More on association analysis

Lecture 13

Association rule mining: problems

1. Extracting frequent itemsets is computationally challenging
2. Issues with the levels of generalization
3. High-confidence rules can be misleading
4. We cannot compute negative associations

1. Discovering frequent itemsets:

Performance Challenge

Frequent Itemset Mining Implementations (FIMI) 2004 challenge

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

Input:

- WebDocs dataset of 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

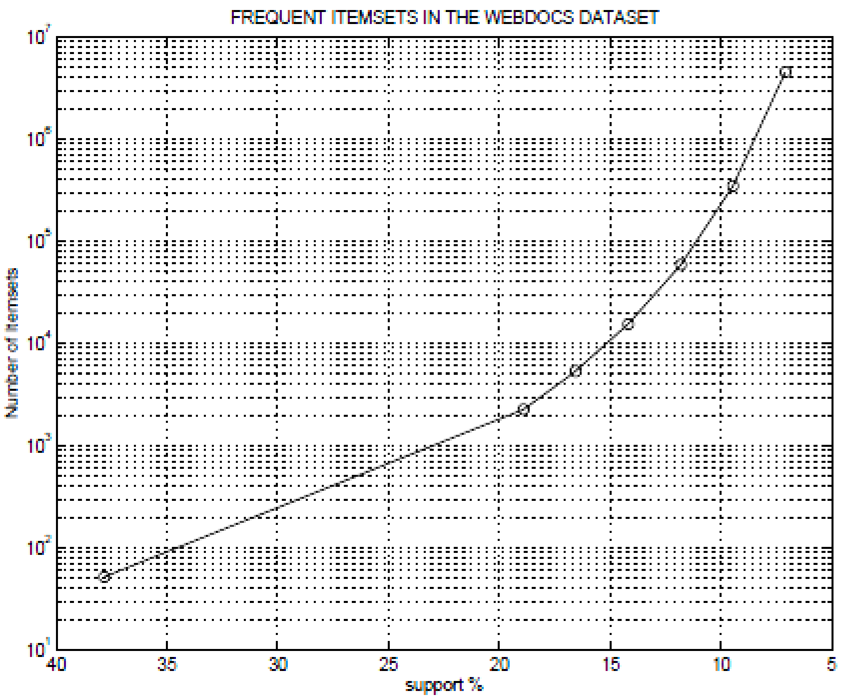
This does not seem too intimidating, right?

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

- When we go below 10% support, the number of frequent itemsets becomes too big
- How big?

So big that we cannot keep in memory all different 2-item combinations, to update their counts

- So we are forced to use high min support threshold and produce only **very** frequent itemsets



