Associations advanced

Lecture 08.02

Frequent itemsets can be very numerous

- We might choose to work with the top frequent itemsets

Frequent items in 5 Shakespeare sonnets

Tag (word) cloud – visualization of the most frequent words:

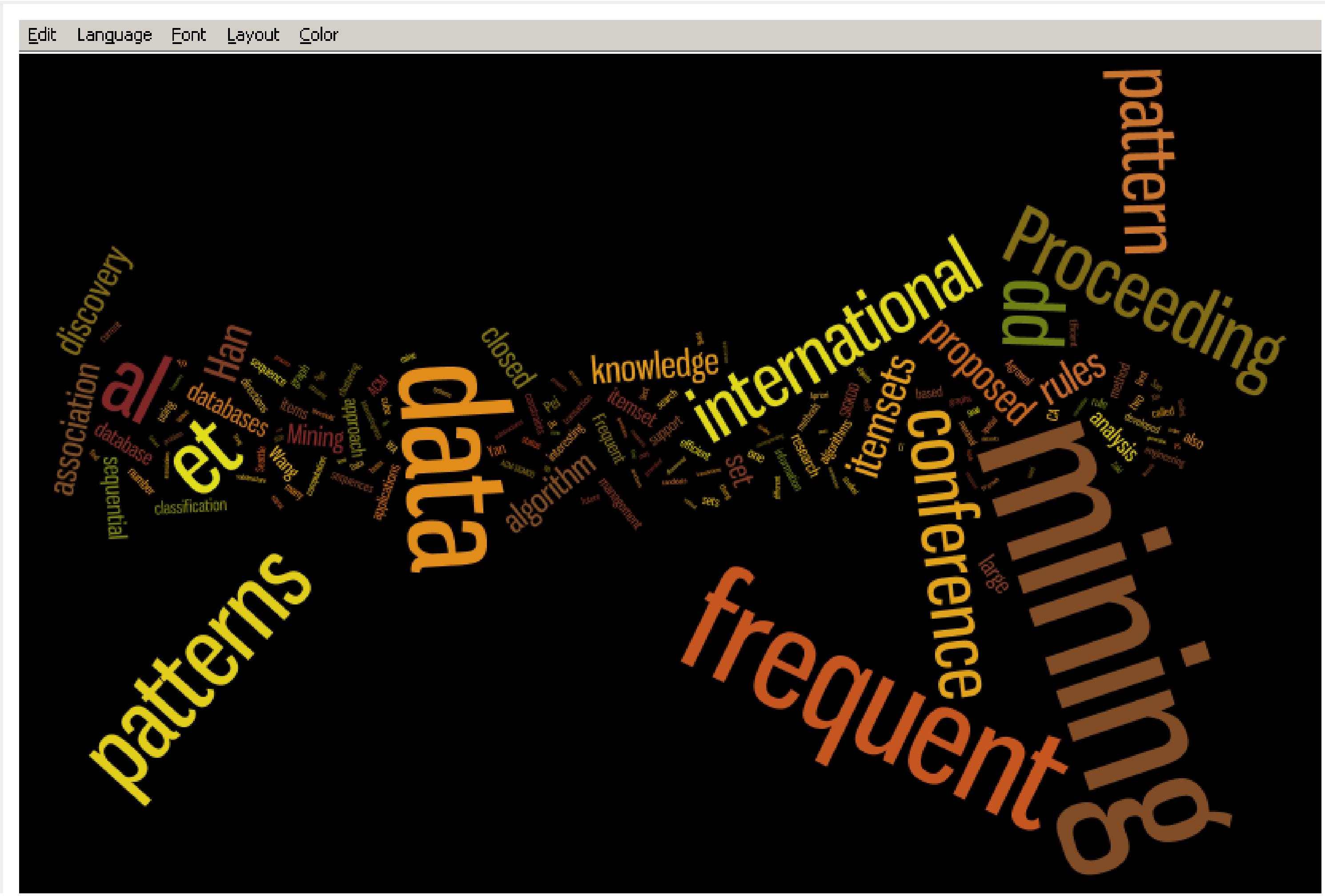
http://www.wordle.net/create

Frequent items in 5 Shakespeare sonnets

- http://www.tagcrowd.com/

Frequent items in papers on frequent pattern mining

Frequent items in papers on frequent pattern mining



Top-frequent itemsets

- Easy to compute
- Not interesting!

- We need to lower the min support threshold to find something non-trivial

Frequent Itemset Mining Implementations (FIMI) 2004 challenge

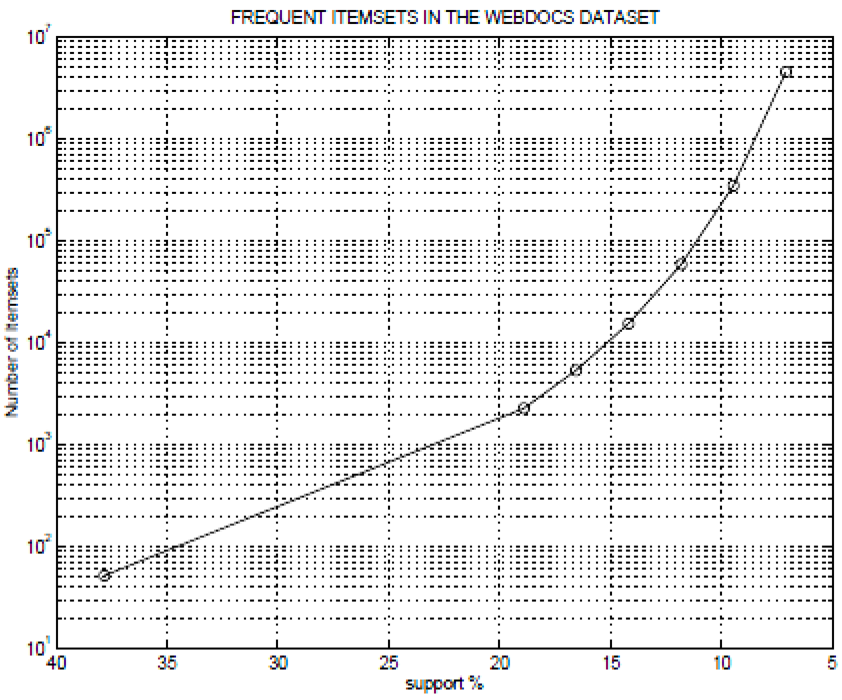
http://fimi.ua.ac.be/data/

- WebDocs dataset is about 5GB
- Each document – transaction, each word - item
- The challenge is to compute all frequent itemsets (word combinations which frequently occur together)
- The number of distinct items (words) = 5,500,000
- The number of transactions (documents) = 2,500,000
- Max items per transaction = 281

We can find the most frequent itemsets with support >= 10%

- These itemsets are trivial word combinations

- When we go to the lower support, the number of frequent itemsets becomes big
- How big? Very big: we cannot keep in memory all different 2-item combinations, to update their counters



How can we find new non-trivial knowledge

- Use confidence?

