Association Rule Mining

Yannis Kotidis

kotidis@aueb.gr

Professor, Department of Informatics

Athens University of Economics and Business

Suggested Reading

- **Data Mining: Concepts and Techniques**, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems) 3rd Edition, by Jiawei Han, Micheline Kamber, Jian Pei (Chapter 6)

- **Mining of Massive Datasets**, 2nd Edition, by Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University (Chapter 6)

Data Mining

- The process of analyzing data to identify patterns or relationships
- Has become a well-established discipline related to Artificial Intelligence and Statistical Analysis
  - Led by advances in computer hardware and our ability to analyze big datasets
    - Data warehousing, BI, Cloud Computing

Association Rule Mining

- Finding frequent patterns (associations) among sets of items in transactional databases
  - Basket data analysis, catalog design, direct mailing,…

- *Basic question:* ***"Which groups or <u>sets of items</u> are customers likely to purchase on a given trip to the store?"***

- Learned patterns or itemsets, sush as {diapers, beers}, are used to construct if-then scenario (probabilistic) rules
  - buys(x, "diapers") → buys(x, "beers") [5%, 60%]

What to do with rule Diapers→Beers ?

- Enhance observed behavior
  - Place products in proximity to further encourage the combined sale
  - Increase the price of diapers but put beer in discount for a combined sale

- Put products at opposite ends of the store to make customers spend more time (and buy more products) at the store

More ideas

- Assume laptops and printers are frequently sold together
  - Place a higher-margin printer near the laptop section
  - Take a soon to be updated software suite and bundle it in an offer with laptops and printers

- See https://www.kdnuggets.com/news/98/n01.html
  - What Wal-Mart might do with Barbie doll → Candy bars association rule

Basic Concepts

- Example: Basket Data analysis
  - Each transaction (basket) is a set of items (e.g. purchased by a customer in a visit)

T1: Milk, Diaper, Chocolate

T2: Diaper, Beer, Meat

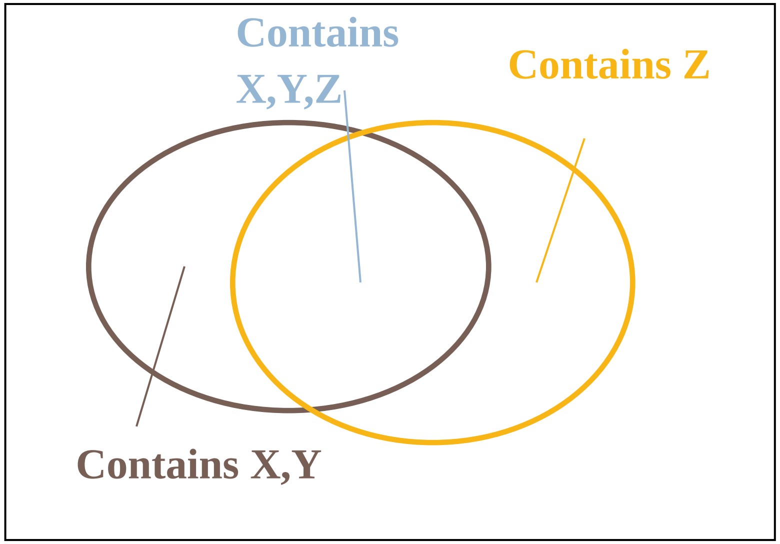
T3: Sugar, Beer, Diaper

…

Inferred rule:

buys(x, "Diaper") → buys(x, "Beer") [5%, 67%]

Support and Confidence



TID	Items
T1	A,C
T2	A,C,D
T3	A,E
T4	D,E,F,G

- Given rule X,Y $\Rightarrow$ Z
- **Support**: probability that a transaction contains {X,Y,Z}
  - s=P[X and Y and Z]
- **Confidence**: probability that a transaction having {X,Y} also contains Z
  - c=P[Z|X,Y]