*That is, the frequent combinations that occur in 250,000 documents or more

We set the min support threshold high to make computation efficient

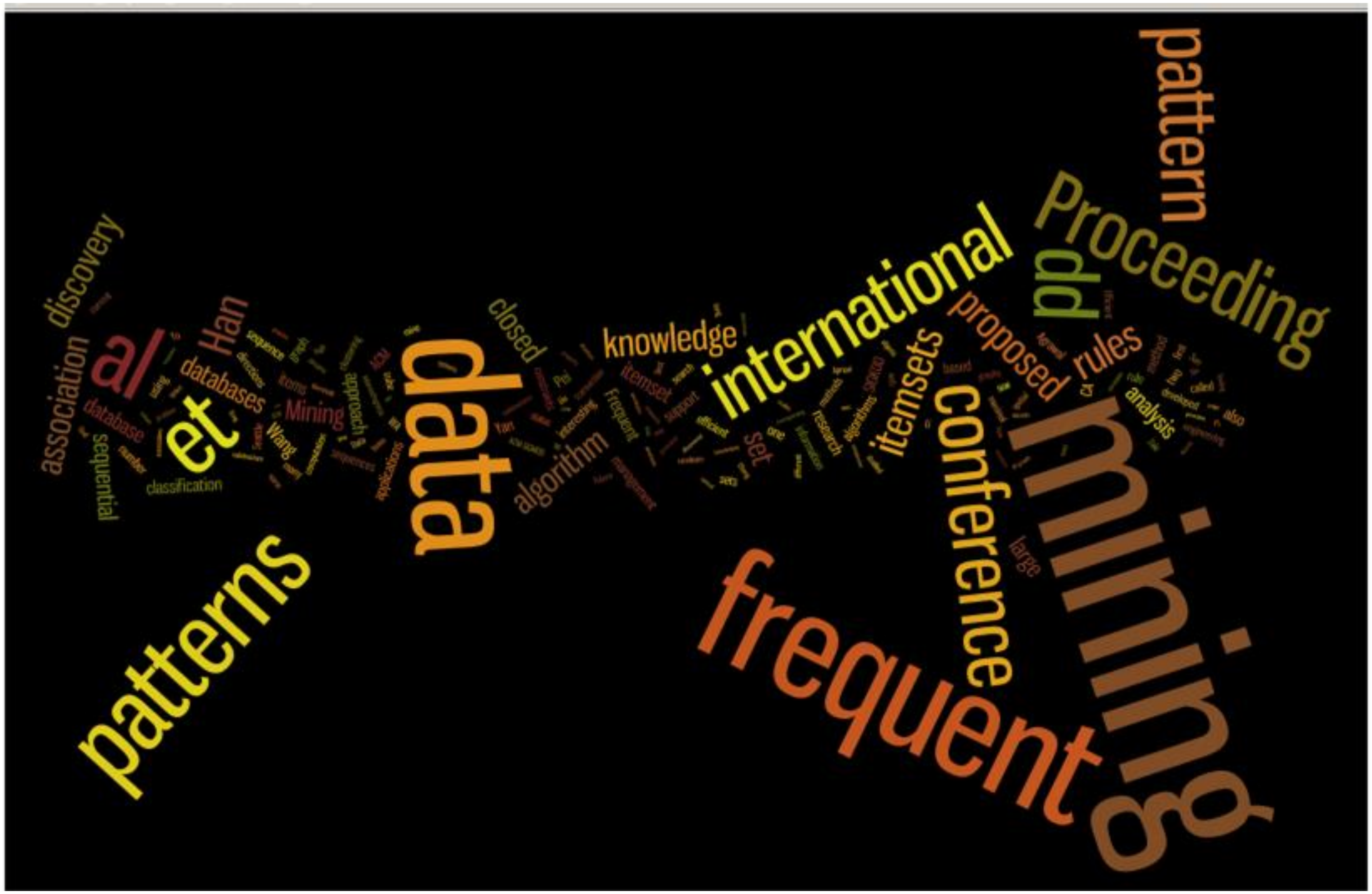
We can discover only **very** frequent datasets