- The confidence is not-antimonotone, so the algorithm cannot prune any item combination and needs to compute confidence for each possible combination of items
- Computationally infeasible

Pitfalls of confidence

- Suppose we managed to rank all possible association rules by confidence
- How good are the top-confidence rules?

Evaluation of association between items: contingency table

- Given an itemset {X, Y}, the information about the relationship between X and Y can be obtained from a contingency table

Contingency table for {X ,Y}

	Y	Y	
X	f_{11}	f_{10}	f_{1+}
X	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	|T|

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

Used to define various measures

Example: tea and coffee

	Coffee	**¬Coffee**	
Tea	150	50	200
¬Tea	750	150	900
	900	200	1100

Example: tea and coffee

	C	**¬C**	
T	150	50	200
¬T	750	150	900
	900	200	1100

- Confidence of rule T $\rightarrow$ C (conditional probability P(C|T)): sup(T and C)/sup (T)=150/200=0.75

This is a top-confidence rule!

Example: tea and coffee

	C	**¬C**	
T	150	50	200
¬T	750	150	900
	900	200	1100

- Confidence of rule $T \rightarrow C$

  $P(C|T)=0.75$

However, $P(C)=900/1100=0.85$

Example: tea and coffee

	C	¬C	
T	150	50	200
¬T	750	150	900
	900	200	1100

- Confidence of rule $T \rightarrow C$ P(C|T)=0.75

However, P(C)=900/1100=0.85

Although confidence is high, the rule is misleading:

P(C| ¬T)=750/900=0.83

The probability that the person drinks coffee is not increased due to the fact that he drinks tea: quite the opposite – knowing that someone is a tea-lover decreases the probability that he is also a coffee-addict

Why did it happen?

	C	**¬C**	
T	150	50	200
¬T	750	150	900
	900	200	1100

- Confidence of rule $T \rightarrow C$ P(C|T)=0.75

Because the support counts are skewed: much more people drink coffee (900) than tea (200)

and confidence takes into account only one-directional conditional probability

We want to evaluate mutual dependence (association, correlation)

- Not top-frequent
- Not top-confident

- Idea: apply statistical independence test

Statistical measure of association (correlation)-*Lift*

- If the appearance of T is statistically independent of appearance of C, then the probability to find them in the same trial (transaction) is P(C)xP(T)
- We expect to find both C and T with support P(C) x P(T) – expected support

- If actual support $P(C \wedge T)$

$P(C \wedge T) = P(C) \times P(T)$ => Statistical independence

$P(C \wedge T) > P(C) \times P(T)$ => Positive association

$P(C \wedge T) < P(C) \times P(T)$ => Negative association

Lift (Interest Factor)

- Measure that takes into account statistical dependence

$$Interest = \frac{P(A \wedge B)}{P(A)P(B)} = \frac{f_{11}/N}{(f_{1+}/N) \times (f_{+1}/N)} = \frac{N \times f_{11}}{f_{1+} \times f_{+1}}$$

- Interest factor compares the frequency of a pattern against a baseline frequency computed under the statistical independence assumption.
- The **baseline** frequency for a pair of mutually independent variables is:

$$\frac{f_{11}}{N} = \frac{f_{1+}}{N} \times \frac{f_{+1}}{N}$$

Or equivalently

$$f_{11} = \frac{f_{1+} \times f_{+1}}{N}$$

Interest Equation

- Fraction f_{11}/N is an estimate for the joint probability P(A,B), while f_{1+}/N and f_{+1}/N are the estimates for P(A) and P(B), respectively.
- If A and B are statistically independent, then $P(A \wedge B)=P(A)\times P(B)$, thus the Interest is 1.

$$I(A,B)\begin{cases} =1, & \text{if } A \text{ and } B \text{ are independent;} \\ >1, & \text{if } A \text{ and } B \text{ are positively correlated;} \\ <1, & \text{if } A \text{ and } B \text{ are negatively correlated.} \end{cases}$$

Example: tea and coffee

	Coffee	**¬Coffee**	
Tea	150	50	200
¬Tea	750	150	900
	900	200	1100

Association Rule: Tea → Coffee

Interest = 150*1100 / (200*900)= 0.92

(< 1, therefore they are negatively correlated – almost independent)

Problems with Lift

- Consider two contingency tables from the same dataset:

Coffee (C) and milk (M)

	C	**¬C**	
M	10,000	1,000	11,000
¬M	1,000	88,000	89,000
	11,000	89,000	100,000

Popcorn (P) and soda (S)

	P	**¬P**	
S	1,000	1,000	2,000
¬S	1,000	97,000	98,000
	2,000	98,000	100,000

Which items are more correlated: M and C or P and S?

Problems with Lift

Coffee (C) and milk (M)

	C	¬C	
M	10,000	1,000	11,000
¬M	1,000	88,000	89,000
	11,000	89,000	100,000

Popcorn (P) and soda (S)

	P	¬P	
S	1,000	1,000	2,000
¬S	1,000	97,000	98,000
	2,000	98,000	100,000

Well,
Lift (M,C) = 8.26
Lift (P,S)=25.00

Problems with Lift

Coffee (C) and milk (M)

	C	¬C	
M	10,000	1,000	11,000
¬M	1,000	88,000	89,000
	11,000	89,000	100,000

Popcorn (P) and soda (S)

	P	¬P	
S	1,000	1,000	2,000
¬S	1,000	97,000	98,000
	2,000	98,000	100,000

Lift (M,C) = 8.26

Lift (P,S)=25.00

Why did that happen?
Because probabilities P(S)= P(P) =0.02 are very low comparing with probabilities P(C) = P(M)= 0.11

By multiplying very low probabilities, we get very-very low expected probability and then any number of items occurring together will be larger than expected

Problems with Lift

Coffee (C) and milk (M)

	C	¬C	
M	10,000	1,000	11,000
¬M	1,000	88,000	89,000
	11,000	89,000	100,000

Popcorn (P) and soda (S)

	P	¬P	
S	1,000	1,000	2,000
¬S	1,000	97,000	98,000
	2,000	98,000	100,000

Lift (M,C) = 8.26
Lift (P,S)=25.00

But most of the items in a large database have very low supports comparing with the total number of transactions

Conclusion: we are dealing with small probability events, where regular statistical methods might not be applicable

More problems with Lift: positive or negative?

