Asszociációs-szabályok, 3. rész

Csima Judit

BME, VIK,
Számítástudományi és Információelméleti Tanszék

2018. május 10.

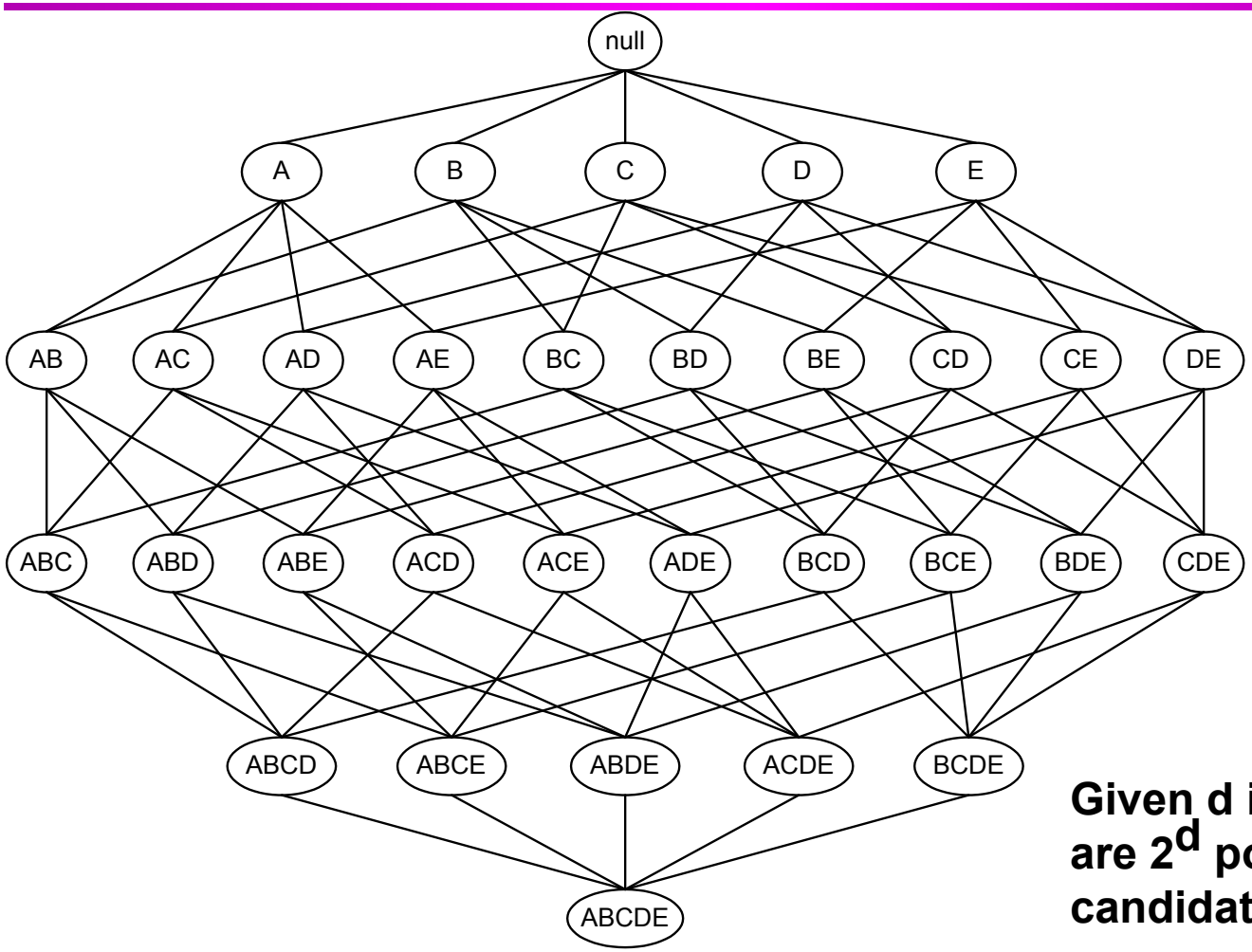
Eddig mi volt?

- Apriori-algoval gyakori elemhalmazok generálása
- a zárt gyakoriak és a hozzájuk tartozó tárolt σ értékekből az összes gyakori és ezek σ-jának meghatározása
- gyakoriak elemhalmazokból a nagy megbízhatóságú szabályok előállítása

Most mi lesz?