Example 1: frequent 1-itemsets in 5 Shakespeare sonnets

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

http://www.tagcrowd.com/

Example 2: Frequent 1-itemsets in papers on frequent pattern mining



http://www.wordle.net/create

Trivial discoveries

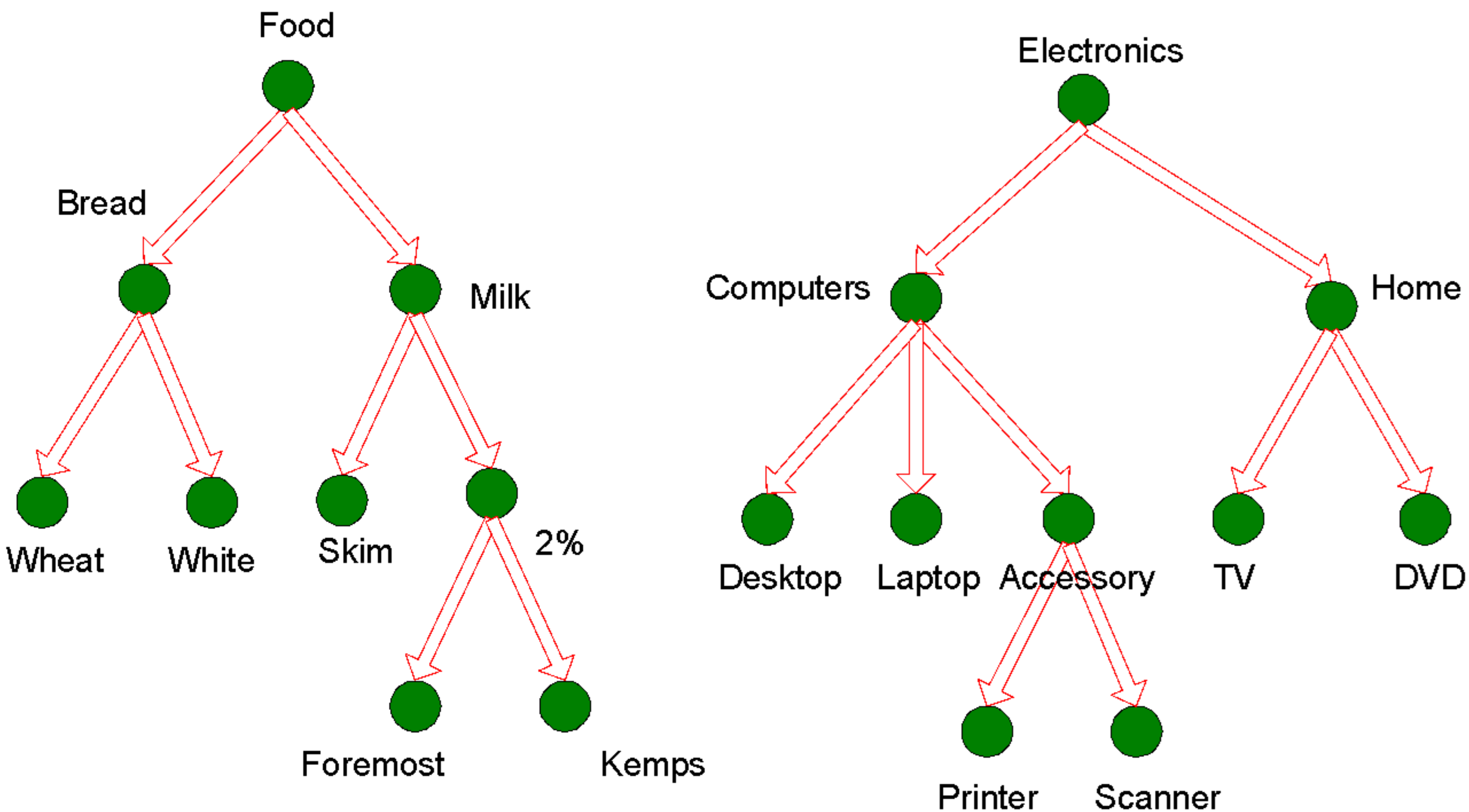
- Frequent itemsets are easily computable only for high min support
- Most rules discovered from these itemsets are trivial and obvious!

- We need to lower the min support threshold to make some non-trivial discoveries
- This leads to computational challenges that have no good solutions so far

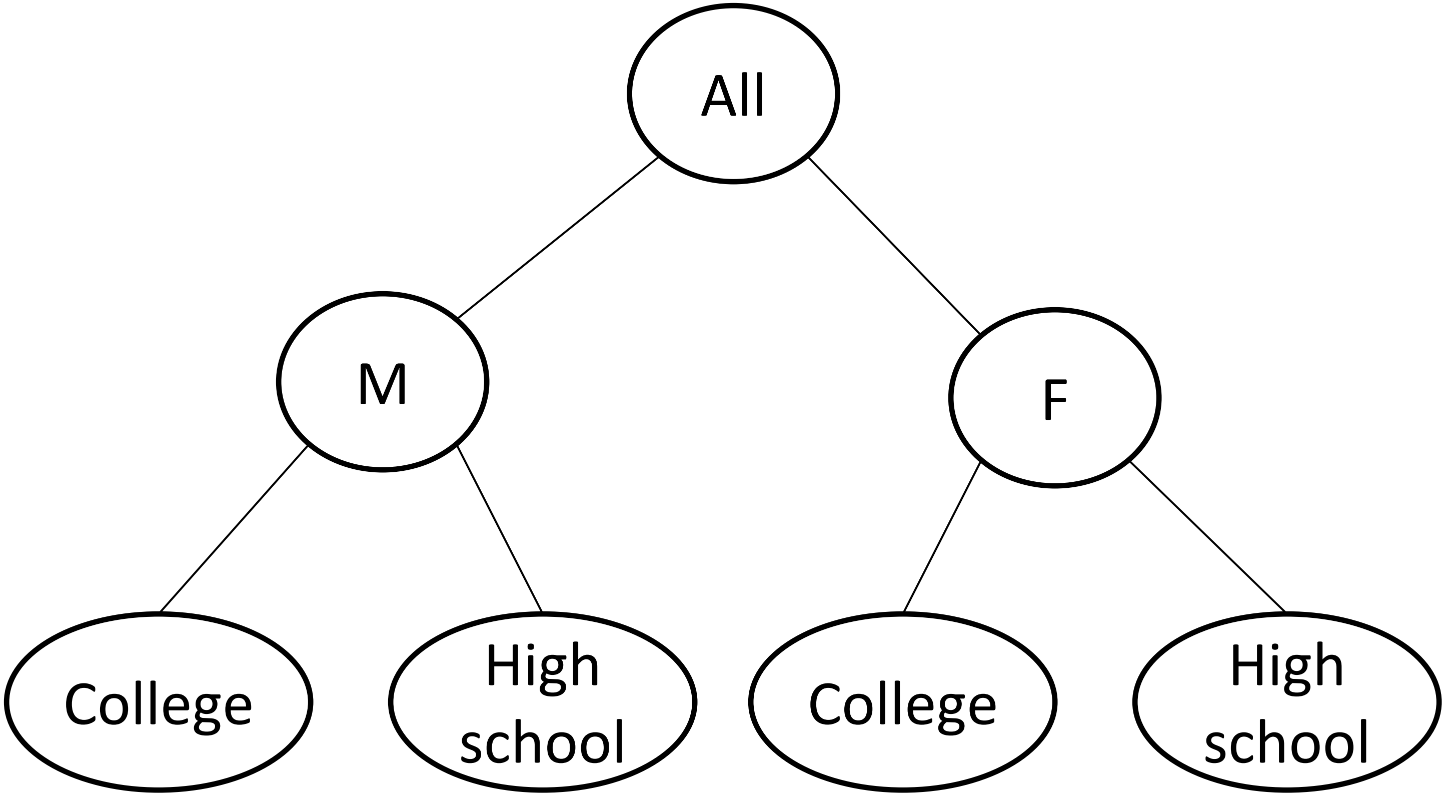
2. Different levels of abstraction:

Simpson's Paradox

Concept hierarchies: items



Concept hierarchies: customers



Hierarchy of groups: strata

How much to generalize?

- Should we consider correlation between milk and bread, between cream and bagels, or between specific labels of cream and bagels?
- Should we consider frequent itemsets for a specific strata separately?

Example 3 (symmetric binary variables)

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

- Compute 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: the customers who bought HDTV are more likely to buy 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

- Compute 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 contradicts the previous conclusion when data from the two customer groups were 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**

Simpson's paradox in real life

- Two examples:
  - Gender bias
  - Medical treatment

Example 4: 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 4: 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 5: 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 5: 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

So what treatment is better?

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

Example 6.
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 fractions, but adding the absolute numbers)

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

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

3. High-confidence rules can be misleading:

Pitfalls of Confidence

High-confidence rules

What does it mean for a rule to have a high confidence?

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} is used to define various rule metrics

	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 ¬Y
f_{01}: support count of ¬X and Y
f_{00}: support count of ¬X and ¬Y

|T|: total number of transactions

Example 7: tea and coffee

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

Example 7: tea and coffee

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

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

This is a top-confidence rule!

Example 7: tea and coffee

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

- Confidence of rule T $\longrightarrow$ C

  P(C|T)=0.75

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

Example 7: tea and coffee

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

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

Although confidence is high, the rule is misleading:

$P(C|\neg T)=750/900=0.83$

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 → 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

Idea: apply statistical independence test

Statistical measure of association (correlation)-*Lift*

- If the purchase of T is statistically independent 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: Rule Interest Factor

Measure that takes into account statistical (in)dependence

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

- Interest factor compares the frequency of a pattern against a baseline frequency computed under the assumption of statistical independence.
- 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.

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

Back to Example 7: 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 – close to 1)

Example 8: 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

Example 9. More problems with Lift:

- 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	400	1,000
	1,000	1000	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

