



COMP3420: Advanced Databases and Data Mining

Advanced association mining



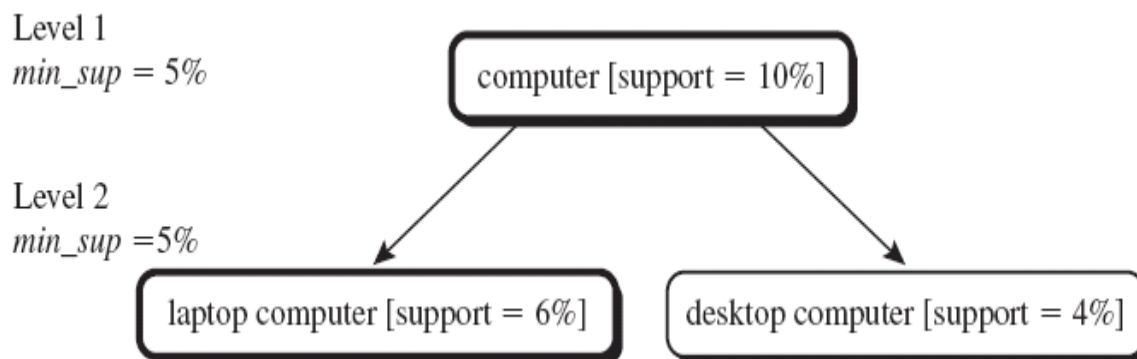
Lecture outline

- Mining various kinds of association rules
 - Multi-level association
 - Multi-dimensional association
 - Quantitative association
 - Interesting correlation patterns
- Constraints based mining
- Interestingness measure: Correlation (Lift)
 - More interestingness measures
- Visualisation of association rules

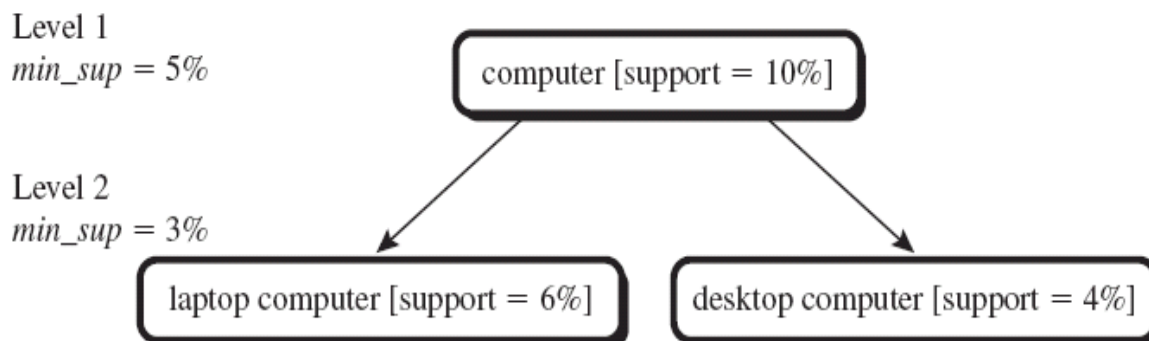
Multi-level association mining (2)

- Items often form hierarchies
- Items at lower levels are expected to have lower support
 - Flexible support setting (uniform, reduced, or group-based (user specific))

Uniform support



Reduced support



Multi-level association mining (3)

- Some rules may be redundant due to *ancestor* relationships between items
- For example:
 $buys(X, \text{'milk'}) \Rightarrow buys(X, \text{'bread'}) \quad [8\%, 70\%]$
 $buys(X, \text{'skim milk'}) \Rightarrow buys(X, \text{'bread'}) \quad [2\%, 72\%]$
 - The first rule is said to be an *ancestor* of the second rule
- A rule is redundant if its support is close to the “expected” value, based on the rule’s ancestor
 - For example, if around 25% of all milk purchased is `skim milk', then the second rule above is redundant, as it has a $\frac{1}{4}$ of the support of the first, more general rule (and similar confidence)

