-SURGELATI] [FRESCHI-GASTRONOMIA] → [FRESCHI-CARNI ROSSE]
  - Support: 5,2 Confidence: 75,5 Lift: 1,217

- Al livello di Settore, non sono state trovate regole aventi Lift maggiore di 1.

- **Infine si crea di dataset finale (per le regole associative**



- **Contiene 5 098 533 record e 7 campi**
-

key	nro_carta	cod_art	cod_categ	cod_subcateg	cod_reparto	cod_settore
01030011731	31403686	2561	009	01	01	01
01030011731	31403686	2545	009	01	01	01
01030011731	31403686	3393	009	03	01	01

Support and Confidence of Multilevel AR

- from specialized to general: support of rules increases (new rules may become valid)
- from general to specialized: support of rules decreases (rules may become not valid, their support falls under the threshold)
- Confidence is not affected

Multi-level Association Rules

- **Why should we incorporate concept hierarchy?**
  - **Rules at lower levels may not have enough support to appear in any frequent itemsets**
  - **Rules at lower levels of the hierarchy are overly specific**
    - ✓ e.g., skim milk → white bread, 2% milk → wheat bread,
      skim milk → wheat bread, etc.

      are indicative of association between milk and bread

Reasoning with Multilevel AR

- Too low level => too many rules and too primitive.
  Example: Apple Melinda ⇒ Colgate Tooth-paste
  It is a curiosity not a behavior

- Too high level => uninteresting rules
  Example: Foodstuff ⇒ Varia
- Redundancy => some rules may be redundant due to "ancestor" relationships between items.
  - A rule is redundant if its support is close to the "expected" value, based on the rule's ancestor.
- Example (milk has 4 subclasses)
  - milk ⇒ wheat bread, [support = 8%, confidence = 70%]
  - 2%-milk ⇒ wheat bread, [support = 2%, confidence = 72%]

Mining Multilevel AR

- Calculate frequent itemsets at each concept level, until no more frequent itemsets can be found
- For each level use Apriori
- A top_down, progressive deepening approach:
  - First find high-level strong rules:
    fresh → bakery [20%, 60%].
  - Then find their lower-level "weaker" rules:
    fruit → bread [6%, 50%].
- Variations at mining multiple-level association rules.
  - Level-crossed association rules:
    fruit → *wheat bread*
  - Association rules with multiple, alternative hierarchies:
    fruit → *Wonder bread*

Reasoning with AR

- **Significance**:
  Example: <1, {a, b}>
  <2, {a} >
  <3, {a, b, c}>
  <4, {b, d}>

  {b} ⇒ {a} has confidence (66%), but is not significant as support({a}) = 75%.

Beyond Support and Confidence

- Example 1: (Aggarwal & Yu, PODS98)

	coffee	not coffee	sum(row)
tea	20	5	25
not tea	70	5	75
sum(col.)	90	10	100

- {tea} => {coffee} has high support (20%) and confidence (80%)
- However, a priori probability that a customer buys coffee is 90%
  - A customer who is known to buy tea is less likely to buy coffee (by 10%)
  - There is a negative correlation between buying tea and buying coffee
  - {~tea} => {coffee} has higher confidence(93%)
  - P(coffe&tea)/p(coffe)*P(Tea)= 20/90*25=0,2/0,225= 0.88

Correlation and Interest

- Two events are independent
  if $P(A \wedge B) = P(A) * P(B)$, otherwise are correlated.
- Interest $= P(A \wedge B) / P(B) * P(A)$
- Interest expresses measure of correlation
  - $= 1 \Rightarrow$ A and B are independent events
  - less than 1 $\Rightarrow$ A and B negatively correlated,
  - greater than 1 $\Rightarrow$ A and B positively correlated.
  - In our example, I(*buy tea* $\wedge$ *buy coffee*)=0.89 i.e. they are negatively correlated.

Computing Interestingness Measure

- **Given a rule X → Y, information needed to compute rule interestingness can be obtained from a contingency table**

Contingency table for X → Y

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	$\|T\|$

f_{11}: support of X and Y
f_{10}: support of X and $\overline{Y}$
f_{01}: support of $\overline{X}$ and Y
f_{00}: support of $\overline{X}$ and $\overline{Y}$

Used to define various measures

- support, confidence, lift, Gini, J-measure, etc.

Statistical-based Measures

- **Measures that take into account statistical dependence**

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

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

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

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

Example: Lift/Interest

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

Association Rule: Tea → Coffee

Confidence= P(Coffee|Tea) = 0.93

but P(Coffee) = 0.93

⇒ Lift = 75/90*80= 1.04 (< 1, therefore is negatively associated)

Drawback of Lift & Interest

	Y	$\overline{Y}$	
X	10	0	10
$\overline{X}$	0	90	90
	10	90	100

	Y	$\overline{Y}$	
X	90	0	90
$\overline{X}$	0	10	10
	90	10	100

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

$$Lift = \frac{0.9}{(0.9)(0.9)} = 1.11$$

Statistical independence:

If P(X,Y)=P(X)P(Y) => Lift = 1

There are lots of measures 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?

