ASSOCIATION MINING II

Evaluation Metrics

- Which rules are interesting?

- Subjective measures:
  - Based on subjective arguments to decide if it reveals interesting information
  - {Butter} => {Bread}: Not interesting
  - {Diapers} => {Bread}: Interesting

- Objective measures:
  - based on statistics computation

Subjective Measures

- Subjective interestingness measures are based on user belief in the data. These measures find patterns interesting if
  - they are unexpected (contradicting user's belief)
  - offer strategic information on which user can act.

- Visualization: allows human beings to interact with the data mining system and interpret and verify rules
- Template-Based: allows users to constrain the type of patterns extracted
- Subjective interest measures: based on domain information such as concept hierarchy or profit margin

Objective Measures

- **Support**
  - Support is an important measure because a rule that has very low support may occur simply by chance.
- **Confidence**
  - Confidence on the other hand measures the reliability of the inference made by the rule.

Objective Measures

- Limitation of Support:
  - some items appear infrequently in their normal settings compared to other items
  - For example: number of times a TV is purchased vs eggs are purchased

If we **increase** the support: patterns containing low occurring items (e.g., TV) will not be extracted

If we **decrease** the support: many uninteresting patterns will be extracted

Objective Measures

- Limitation of Confidence:
  - is more subtle
  - Better demonstration through an example
- Consider rule: R = {Tea} => {Coffee}

	Coffee	¬Coffee	
Tea	150	50	200
¬Tea	650	150	800
	800	200	1000

Support(R): 15%
Confidence(R): 75%

→ *Tea drinkers tend to also drink coffee*

Confidence does not look at support of rule consequent (i.e., coffee) -> fake patterns

Support{Coffee} = 80%
Probability of drinking coffee is 80%
Probability of drinking coffee knowing that the person drinks tea is 75%

→ *Drinking tea reduces the probability of drinking coffee*

Lift

The lift, which we also call interest factor, measures the ratio of the deviation of s(A,B) from the support of A **and the support of B when they are independent**.

actual co-occurrence of A and B

$$\text{Lift}(A, B) = \frac{s(A, B)}{P(A) \times P(B)}$$

co-occurrence you would expect under independence

Lift

$$\text{support}(B) = P(B) = \frac{N_B}{N}$$

$$\text{Lift}(A,B) = \frac{s(A,B)}{P(A) \times P(B)} = \frac{\frac{s(A,B)}{s(A)}}{s(B)} = \frac{\text{confidence}(A \to B)}{\text{support}(B)}$$

- If Lift < 1: the occurrence of A is negatively correlated with the occurrence of B
- If Lift > 1: A and B are positively correlated
- If Lift = 1: A and B are independent

Lift(Tea => Coffee) = conf(Tea => Coffee) /Support(Coffee)
= 75%/80% = 0.937
=> Slight negative correlation

Correlation Analysis

$$\phi = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$$

- Correlation factor is in the range [-1, 1]
- ϕ = -1: Perfect negative correlation
- ϕ = +1: Perfect positive correlation
- ϕ = 0: No correlation

	Coffee	¬Coffee	
Tea	150	50	200
¬Tea	650	150	800
	800	200	1000

ϕ = (150*150-650*50) / sqrt(200*800*800*200)
= -0.0625

Correlation Analysis - Limitation

- Consider text mining application

	p	$\bar{p}$	
q	880	50	930
$\bar{q}$	50	20	70
	930	70	1000

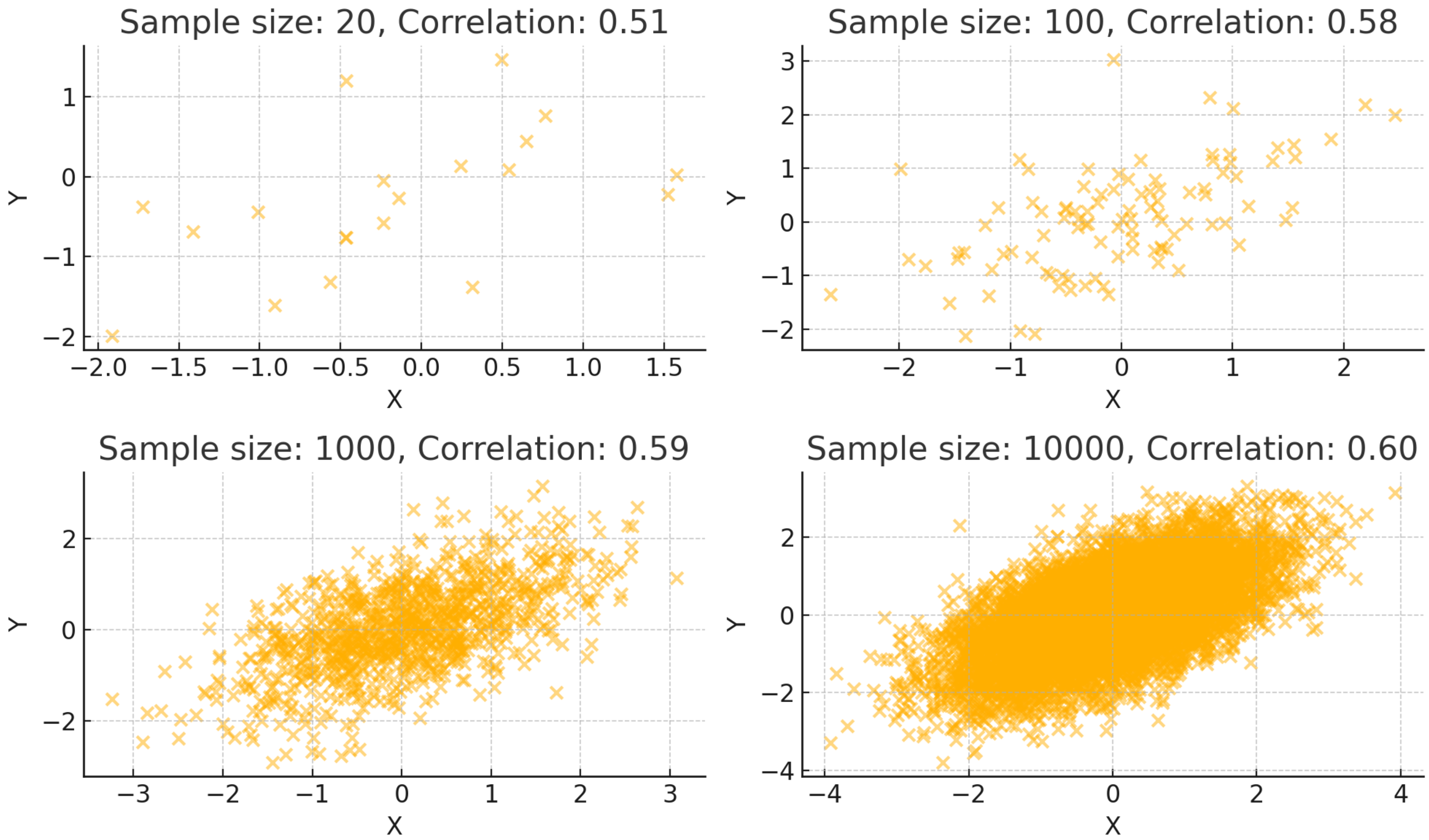
	r	$\bar{r}$	
s	20	50	70
$\bar{s}$	50	880	930
	70	930	1000