Multi-dimensional association mining

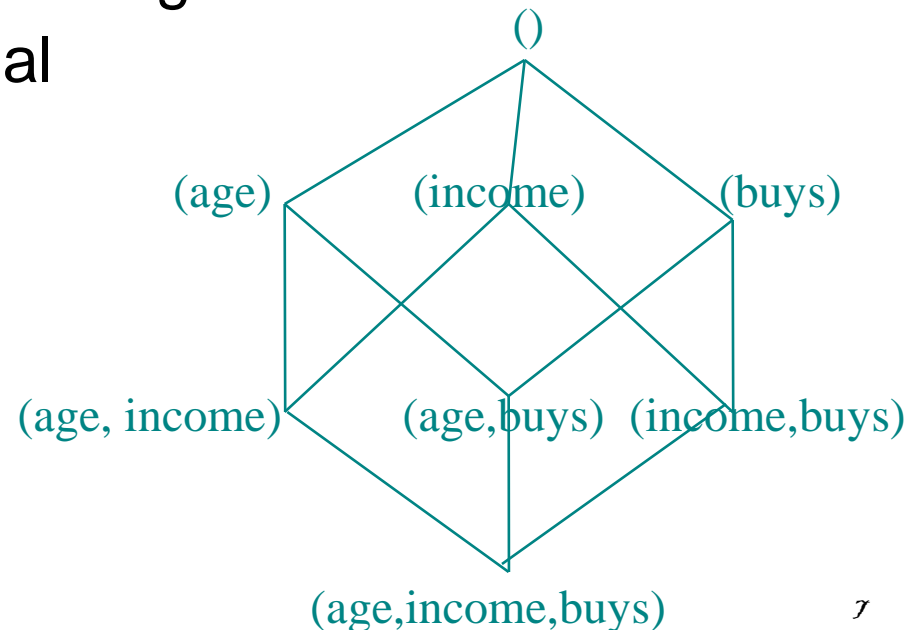
- Single-dimensional rules: $buys(X, \text{'milk'}) \Rightarrow buys(X, \text{'bread'})$
- Multi-dimensional rules: Two or more dimensions or predicates
 - Inter-dimension association rules (*no repeated predicates*):
 $age(X, \text{'19-25'}) \text{ and } occupation(X, \text{'student'}) \Rightarrow buys(X, \text{'coke'})$
 - Hybrid-dimension association rules (*repeated predicates*):
 $age(X, \text{'19-25'}) \text{ and } buys(X, \text{'popcorn'}) \Rightarrow buys(X, \text{'coke'})$
- Categorical Attributes: finite number of possible values, no ordering among values (data cube approach)
- Quantitative Attributes: numeric, implicit ordering among values (discretisation, clustering, etc.)

Quantitative association mining

- Techniques can be categorised by how numerical attributes, such as *age* or *income*, are treated
- Static discretisation based on predefined concept hierarchies (data cube methods)
- Dynamic discretisation based on data distribution
 - A_{quant1} and $A_{quant2} \Rightarrow A_{cat}$
 - Example: $age(X, '19-25')$ and $income(X, '40K-60K') \Rightarrow buys(X, 'HDTV')$
- For quantitative rules, do discretisation such that (for example) the confidence of the rules mined is maximised

Static discretisation of quantitative attributes

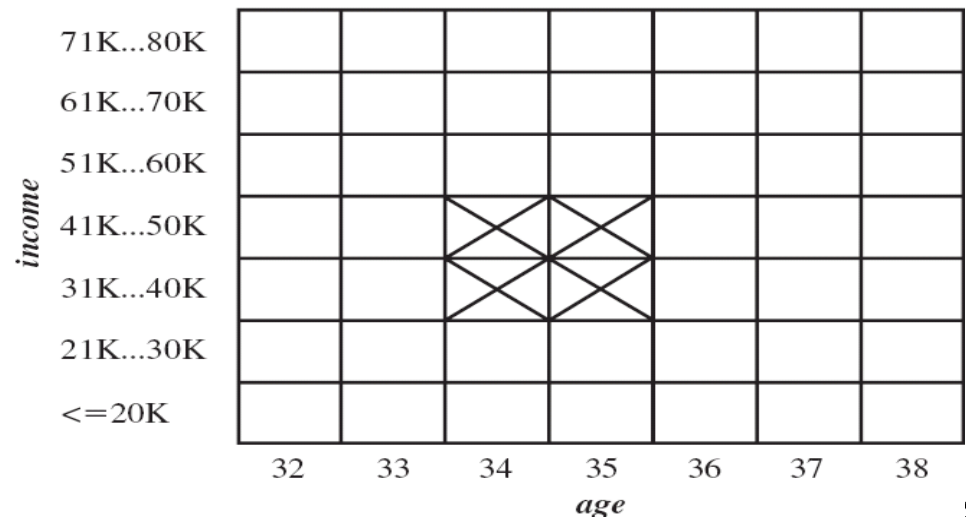
- Discretised prior to mining using concept hierarchy
- Numeric values are replaced by ranges
- In relational database, finding all frequent k -item sets will require k database scans
- Data cube is well suited for mining
- The cells of an n -dimensional cuboid correspond to the item or *predicate sets*
- Mining from data cubes can be much faster