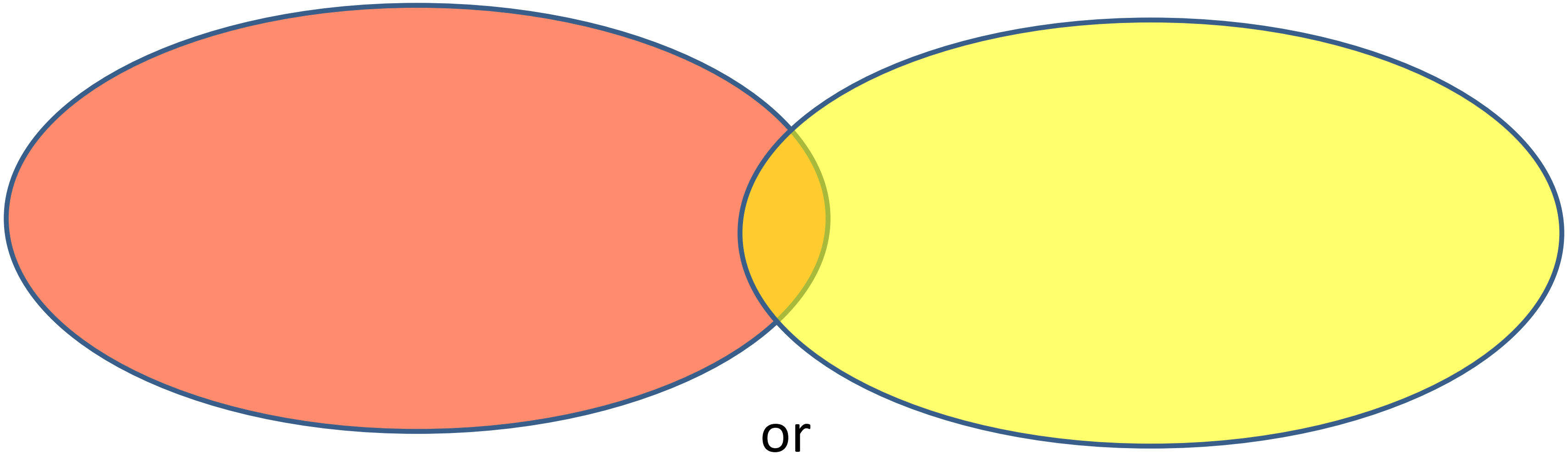
But the relationship between C and M is the same in both datasets

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

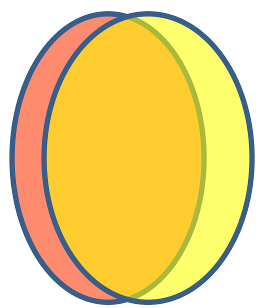
Such transactions are called *null-transactions* with respect to C and M

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

Which items are more correlated?



or

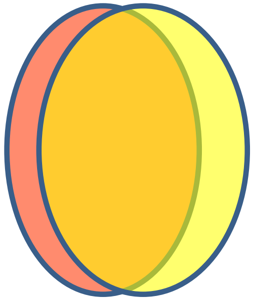


The areas correspond to support counts

Possible null-invariant measure: Jaccard index

Jaccard index: intersection/union

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



Back to Example 8

- 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?

Lift (M,C) = 8.26

Lift (P,S)=25.00

Jaccard on Example 8

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

JI (C,M) = 10000/(11000 + 11000 - 10000) =0.83

JI (P,S) = 1000/(2000 + 2000 - 1000) = 0.33

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

Back to Example 9: 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: JI (C,M) = 400/(1000+1000-400) =0.25

DB2: JI (C,M) = 400/(1000+1000-400) = 0.25

DB1: Lift = 8.0 (positive correlation)

DB2: Lift = 0.8 (negative correlation)

Computational challenge

- It seems that 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 dataset?
- All null-invariant measures are non-antimonotone
- No efficient solution so far

4. Some research on negative associations

Flipping correlations

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 associations/correlations can be useful:
  - To identify competing items: absence of Blu ray and DVD player in the same transaction
  - To find rare important events: {Fire=yes} is frequent, but {Fire=yes, Alarm=On} is infrequent → faulty alarm?

