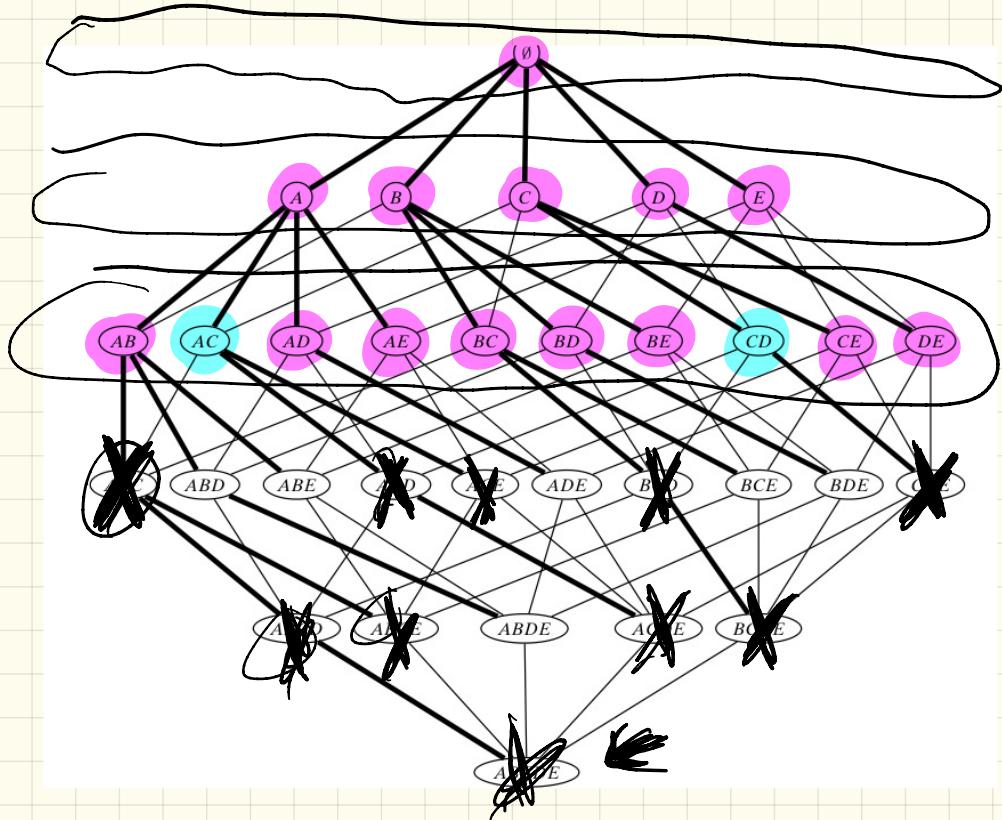


Apriori: Algo & Rule Mining

Obs: Let X, Y be itemsets s.t $X \subseteq Y$
 $\text{sup}(X) \geq \text{sup}(Y)$

① X is frequent $\Rightarrow Y \subseteq X \Rightarrow Y$ frequent

② X is not frequent $\Rightarrow \forall Y \supseteq X \quad Y$ is not frequent



Algorithm 8.2: Algorithm APRIORI

APRIORI (D, I, minsup):

- 1 $\mathcal{F} \leftarrow \emptyset$
- 2 $\mathcal{C}^{(1)} \leftarrow \{\emptyset\}$ // Initial prefix tree with single items
- 3 **foreach** $i \in I$ **do** Add i as child of \emptyset in $\mathcal{C}^{(1)}$ with $\text{sup}(i) \leftarrow 0$
- 4 $k \leftarrow 1$ // k denotes the level
- 5 **while** $\mathcal{C}^{(k)} \neq \emptyset$ **do**
- 6 **COMPUTESUPPORT** ($\mathcal{C}^{(k)}, D$)
- 7 **foreach** leaf $X \in \mathcal{C}^{(k)}$ **do**
- 8 **if** $\text{sup}(X) \geq \text{minsup}$ **then** $\mathcal{F} \leftarrow \mathcal{F} \cup \{(X, \text{sup}(X))\}$
- 9 **else** remove X from $\mathcal{C}^{(k)}$
- 10 $\mathcal{C}^{(k+1)} \leftarrow \text{EXTENDPREFIXTREE}(\mathcal{C}^{(k)})$
- 11 $k \leftarrow k + 1$
- 12 **return** $\mathcal{F}^{(k)}$