#	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)	$\max\Big(P(A,B)\log(\frac{P(B\mid A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\mid A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\mid B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\mid B)}{P(\overline{A})})\Big)$
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)	$\max\Big(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\Big)$
13	Conviction (V)	$\max\Big(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\Big)$
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)	$\max\Big(\frac{P(B\mid A)-P(B)}{1-P(B)}, \frac{P(A\mid B)-P(A)}{1-P(A)}\Big)$
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)	$\sqrt{P(A,B)}\max(P(B\mid A)-P(B), P(A\mid B)-P(A))$

Domain dependent measures

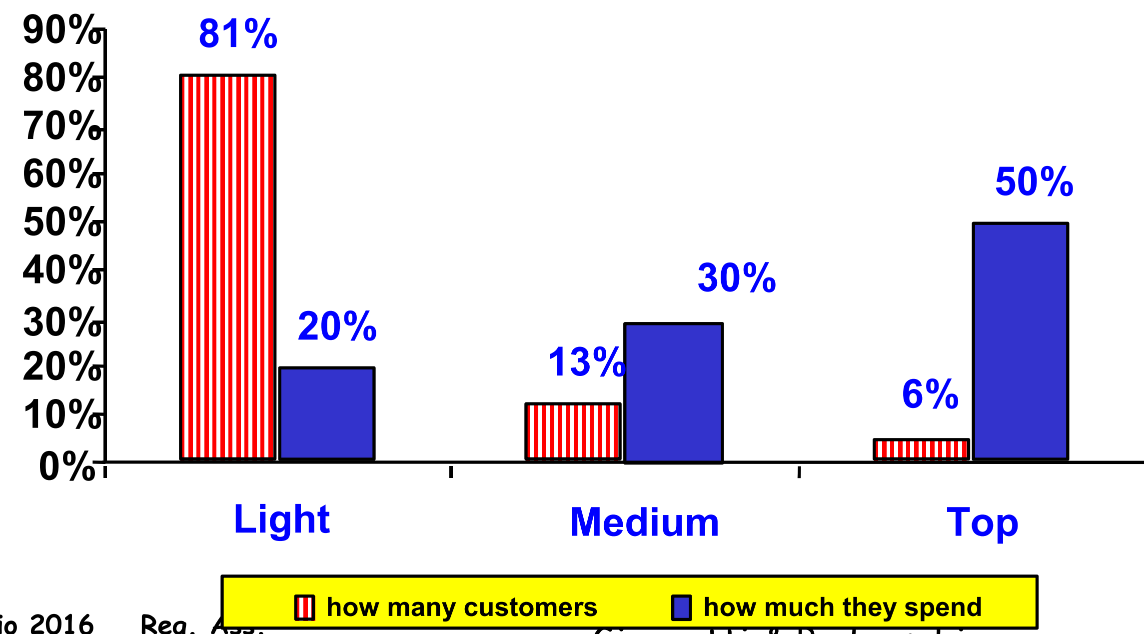
- **Together with support, confidence, interest, …, use also (in post-processing) domain-dependent measures**
- **E.g., use rule constraints on rules**
- Example: take only rules which are significant with respect their economic value
- **sum(LHS)+ sum(RHS) > 100**

Conclusions

- Association rule mining
  - probably the most significant contribution from the database community to KDD
  - A large number of papers have been published
- Many interesting issues have been explored
- An interesting research direction
  - Association analysis in other types of data: spatial data, multimedia data, time series data, etc.

Conclusion (2)

- MBA is a key factor of success in the competition of supermarket retailers.
- Knowledge of customers and their purchasing behavior brings potentially huge added value.