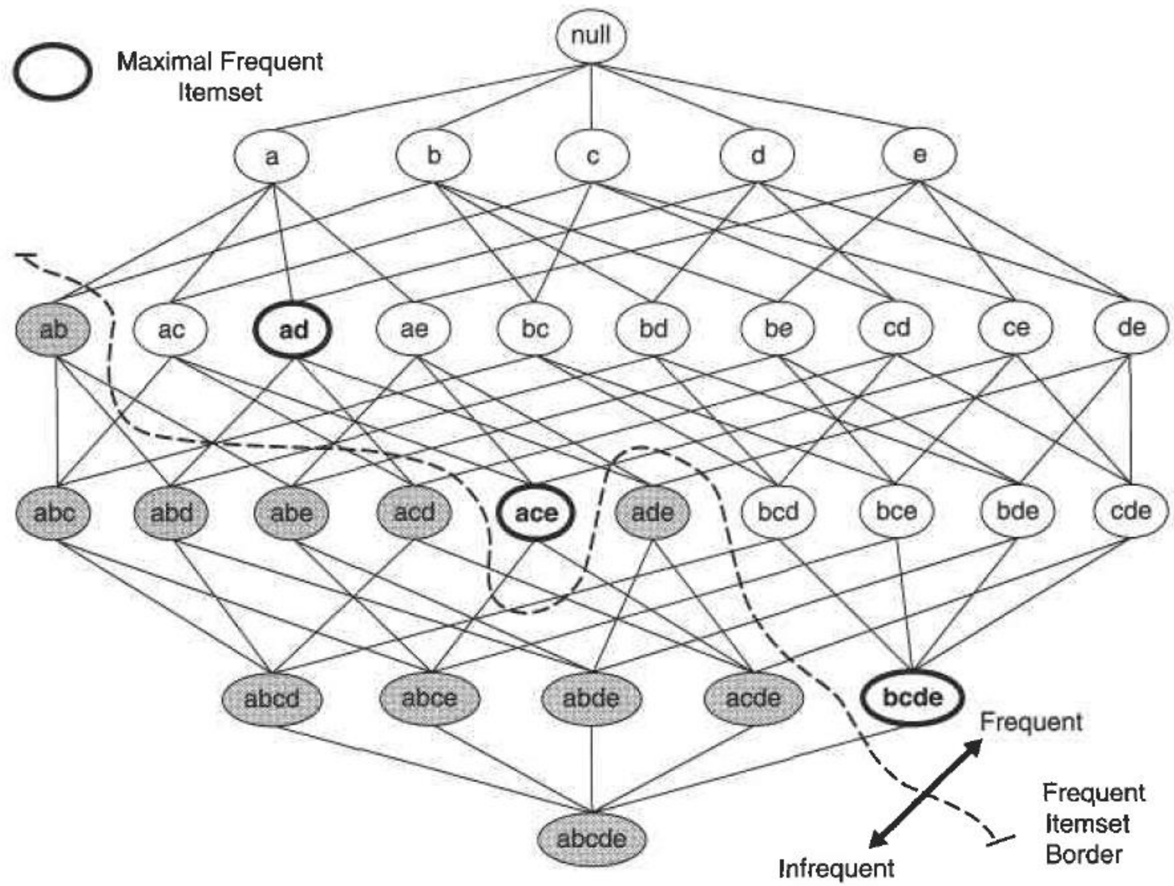
Discovering negative patterns

- 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 ≥ *minConf*

Example: Tea → ! Coffee

- Each transaction now contains all the *d* items: some present some absent
- 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

Infrequent itemsets: all minus frequent



Challenging task

- Positive associations can be extracted only for high-levels of support. Then the set of all frequent itemsets is manageable
- To compute negative associations, the complement to all frequent itemsets is exponentially large, and cannot even be efficiently enumerated!