Let minimum support 50%, and minimum confidence 50%, we have

$$A \Rightarrow C \;\; (50\%, 66.6\%)$$

$$C \Rightarrow A \;\; (50\%, 100\%)$$

Problem formulation

- **Given**
  - a set of 'market baskets' (=binary matrix, of N rows/baskets and M columns/products)
  - min-support 's' and
  - min-confidence 'c'
- **Find**
  - all the rules with:

support ≥ s & confidence ≥ c

Tid	Diaper	Meat	Milk	Beer
1	1	0	1	1
2	1	1	0	0
3	1	1	0	0
4	0	1	1	0

From rules to itemsets

- First, find frequent itemsets
  - e.g. {X,Y,Z}
  - "Frequent" means support $\geq$ s (min-support)
- Once we have a 'frequent itemset', we can find out the qualifying rules easily (how?)

$$\text{Support}(X,Y \rightarrow Z) = \text{Freq}(\{X,Y,Z\})$$

$$\text{Conf}(X,Y \rightarrow Z) = P[Z|X,Y] = P[X,Y,Z]/P[X,Y]$$
$$= \text{Freq}(\{X,Y,Z\}) \;/\; \text{Freq}(\{X,Y\})$$

- Thus, let's focus on how to find frequent itemsets

Brute-force Frequent Itemsets Counting

- Scan database once; maintain 2^M-1 counters
  - One counter for each of {A}, {B}, {C}, …., {A,B}, {A,C}, {A,D}, … {B,C}, {B,D}, {B,E},… {A,B,C}, …
- Example (M=3, 2^3-1=7 possible itemsets)

Itemset	Counter
{A}	0 +1
{B}	0 +1
{C}	0
{A,B}	0 +1
{A,C}	0
{B,C}	0
{A,B,C}	0

Increase counters of itemsets contained in the basket

Basket 1: A,B

Brute-force Frequent Itemsets Counting

- Scan database once; keep 2^M-1 counters
  - One counter for each of {A}, {B}, {C}, …., {A,B}, {A,C}, {A,D}, … {B,C}, {B,D}, {B,E},… {A,B,C}, …
- Example (M=3)

Itemset	Counter
{A}	1
{B}	1 +1
{C}	0
{A,B}	1
{A,C}	0
{B,C}	0
{A,B,C}	0

Basket 1: A,B
Basket 2: B

Brute-force Frequent Itemsets Counting

- Scan database once; keep 2^M-1 counters
  - One counter for each of {A}, {B}, {C}, …., {A,B}, {A,C}, {A,D}, … {B,C}, {B,D}, {B,E},… {A,B,C}, …
- Example (M=3)

Itemset	Counter
{A}	1
{B}	2 +1
{C}	0 +1
{A,B}	1
{A,C}	0
{B,C}	0 +1
{A,B,C}	0

Basket 1: A,B
Basket 2: B
Basket 3: B,C

Brute-force Frequent Itemsets Counting

- Scan database once; keep 2^M-1 counters
  - One counter for each of {A}, {B}, {C}, …., {A,B}, {A,C}, {A,D}, … {B,C}, {B,D}, {B,E},… {A,B,C}, …
- Example (M=3)

Itemset	Counter
{A}	1
{B}	3
{C}	1
{A,B}	1
{A,C}	0
{B,C}	1
{A,B,C}	0

Basket 1: A,B
Basket 2: B
Basket 3: B,C

Brute-force Frequent Itemsets Counting

- Scan database once; keep 2^M-1 counters
  - One counter for each of {A}, {B}, {C}, …., {A,B}, {A,C}, {A,D}, … {B,C}, {B,D}, {B,E},… {A,B,C}, …
- Example (M=3)

Itemset	Counter
{A}	2
{B}	4
{C}	1
{A,B}	2
{A,C}	0
{B,C}	1
{A,B,C}	0

Basket 1: A,B
Basket 2: B
Basket 3: B,C
Basket 4: A,B

Brute-force Frequent Itemsets Counting

- Scan database once; keep 2^M-1 counters
  - One counter for each of {A}, {B}, {C}, …., {A,B}, {A,C}, {A,D}, … {B,C}, {B,D}, {B,E},… {A,B,C}, …
- Example (M=3)

Itemset	Counter
{A}	3
{B}	4
{C}	1
{A,B}	2
{A,C}	0
{B,C}	1
{A,B,C}	0

A→B [Support = ? , Confident = ?]