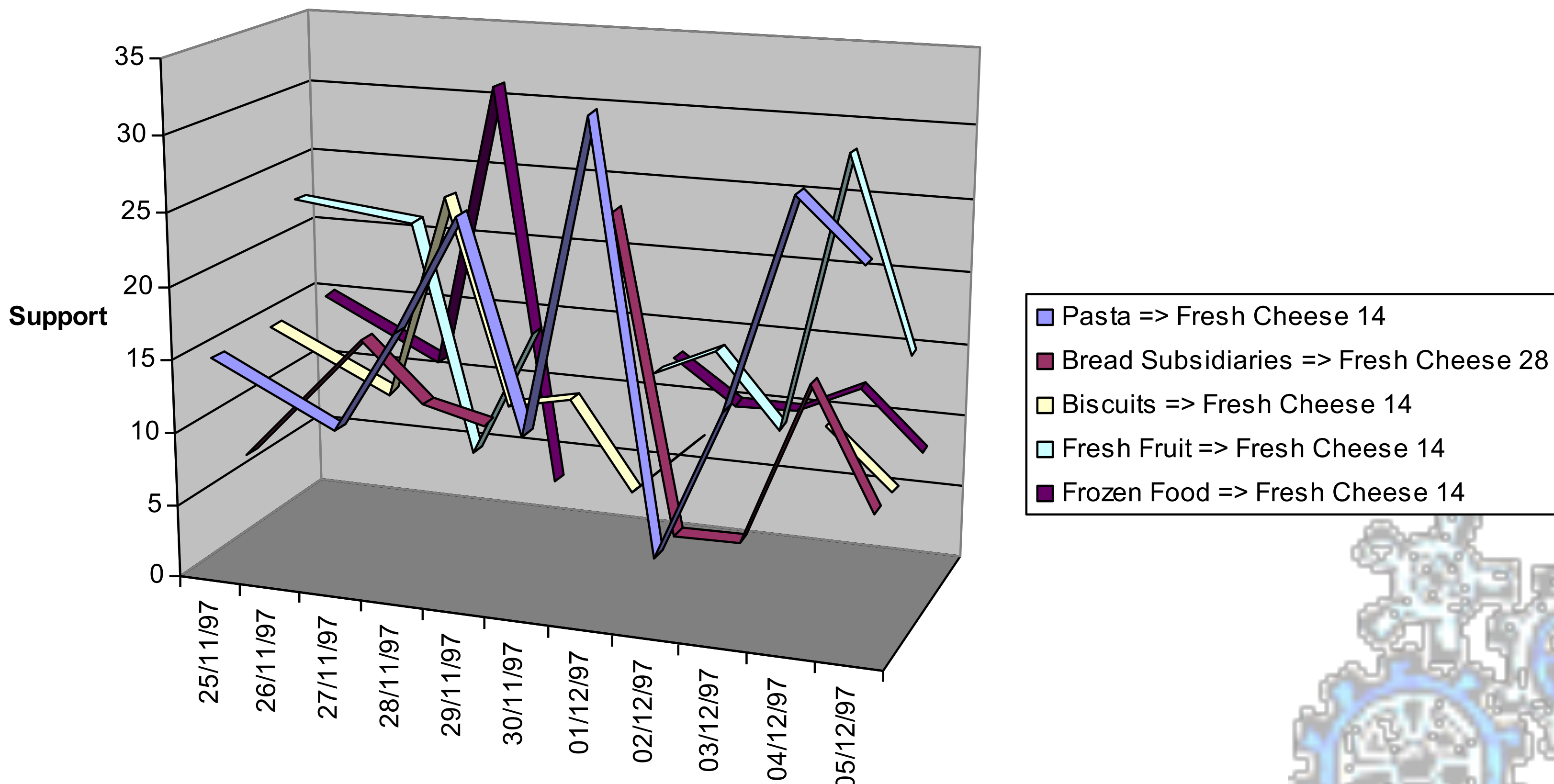
Which tools for market basket analysis?

- Association rule are needed but insufficient

- Market analysts ask for business rules:
  - Is supermarket assortment adequate for the company's target class of customers?
  - Is a promotional campaign effective in establishing a desired purchasing habit?

Business rules: temporal reasoning on AR

- Which rules are established by a promotion?
- How do rules change along time?



Association rules - module2 Examples

Association Rules in Web Miming
AR & Atherosclerosis prevention study
Moviegoer Data bases

MBA in Web Usage Mining

- **Association Rules in Web Transactions**
  - **discover affinities among sets of Web page references across user sessions**

- **Examples**
  - **60% of clients who accessed** `/products/`**, also accessed** `/products/software/webminer.htm`
  - **30% of clients who accessed** `/special-offer.html`**, placed an online order in** `/products/software/`
  - **Actual Example from IBM official Olympics Site:**
    - ✓ {Badminton, Diving} ==> {Table Tennis} [conf = 69.7%, sup = 0.35%]

- **Applications**
  - **Use rules to serve dynamic, customized contents to users**
  - **prefetch files that are most likely to be accessed**
  - **determine the best way to structure the Web site (site optimization)**
  - **targeted electronic advertising and increasing cross sales**

Web Usage Mining: Example

- Association Rules From Cray Research Web Site

Conf	supp	Association Rule
82.8	3.17	/PUBLIC/product-info/T3E ===> /PUBLIC/product-info/T3E/CRAY_T3E.html
90	0.14	/PUBLIC/product-info/J90/J90.html, /PUBLIC/product-info/T3E ===> /PUBLIC/product-info/T3E/CRAY_T3E.html
97.2	0.15	/PUBLIC/product-info/J90, /PUBLIC/product-info/T3E/CRAY_T3E.html, /PUBLIC/product-info/T90, ===> /PUBLIC/product-info/T3E, /PUBLIC/sc.html

- Design “suggestions”
  - from rules 1 and 2: there is something in J90.html that should be moved to th page /PUBLIC/product-info/T3E (why?)