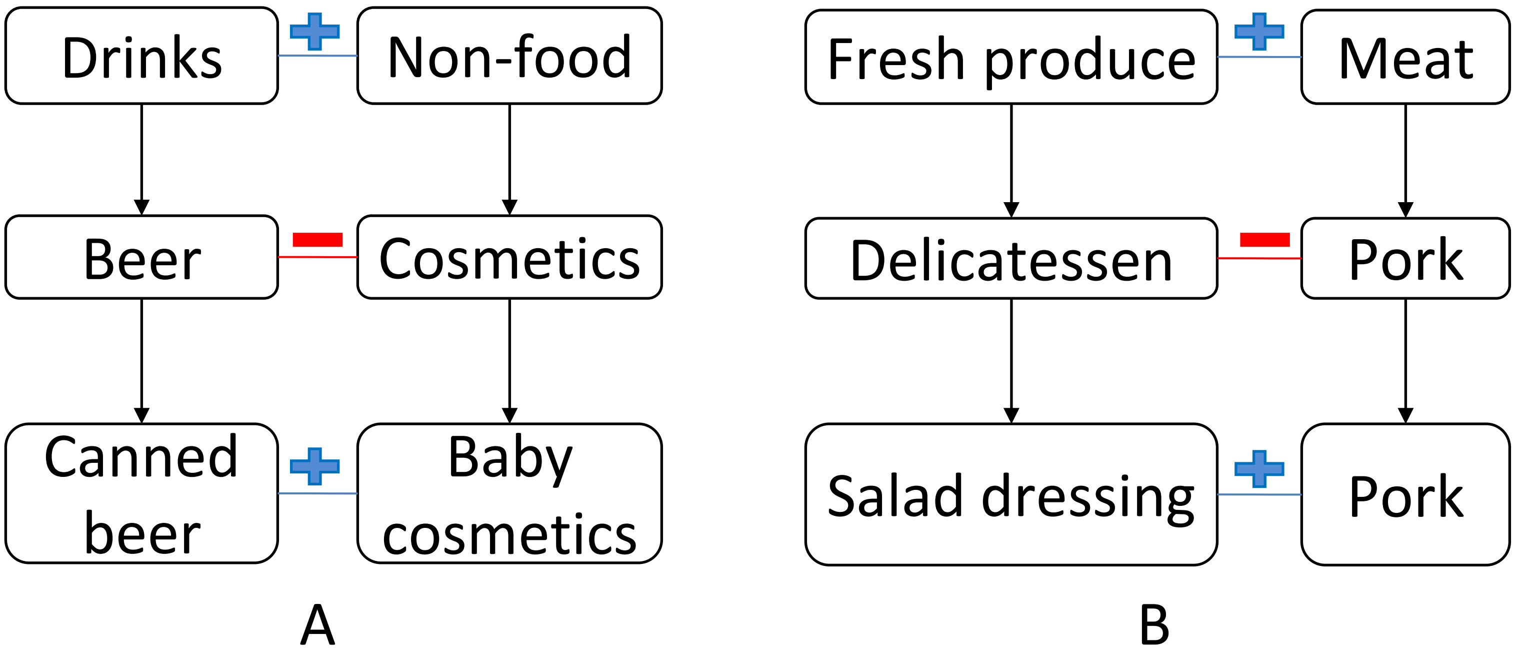
- Maybe we can find only some interesting negative associations?

Flipping patterns

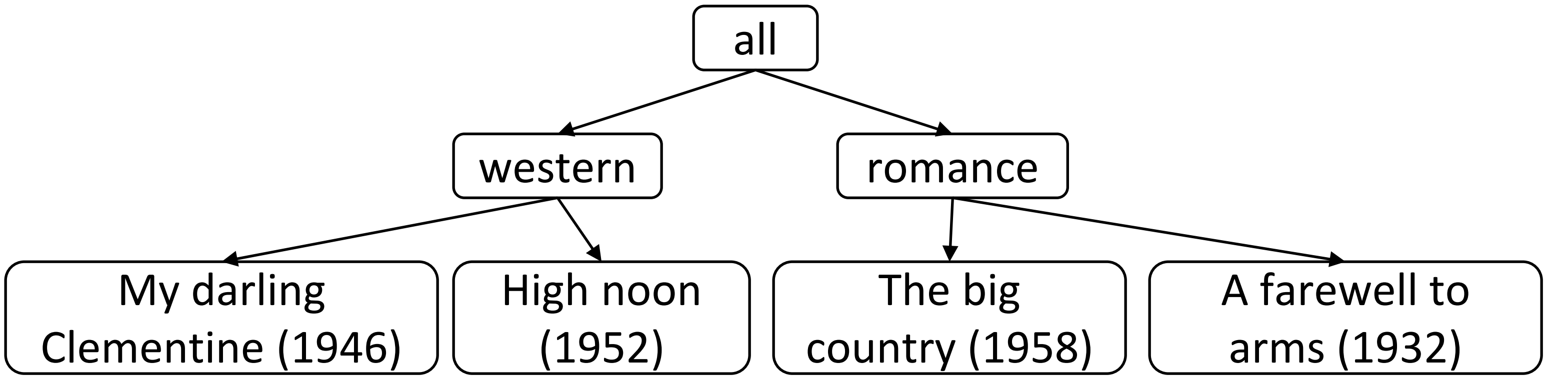
- Flipping correlations are extracted from the datasets with concept hierarchies
- The pattern is flipping if it has positive correlation between items which is accompanied by the negative correlation between their minimal generalizations, and vice versa