- Consider two contingency tables for C and M from 2 different datasets:

Dataset 1

	C	¬C	
M	400	600	1,000
¬M	600	18,400	19,000
	1,000	19,000	20,000

Dataset 2

	C	¬C	
M	400	600	1,000
¬M	600	1,300	1,900
	1,000	1,900	2,000

According to definition of Lift:

DB1: expected (M and C)=1000/20000 x 1000/20000 =0.0025
actual (M and C)=400/20000 = 0.02
Lift = 8.0 (positive correlation)

?

DB2: expected (M and C)=1000/2000 x 1000/2000 =0.25
actual (M and C)=400/2000 = 0.2
Lift = 0.8 (negative correlation)

More problems with Lift: positive or negative?

Dataset 1

	C	¬C	
M	400	600	1,000
¬M	600	18,400	19,000
	1,000	19,000	20,000

Dataset 2

	C	¬C	
M	400	600	1,000
¬M	600	1,300	1,900
	1,000	1,900	2,000

But nothing has changed in connections between C and M

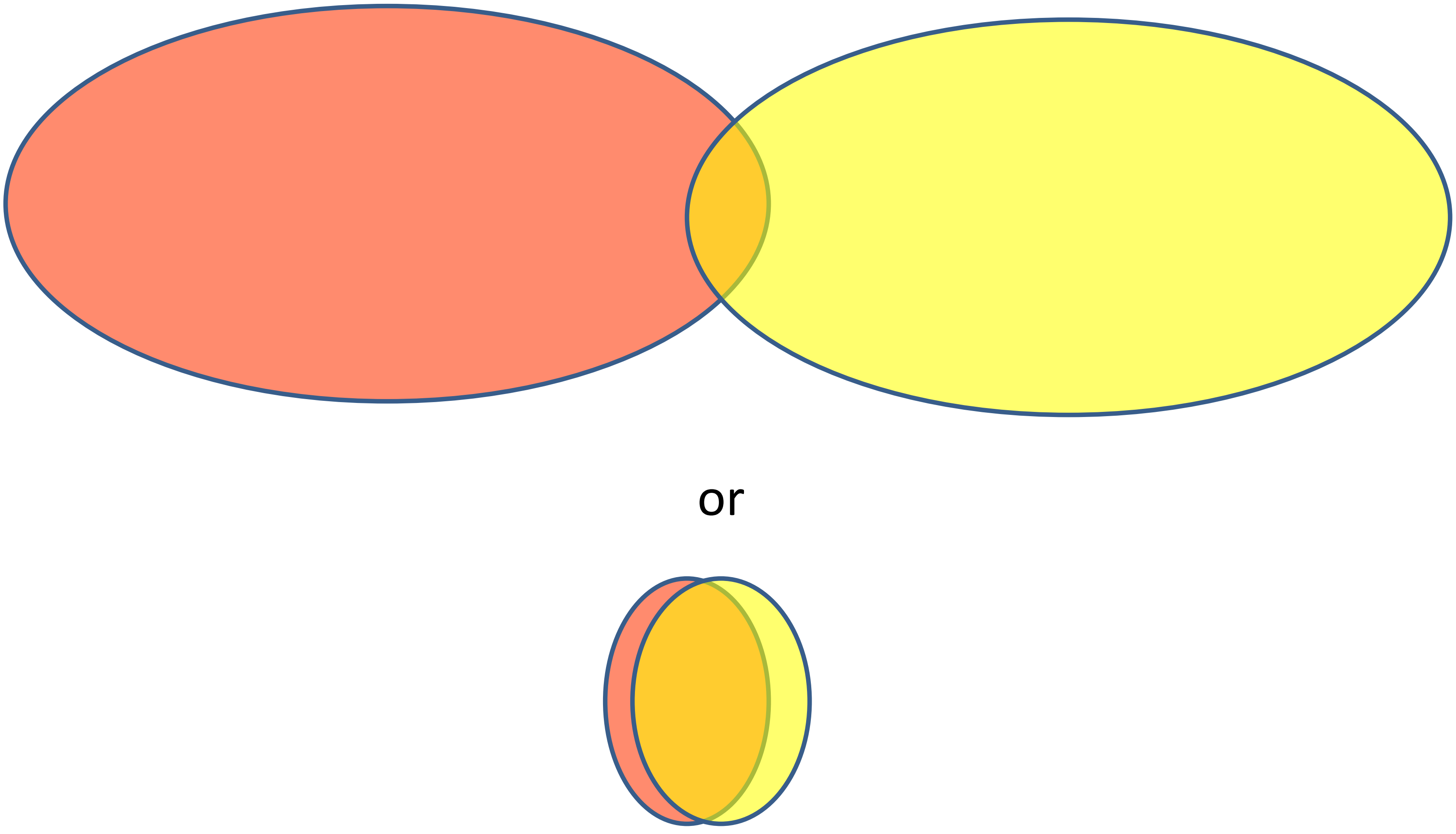
The changes are in the count of transactions which do not contain neither C nor M.

Such transactions are called *null-transactions* with respect to C and M

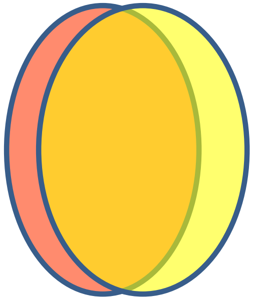
We want the measure which does not depend on null-transactions: null-transaction invariant. Which depends only on counts of items in the current itemset

What are we looking for?