MBA in Text / Web Content Mining

- Documents Associations
  - Find (content-based) associations among documents in a collection
  - Documents correspond to items and words correspond to transactions
  - Frequent itemsets are groups of docs in which many words occur in common

	Doc 1	Doc 2	Doc 3	. . .	Doc n
business	5	5	2	. . .	1
capital	2	4	3	. . .	5
fund	0	0	0	. . .	1
⋮	⋮	⋮	⋮	⋮ ⋮ ⋮	⋮
invest	6	0	0	. . .	3

- Term Associations
  - Find associations among words based on their occurrences in documents
  - similar to above, but invert the table (terms as items, and docs as transactions)

Atherosclerosis prevention study

2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC)

Atherosclerosis prevention study:

- The STULONG 1 data set is a real database that keeps information about the study of the development of atherosclerosis risk factors in a population of middle aged men.
- Used for Discovery Challenge at PKDD 00-02-03-04

Atherosclerosis prevention study:

- Study on 1400 middle-aged men at Czech hospitals
  - Measurements concern development of cardiovascular disease and other health data in a series of exams
- The aim of this analysis is to look for associations between medical characteristics of patients and death causes.
- Four tables
  - Entry and subsequent exams, questionnaire responses, deaths

The input data

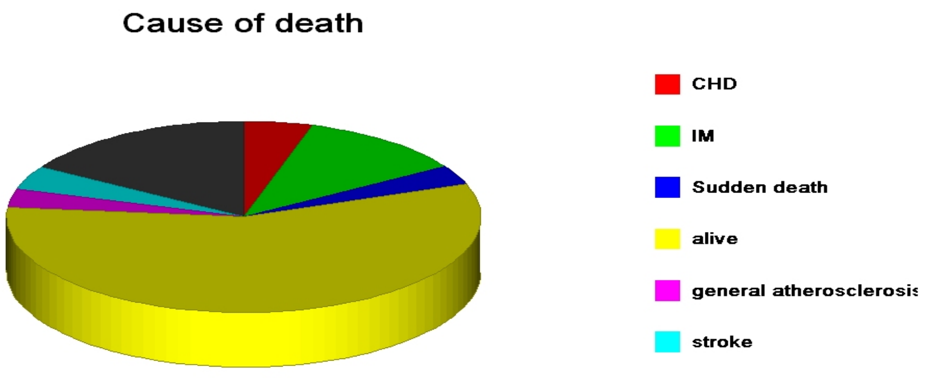
Data from Entry and Exams		
General characteristics	**Examinations**	**habits**
Marital status	Chest pain	Alcohol
Transport to a job	Breathlesness	Liquors
Physical activity in a job	Cholesterol	Beer 10
Activity after a job	Urine	Beer 12
Education	Subscapular	Wine
Responsibility	Triceps	Smoking
Age		Former smoker
Weight		Duration of smoking
Height		Tea
		Sugar
		Coffee

The input data

DEATH CAUSE	PATIENTS	%
myocardial infarction	80	20.6
coronary heart disease	33	8.5
stroke	30	7.7
other causes	79	20.3
sudden death	23	5.9
unknown	8	2.0
tumorous disease	114	29.3
general atherosclerosis	22	5.7
TOTAL	389	100.0

Data selection

- When joining "Entry" and "Death" tables we implicitely create a new attribute "Cause of death", which is set to "alive" for subjects present in the "Entry" table but not in the "Death" table.
- We have only 389 subjects in death table.



The prepared data

Patient	General characteristics		Examinations		Habits		Cause of death
	Activity after work	**Education**	**Chest pain**	**…**	**Alcohol**	**…..**	
1	moderate activity	university	not present		no		Stroke
2	great activity		not ischaemic		occasionally		myocardial infarction
3	he mainly sits		other pains		regularly		tumorous disease
……	……..	……..	………..	..	…	……	alive
389	he mainly sits		other pains		regularly		tumorous disease

Descriptive Analysis/ Subgroup Discovery /Association Rules

Are there strong relations concerning death cause?

General characteristics (?) ⇒ Death cause (?)

Examinations (?) ⇒ Death cause (?)

Habits (?) ⇒ Death cause (?)

Combinations (?) ⇒ Death cause (?)

Example of extracted rules

- Education(university) & Height<176-180> ⇒Death cause (tumouros disease), *16 ; 0.62*
- **It means that on tumorous disease have died 16, i.e. 62% of patients with university education and with height 176-180 cm.**

Example of extracted rules

- **Physical activity in work(he mainly sits) & Height<176-180> ⇒ Death cause (tumouros disease),** ***24; 0.52***
- **It means that on tumorous disease have died 24 i.e. 52% of patients that mainly sit in the work and whose height is 176-180 cm.**