Basket 1: A,B
Basket 2: B
Basket 3: B,C
Basket 4: A,B
Basket 5: A

Brute-force Frequent Itemsets Counting

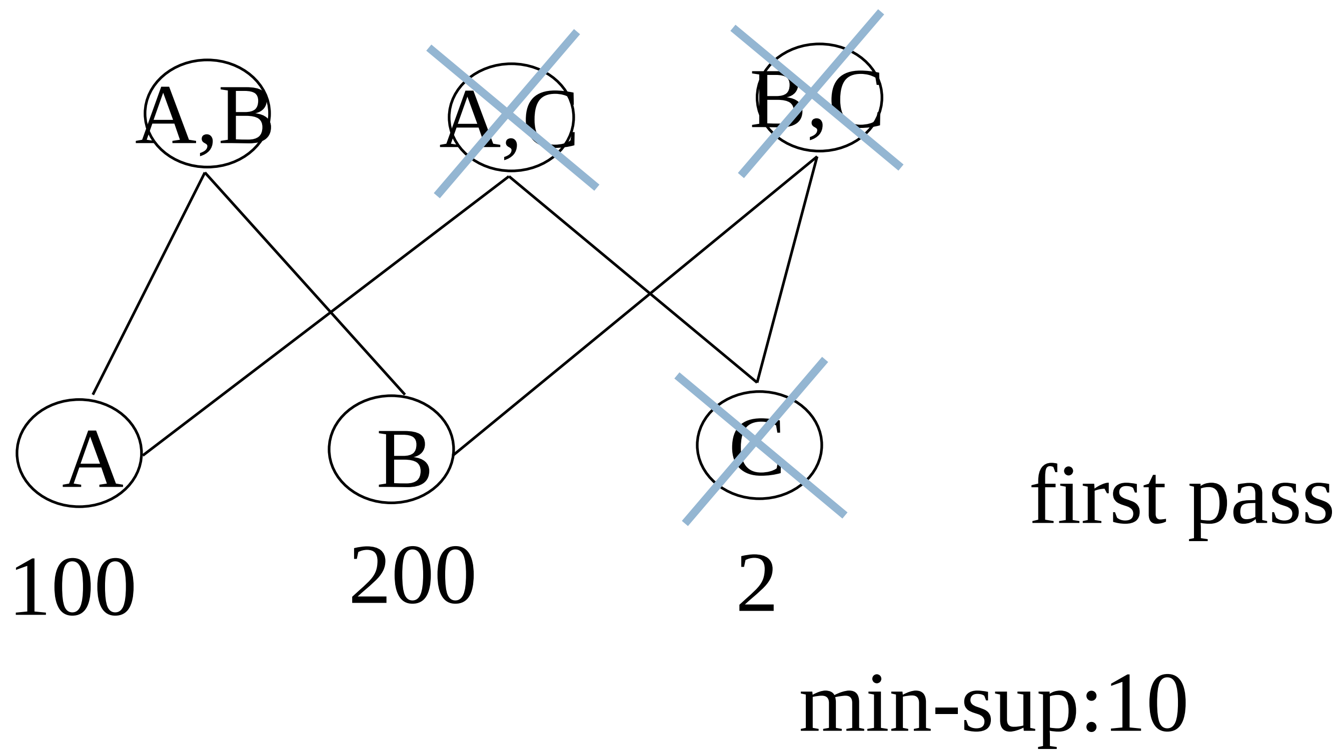
- Scan database once; keep 2^M-1 counters
  - One counter for each of {A}, {B}, {C}, …., {A,B}, {A,C}, {A,D}, … {B,C}, {B,D}, {B,E},… {A,B,C}, …
- Drawback?
  - For M=1000 products, 2^{1000} is prohibitive...
  - E.g. 16GB RAM (=2^{34} bits) stores 2^{29} counters using 32=2^5 bit integers
- Improvement?
  - Scan the db M times, looking for 1-, 2-, etc itemsets

Assume three products/items A,B and C (M=3)

A	B	C	
100	200	2	first pass

min-sup:10

Move on



Anti-monotonicity property

- If an itemset fails to be frequent, so will every superset of it
  - hence all supersets can be pruned
- A subset of a frequent itemset must also be a frequent itemset
  - i.e., if $\{AB\}$ is a frequent itemset, both $\{A\}$ and $\{B\}$ should be a frequent itemset
- Sketch of the (famous!) 'a-priori' algorithm
  - Let $L(i-1)$ be the set of large (=frequent) itemsets with $i-1$ elements
  - Let $C(i)$ be the set of candidate itemsets (of size i)