Dynamic discretisation of quantitative attributes

- Mapping of pairs of quantitative attributes into a 2-dimensional grid, such that categorical attribute conditions are satisfied
- The grid is then searched for clusters of points from which association rules are generated

- For example:
 $age(X, '34-35')$ and
 $income(X, '31K-50K')$
 $\Rightarrow buys(X, 'HDTV')$



Interestingness measure: Correlation (lift)

- Example: *Play basketball* \Rightarrow *Eat cereal* [40%, 66.67%] is misleading
 - If overall 75 % of all students eat cereal
 - *Play basketball* \Rightarrow *Not eat cereal* [20%, 33.3%] is more accurate, although with lower support and confidence
- Measure of dependent / correlated events: *Lift*

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

$$lift = \frac{conf(A \rightarrow B)}{support(B)}$$

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

$$lift(B, C) = \frac{2000/5000}{3000/5000 * 3750/5000} = 0.89 \quad lift(B, \neg C) = \frac{1000/5000}{3000/5000 * 1250/5000} = 1.33$$

More interestingness measures

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	$\frac{P(A,B)P(\bar{A},\bar{B}) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(\bar{A},\bar{B}) + P(A,\bar{B})P(\bar{A},B)}$
Y	Yule's Y	-1 ... 1	$\frac{\sqrt{P(A,B)P(\bar{A},\bar{B})} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(\bar{A},\bar{B})} + \sqrt{P(A,\bar{B})P(\bar{A},B)}}$
k	Cohen's	-1 ... 1	$\frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$
PS	Piatetsky-Shapiro's	-0.25 ... 0.25	$P(A, B) - P(A)P(B)$
F	Certainty factor	-1 ... 1	$\max\left(\frac{P(B A) - P(B)}{1 - P(B)}, \frac{P(A B) - P(A)}{1 - P(A)}\right)$
AV	added value	-0.5 ... 1	$\max(P(B A) - P(B), P(A B) - P(A))$
K	Klogsen's Q	-0.33 ... 0.38	$\sqrt{P(\bar{A}, \bar{B})} \max(P(B A) - P(B), P(A B) - P(A))$
g	Goodman-kruskal's	0 ... 1	$\frac{\sum_j \max_k P(A_j, B_k) + \sum_k \max_j P(A_j, B_k) - \max_j P(A_j) - \max_k P(B_k)}{2 - \max_j P(A_j) - \max_k P(B_k)}$
M	Mutual Information	0 ... 1	$\frac{\sum_i \sum_j P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i) \log P(A_i) \log P(A_i), -\sum_i P(B_i) \log P(B_i) \log P(B_i))}$
J	J-Measure	0 ... 1	$\max(P(A, B) \log\left(\frac{P(B A)}{P(B)}\right) + P(\bar{A}\bar{B}) \log\left(\frac{P(\bar{B} \bar{A})}{P(\bar{B})}\right), \\ P(A, B) \log\left(\frac{P(A B)}{P(A)}\right) + P(\bar{A}\bar{B}) \log\left(\frac{P(\bar{A} \bar{B})}{P(\bar{A})}\right))$
G	Gini index	0 ... 1	$\max(P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] - P(B)^2 - P(\bar{B})^2, \\ P(B)[P(A B)^2 + P(\bar{A} B)^2] + P(\bar{B})[P(A \bar{B})^2 + P(\bar{A} \bar{B})^2] - P(A)^2 - P(\bar{A})^2)$
s	support	0 ... 1	$P(A, B)$
c	confidence	0 ... 1	$\max(P(B A), P(A B))$
L	Laplace	0 ... 1	$\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$
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 ... ∞	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(\bar{A},B)P(A,\bar{B})}$
V	Conviction	0.5 ... ∞	$\max\left(\frac{P(A)P(\bar{B})}{P(\bar{A}\bar{B})}, \frac{P(B)P(\bar{A})}{P(\bar{B}\bar{A})}\right)$
λ	lift	0 ... ∞	$\frac{P(A,B)}{P(A)P(B)}$
S	Collective strength	0 ... ∞	$\frac{P(A,B) + P(\bar{A}\bar{B})}{P(A)P(B) + P(\bar{A})P(\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A,B) - P(\bar{A}\bar{B})}$
χ^2	χ^2	0 ... ∞	$\sum_i \frac{(P(A_i) - E_i)^2}{E_i}$