$$\phi(p, q) = \phi(r, s) = 0.232$$

- The correlation coefficient puts equal importance to both co-presence and co-absence of items in a transaction
- Suitable for analyzing symmetric binary variables

Correlation Analysis - Limitation

- It does not remain invariant when there are proportional change to the sample size



#	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)	$\max\Big(P(A,B)\log(\frac{P(B\vert A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\vert A)}{P(\overline{B})}), P(A,B)\log(\frac{P(A\vert B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\vert B)}{P(\overline{A})})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\vert A)^2+P(\overline{B}\vert A)^2]+P(\overline{A})[P(B\vert\overline{A})^2+P(\overline{B}\vert\overline{A})^2]-P(B)^2-P(\overline{B})^2, P(B)[P(A\vert B)^2+P(\overline{A}\vert B)^2]+P(\overline{B})[P(A\vert\overline{B})^2+P(\overline{A}\vert\overline{B})^2]-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\vert A), P(A\vert B))$
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)	$\max\Big(\frac{P(B\vert A)-P(B)}{1-P(B)}, \frac{P(A\vert B)-P(A)}{1-P(A)}\Big)$
18	Added Value (AV)	$\max(P(B\vert A)-P(B), P(A\vert B)-P(A))$
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)	$\sqrt{P(A,B)}\max(P(B\vert A)-P(B), P(A\vert B)-P(A))$

- The best metric to use for a given application domain is usually unknown
- There are several **properties** that need to be considered when we analyze a measure.
- One important property is the sensitivity of a measure to row and column scaling operations

Property under Row/Column Scaling

Grade-Gender Example (Mosteller, 1968):

the relationship between the gender of a student and the grade obtained for a particular course

	Male	Female	
High	2	3	5
Low	1	4	5
	3	7	10

	Male	Female	
High	4	30	34
Low	2	40	42
	6	70	76

2x 10x

Underlying association should be independent of the relative number of male and female students in the samples

Some intuitively appealing measures can be sensitive to scaling. Some are not, such as the odds ratio.