Ο αλγόριθμος είναι εκτός ύλης

The A-priori Algorithm

Compute L(1), by scanning the database.

repeat, for i=2,3...,

- **'join'** L(i-1) with itself, to generate C(i)
  - two itemset in L(i-1) can be joined, if they agree on their first *i-2* elements (i.e. all but the last)
- **prune** the itemsets of C(i) (how?)
- **scan** the db, finding the counts of the C(i) itemsets – those that reach or exceed threshold are placed in L(i)

unless L(i) is empty, repeat the loop

Ο αλγόριθμος είναι εκτός ύλης

notation for itemset {a,c,e}

notation for itemset {b,c,d}

An Example

- L_3={*abc, abd, acd, ace, bcd*}
- Self-joining: $L_3 \bowtie L_3$ *to obtain candidates for* C_4
  - *abcd is produced* from *ab*c and *ab*d
  - *acde* is *produced* from *ac*d and *ac*e
- Pruning:
  - *acde* is removed because *ade* is not in L_3
- C_4={*abcd*}

Ο αλγόριθμος είναι εκτός ύλης

Note on Self-joining $L_1 \bowtie L_1$

- *The result is essentially a Cartesian Product (**x**)*
- *For example:*
  - $L_1=\{a, b, c, d, e\}$
  - $C_2 = L_1 \times L_1 = \{ab, ac, ad, ae, bc, bd, be, cd, ce, de\}$
- *No pruning possible (why?)*

Ο αλγόριθμος είναι εκτός ύλης

Example 2

Min Support = 2 (50%)

Database D

TID	Items
100	A,C,D
200	B,C,E
300	A,B,C,E
400	B E

Scan D →

C_1

itemset	sup.
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

→

L_1

itemset	sup.
{A}	2
{B}	3
{C}	3
{E}	3

C_2

itemset
{A,B}
{A,C}
{A,E}
{B,C}
{B,E}
{C,E}

Scan D ←

C_2

itemset	sup
{A,B}	1
{A,C}	2
{A,E}	1
{B,C}	2
{B,E}	3
{C,E}	2

←

L_2

itemset	sup
{A,C}	2
{B,C}	2
{B,E}	3
{C,E}	2

C_3

itemset
{B,C,E}

Scan D →

L_3

itemset	sup
{B,C,E}	2

Εντός ύλης!

Generate Rules

Min Support = 2 (50%)

L_1

itemset	sup.
{A}	2
{B}	3
{C}	3
{E}	3

B→C [Support = ?, Confidence = ?]

L_2

itemset	sup
{A,C}	2
{B,C}	2
{B,E}	3
{C,E}	2

L_3

itemset	sup
{B,C,E}	2

Generate Rules

Min Support = 2 (50%)

L_1

itemset	sup.
{A}	2
{B}	3
{C}	3
{E}	3

B→C [Support = 2/4, Confidence = ?]

L_2

itemset	sup
{A,C}	2
{B,C}	$\boxed{2}$
{B,E}	3
{C,E}	2

L_3

itemset	sup
{B,C,E}	2

Generate Rules

Min Support = 2 (50%)

L_1

itemset	sup.
{A}	2
{B}	3
{C}	3
{E}	3

B→C [Support = 2/4, Confidence = (2/4)/(3/4)=2/3]

Recall that Confidence = P[C|B] = P[B,C]/P[B]

L_2

itemset	sup
{A,C}	2
{B,C}	2
{B,E}	3
{C,E}	2

L_3

itemset	sup
{B,C,E}	2

From Itemsets to Association Rules

- Itemset {B,C,E} is frequent (support=50%)
- Consider rule B,C $\rightarrow$E
  - Support(B,C $\rightarrow$E) = P[B,C,E] = 50%
  - Confidence(B,C $\rightarrow$E) = P[B,C,E]/P[B,C]=2/2=100%
- Thus : B,C$\rightarrow$E [50%,100%]
- More rules?
- Also look at L_2

MIN-SUPPORT = 50%
MIN-CONFIDENCE=90%

Exercise 3

- Frequent Itemsets
  - {A,B,C} support = 50%, {A,B} support = 50%, {A,C} support=80%, {B,C} support = 80%, {A}=90%, {B}=90%, {C}=90%
- A,B→C [50%, 100%] (OK, exceeds thresholds)
- Reject the following (confidence < 90%)
  - A,C→B [50%, 62.5%]
  - B,C→A [50%, 62.5%]
  - A→B [50% , 55.5%]
    - (also B→A, A→C, C→A, B→C, C→B)