The area corresponds to support counts



Possible null-invariant measure 1: Jaccard index

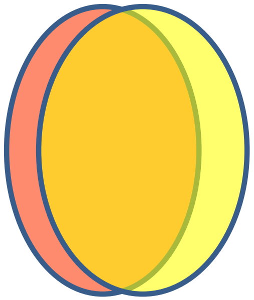


Jaccard index: intersection/union

JI (A, B) = sup (A and B)/[sup(A)+sup(B)-sup(A and B)]

Possible null-invariant measure 2: Kulczynsky

Kulczynsky: arithmetic mean of conditional probabilities



Kulc (A, B) = [P(A|B)+P(B|A)]/2

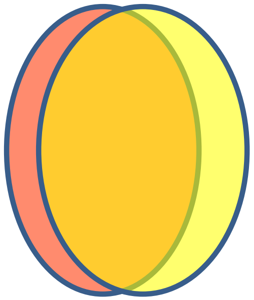
In terms of support counts:

Kulc(A,B) = ½ [sup (A and B)/sup (A) + sup (A and B)/sup (B)]

Possible null-invariant measure 3: Cosine

Cosine: geometric mean of conditional probabilities

Cos (A, B) = sqrt[P(A|B) x P(B|A)]



In terms of support counts:

Cos (A,B) = sup (A and B)/sqrt [sup (A) x sup (B)]

Kulc on the same dataset

- Consider two contingency tables from the same dataset:

Coffee (C) and milk (M)

	C	**¬C**	
M	10,000	1,000	11,000
¬M	1,000	88,000	89,000
	11,000	89,000	100,000

Popcorn (P) and soda (S)

	P	**¬P**	
S	1,000	1,000	2,000
¬S	1,000	97,000	98,000
	2,000	98,000	100,000

Which items are more correlated: M and C or P and S?

Kulc on the same dataset

Coffee (C) and milk (M)

	C	¬C	
M	10,000	1,000	11,000
¬M	1,000	88,000	89,000
	11,000	89,000	100,000

Popcorn (P) and soda (S)

	P	¬P	
S	1,000	1,000	2,000
¬S	1,000	97,000	98,000
	2,000	98,000	100,000

Kulc (C,M) = ½ *(10000/11000+10000/11000) =0.91

Kulc (P,S) = ½ *(1000/2000+1000/2000) = 0.5

Lift (M,C) = 8.26
Lift (P,S)=25.00

Kulc on two datasets: positive or negative?

Dataset 1

	C	¬C	
M	400	600	1,000
¬M	600	18,400	19,000
	1,000	19,000	20,000

Dataset 2

	C	¬C	
M	400	600	1,000
¬M	600	1,300	1,900
	1,000	1,900	2,000

DB1: Kulc (C,M) = ½ *(400/1000+400/1000) =0.4

DB2: Kulc (C,M) = ½ *(400/1000+400/1000) = 0.4

DB1: Lift = 8.0 (positive correlation)

DB2: Lift = 0.8 (negative correlation)

Problems begin

- We found decent null-invariant measures to evaluate the quality of associations (correlations) between items

- The problem: how do we extract top-ranked correlations from large transactional database?
- All null-invariant measures are non-antimonotone
- This is the area of current research

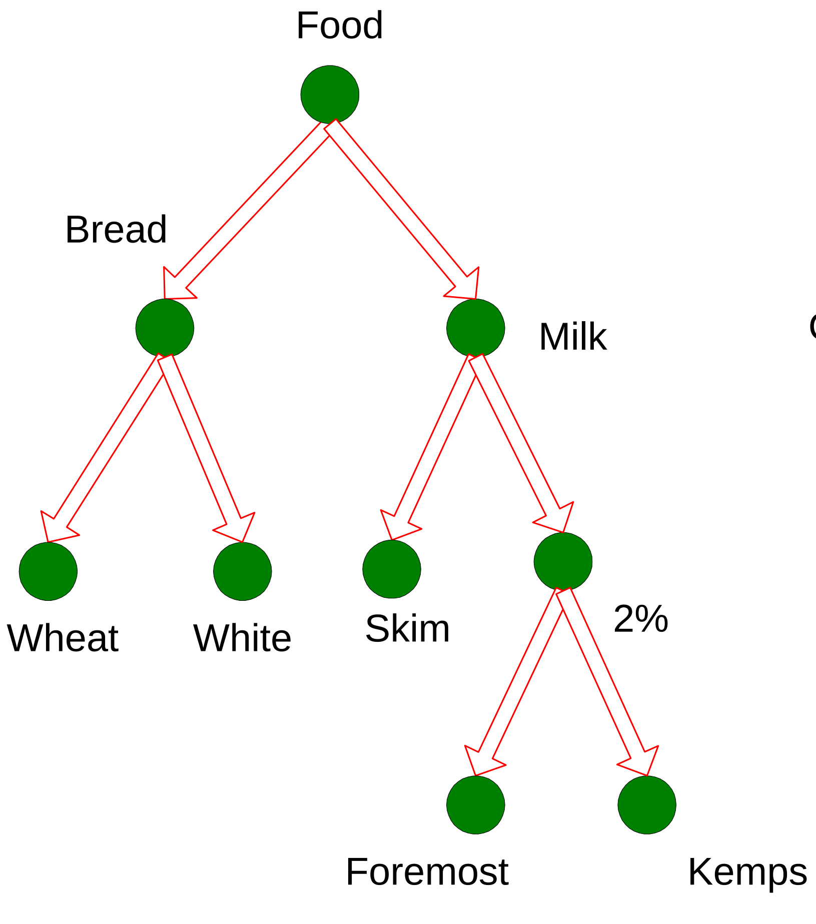
We were able to discover interesting strong correlations with low supports

DBLP AUTHORS	{*Steven M. Beitzel, Eric C. Jensen*}	25	1.00
	{*In-Su Kang, Seung-Hoon Na*}	20	0.98
	{*Ana Simonet, Michel Simonet*}	16	0.94
	{*Caetano Traina Jr., Agma J. M. Traina*}	35	0.92
	{*Claudio Carpineto, Giovanni Romano*}	15	0.91
COMMUNITIES	{*People with social security income:* $>$ 80%, *Age* $\geq$ 65: $>$ 80%}	47	0.76
	{*Large families* ($\geq$ 6): $\leq$ 20%, *White:* $>$ 80%}	1017	0.75
	{*In dense housing* ($\geq$ 1 *per room*): $>$ 80%, *Hispanic:* $>$ 80%, *Large families* ($\geq$ 6): $>$ 80%}	53	0.64
	{*People with Bachelor or higher degree:* $>$ 80%, *Median family income: very high* }	60	0.63
	{*People with investment income:* $>$ 80%, *Median family income: very high* }	66	0.61

***Efficient mining of top correlated patterns based on null-invariant measures by S. Kim et al., 2011**

ASSOCIATIONS ACROSS CONCEPT HIERARCHIES

Items: levels of abstraction



How much to generalize?