- Apriori algo helyett más módszerek a gyakori elemhalmazok megtalálására:
  - általános stratégiák az elemhalmazok hálójának bejárására
  - Eclat algo

Frequent Itemset Generation



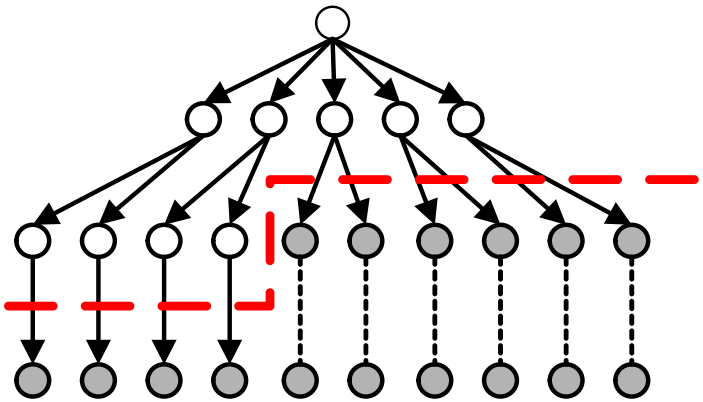
Given d items, there are 2^d possible candidate itemsets

Általános stratégiák a háló bejárására

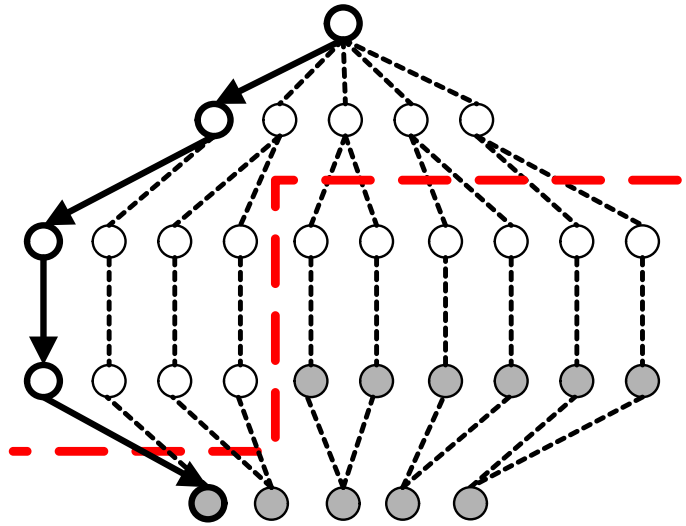
- az Apriori algo lényegében egy szélességi bejárást valósít meg
- más stratégiák:
  - mélységi bejárás
  - ekvivalencia-osztályok szerinti bejárás
- mindegyik esetben alkalmazzuk az Apriori-elvet: ha egy EH nem gyakori, akkor egyetlen olyan halmaz sem gyakori, aki őt tartalmazza vagy (ami ugyanez): ha egy elemhalmaz gyakori, akkor minden része is az

Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice
  - Breadth-first vs Depth-first



