

School of Information Technology and Electrical Engineering
INFS4203/7203 – Data Mining

Tutorial 2 Mining Association Rules

Question 1.

A database has four transactions. Let $\text{min_sup} = 60\%$ and $\text{min_conf} = 80\%$.

TID	Date	Items bought
120	5/10/02	{K, A, D, B}
230	5/10/02	{D, A, C, E, B}
340	15/10/02	{C, A, B, E}
450	20/10/02	{B, A, D}

- (a) Find all frequent itemsets using Apriori and FP-Growth algorithm respectively (show all intermediate results).
- (b) List all of the strong association rules (with support s and confidence c) matching the following meta rule, where X is a variable representing customers, and item_i denotes variables representing items (e.g., “A”, “B”, etc.):

$$\forall X \in \text{transaction}, \mathbf{buys}(X, \text{item}_1) \wedge \mathbf{buys}(X, \text{item}_2) \Rightarrow \mathbf{buys}(X, \text{item}_3) [s, c]$$

ANSWER 1 - (a)

Apriori Algorithm

1.a.i Apriori

Given a database:

TId	Date	Items Bought
100	15/10/01	{K, A, D, B}
200	15/10/01	{D, A, C, E, B}
300	19/10/01	{C, A, B, E}
400	22/10/01	{B, A, D}

When using Apriori to find large itemsets, we are not interested in unrelated attributes:

TId	Items Bought
100	{A, B, D, K}
200	{A, B, C, D, E}
300	{A, B, C, E}
400	{A, B, D}

We now compute the support for each of the candidate 1-itemsets:

C_1	
Itemset	Support
{A}	4
{B}	4
{C}	2
{D}	3
{E}	2
{K}	1

Then select the itemsets with sufficient support:

L_1	
Itemset	Support
{A}	4
{B}	4
{D}	3

And now compute a self-join to produce the candidate 2-itemsets:

C_2	
Itemset	Support
{A, B}	4
{A, D}	3
{B, D}	3

Then, select the itemsets with sufficient support:

L_2	
Itemset	Support
{A, B}	4
{A, D}	3
{B, D}	3

And compute a self-join to produce the candidate 3-itemsets:

C_3	
Itemset	Support
{A, B, D}	3

Then we select the itemsets with sufficient support:

L_3	
Itemset	Support
{A, B, D}	3

If we compute a self-join, we have no candidate 4-itemsets.

Thus, all itemsets with minimum support (as per Apriori) are:

$\{\{A\}, \{B\}, \{D\}, \{A, B\}, \{A, D\}, \{B, D\}, \{A, B, D\}\}$

Answer for the FP Growth Algorithm:

Given a database:

TId	Date	Items Bought
100	15/10/01	{K, A, D, B}
200	15/10/01	{D, A, C, E, B}
300	19/10/01	{C, A, B, E}
400	22/10/01	{B, A, D}

When using FP-Growth to find large itemsets, we are not interested in un-related attributes:

TId	Items Bought
100	{A, B, D, K}
200	{A, B, C, D, E}
300	{A, B, C, E}
400	{A, B, D}

We begin the FP-Growth algorithm by scanning the database and sorting each of the possible 1-itemsets by support, to produce a structure called a

header table:

Header Table		
Item Id	Support	Node Link
A	4	•
B	4	•
D	3	•
C	2	•
E	2	•
K	1	•

We are only interested in frequent items, so we discard those items not needed, giving header table:

Header Table		
Item Id	Support	Node Link
A	4	•
B	4	•
D	3	•

and database:

TId	(Sorted) Frequent Items
100	{A, B, D}
200	{A, B, D}
300	{A, B}
400	{A, B, D}

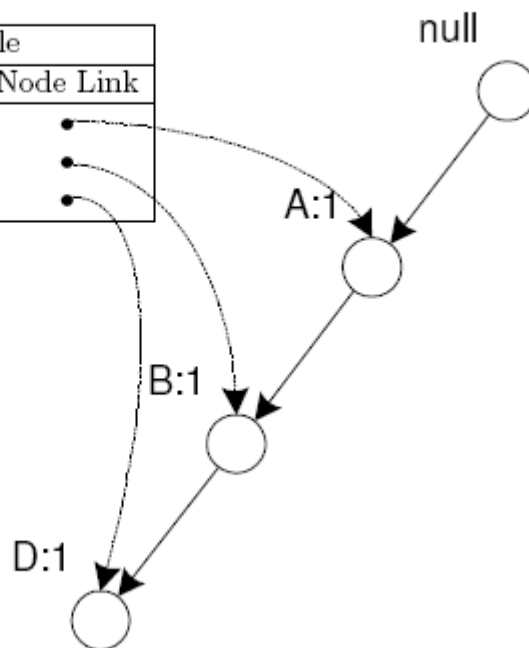
We now scan the database a second time, constructing the FP-tree as we go. Initially, we have the empty tree:

Header Table		
Item Id	Support	Node Link
A	4	•
B	4	•
D	3	•

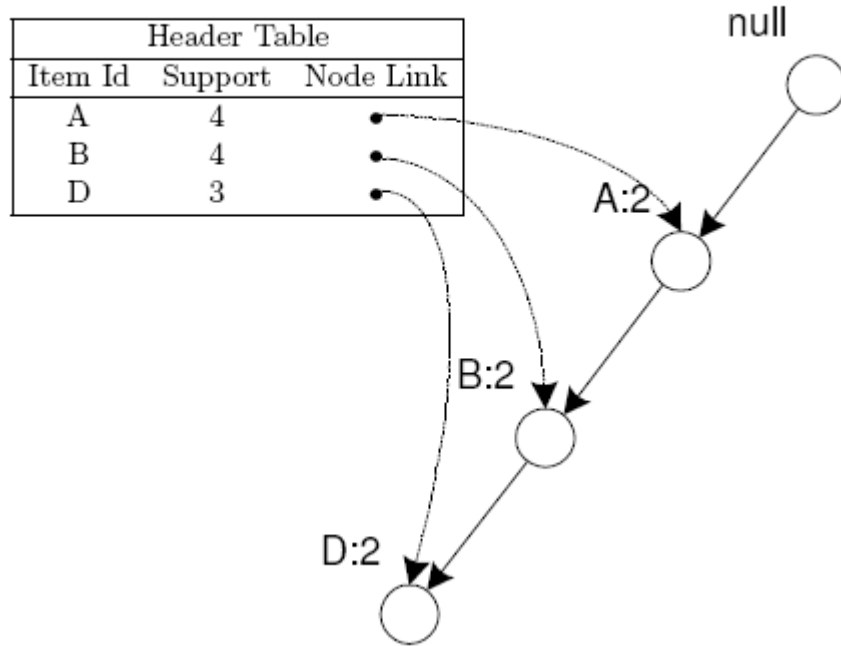


After transaction 100:

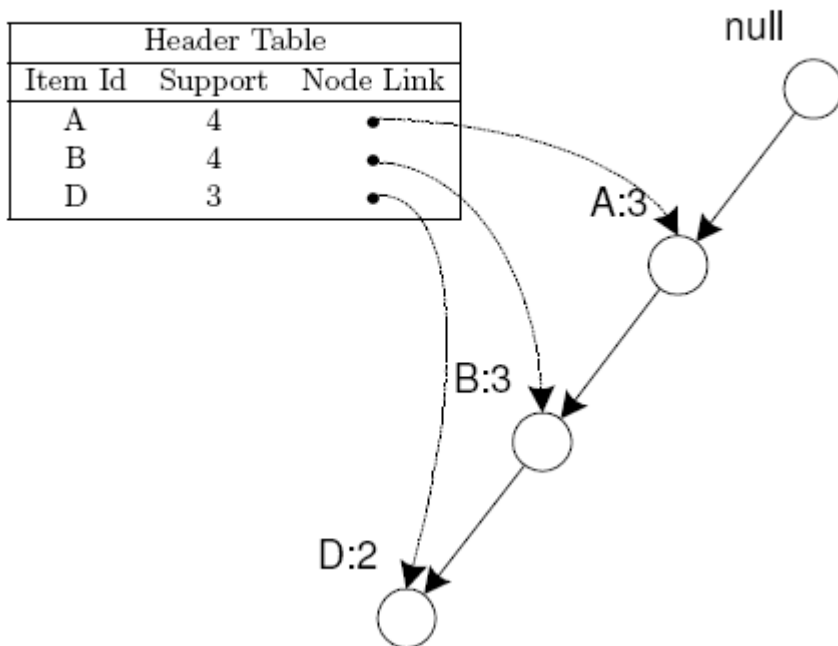
Header Table		
Item Id	Support	Node Link
A	4	•
B	4	•
D	3	•