Property under Variable Permutation

	B	$\overline{B}$
A	p	q
$\overline{A}$	r	s

$\Rightarrow$

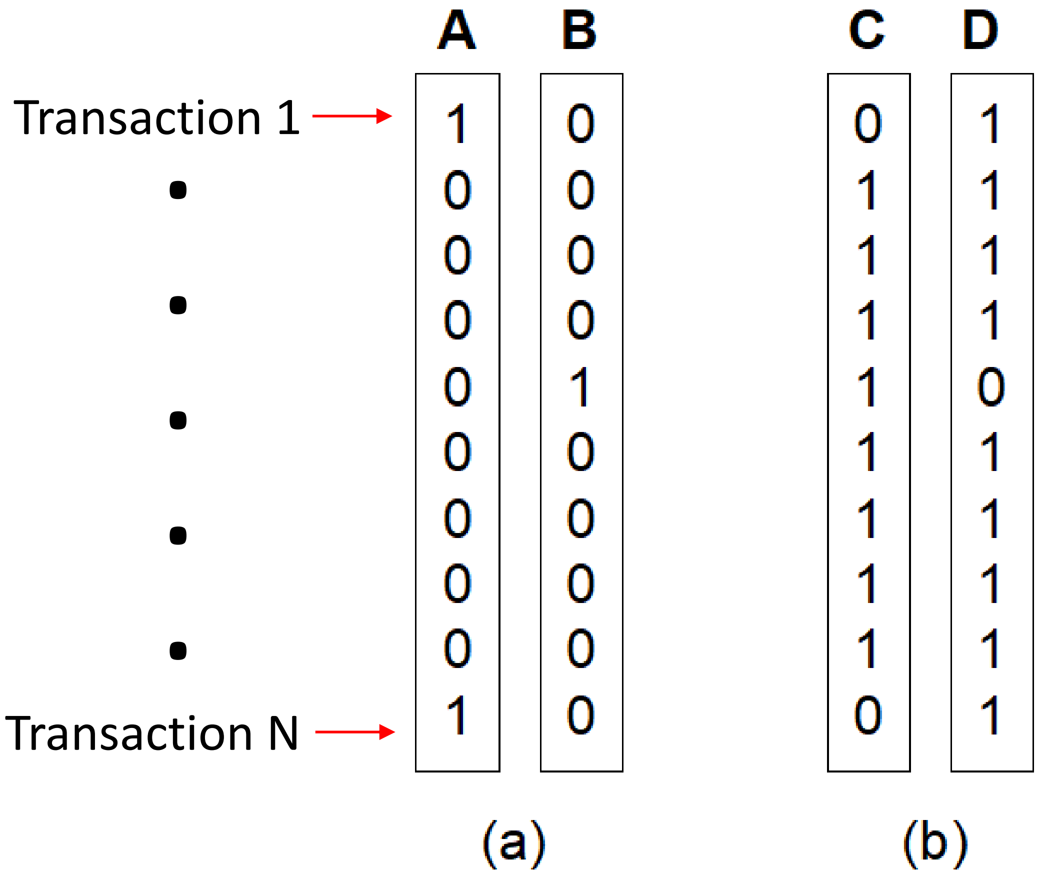
	A	$\overline{A}$
B	p	r
$\overline{B}$	q	s

Does M(A => B) = M(B => A)?

Symmetric measures: support, lift, collective strength, cosine, Jaccard, …

Asymmetric measures: confidence, conviction, Laplace, J-measure, …

Property under Inversion Operation



- have very little association between them.

both C and D co-occur together more frequently, their corr coefficients are still the same as before

flipping the 0's (absence) to become 1's (presence)

If invariant under inversion operation =>
Not suitable for asymmetric data

In other words: distinguish between symmetric binary measures, which are invariant under the inversion operation, from asymmetric binary measures.

Property under Null Addition

	B	$\overline{B}$
A	p	q
$\overline{A}$	r	s

$\Rightarrow$

	B	$\overline{B}$
A	p	q
$\overline{A}$	r	$s + k$

What is the effect of adding more records that do not contain either property?

Invariant measures: cosine, Jaccard, …

Non-invariant measures: interest factor, correlation, odds ratio, …

This property is useful for domains having sparse data sets, where co-presence of items is more important than co-absence

Quantitative Rules

- General association rules where both the left-hand and the right-hand sides of the rule should be **categorical** (nominal or discrete) attributes

- Quantitative rules: at least one attribute (left or right) must involve a **numerical** attribute.

Categorical Attributes

- Categorical attributes are transformed into items

Gender	Level of Education	State	Chat Online	Shop Online
Female	Graduate	Illinois	Yes	No
Male	College	California	Yes	Yes
Male	High School	Michigan	No	Yes

Gender	Lvl-Educ-Graduate	Lvl-Educ-HighSchool	Lvl-Educ-College	---	Shop Online-yes	Shop-online-no
Female	1	0	0		0	1
Male	0	0	1		1	0
Male	0	1	0		1	

Issues

- Some values may not be frequent enough:
  - Lowering support does not help
  - Group related values

    Instead of creating 50 columns, one for each for state

    Create columns for: Midwest, pacific Northwest, Southwest, East Coast
- Some values are very frequent to an extent they don't bring new information but result in a large number of rules
  - Remove these attributes

    Own-Computer = yes: is present 85% of the time.
- To avoid generating too many candidate sets: use only one attribute from each group generated from the same original attribute:

    Ignore itemsets such as {gender-female, gender-male}

Continuous Attributes

- Discretization

Gender	...	Age	Annual Income	No of hours spent online per week	No of email accounts	Privacy Concern
Female	...	26	90K	20	4	Yes
Male	...	51	135K	10	2	No
Male	...	29	80K	10	3	Yes
Female	...	45	120K	15	3	Yes
Female	...	31	95K	20	5	Yes
Male	...	25	55K	25	5	Yes
Male	...	37	100K	10	1	No
Male	...	41	65K	8	2	No
Female	...	26	85K	12	1	No
...	...	...	...	...	...	...

Male	Female	... ...	Age < 13	Age $\in [13, 21)$	Age $\in [21, 30)$	... ...	Privacy $=$ Yes	Privacy $=$ No
0	1	...	0	0	1	...	1	0
1	0	...	0	0	0	...	0	1
1	0	...	0	0	1	...	1	0
0	1	...	0	0	0	...	1	0
0	1	...	0	0	0	...	1	0
1	0	...	0	0	1	...	1	0
1	0	...	0	0	0	...	0	1
1	0	...	0	0	0	...	0	1
0	1	...	0	0	1	...	0	1
...	...	...	...	...	...	...	...	...