- Should we consider correlation between milk and bread, between cream and bagels, or between specific labels of cream and bagels?

- The correlation between specific items can be hard to find because of the low support
- The correlation between more general itemsets can be very low, despite that the support is high

Multi-level Association Rules

- Generate frequent patterns at highest level first.
- Then, generate frequent patterns at the next highest level, and so on, decreasing minsupport threshold
- Issues:
  - May miss some potentially interesting **cross-level** association patterns. E.g.

    skim milk $\rightarrow$ white bread,

    2% milk $\rightarrow$ white bread,

    skim milk $\rightarrow$ white bread

    might not survive because of low support, but

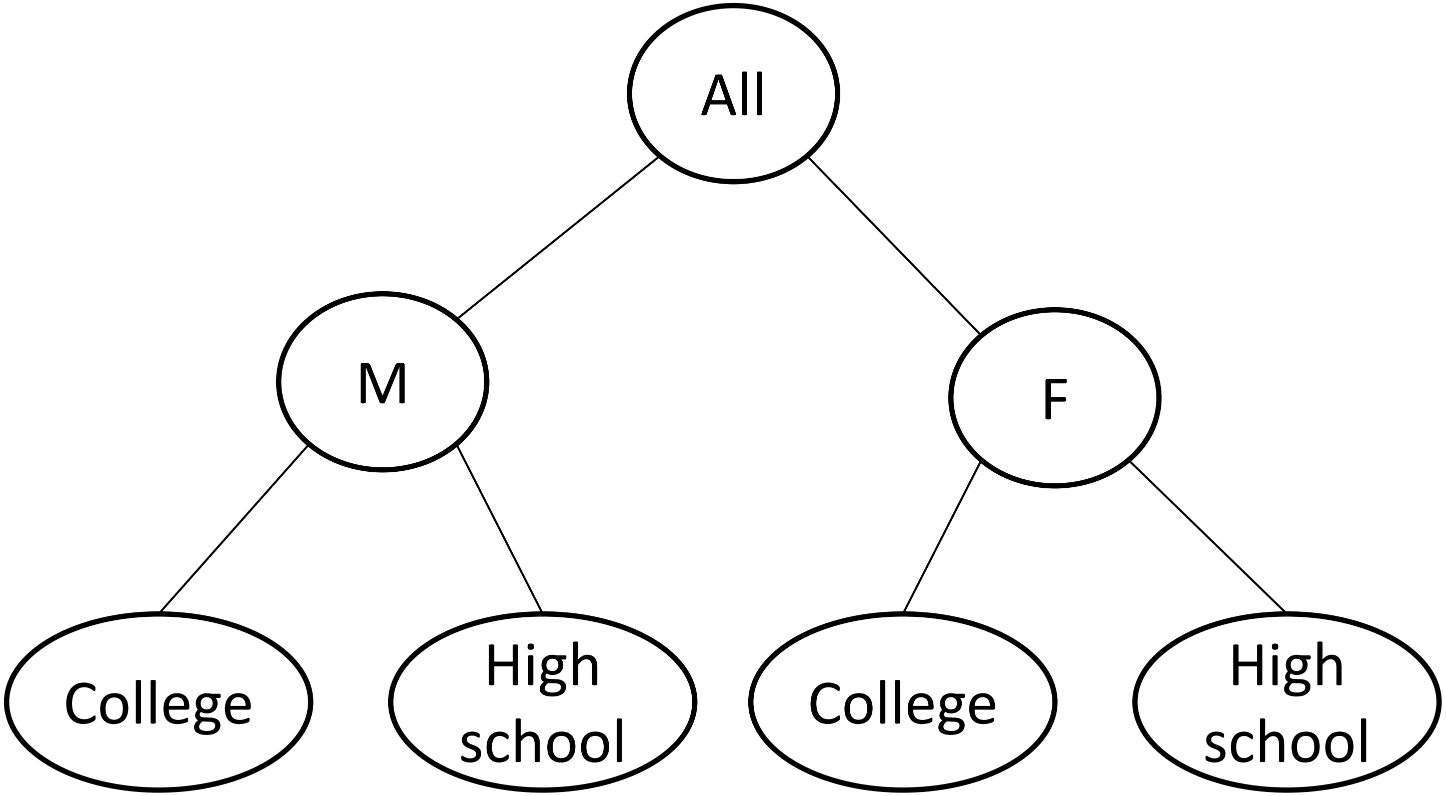
    milk $\rightarrow$ white bread

    could.

    However, we don't generate a cross-level itemset such as

    {milk, white bread}

Customers also may have hierarchies



Hierarchy of groups: strata

Example (symmetric binary variables)

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	99	81	180
No	54	66	120
	153	147	300

- What's the confidence of the following rules:
  (rule 1) {HDTV=Yes} → {Exercise machine = Yes}
  (rule 2) {HDTV=No} → {Exercise machine = Yes} ?

Confidence of rule 1 = 99/180 = 55%

Confidence of rule 2 = 54/120 = 45%

Conclusion: there is a positive correlation between buying HDTV and buying exercise machines

What if we look into more specific groups

Customer Group	Buy HDTV	Buy Exercise Machine		Total
		Yes	No	
College Students	Yes	1	9	10
	No	4	30	34
Working Adult	Yes	98	72	170
	No	50	36	86

- What's the confidence of the rules for each strata:
  (rule 1) {HDTV=Yes} → {Exercise machine = Yes}
  (rule 2) {HDTV=No} → {Exercise machine = Yes} ?

College students:

Confidence of rule 1 = 1/10 = 10%

Confidence of rule 2 = 4/34 = 11.8%

Working Adults:

Confidence of rule 1 = 98/170 = 57.7%

Confidence of rule 2 = 50/86 = 58.1%

The rules suggest that, for each group, customers who don't buy HDTV are more likely to buy exercise machines, which contradict the previous conclusion when data from the two customer groups are pooled together.