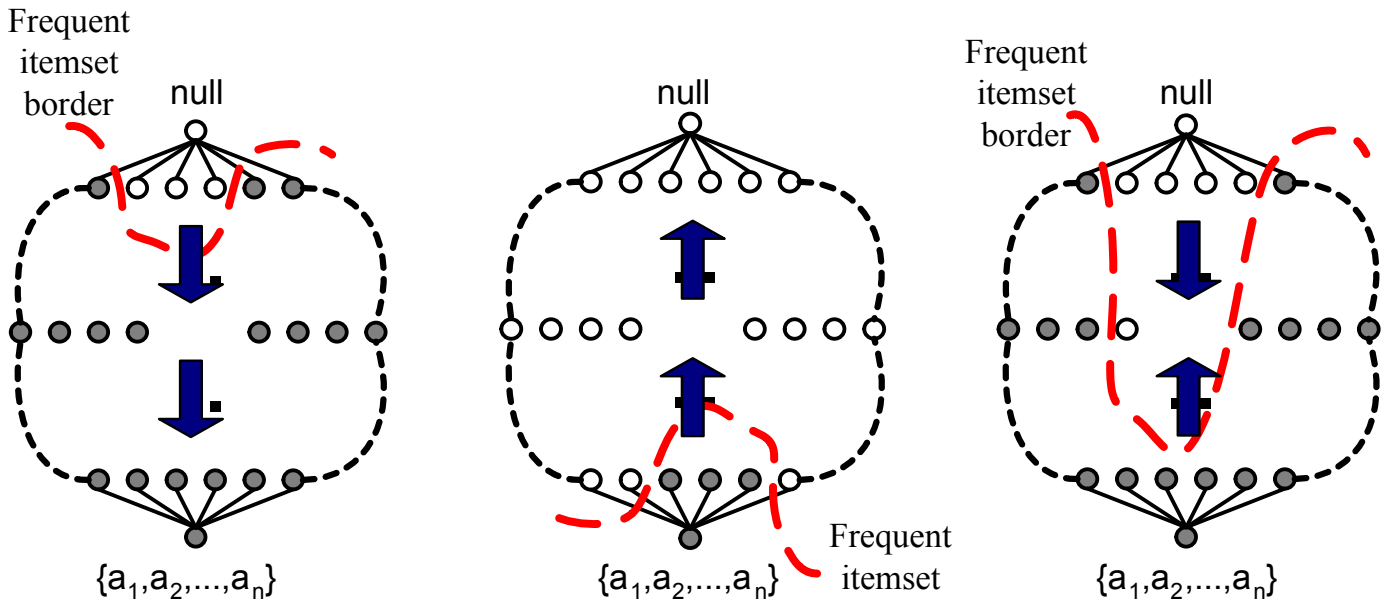
(a) Breadth first



(b) Depth first

Alternative Methods for Frequent Itemset Generation

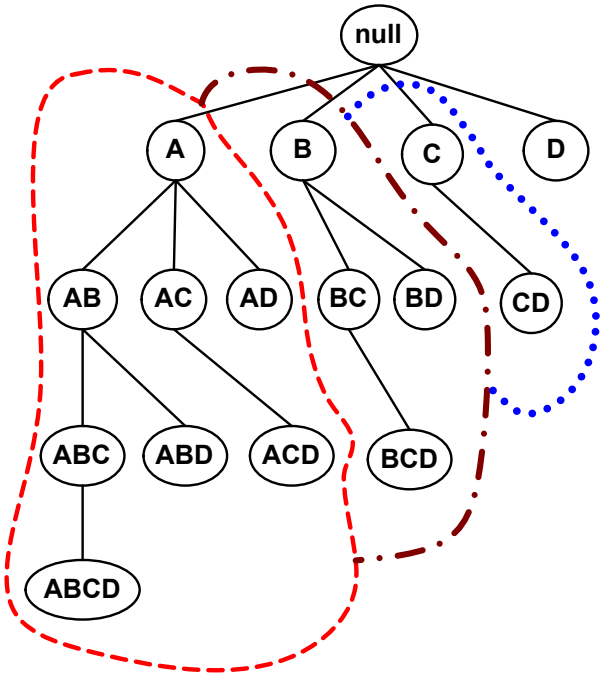
- Traversal of Itemset Lattice
  - General-to-specific vs Specific-to-general



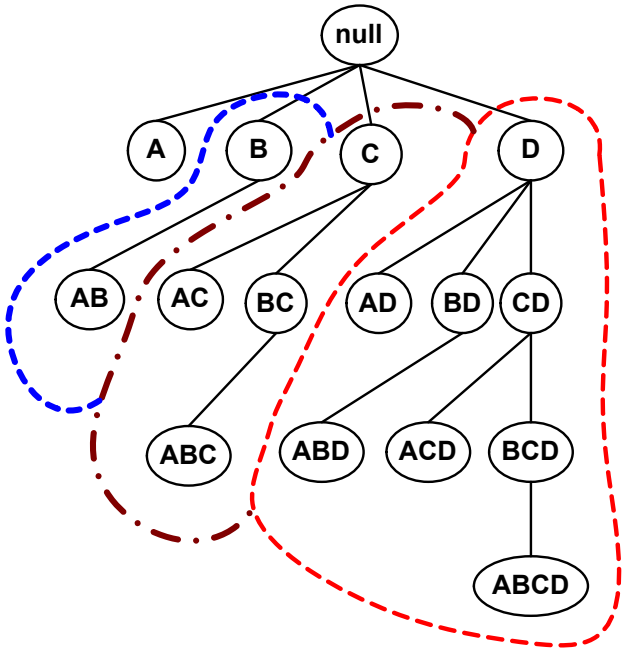
(a) General-to-specific (b) Specific-to-general (c) Bidirectional

Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice
  - Equivalent Classes



(a) Prefix tree



(b) Suffix tree

ECLAT algo

- más szisztéma
- nem azt írjuk fel, hogy melyik tranzakciókban mik az elemek, hanem azt, hogy írjuk fel az egyes elemekről, hogy melyik tranzakciókban vannak benne
- ezt vertikális felírásnak is nevezik

ECLAT

- For each item, store a list of transaction ids (tids)