Continuous Attributes

Age => Age < 13, Age in [13, 21), Age in [21, 30), …

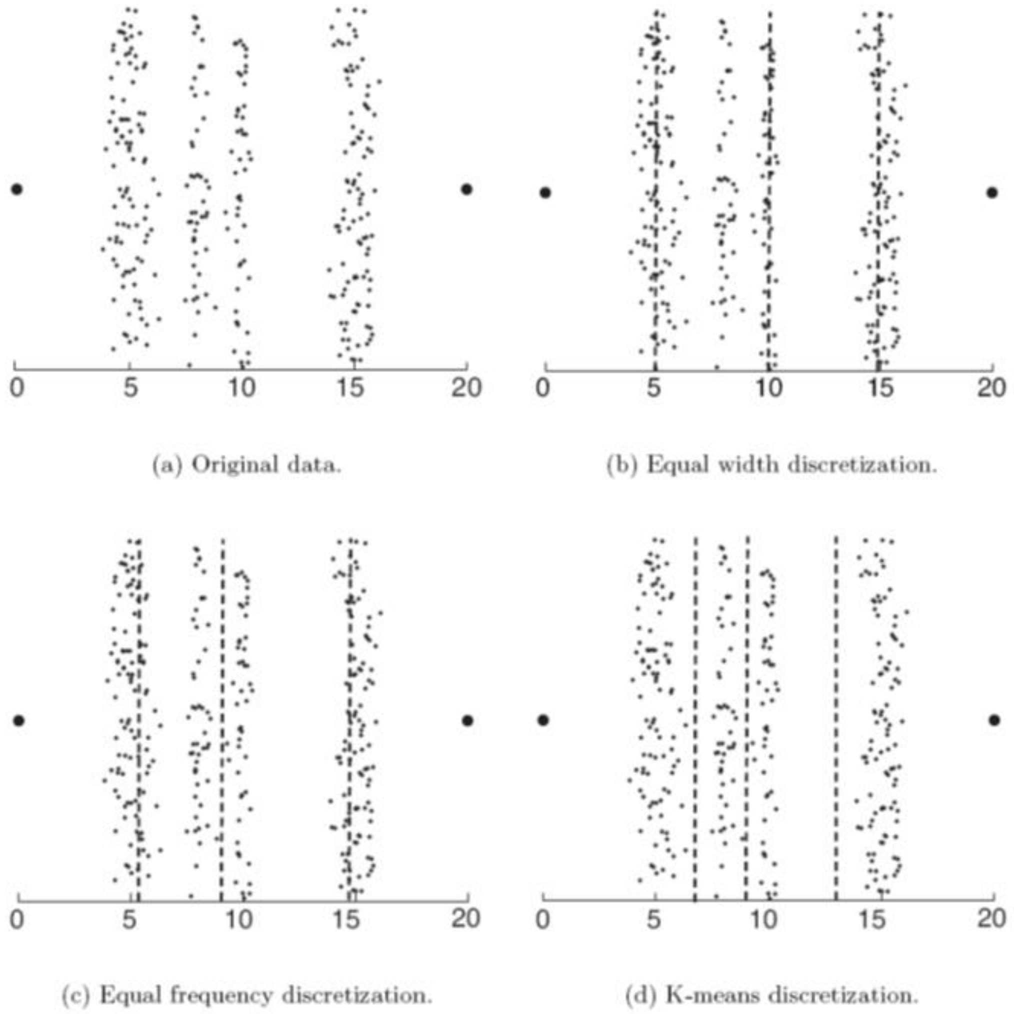
- Interval width: affects support and confidence of generated rules
- If interval too wide: may lose patterns because of lack of confidence
- If interval too narrow: may lose patterns because of lack of support

supp(age in [21..30]) = 25%
supp(age in [31..40]) = 30%
supp(age in [21..40]) = 55%

conf(age in [21..30] => income in [40k..50k]) = supp(income in [40k..50k], age in [21..30]) / supp(age in [21..30])
conf(age in [21..40] => income in [40k..50k]) = supp(income in [40k..50k], age in [21..40]) / supp(age in [21..40])

How to determine interval width?

- Equal width intervals
- Equal depth intervals
- Clustering



Example – Customer profiles

People

RecordID	Age	Married	NumCars
100	23	No	1
200	25	Yes	1
300	29	No	0
400	34	Yes	2
500	38	Yes	2

(minimum support = 40%, minimum confidence = 50%)

Rules (Sample)	Support	Confidence
$\langle$Age: 30..39$\rangle$ and $\langle$Married: Yes$\rangle$ $\Rightarrow$ $\langle$NumCars: 2$\rangle$	40%	100%
$\langle$NumCars: 0..1$\rangle$ $\Rightarrow$ $\langle$Married: No$\rangle$	40%	66.6%

Example – Congressional voting records

Attribute	Values
Party affiliation	Democrat/Republican
Handicapped Infants	Yes/No
Water project cost sharing	Yes/No
Budget resolution	Yes/No
Physician fee freeze	Yes/No
Immigration	Yes/No
⋮	⋮
Aid to Nicaragua	Yes/No
Education Spending	Yes/No

Attribute
Party-affiliation-Democrat
Party-affiliation-Republican
Handicapped-Infants-yes
Handicapped-Infants-no
Water project cost sharing - yes
Water project cost sharing - no
Budget resolution-yes
Budget resolution-no
Immigration-yes
⋮