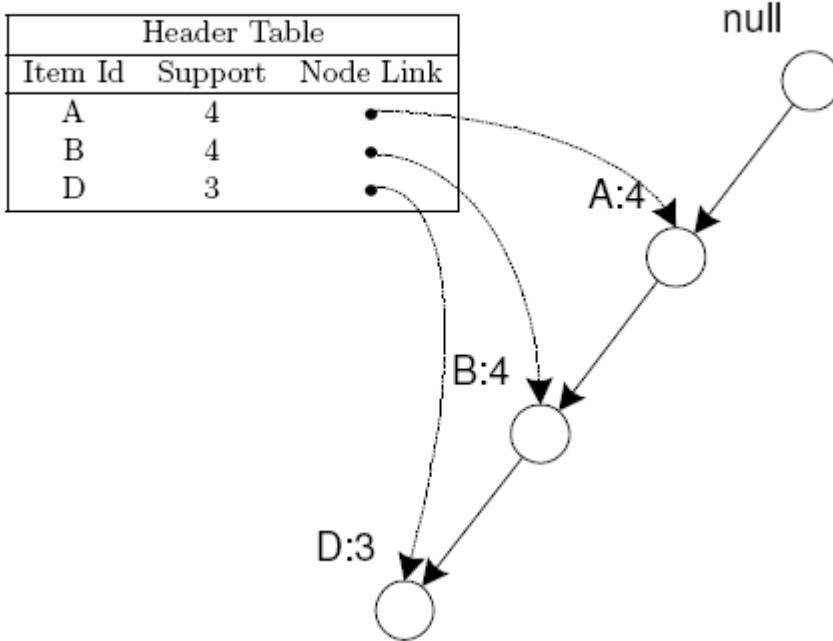
After transaction 200:



After transaction 300:



After transaction 400:



We now traverse the tree for frequent items, that contain each of the elements of the header table.

Because this is an FP-tree with a single branch, we can mine all large itemsets by simply outputting all the combinations of items along the path. All such large itemsets are as follows:

$\{\{A\}, \{B\}, \{D\}, \{A, B\}, \{A, D\}, \{B, D\}, \{A, B, D\}\}$

Alternately, we could have run the algorithm naïvely, giving the following results of the depth first recursive process:

Item	Conditional Pattern Base	Conditional FP-tree	Generated itemsets
D	$\{\langle A:3, B:3, D:3 \rangle\}$	$\{\langle A:3, B:3, D:3 \rangle\} D$	$\{\{D\}\}$
B D	$\{\langle A:3, B:3 \rangle\}$	$\{\langle A:3 \rangle\} BD$	$\{\{B, D\}\}$
A BD	$\{\langle A:3 \rangle\}$	$\{\} ABD$	$\{\{A, B, D\}\}$
A D	$\{\langle A:3 \rangle\}$	$\{\} AD$	$\{\{A, D\}\}$
B	$\{\langle A:4, B:4 \rangle\}$	$\{\langle A:4 \rangle\} B$	$\{\{B\}\}$
A B	$\{\langle A:4 \rangle\}$	$\{\} AB$	$\{\{A, B\}\}$
A	$\{\langle A:4 \rangle\}$	$\{\} A$	$\{\{A\}\}$

And thus we do indeed conclude the same set of frequent itemsets:

$\{\{A\}, \{B\}, \{D\}, \{A, B\}, \{A, D\}, \{B, D\}, \{A, B, D\}\}$

ANSWER 1 - (b)

There are two resulting rules.

Given the following meta-rule:

$$\forall X \in \text{transaction} \bullet \text{buys}(X, \text{item}_1) \wedge \text{buys}(X, \text{item}_2) \Rightarrow \text{buys}(X, \text{item}_3)[s, c]$$

Assuming item_1 , item_2 and item_3 are distinct, we can find bindings for these with support s and confidence c , by testing combinations of S for each large itemset, \mathcal{L} (such that $S \subset \mathcal{L}$) where $S \Rightarrow \mathcal{L} - S$ has the minimum confidence.