Criticism on high conf/support

- Example 1: (Aggarwal & Yu, PODS98)
  - Among 5000 students
    - 3000 play basketball
    - 3750 eat cereal
    - 2000 both play basket ball and eat cereal
- *Compare the following two rules*
  - *play basketball* $\Rightarrow$ *eat cereal* [40%, 66.7]
  - *play basketball* $\Rightarrow$ *not eat cereal* [20%, 33.3%]

2000/5000

2000/3000

	basketball	not basketball	sum(row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum(col.)	3000	2000	5000

Strong Rules Are Not Necessarily Interesting

- *play basketball* $\Rightarrow$ *eat cereal* [40%, 66.7%] is misleading because the overall percentage of students eating cereal is 75% which is higher than 66.7%.
- *play basketball* $\Rightarrow$ *not eat cereal* [20%, 33.3%] is more interesting, although with lower support and confidence

	basketball	not basketball	sum(row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum(col.)	3000	2000	5000

Criticism to Support and Confidence (Cont.)

- Example 2:
  - X and Y: positively correlated,
  - X and Z, negatively related
  - support and confidence of X→Z dominates
- We need a measure of dependent or correlated events

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

Rule	Support	Confidence
X=>Y	25%	50%
X=>Z	37,50%	75%

Lift of an Association Rule

- Lift(X→Y) = P(X and Y)/(P(X)*P(Y))
  - P(X and Y) = support observed in the dataset
  - P(X)*P(Y) = expected support if X and Y were independent
  - Lift(X→Y)>1 suggests that X&Y appear together more often that expected. Thus, the occurrence of X has a positive effect on the occurrence of Y

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

$$\mathrm{Lift}(X \to Y) = \frac{\frac{2}{8}}{\frac{4}{8} * \frac{2}{8}} = 2$$

observed=25%

expected=12.5%

  - In some cases rare items may produce rules with very high values of lift

Lift of an Association Rule

- Lift(X→Y) = P(X and Y)/(P(X)*P(Y))
  - P(X and Y) = support observed in the dataset
  - P(X)*P(Y) = expected support if X and Y were independent
  - Lift(X→Y)>1 suggests that X&Y appear together more often that expected. Thus, the occurrence of X has a positive effect on the occurrence of Y

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

observed=37.5%

$$\mathrm{Lift}(X \rightarrow Z) = \frac{\frac{3}{8}}{\frac{4}{8} * \frac{7}{8}} = 0.86$$

expected=43.75%

  - In some cases rare items may produce rules with very high values of lift

Lift of an Association Rule

- Lift(X$\rightarrow$Y) = P(X and Y)/(P(X)*P(Y))
  - P(X and Y) = support observed in the dataset
  - P(X)*P(Y) = expected support if X and Y were independent
  - Lift(X$\rightarrow$Y)>1 suggests that X&Y appear together more often that expected. Thus, the occurrence of X has a positive effect on the occurrence of Y

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

Itemset	Support	Lift
{X,Y}	25%	2.00
{X,Z}	37.5%	0.86
{Y,Z}	12.5%	0.57

- In some cases rare items may produce rules with very high values of lift

Rules with multiple items in the antecedent

- Lift(**A**→B) = P(**A** and B)/(P(**A**)*P(B))
  - **A** in this formula can be a **set** of items
- Example:

Assume rule X,Y→Z

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

$$\text{Lift}(X, Y \rightarrow Z) = \frac{\frac{1}{8}}{\frac{2}{8} * \frac{7}{8}} = 0.57$$

Back to the student's survey

- *play basketball* $\Rightarrow$ *eat cereal* [40%, 66.7%]
  - Lift = (2000/5000)/((3000/5000)*(3750/5000)) = 0.89 < 1
- *play basketball* $\Rightarrow$ *not eat cereal* [20%, 33.3%]
  - Lift = (1000/5000)/((3000/5000)*(1250/5000)) = 1.33 > 1

	basketball	not basketball	sum(row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum(col.)	3000	2000	5000