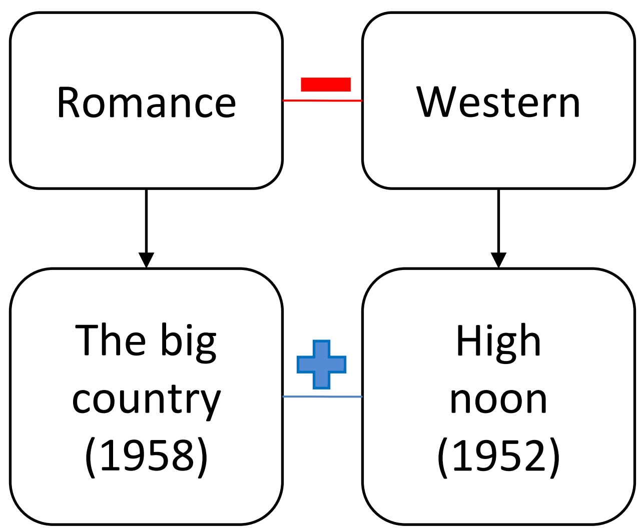
Marina Barsky, Sangkyum Kim, Tim Weninger, Jiawei Han:
Mining Flipping Correlations from Large Datasets with Taxonomies.
Proc. VLDB Endow. 5(4): 370-381 (2011)

Example from Groceries dataset

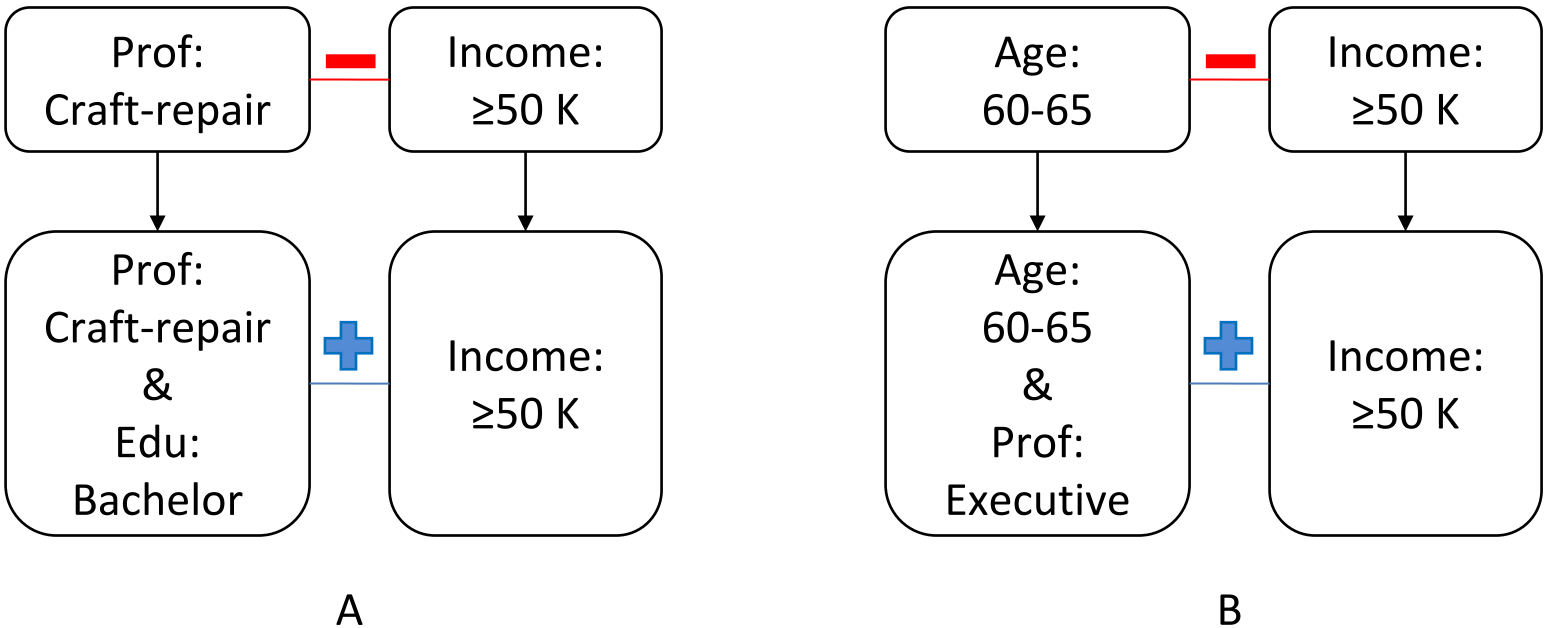


Examples from Movie rating dataset





Examples from US census dataset



Examples from medical papers dataset