

Apriori Algorithm for Association Rule Mining

CSED, TIET



Apriori Algorithm

- Apriori algorithm is developed by two Indians Rakesh Agarwal and Ramakrishnan Shrikant in 1994.

- Apriori algorithm works in following two phases:

Phase 1: Identification of Frequent Item sets.

Phase 2: Generation of Association Mining Rules.

Phase 1: Identification of Frequent Item sets

- In this phase we generate frequent item sets from transactional database that qualifies the criteria of minimum support and confidence. But, it is necessary that the items in each transaction are listed in ascending order (or sorted order).
- This phase generates frequent item sets in the following steps:
 1. Identify Candidate One Itemset C_1 - i.e. all the items present in the dataset.
 2. Identify Frequent One Itemset L_1 - i.e. all one items having frequency greater than or equal to threshold value of support.
 3. Identify Candidate 2-Itemsets C_2 - which is generated as L_1 join L_1 i.e. join of L_1 with L_1 .
 4. Identify Frequent 2-Itemsets L_2 - all 2-Itemsets having frequency greater than or equal to threshold value of support.
 5. Similarly, repeat steps 3 and 4 to generate L_3, L_4, \dots

Phase 1: Identification of Frequent Item sets

- In general, each k^{th} candidate Itemset (C_k) is produced through $L_{k-1} \text{ JOIN } L_{k-1}$. For instance, C_3 is produced through $L_2 \text{ JOIN } L_2$.
- Two item sets are joinable if their first $k-2$ items are same. So, in case of C_3 , the first item in L_2 are same. Similarly for C_4 , first two items in L_3 should be same, and so on.
- In case of C_2 , there is no such requirement because $k-2$ is 0.

C1 (Candidate 1-itemsets)

List all individual items present in the dataset.

L1 (Frequent 1-itemsets)

Keep only those items from C1 whose occurrence count \geq minimum support.

C2 (Candidate 2-itemsets)

Form all possible pairs by combining items in L1 ($L1 \times L1$).

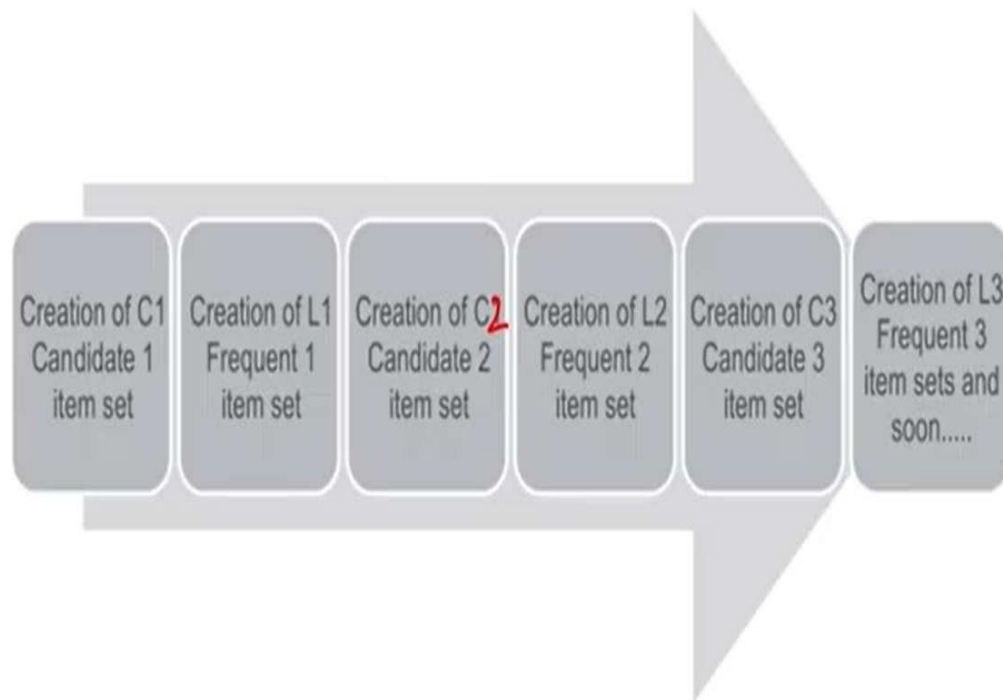
L2 (Frequent 2-itemsets)

Keep only those pairs whose occurrence count \geq minimum support.