Correlation is reversed at different levels of generalization

At a more general level of abstraction:

{HDTV=Yes} → {Exercise machine = Yes}

College students:

{HDTV=No} → {Exercise machine = Yes}

Working Adults:

{HDTV=No} → {Exercise machine = Yes}

This is called Simpson's Paradox

Importance of Stratification

- The lesson here is that proper stratification is needed to avoid generating spurious patterns resulting from Simpson's paradox.

For example

- Market basket data from a major supermarket chain should be stratified according to store locations, while
- Medical records from various patients should be stratified according to confounding factors such as age and gender.

Explanation of Simpson's paradox

- Lisa and Bart are programmers, and they fix bugs for two weeks

	Week 1	Week 2	Both weeks
Lisa	60/100	1/10	**61/110**
Bart	**9/10**	**30/100**	39/110

Who is more productive: Lisa or Bart?

Explanation of Simpson's paradox

	Week 1	Week 2	Both weeks
Lisa	60/100	1/10	**61/110**
Bart	**9/10**	**30/100**	39/110

If we consider productivity for each week, we notice that the samples are of a very different size

The work should be judged from an equal sample size, which is achieved when the numbers of bugs each fixed are added together

Explanation of Simpson's paradox

	Week 1	Week 2	Both weeks
Lisa	60/100	1/10	**61/110**
Bart	**9/10**	**30/100**	39/110

Simple algebra of fractions shows that even though

a1/A > b1/B
c1/C > d1/D

(a1+c1)/(A+C) can be smaller than (b1+d1)/(B+D) !

This may happen when the sample sizes A, B, C, D are skewed (Note, that we are not adding two inequalities, but adding the absolute numbers)

Simpson’s paradox in real life

- Two examples:
  - Gender bias
  - Medical treatment

Example 1: Berkeley gender bias case

Admitted to graduate school at University of California, Berkeley (1973)

	Admitted	Not admitted	Total
Men	3,714	4,727	8,441
Women	1,512	2,808	4,320

- What's the confidence of the following rules:
  (rule 1) {Man=Yes} → {Admitted= Yes}
  (rule 2) {Man=No} → {Admitted= Yes} ?

Confidence of rule 1 = 3714/8441= 44%

Confidence of rule 2 = 1512/4320 = 35%

Conclusion: bias against women applicants

Example 1: Berkeley gender bias case

Stratified by the departments

	Men		Women	
Dept.	Total	Admitted	Total	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

In most departments,
the bias is towards women!

Example 2: Kidney stone treatment

Success rates of 2 treatments for kidney stones

Treatments	Success	Not success	Total
A*	273	77	350
B**	289	61	350

- What's the confidence of the following rules:
  (rule 1) {treatment=A} → {Success= Yes}
  (rule 2) {treatment=B} → {Success = Yes} ?

(A) Confidence of rule 1 = 273/350= 78%

(B) Confidence of rule 2 = 289/350 = **83**%

Conclusion: treatment B is better

*Open procedures (surgery)

** Percutaneous nefrolithotomy (removal through a small opening)

Example 2: Kidney stone treatment

Success rates of 2 treatments for kidney stones

	Treatment A	Treatment B
Small stones	**93% (81/87)**	87%(234/270)
Large stones	**73%(192/263)**	69%(55/80)
Both	78%(273/350)	**83% (289/350)**

Treatment A is better for both small and large stones,
But treatment B is more effective if we add both groups together

Implications in decision making

- Which data should we consult when choosing an action: the aggregated or stratified?
- Kidney stones: if you know the size of the stone, choose treatment A, if you don't – treatment B?

Implications in decision making

- Which data should we consult when choosing an action: the aggregated or stratified?

- The common sense: the treatment which is preferred under both conditions should be preferred when the condition is unknown

Implications in decision making

- Which data should we consult when choosing an action: the aggregated or stratified?

- If we always choose to use the stratified data, we can partition strata further, into groups by eye color, age, gender, race … These arbitrary hierarchies can produce opposite correlations, and lead to wrong choices

Implications in decision making

- Which data should we consult when choosing an action: the aggregated or stratified?

- Conclusion: data should be consulted with care and the understanding of the underlying story about the data is required for making correct decisions

From: Judea Pearl. *Causality: Models, Reasoning, and Inference*

NEGATIVE ASSOCIATIONS

Negative association rules

- The methods for association learning were based on the assumption that the presence of an item is more important than its absence (asymmetric binary attributes)