Mining interesting correlation patterns

- Flexible support
 - Some items might be very rare but are valuable (like diamonds)
 - Customise $support_{min}$ specification and application
- Top- k frequent patterns
 - It can be hard to specify $support_{min}$, but top- k rules with $length_{min}$ are more desirable
 - Achievable using special data structures, like Frequent-Pattern (FP) tree
 - Dynamically raise $support_{min}$ during FP-tree construction phase, and select most promising to mine

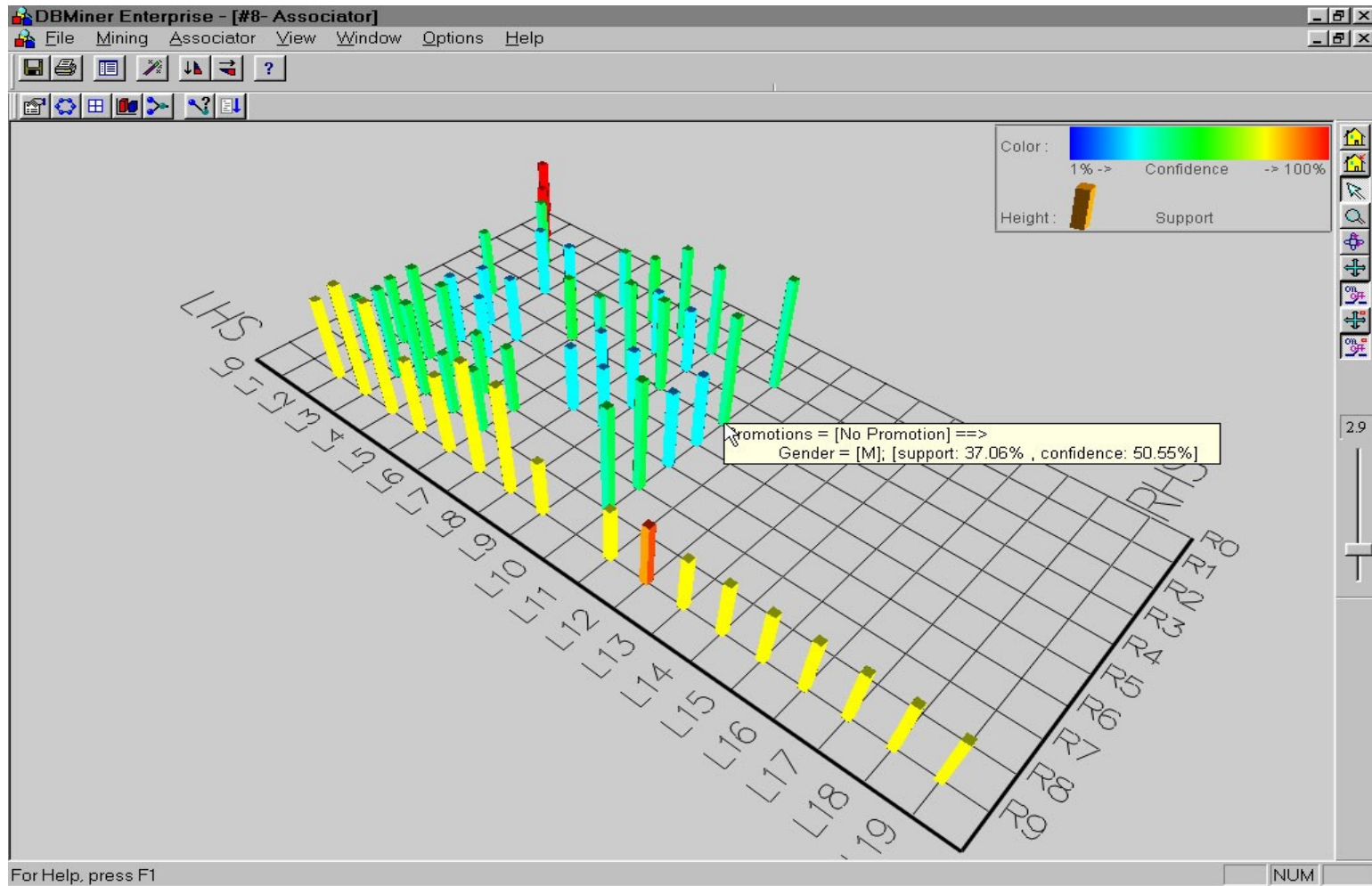
Constraint based mining

- Finding *all* the frequent rules or patterns in a database autonomously is unrealistic
 - The rules / patterns could be too many and not focussed
- Data mining should be an *interactive* process
- The user directs what should 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 (and what not)
 - System optimisation: explores such constraints for efficient mining