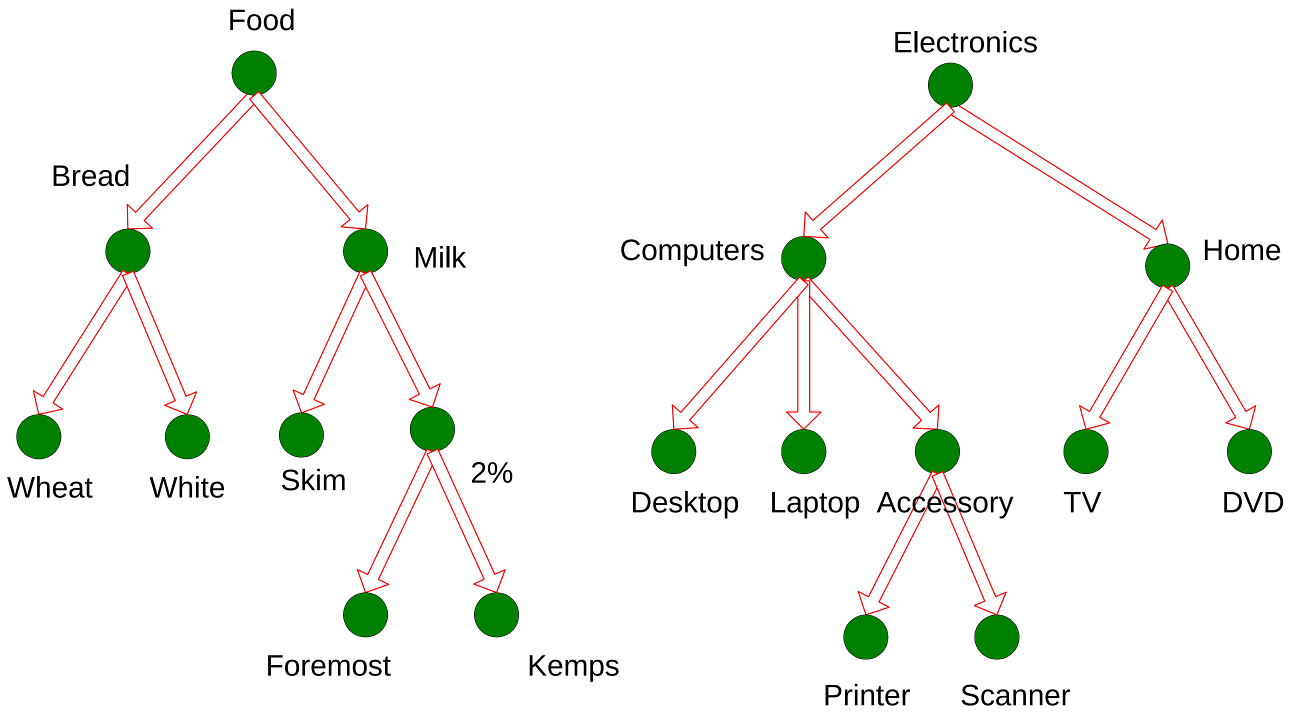
Example – Congressional voting records

Association Rule	*Confidence*
{Budget resolution = no, MX-missile = no, aid to El Salvador = yes} => {Republican}	91%
{Budget resolution = yes, MX-missile = yes, aid to El Salvador = no} => {Democrat}	97.5%
{crime=yes, right to sue = yes, physician fee freeze= yes} => {Republican}	93.5%
{crime=no, right to sue = no, physician fee freeze= no} => {Democrat}	100%

Concept Hierarchy

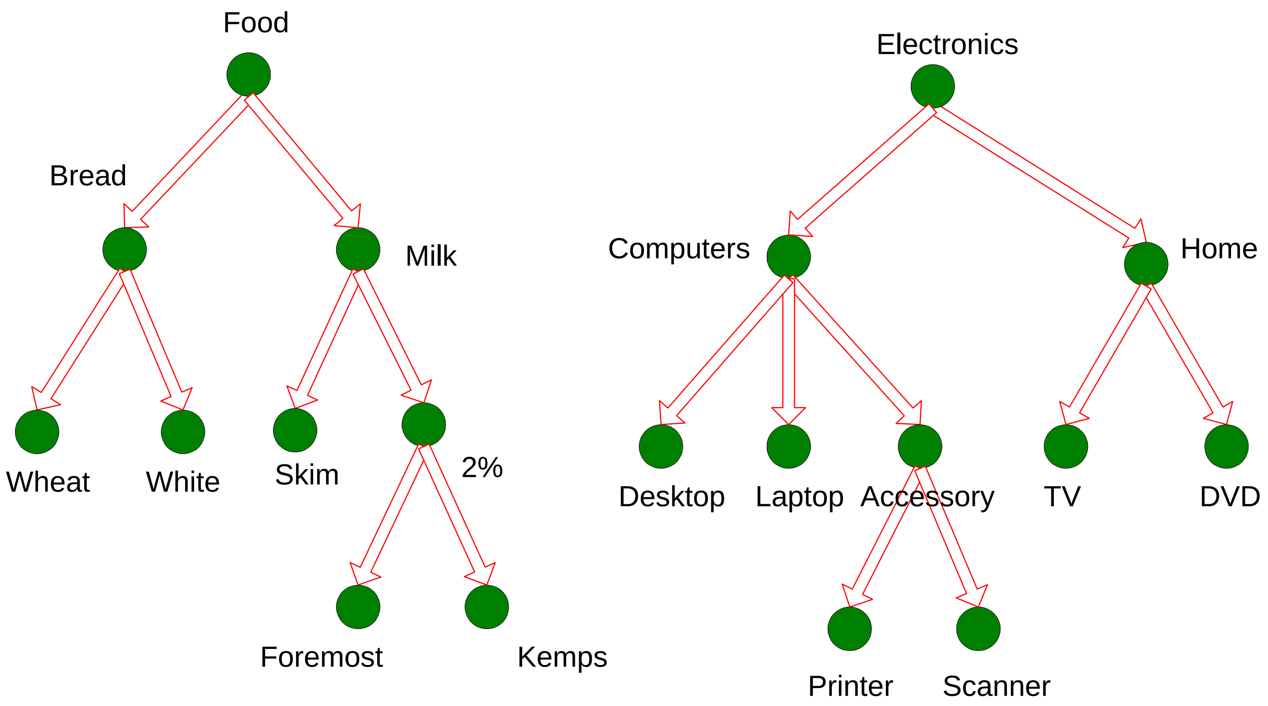
--Multi-level Association Rules

- Concept hierarchies are based on domain knowledge
  - Milk, eggs, cheese: food concept
  - Video games, tv, dvds: electronic concepts