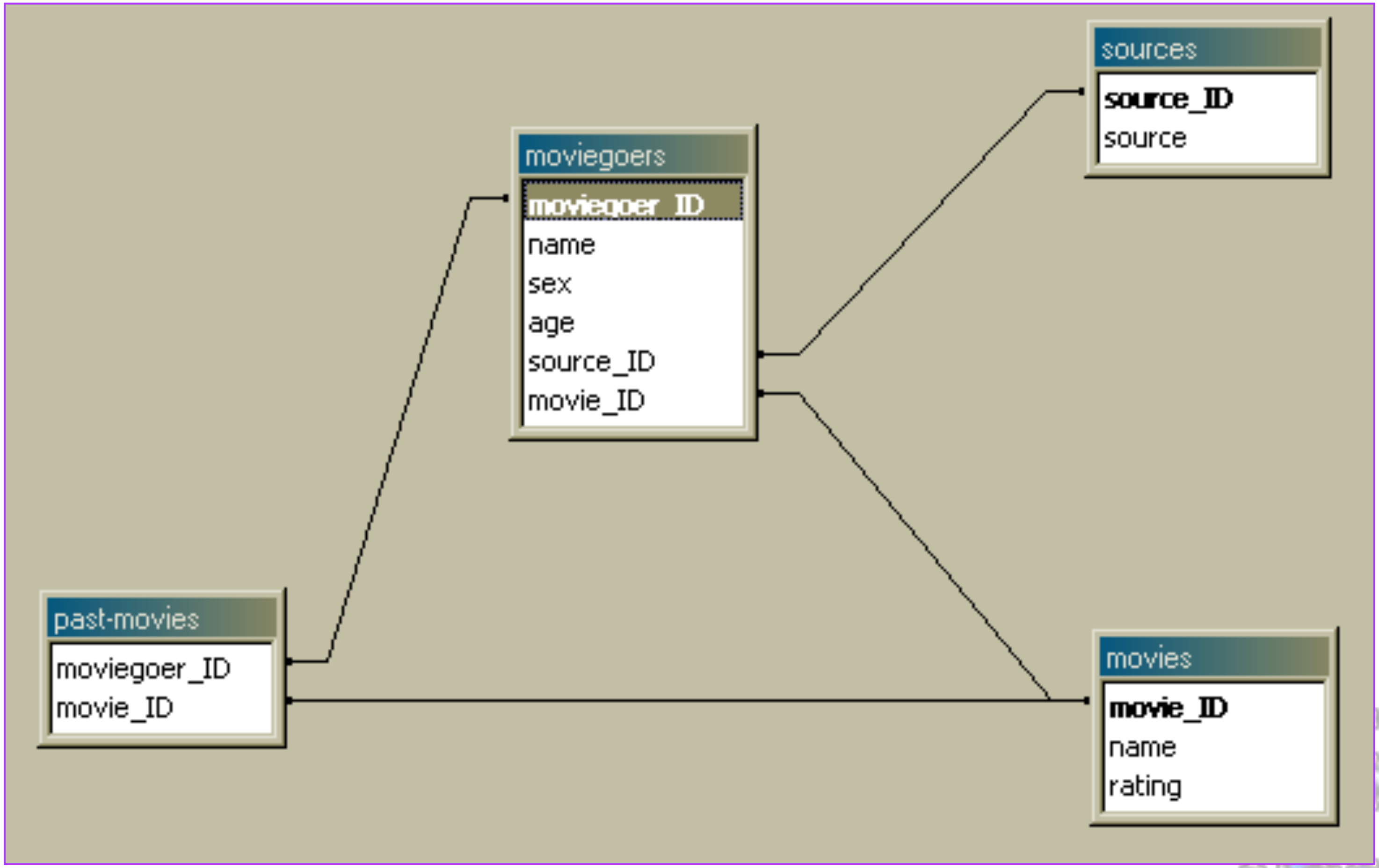
Example of extracted rules

- **Education(university) & Height<176-180> ⇒Death cause (tumouros disease),**
  16; 0.62; *+1.1;*
- **the** relative frequency **of patients who died on tumorous disease among patients with university education and with height 176-180 cm is 110 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients**

Association rules - Examples

Association Rules in Web Miming
AR & Atherosclerosis prevention study
Moviegoer Data bases

Example2: Moviegoer Database



Example: Moviegoer Database

```
SELECT moviegoers.name, moviegoers.sex, moviegoers.age,
sources.source, movies.name
FROM movies, sources, moviegoers
WHERE sources.source_ID = moviegoers.source_ID AND
      movies.movie_ID = moviegoers.movie_ID
ORDER BY moviegoers.name;
```

moviegoers.name	sex	age	source	movies.name
Amy	f	27	Oberlin	Independence Day
Andrew	m	25	Oberlin	12 Monkeys
Andy	m	34	Oberlin	The Birdcage
Anne	f	30	Oberlin	Trainspotting
Ansje	f	25	Oberlin	I Shot Andy Warhol
Beth	f	30	Oberlin	Chain Reaction
Bob	m	51	Pinewoods	Schindler's List
Brian	m	23	Oberlin	Super Cop
Candy	f	29	Oberlin	Eddie
Cara	f	25	Oberlin	Phenomenon
Cathy	f	39	Mt. Auburn	The Birdcage
Charles	m	25	Oberlin	Kingpin
Curt	m	30	MRJ	T2 Judgment Day
David	m	40	MRJ	Independence Day
Erica	f	23	Mt. Auburn	Trainspotting

Example: Moviegoer Database

- Association Rules
  - market basket analysis (MBA): “which movies go together?”
  - need to create “transactions” for each moviegoer containing movies seen by that moviegoer:

name	TID	Transaction
Amy	001	{Independence Day, Trainspotting}
Andrew	002	{12 Monkeys, The Birdcage, Trainspotting, Phenomenon}
Andy	003	{Super Cop, Independence Day, Kingpin}
Anne	004	{Trainspotting, Schindler's List}
...	...	...