Horizontal Data Layout

TID	Items
1	A,B,E
2	B,C,D
3	C,E
4	A,C,D
5	A,B,C,D
6	A,E
7	A,B
8	A,B,C
9	A,C,D
10	B

Vertical Data Layout

A	B	C	D	E
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

TID-list

ECLAT algo

- DFS-sel járjuk be az elemhalmazok hálóját
- a példában legyen a gyakorisági-küszöb 2
- ekkor E gyakori
- nézzük meg E gyerekeit: DE, CE, BE, AE gyakoriságai mik?
- pl. DE gyakorisága D és E oszlopának metszetének magassága
- hasonlóan kapható a többi kételemű gyakorisága is

Továbblépés DFS-sel

- amelyik elemhalmazról éppen kiderült, hogy gyakori, arról tudom az őt tartalmazó tranzakciók halmazát
- az egy elemű bővítések gyakorisága ezen oszlop és a bővítő elem oszlopának metszetéből számolható

ECLAT összefoglalás

- nem gyakori egy-eleműek kidobálása
- vertikális felírás elkészítése
- DFS a fenti módon, a hálót reprezentáló gráfban az éllistában a csúcsok gyakoriság szerint csökkenően (ez gyorsítja a nem gyakoriak felismerését)
- bővülő elemhalmazok gyakorisága oszlopmetszet alapján

Milyen szabályokat akarok?

- eddig: supp és conf legyen magas
- ezekhez min_sup és min_conf küszöbök
- ezek beállítása nehéz
  - ha magasak, akkor esetleg érdekes szabályok is kiesnek
  - ha alacsonyak, akkor túl sok szabály marad bent, nehéz válogatni a tényleg jókat

Érdekes szabályok keresése

- a sok szabály közül, amire supp és conf elég nagy kiválogatni azokat, amik tényleg érdekesek:
  - váratlanok
  - hasznot hozhatnak
- ezek (mechanikus algoval) megfoghatatlan fogalmak
- megoldások:
  - valami ember válogassa ki az előszűrt szabályokból az érdekeseket (ez nem járható út igazán)
  - valami szakértő előszűri, hogy milyen szabályokat keresünk: pl. A és B termékcsoport között van-e valami asszociációs összefüggés)
- supp és conf-on kívül valami más, ami méri valahogyan az érdekességet

Computing Interestingness Measure

- Given a rule $X \rightarrow Y$, information needed to compute rule interestingness can be obtained from a contingency table

Contingency table for $X \rightarrow Y$

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	$\|T\|$

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

Used to define various measures

- support, confidence, lift, Gini, J-measure, etc.

Drawback of Confidence

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence= P(Coffee|Tea) = 0.75

but P(Coffee) = 0.9

⇒ Although confidence is high, rule is misleading

⇒ P(Coffee|$\overline{\text{Tea}}$) = 0.9375

Lift-mutató, motiváció

- az előző fólia mutatja, hogy a conf és supp nem elég
- lehet, hogy egy elég jó támogatottságú, nagyon magas megbízhatóságú szabály teljesen butaság
- próbáljuk valahogy kizárni az előző fólián látható jelenséget
- hasonlítsuk össze az $X \to Y$ szabály conf-ját a Y relatív gyakoriságával (gyakoribb-e X mellett Y, mint általában?)

Lift-mutató

- $Lift(X \to Y) = \frac{conf(X \to Y)}{\frac{\sigma(Y)}{n}}$, ahol n a tranzakciók száma
- ez uaz, mint $\frac{\sigma(X \cup Y)}{\sigma(X)} \cdot \frac{n}{\sigma(Y)} = \frac{supp(X \cup Y)}{supp(X) \cdot supp(Y)}$
- ez igazából X és Y előfordulásának függetlenségét méri
- ha $Lift(X \to Y) = 1$ az azt jelenti, hogy függetlenek
- ha $Lift(X \to Y) > 1$ az azt jelenti, hogy Y gyakoribb X mellett, mint általában, ez érdekel minket

Mindenféle mérőszámok

- persze Lift sem mindenható, simán lehet olyan szabály, amire supp, conf és Lift is jó, de mégis butaság
- sok más mérőszám szabályok jóságára (következő fólia, de csak illusztráció!)
- általában sup, conf és vmi Lift-szerű, függetlenséget mérő mérték

There are lots of measures proposed in the literature

Some measures are good for certain applications, but not for others

What criteria should we use to determine whether a measure is good or bad?

What about Apriori-style support based pruning? How does it affect these measures?

#	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)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left. P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
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)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
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)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$