Concept Hierarchy

- *Items in lower level of the tree may have low support, grouping increases their support*
  - *Example: sale of scanners may be low but sale of accessories is high*



- *Consider lower level items only: rules tend to be overly specific*
  - *Example:*
    - *skim milk => wheat bread*
    - *2% milk => wheat bread*
    - *2% milk => white bread*

Too specific

More interested in

Milk => bread

Concept Hierarchy

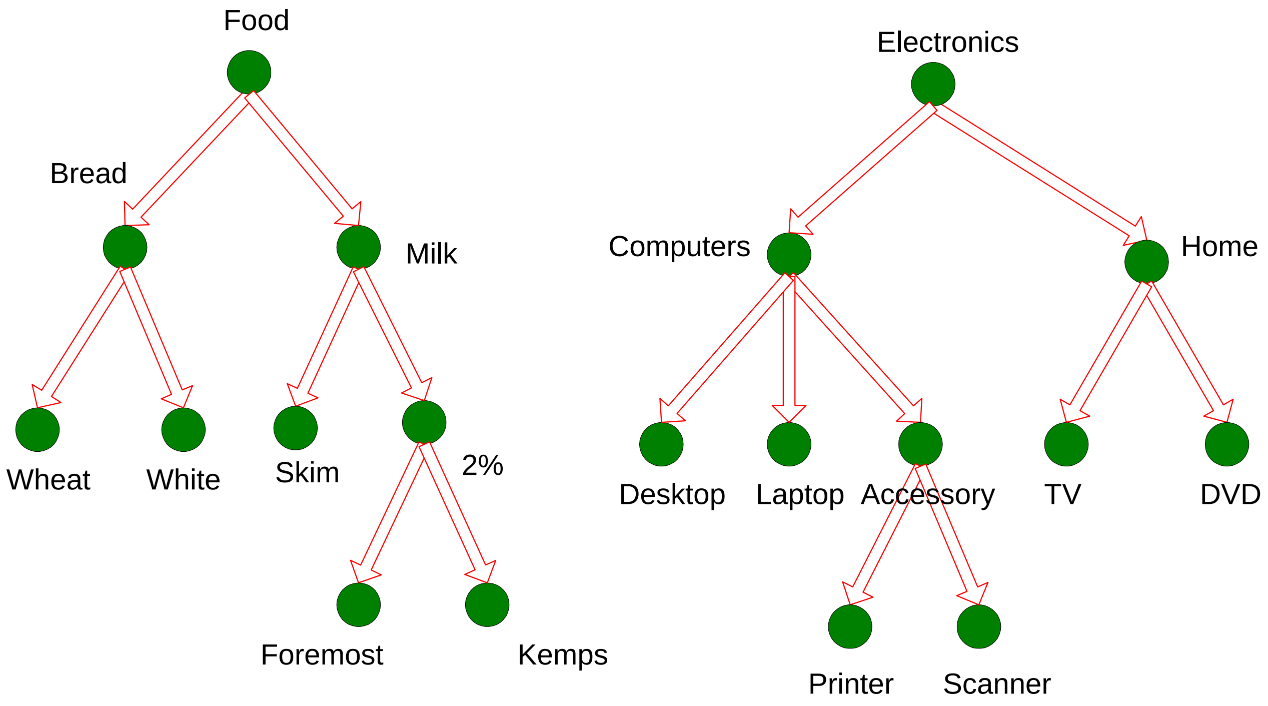
- *Consider higher level items only: rules tend to be too general*

  Example:

  Electronics => Food

Overgeneralizing.
More interested in

DVD => 2% Milk

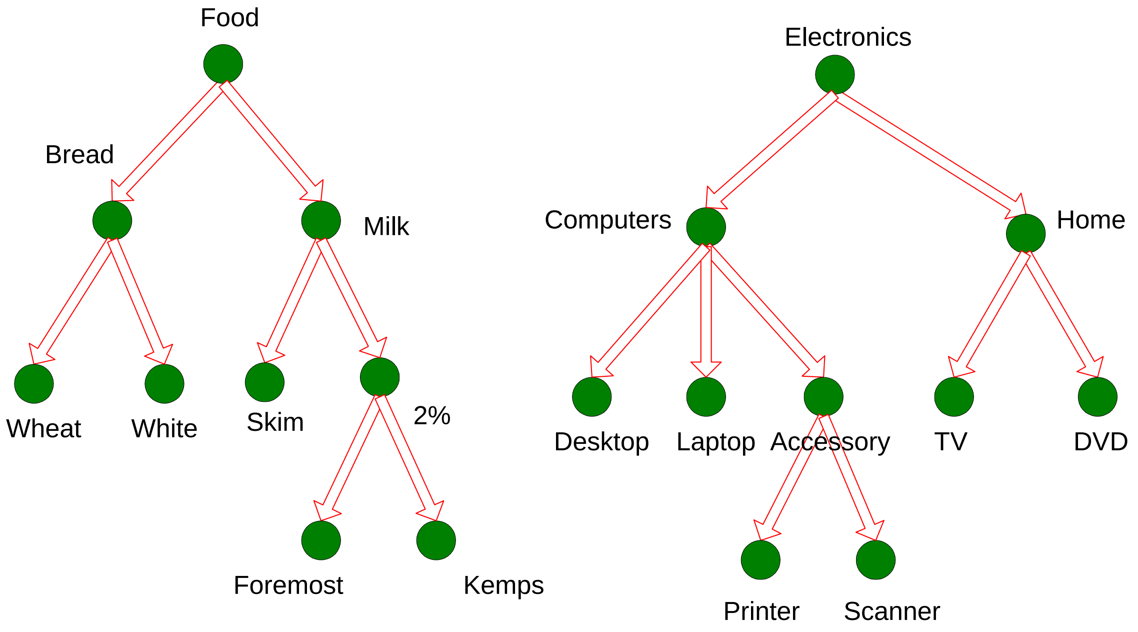


Handling Concept Hierarchy
--Multi-level Association Rules

Approach 1: Extend lower level items by parents in hierarchy

{2% milk, wheat bread}
becomes:
{2% milk, milk, wheat bread, bread, Food}

{Foremost milk, wheat bread}
becomes:
{Foremost milk, 2% milk, milk, wheat bread, bread, Food}



Handling Concept Hierarchy

--Multi-level Association Rules

- *Min support choice: Items in higher levels have higher support.*
  - *If threshold too high => generate rules involving higher level items only*
  - *If support too low => generate too many patterns*

- *Concept hierarchy increases the computation time:*
  - *Increasing number of items*
  - *Increasing transactions width*

- *Concept hierarchy may produce redundant rules and itemsets*

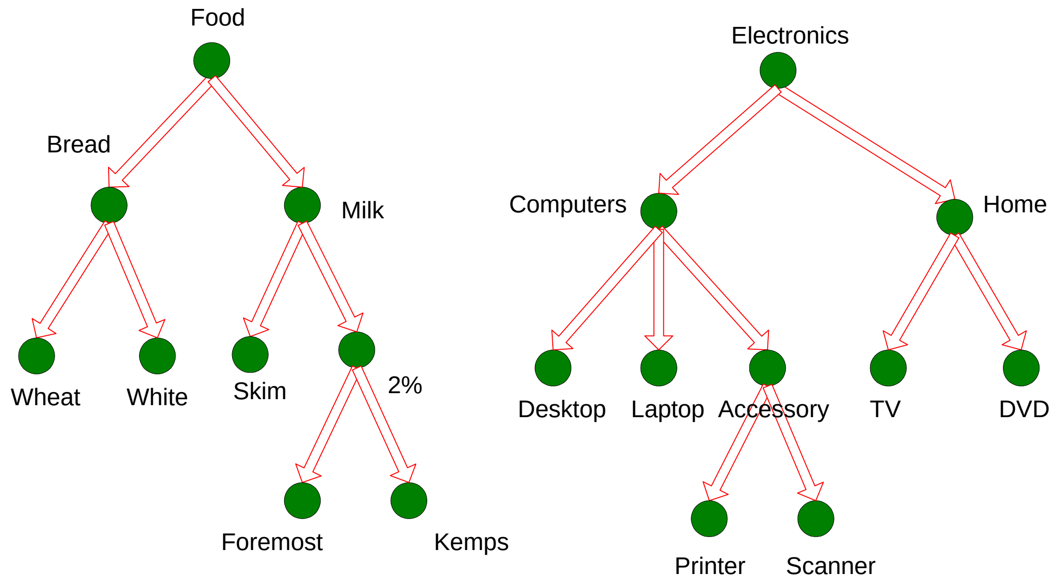
{Skim milk, milk, food}

Since hierarchy is known =>
eliminate redundant itemsets
during frequent itemset generation