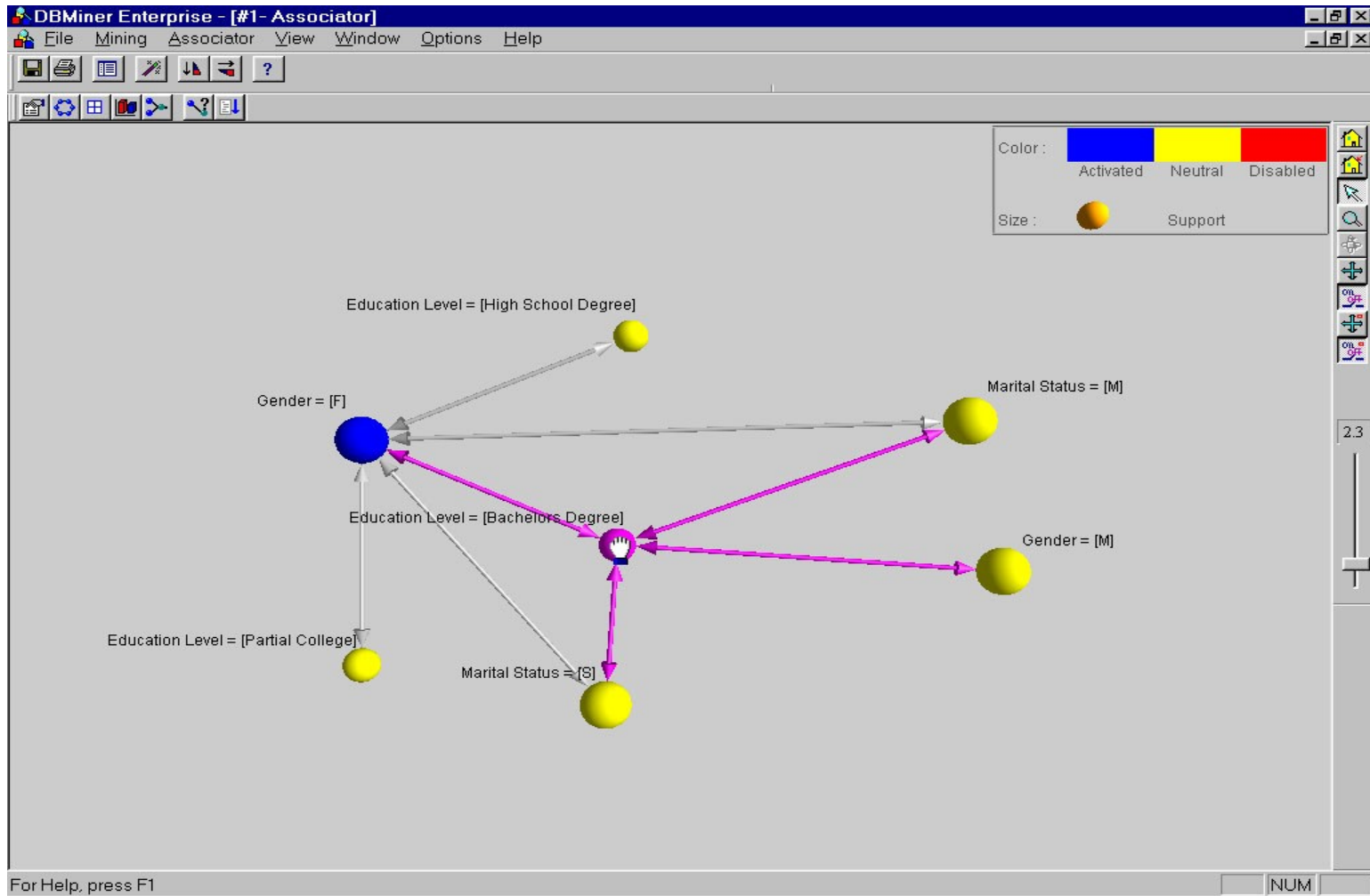
Constraints in data mining

- Knowledge type constraint
 - Correlation, association, etc.
- Data constraint (use SQL like queries)
 - For example: *Find product pairs sold frequently in both stores in Sydney and Melbourne*
- Dimension / level constraint
 - In relevance to region, price, brand, customer category, etc.
- Rule or pattern constraint
 - Small sales (price < \$10) trigger big sales (sum > \$200)
- Interestingness constraint
 - Strong rules only: $\text{support}_{\min} > 3\%$, $\text{confidence}_{\min} > 75\%$