  - may result in association rules such as:

```
{“Phenomenon”, “The Birdcage”} ==> {“Trainspotting”}
{“Trainspotting”, “The Birdcage”} ==> {sex = “f”}
```

Example: Moviegoer Database

- Sequence Analysis
  - similar to MBA, but order in which items appear in the pattern is important
  - e.g., people who rent “The Birdcage” during a visit tend to rent “Trainspotting” in the next visit.

Sequential Patterns

Sequential / Navigational Patterns

- Sequential patterns add an extra dimension to frequent itemsets and association rules - time.
  - Items can appear before, after, or at the same time as each other.
  - General form: "x% of the time, when A appears in a transaction, B appears within z transactions."
    - ✓ note that other items may appear between A and B, so sequential patterns do not necessarily imply consecutive appearances of items (in terms of time)
- Examples
  - Renting "Star Wars", then "Empire Strikes Back", then "Return of the Jedi" in that order
  - Collection of ordered events within an interval
  - Most sequential pattern discovery algorithms are based on extensions of the Apriori algorithm for discovering itemsets
- Navigational Patterns
  - they can be viewed as a special form of sequential patterns which capture navigational patterns among users of a site
  - in this case a session is a consecutive sequence of pageview references for a user over a specified period of time

Mining Sequences - Example

Customer-sequence

CustId	Video sequence
1	{(C), (H)}
2	{(AB), (C), (DFG)}
3	{(CEG)}
4	{(C), (DG), (H)}
5	{(H)}

Sequential patterns with support > 0.25

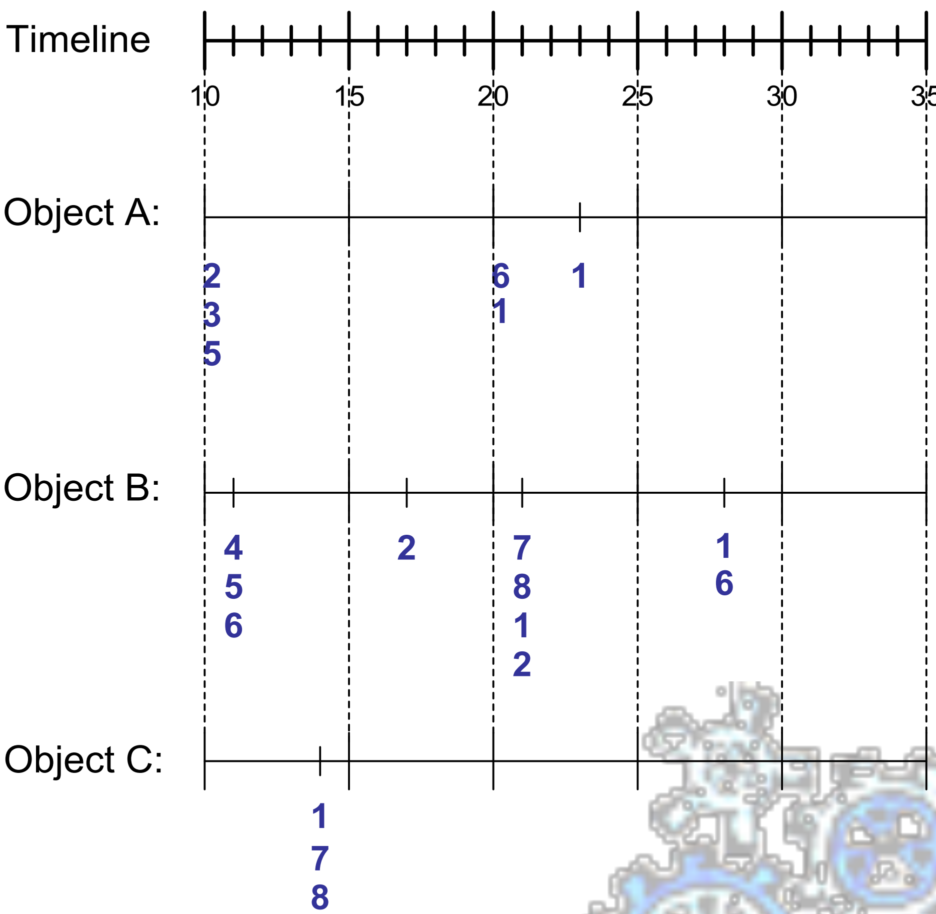
{(C), (H)}

{(C), (DG)}

Sequence Data

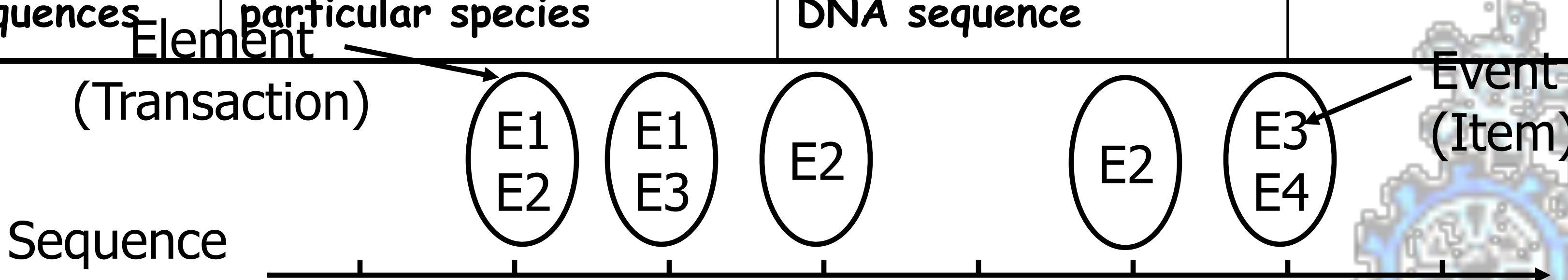
Sequence Database:

Object	Timestamp	Events
A	10	2, 3, 5
A	20	6, 1
A	23	1
B	11	4, 5, 6
B	17	2
B	21	7, 8, 1, 2
B	28	1, 6
C	14	1, 8, 7



Examples of Sequence Data

Sequence Database	Sequence	Element (Transaction)	Event (Item)
Customer	Purchase history of a given customer	A set of items bought by a customer at time t	Books, diary products, CDs, etc
Web Data	Browsing activity of a particular Web visitor	A collection of files viewed by a Web visitor after a single mouse click	Home page, index page, contact info, etc
Event data	History of events generated by a given sensor	Events triggered by a sensor at time t	Types of alarms generated by sensors
Genome sequences	DNA sequence of a particular species	An element of the DNA sequence	Bases A,T,G,C



Formal Definition of a Sequence

- A sequence is an ordered list of elements (transactions)

$$s = \langle e_1 \, e_2 \, e_3 \, \ldots \rangle$$

  - Each element contains a collection of events (items)

$$e_i = \{i_1, \, i_2, \, \ldots, \, i_k\}$$

  - Each element is attributed to a specific time or location
- Length of a sequence, $|s|$, is given by the number of elements of the sequence
- A k-sequence is a sequence that contains k events (items)

Examples of Sequence

- **Web sequence:**

  < {Homepage} {Electronics} {Digital Cameras} {Canon Digital Camera} {Shopping Cart} {Order Confirmation} {Return to Shopping} >

- **Sequence of initiating events causing the nuclear accident at 3-mile Island:**
  (http://stellar-one.com/nuclear/staff_reports/summary_SOE_the_initiating_event.htm)

  < {clogged resin} {outlet valve closure} {loss of feedwater} {condenser polisher outlet valve shut} {booster pumps trip} {main waterpump trips} {main turbine trips} {reactor pressure increases}>

- **Sequence of books checked out at a library:**

  <{Fellowship of the Ring} {The Two Towers} {Return of the King}>

Formal Definition of a Subsequence

- A sequence $<a_1\ a_2\ \dots\ a_n>$ is contained in another sequence $<b_1\ b_2\ \dots\ b_m>$ $(m \geq n)$ if there exist integers $i_1 < i_2 < \dots < i_n$ such that $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i1}$, ..., $a_n \subseteq b_{in}$

Data sequence	Subsequence	Contain?
< {2,4} {3,5,6} {8} >	< {2} {3,5} >	Yes
< {1,2} {3,4} >	< {1} {2} >	No
< {2,4} {2,4} {2,5} >	< {2} {4} >	Yes

- The support of a subsequence w is defined as the fraction of data sequences that contain w
- A *sequential pattern* is a frequent subsequence (i.e., a subsequence whose support is ≥ *minsup*)

Sequential Pattern Mining: Definition

- Given:
  - a database of sequences
  - a user-specified minimum support threshold, *minsup*
- Task:
  - Find all subsequences with support ≥ *minsup*

Extracting Sequential Patterns

- **Given n events:** $i_1, i_2, i_3, \ldots, i_n$
- **Candidate 1-subsequences:**

  $<\{i_1\}>, <\{i_2\}>, <\{i_3\}>, \ldots, <\{i_n\}>$
- **Candidate 2-subsequences:**

  $<\{i_1, i_2\}>, <\{i_1, i_3\}>, \ldots, <\{i_1\} \{i_1\}>, <\{i_1\} \{i_2\}>, \ldots, <\{i_{n-1}\} \{i_n\}>$
- **Candidate 3-subsequences:**

  $<\{i_1, i_2 , i_3\}>, <\{i_1, i_2 , i_4\}>, \ldots, <\{i_1, i_2\} \{i_1\}>, <\{i_1, i_2\} \{i_2\}>, \ldots,$

  $<\{i_1\} \{i_1 , i_2\}>, <\{i_1\} \{i_1 , i_3\}>, \ldots, <\{i_1\} \{i_1\} \{i_1\}>, <\{i_1\} \{i_1\} \{i_2\}>, \ldots$