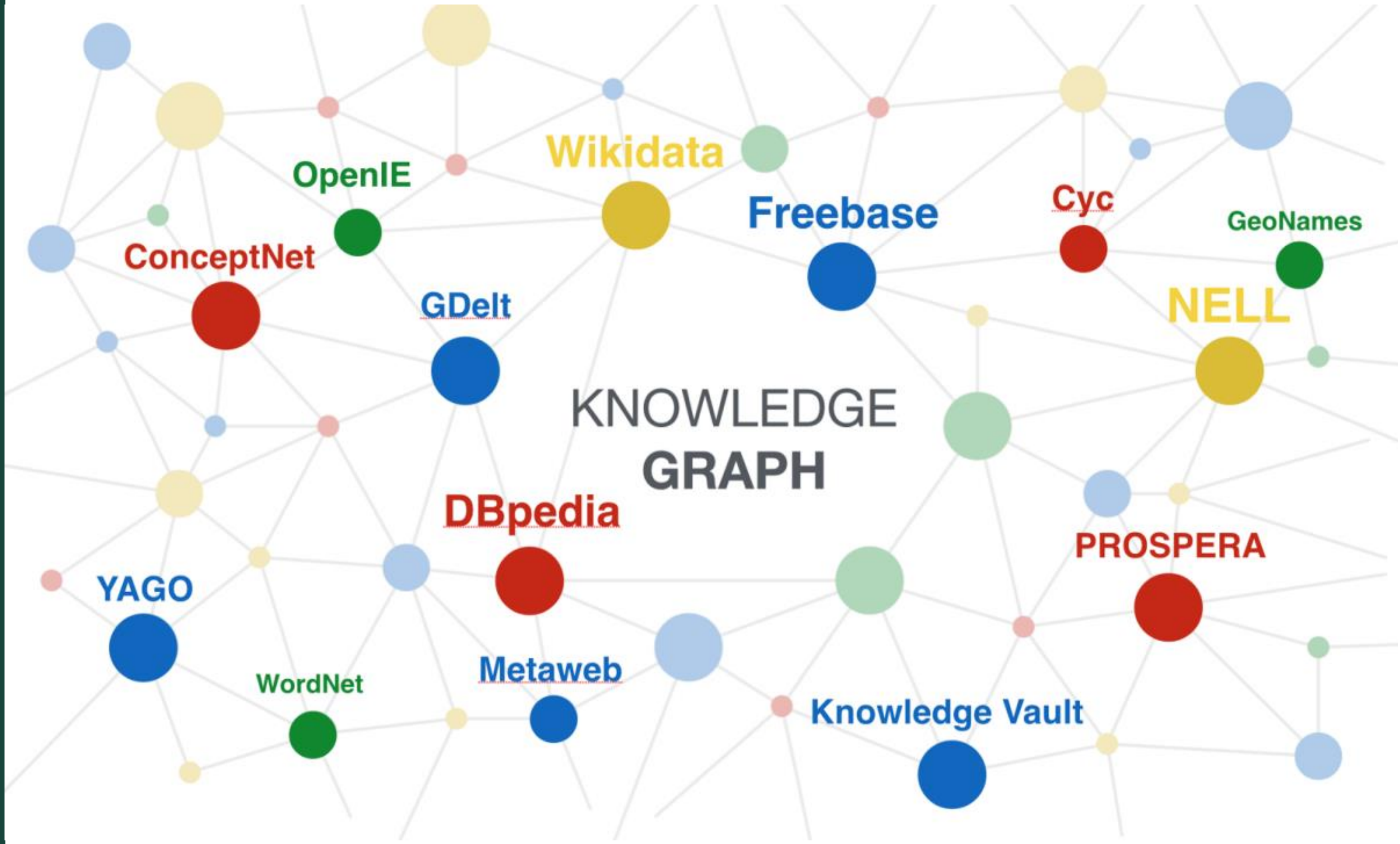
Handling Concept Hierarchy
--Multi-level Association Rules

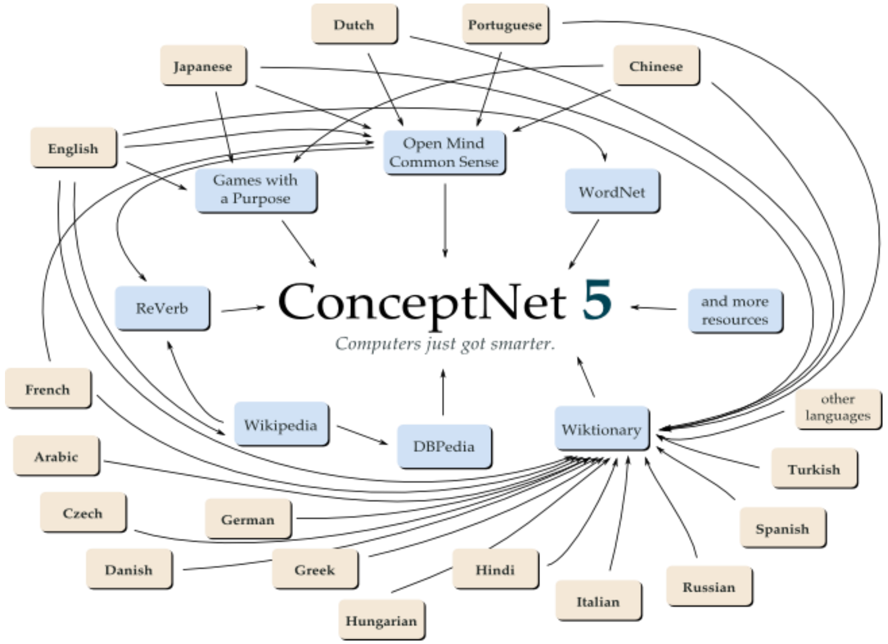
Approach 2:
Generate frequent items from higher levels first
Generate frequent items from the next level, and so on…

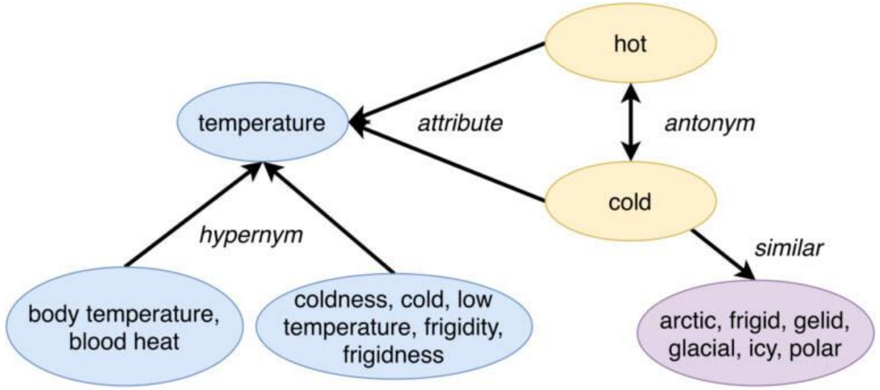
Increases Input/Output since more passes are needed
May miss patterns across different levels



Knowledge Graphs







WordNet

Applications

- Market basket analysis
- Medical diagnosis
- Protein sequences
- Census data: education, health, transport, funds, public businesses
- CRM of credit card business