Which associations are interesting?

Lecture 16

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:

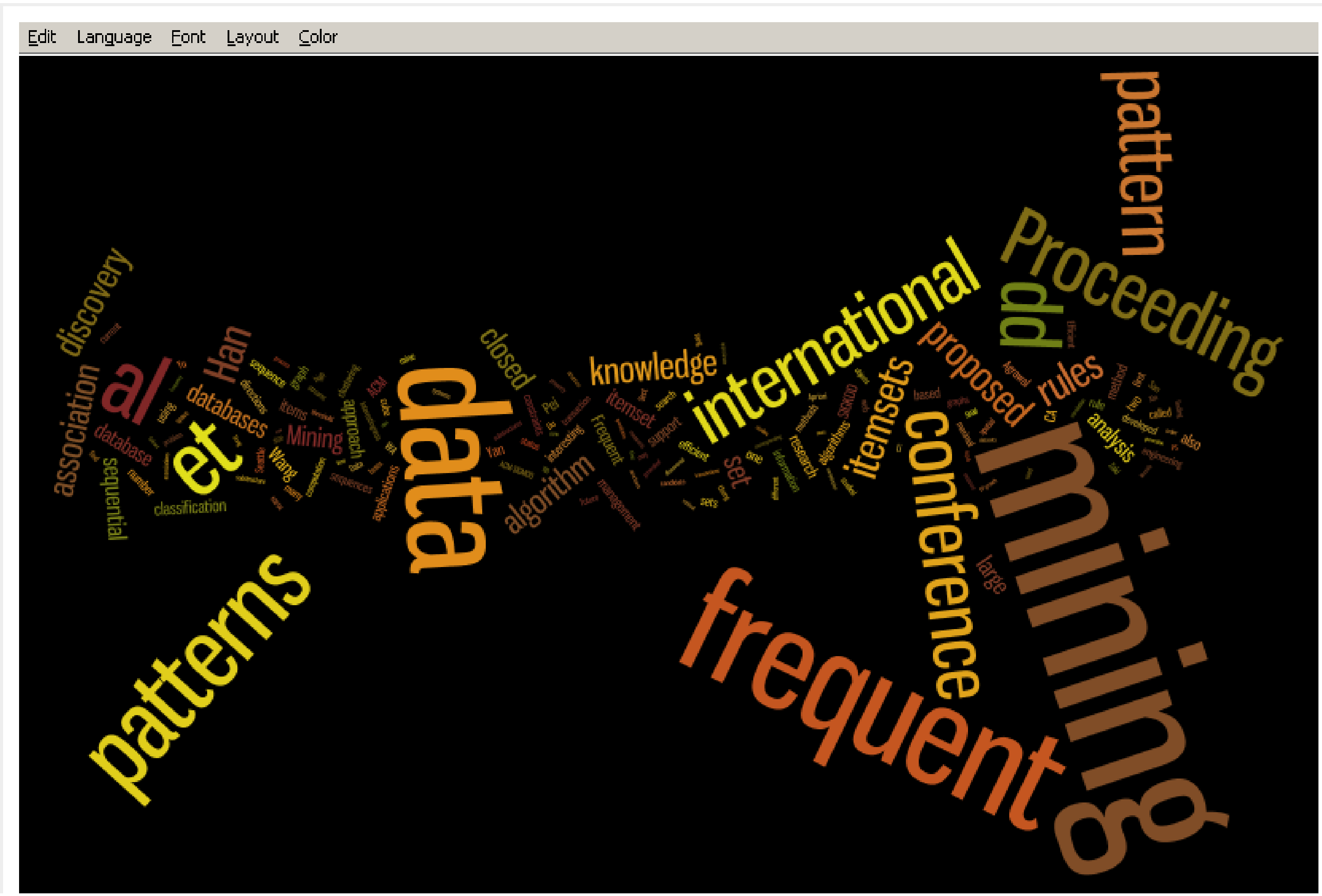
http://www.wordle.net/create

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

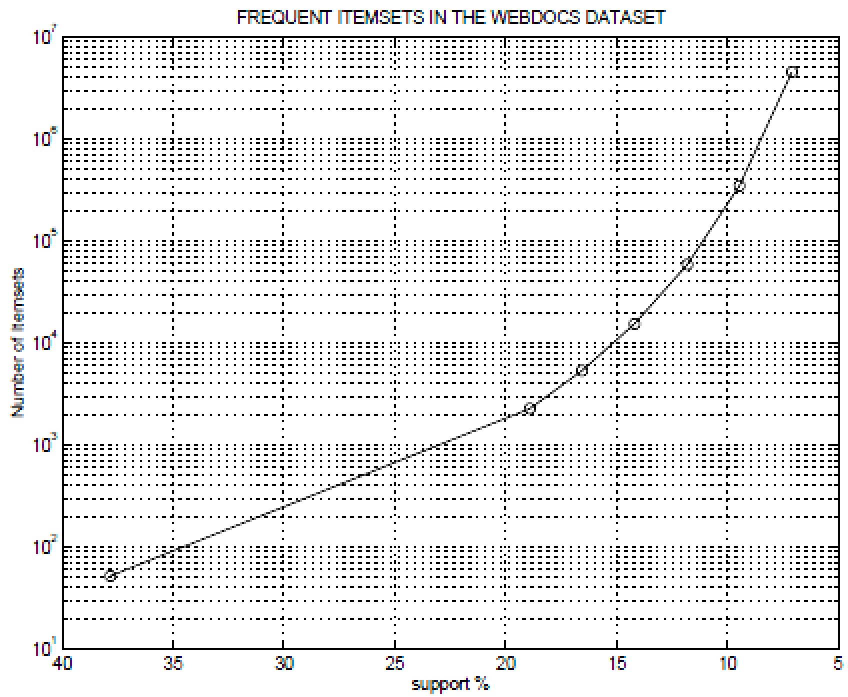
Frequent Itemset Mining Implementations (FIMI) 2004 challenge