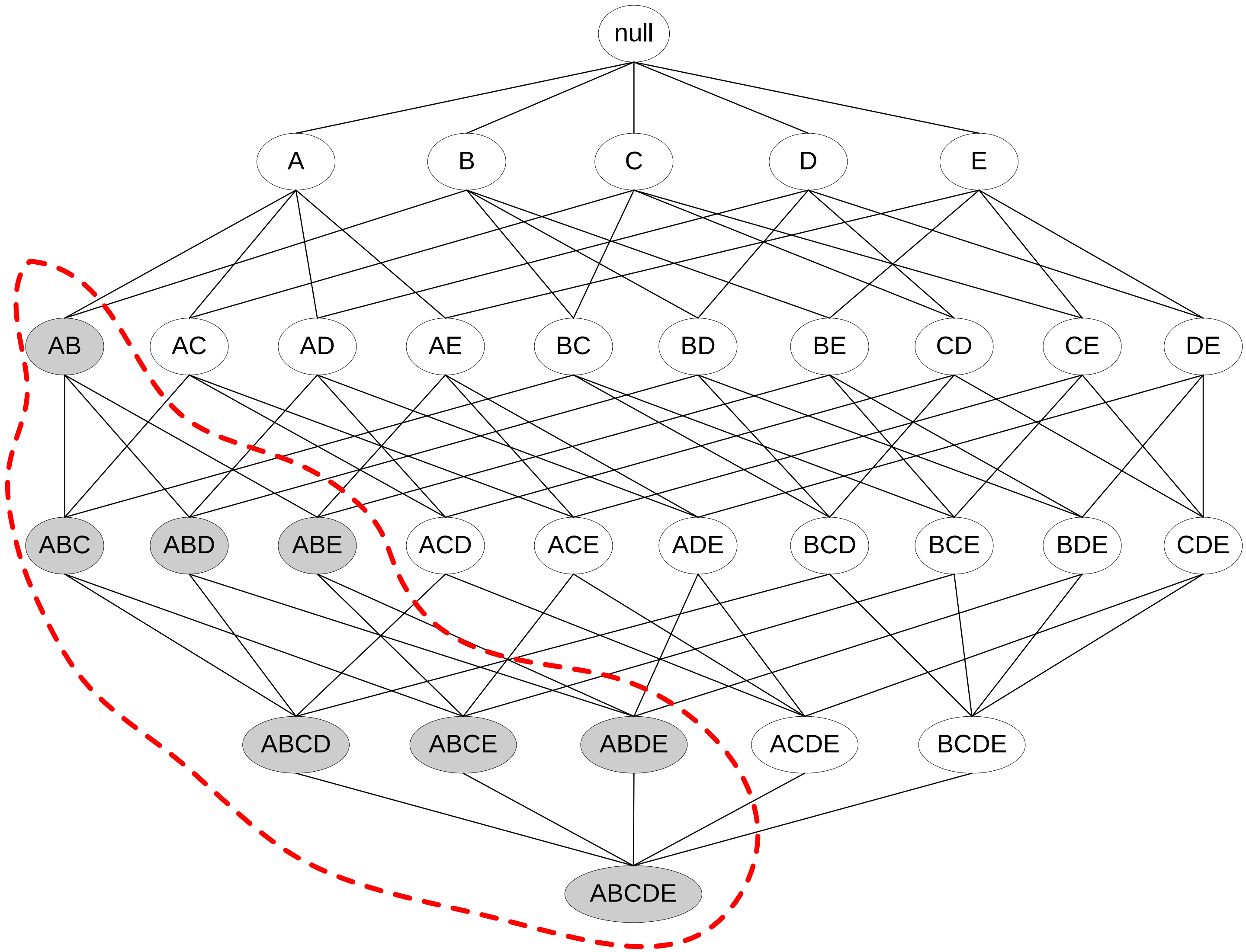
- The negative correlations can be useful:
  - To identify competing items: absence of Blu ray and DVD player in the same transaction
  - To find rare important events: rare occurrence {Fire=yes, Alarm=On}

Mining negative patterns

- Negative itemset: a frequent itemset where at least one item is negated
- Negative association rule: an association rule between items in a negative itemset with confidence $\geq$ *minConf*

- If a regular itemset is infrequent due to the low count of some item, it is frequent if we consider the negation (absence) of a corresponding item

Negative patterns = non-positive



Challenging task

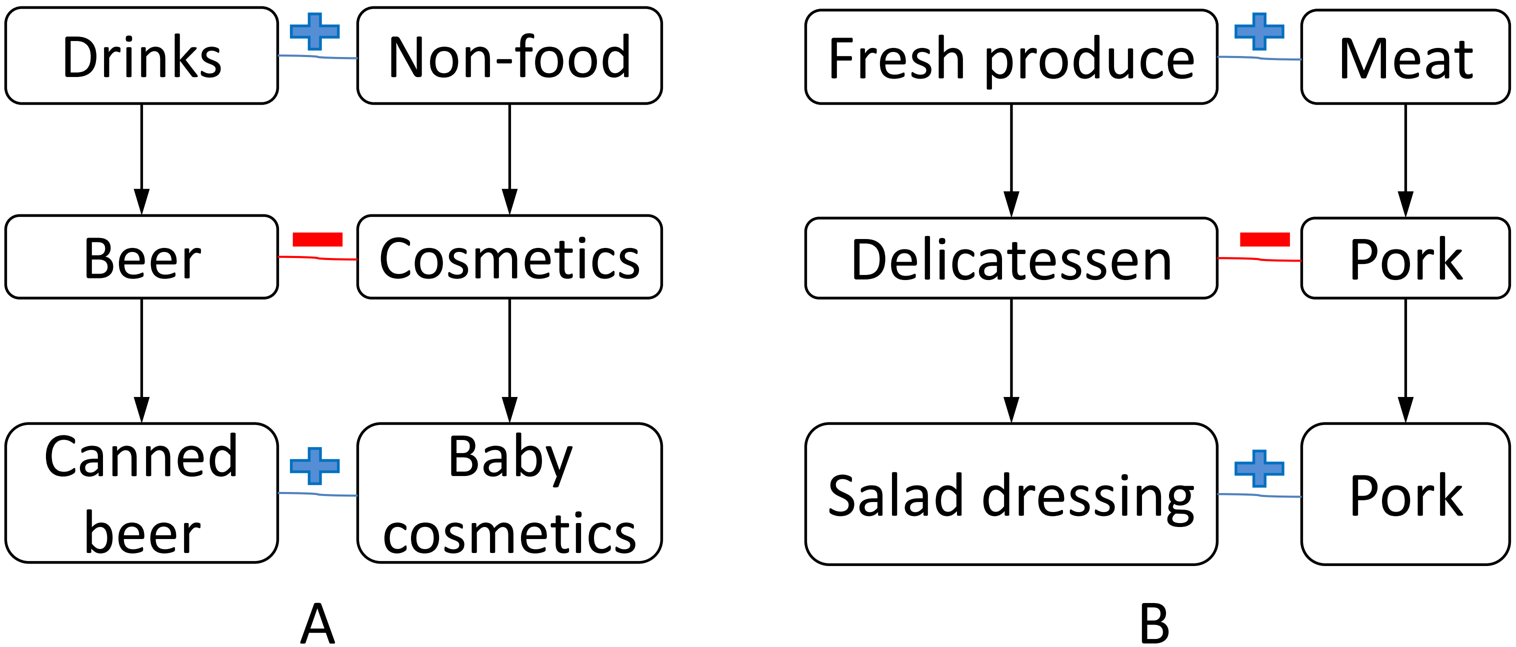
- Positive associations can be extracted only for high-levels of support. Then the set of all frequent itemsets is manageable
- In this case, the complement to all frequent itemsets is exponentially large, and cannot be efficiently enumerated

- But do we need all negative associations?