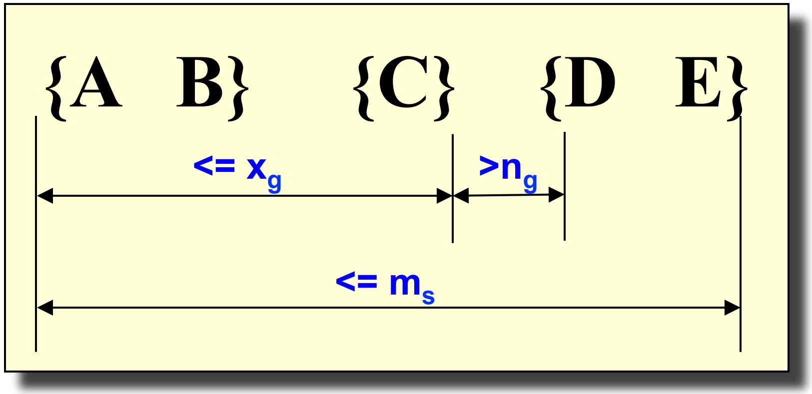
Generalized Sequential Pattern (GSP)

- Step 1:
  - **Make the first pass over the sequence database D to yield all the 1-element frequent sequences**
- Step 2:

Repeat until no new frequent sequences are found

  - Candidate Generation:
    - ✓ Merge pairs of frequent subsequences found in the (k-1)*th* pass to generate candidate sequences that contain k items
  - Candidate Pruning:
    - ✓ Prune candidate *k*-sequences that contain infrequent *(k-1)*-subsequences
  - Support Counting:
    - ✓ Make a new pass over the sequence database D to find the support for these candidate sequences
  - Candidate Elimination:
    - ✓ Eliminate candidate *k*-sequences whose actual support is less than *minsup*

Timing Constraints (I)



x_g: max-gap

n_g: min-gap

m_s: maximum span

$x_g = 2$, $n_g = 0$, $m_s = 4$

Data sequence	Subsequence	Contain?
< {2,4} {3,5,6} {4,7} {4,5} {8} >	< {6} {5} >	Yes
< {1} {2} {3} {4} {5}>	< {1} {4} >	No
< {1} {2,3} {3,4} {4,5}>	< {2} {3} {5} >	Yes
< {1,2} {3} {2,3} {3,4} {2,4} {4,5}>	< {1,2} {5} >	No

References - Association rules

- R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. SIGMOD'93, 207-216, Washington, D.C.
- R. Agrawal and R. Srikant. Fast algorithms for mining association rules. VLDB'94 487-499, Santiago, Chile.
- R. Agrawal and R. Srikant. Mining sequential patterns. ICDE'95, 3-14, Taipei, Taiwan.
- R. J. Bayardo. Efficiently mining long patterns from databases. SIGMOD'98, 85-93, Seattle, Washington.
- S. Brin, R. Motwani, and C. Silverstein. Beyond market basket: Generalizing association rules to correlations. SIGMOD'97, 265-276, Tucson, Arizona..
- D.W. Cheung, J. Han, V. Ng, and C.Y. Wong. Maintenance of discovered association rules in large databases: An incremental updating technique. ICDE'96, 106-114, New Orleans, LA..
- T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. SIGMOD'96, 13-23, Montreal, Canada.
- E.-H. Han, G. Karypis, and V. Kumar. Scalable parallel data mining for association rules. SIGMOD'97, 277-288, Tucson, Arizona.
- J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. VLDB'95, 420-431, Zurich, Switzerland.
- M. Kamber, J. Han, and J. Y. Chiang. Metarule-guided mining of multi-dimensional association rules using data cubes. KDD'97, 207-210, Newport Beach, California.
- M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. CIKM'94, 401-408, Gaithersburg, Maryland.
- R. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained associations rules. SIGMOD'98, 13-24, Seattle, Washington.
- B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. ICDE'98, 412-421, Orlando, FL.
- J.S. Park, M.S. Chen, and P.S. Yu. An effective hash-based algorithm for mining association rules. SIGMOD'95, 175-186, San Jose, CA.
- S. Ramaswamy, S. Mahajan, and A. Silberschatz. On the discovery of interesting patterns in association rules. VLDB'98, 368-379, New York, NY.
- S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. SIGMOD'98, 343-354, Seattle, WA.

References - Association rules

- A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. VLDB'95, 432-443, Zurich, Switzerland.
- C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. VLDB'98, 594-605, New York, NY.
- R. Srikant and R. Agrawal. Mining generalized association rules. VLDB'95, 407-419, Zurich, Switzerland.
- R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. SIGMOD'96, 1-12, Montreal, Canada.
- R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. KDD'97, 67-73, Newport Beach, California.
- D. Tsur, J. D. Ullman, S. Abitboul, C. Clifton, R. Motwani, and S. Nestorov. Query flocks: A generalization of association-rule mining. SIGMOD'98, 1-12, Seattle, Washington.
- B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. ICDE'98, 412-421, Orlando, FL.
- R.J. Miller and Y. Yang. Association rules over interval data. SIGMOD'97, 452-461, Tucson, Arizona.
- J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series database. ICDE'99, Sydney, Australia.
- F. Giannotti, G. Manco, D. Pedreschi and F. Turini. Experiences with a logic-based knowledge discovery support environment. In Proc. 1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (SIGMOD'99 DMKD). Philadelphia, May 1999.
- F. Giannotti, M. Nanni, G. Manco, D. Pedreschi and F. Turini. Integration of Deduction and Induction for Mining Supermarket Sales Data. In Proc. PADD'99, Practical Application of Data Discovery, Int. Conference, London, April 1999.