Clearly, since the meta-rule contains three variables, only subsets of large itemsets \mathcal{L} where $|\mathcal{L}| = 3$, need be considered. Thus, we have the following combinations:

	Bindings	Rule Support	Guard Support	Rule Confidence
1	$\text{item}_1 = A, \text{item}_2 = B, \text{item}_3 = D$	3	4	0.75
2	$\text{item}_1 = A, \text{item}_2 = D, \text{item}_3 = B$	3	3	1.00
3	$\text{item}_1 = B, \text{item}_2 = D, \text{item}_3 = A$	3	3	1.00
4	$\text{item}_1 = B, \text{item}_2 = A, \text{item}_3 = D$	3	4	0.75
5	$\text{item}_1 = D, \text{item}_2 = A, \text{item}_3 = B$	3	3	1.00
6	$\text{item}_1 = D, \text{item}_2 = B, \text{item}_3 = A$	3	3	1.00

Note that because the conjunction in the meta-rule is commutative, rules 4-6 above are equivalent to rules 1-3 respectively, and so we have the following two rules that satisfy the minimum support and confidence:

$$\forall X \in \text{transaction} \bullet \text{buys}(X, A) \wedge \text{buys}(X, D) \Rightarrow \text{buys}(X, B)[0.75, 1.00]$$

$$\forall X \in \text{transaction} \bullet \text{buys}(X, B) \wedge \text{buys}(X, D) \Rightarrow \text{buys}(X, A)[0.75, 1.00]$$

If we do not assume that the distinct variables in the meta-rule represent distinct objects and permit equivalent rules, we would then have the following instantiations of the meta-rule:

$\forall X \in \text{transaction} \bullet \text{buys}(X,A) \wedge \text{buys}(X,A) \Rightarrow \text{buys}(X,A)[1.00, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,A) \wedge \text{buys}(X,A) \Rightarrow \text{buys}(X,B)[1.00, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,A) \wedge \text{buys}(X,B) \Rightarrow \text{buys}(X,A)[1.00, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,A) \wedge \text{buys}(X,B) \Rightarrow \text{buys}(X,B)[1.00, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,A) \wedge \text{buys}(X,D) \Rightarrow \text{buys}(X,A)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,A) \wedge \text{buys}(X,D) \Rightarrow \text{buys}(X,B)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,A) \wedge \text{buys}(X,D) \Rightarrow \text{buys}(X,D)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,B) \wedge \text{buys}(X,A) \Rightarrow \text{buys}(X,A)[1.00, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,B) \wedge \text{buys}(X,A) \Rightarrow \text{buys}(X,B)[1.00, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,B) \wedge \text{buys}(X,B) \Rightarrow \text{buys}(X,A)[1.00, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,B) \wedge \text{buys}(X,B) \Rightarrow \text{buys}(X,B)[1.00, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,B) \wedge \text{buys}(X,D) \Rightarrow \text{buys}(X,A)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,B) \wedge \text{buys}(X,D) \Rightarrow \text{buys}(X,B)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,B) \wedge \text{buys}(X,D) \Rightarrow \text{buys}(X,D)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,D) \wedge \text{buys}(X,A) \Rightarrow \text{buys}(X,A)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,D) \wedge \text{buys}(X,A) \Rightarrow \text{buys}(X,B)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,D) \wedge \text{buys}(X,A) \Rightarrow \text{buys}(X,D)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,D) \wedge \text{buys}(X,B) \Rightarrow \text{buys}(X,A)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,D) \wedge \text{buys}(X,B) \Rightarrow \text{buys}(X,B)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,D) \wedge \text{buys}(X,B) \Rightarrow \text{buys}(X,D)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,D) \wedge \text{buys}(X,D) \Rightarrow \text{buys}(X,A)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,D) \wedge \text{buys}(X,D) \Rightarrow \text{buys}(X,B)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X,D) \wedge \text{buys}(X,D) \Rightarrow \text{buys}(X,D)[0.75, 1.00]$

But, clearly, if the user were to supply the meta-rule, he or she would have intended the variables be distinct and that logically equivalent rules be excluded, so indeed the following two rules would be the meaningful output of any association rule miner that was supplied the given meta-rule:

$\forall X \in \text{transaction} \bullet \text{buys}(X, A) \wedge \text{buys}(X, D) \Rightarrow \text{buys}(X, B)[0.75, 1.00]$
 $\forall X \in \text{transaction} \bullet \text{buys}(X, B) \wedge \text{buys}(X, D) \Rightarrow \text{buys}(X, A)[0.75, 1.00]$