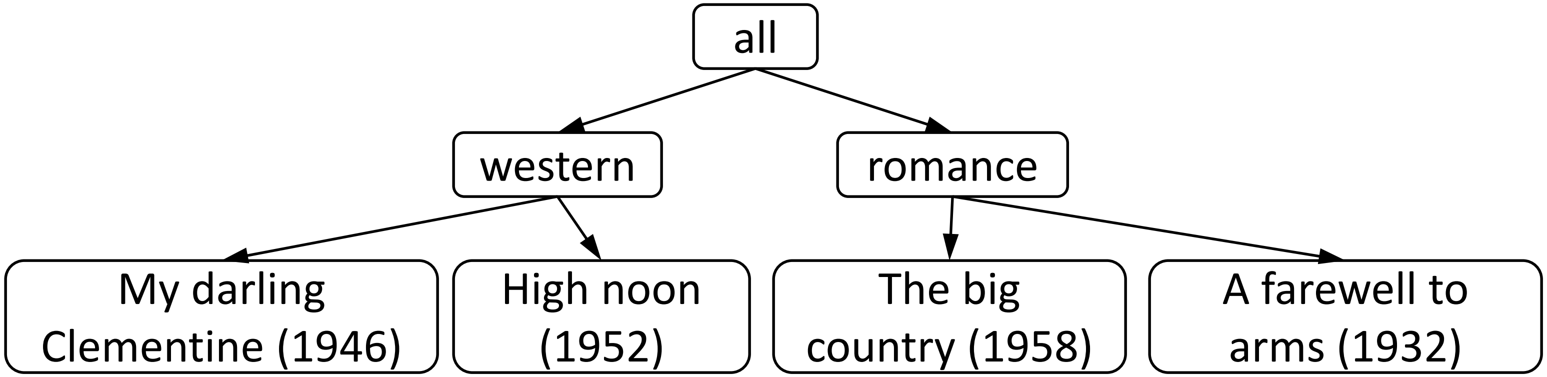
Flipping patterns

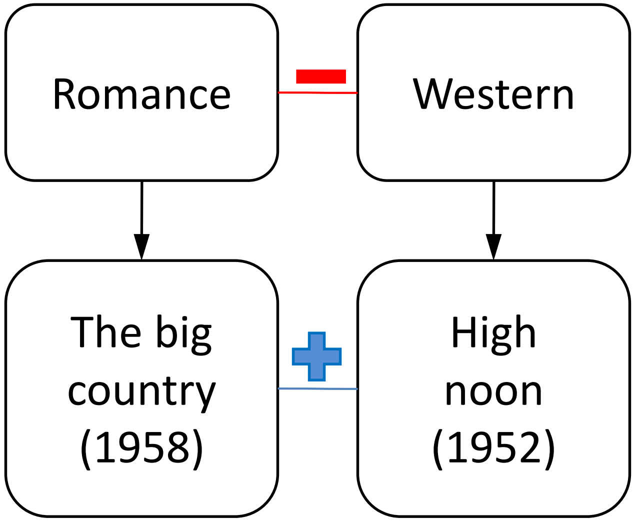
- Flipping patterns are extracted from the datasets with concept hierarchies
- The pattern is interesting if it has positive correlation between items which is accompanied by the negative association of their minimal generalizations, and vice versa
- We call such patterns *flipping patterns*

Example from Groceries dataset

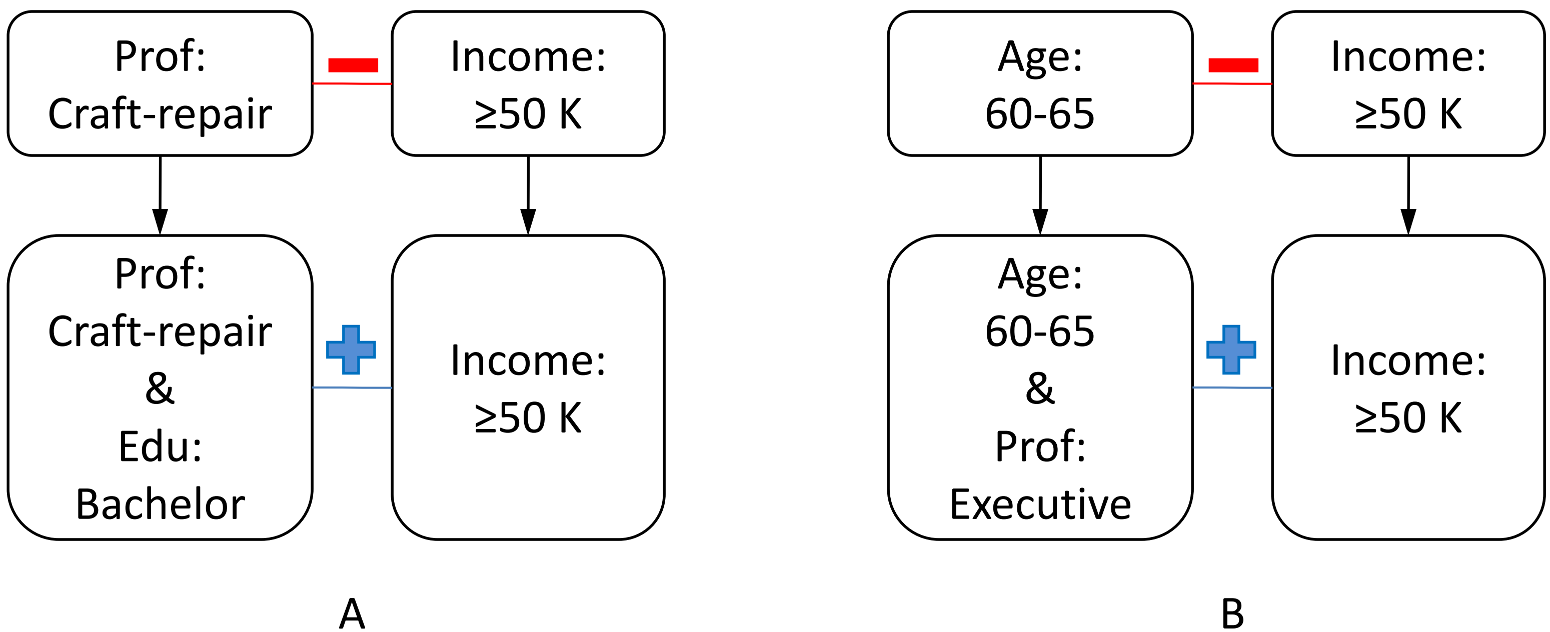


Examples from Movie rating dataset





Examples from US census dataset



Examples from medical papers dataset