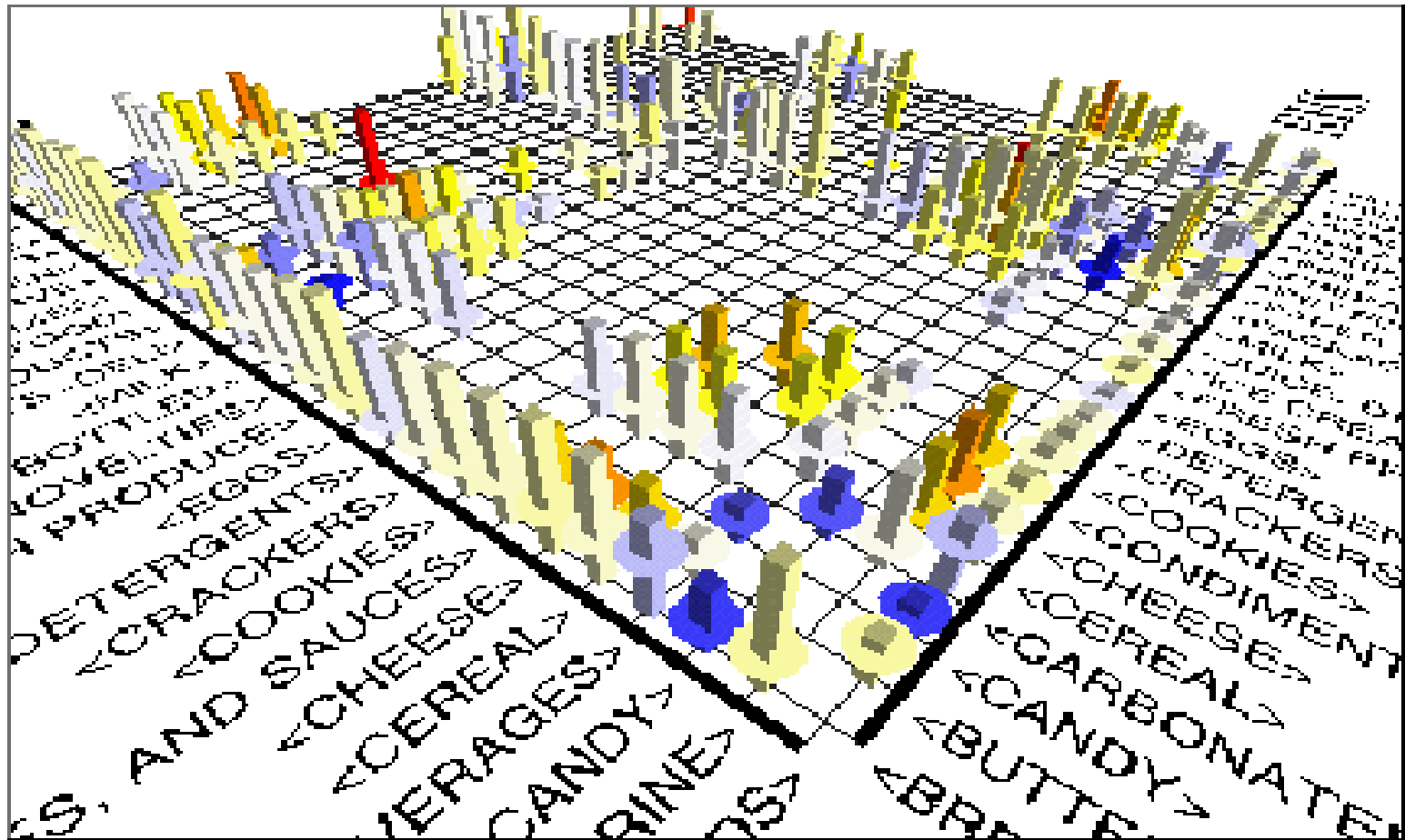
Visualisation of association rules



Visualisation of association rules (2)



Visualisation of association rules (3)





What now.. things to do

- Read chapter 5 in text book (on Mining Frequent Patterns, Associations and Correlations)
- Work on task 2 of assignment 2