COMPUTESUPPORT ($C^{(k)}, D$):

- 13 **foreach** $(t, i(t)) \in D$ **do**
- 14 **foreach** k -subset $X \subseteq i(t)$ **do**
- 15 **if** $X \in C^{(k)}$ **then** $\text{sup}(X) \leftarrow \text{sup}(X) + 1$

EXTENDPREFIXTREE ($C^{(k)}$):

- 16 **foreach** leaf $X_a \in C^{(k)}$ **do**
- 17 **foreach** leaf $X_b \in \text{SIBLING}(X_a)$, such that $b > a$ **do**
- 18 $X_{ab} \leftarrow X_a \cup X_b$
 // prune candidate if there are any infrequent subsets
- 19 **if** $X_j \in C^{(k)}$, for all $X_j \subset X_{ab}$, such that $|X_j| = |X_{ab}| - 1$ **then**
- 20 Add X_{ab} as child of X_a with $\text{sup}(X_{ab}) \leftarrow 0$
- 21 **if no extensions from X_a then**
- 22 remove X_a , and all ancestors of X_a with no extensions, from $C^{(k)}$
- 23 **return** $C^{(k)}$

Complexity:

Worst case for work: $O(|I||D| \cdot 2^{|I|})$

I/O cost is $O(|I|)$

1 DB scan per level

What is the max # of levels?

Ans: $|I|$

\Rightarrow total # of DB scans is
proportate to $|I|$

$\Rightarrow O(|I|)$

Additional improvements:

- Eclat - speeding up support comp
- dEclat - further speed up support comp
- FP growth algo -
FPtree - compressed version of D

Generating Association rules

problem² rule mining

Given frequent itemset F

minimum confidence minconf

([0, 1])

Find all frequent and strong rules

$X \in F$
 $Y \subset X$ w/ $Y \neq \emptyset$

$Y \rightarrow X \setminus Y$ ← this rule is frequent

$$\text{sup}(Y \rightarrow X \setminus Y) = \text{sup}(Y \cup X \setminus Y) = \text{sup}(X) \geq \min \text{sup}$$

Subproblem

2a. check conf

2b. generate rules (algorithmically)

Check conf

if $\text{conf}(X \rightarrow Y) < c$

Let $W \subset X \quad Z = X \cup Y$

$$\frac{\text{sup}(Z)}{\text{sup}(W)} \leq \frac{\text{sup}(Z)}{\text{sup}(X)} \quad (\text{sup}(W) \geq \text{sup}(X))$$

$$\text{conf}(W \rightarrow Z \setminus W) = \text{conf}(X \rightarrow Z \setminus X) = \text{conf}(X \rightarrow Y) < c$$

\Rightarrow avoid checking subsets of X for conf associations

Algorithm 8.6: Algorithm ASSOCIATIONRULES

ASSOCIATIONRULES (\mathcal{F} , minconf):

```
1 foreach  $Z \in \mathcal{F}$ , such that  $|Z| \geq 2$  do
2    $\mathcal{A} \leftarrow \{X \mid X \subset Z, X \neq \emptyset\}$ 
3   while  $\mathcal{A} \neq \emptyset$  do
4      $X \leftarrow$  maximal element in  $\mathcal{A}$ 
5      $\mathcal{A} \leftarrow \mathcal{A} \setminus X$  // remove  $X$  from  $\mathcal{A}$ 
6      $c \leftarrow sup(Z)/sup(X)$ 
7     if  $c \geq minconf$  then
8       print  $X \rightarrow Y, sup(Z), c$ 
9     else
10       $\mathcal{A} \leftarrow \mathcal{A} \setminus \{W \mid W \subset X\}$  // remove subsets of  $X$  from  $\mathcal{A}$ 
```

$$\Sigma_{ABDE} = Z$$

make a set A of all antecedents (LHS of rule)

$$\mathcal{A} = \{ABD(3), ABE(4), ADE(3), BDE(3), \cancel{AB}(4), AD(3), \cancel{AE}(4), \\ BD(4), \cancel{BE}(3), DE(3), \cancel{A}(4), \cancel{B}(6), D(4), \cancel{E}(3)\}$$

for each antecedent $X \in A$
form rule $X \rightarrow Z \setminus X$
compute conf of rule

above min conf output

else below cull all subsets of X

$$e.g., \text{conf}(ABE \rightarrow D) = \frac{3}{7} < 1$$

remove:

$$A, B, E$$

$$AB, AE, BE$$