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 minsupp=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: that 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:

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

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 $T \rightarrow C$ $P(C|T)=0.75$

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

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

- Interest factor compares the frequency of a pattern against a baseline frequency computed under the 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 $\rightarrow$ 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

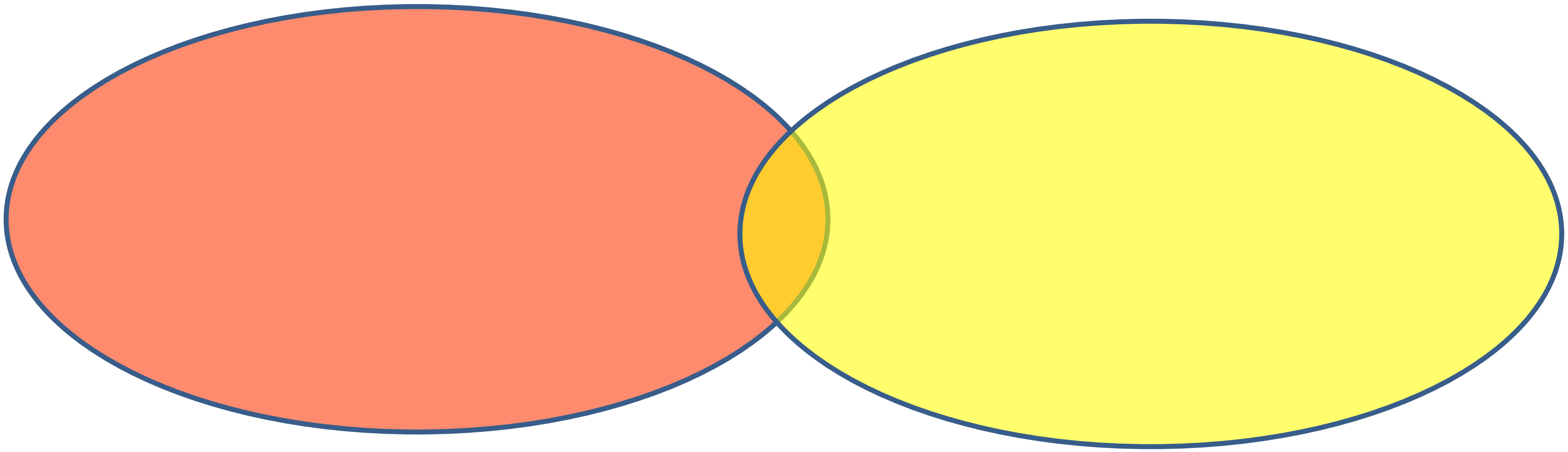
The changes are in the count of transactions which do not contain neither C nor M.

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

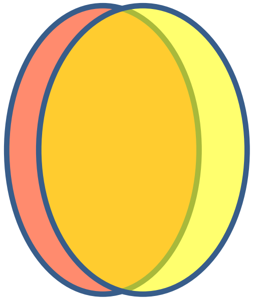
We want the measure which does not depend on null-transactions: null-transaction invariant. Which depends only on counts of items in the current itemset

What are we looking for?

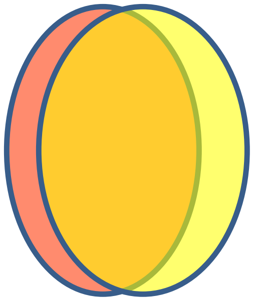
The area corresponds to support counts



or



Possible null-invariant measure 1: Jaccard index



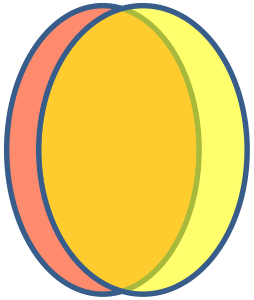
Jaccard index: intersection/union

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

$$\boxed{\text{Kulc}(A, B) = [P(A|B)+P(B|A)]/2}$$



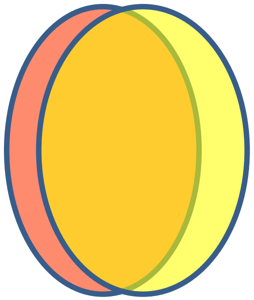
In terms of support counts:

$$\text{Kulc}(A,B) = \frac{1}{2}\left[\text{sup}(A \text{ and } B)/\text{sup}(A) + \text{sup}(A \text{ and } B)/\text{sup}(B)\right]$$

Possible null-invariant measure 3: Cosine

Cosine: geometric mean of conditional probabilities

$\boxed{\text{Cos (A, B) = sqrt[P(A|B) x P(B|A)]}}$



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?
- This is the area of the 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**