Repeat

Use L2 to generate C3, then L3, and continue until no more frequent itemsets are found.

Final Process of Phase-1



Phase 2: Generation of Association Mining Rules

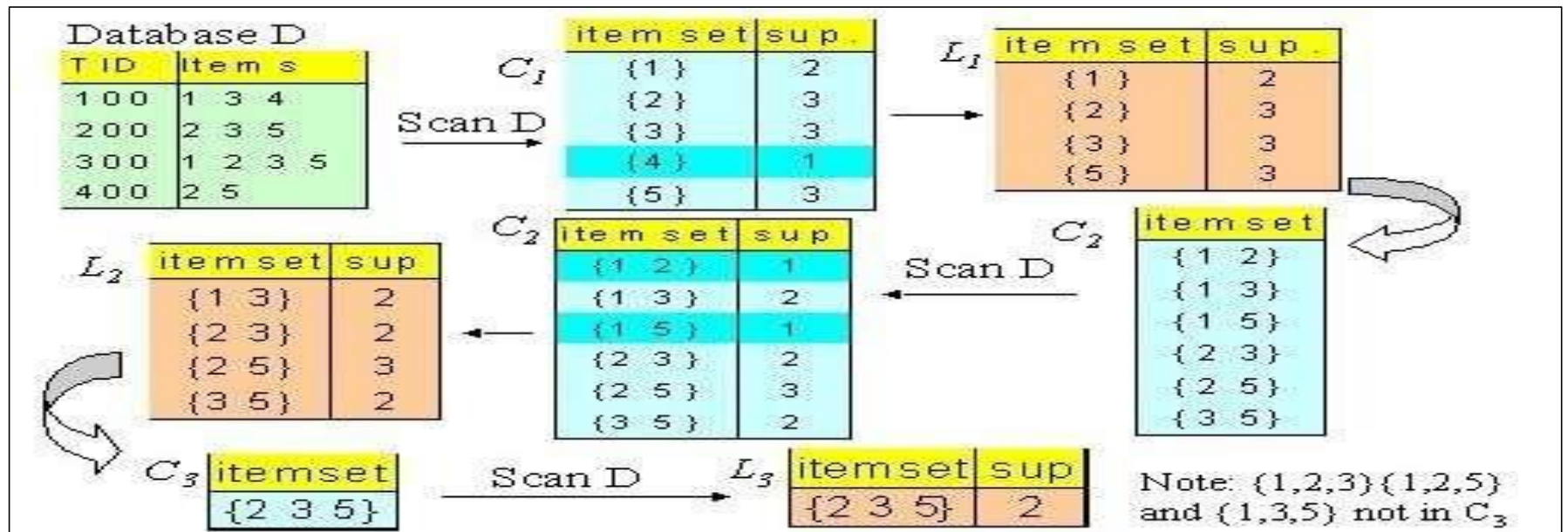
- In this phase, association mining rules are generated.
- This works in following phase:
 1. For each itemset L in the last frequent itemsets, generate all possible non null subsets S.
 2. Generate all possible rules $s \rightarrow L-s$ for $s \in S$.
 3. Compute confidence of each rule.
 4. Add all the rules in the final set that qualifies the minimum confidence criteria.
 5. For all the added rules, add rules that are implicitly generated from them and satisfy the minimum confidence threshold criteria.

Apriori Algorithm- Example 1

Find association rules with minimum support of 50% and confidence of 75%

Database D	
T ID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Example 1-Solution (Phase 1)



Example 1-Solution (Phase 2)

- The 3- frequent item sets generated in phase 1 are (2,3,5)
- Confidence for each possible rule are:

$$\text{confidence}(2 \rightarrow 3,5) = \frac{n(2,3,5)}{n(2)} = \frac{2}{3} = 66.7\%$$

$$\text{confidence}(3,5 \rightarrow 2) = \frac{n(2,3,5)}{n(3,5)} = \frac{2}{2} = 100\%$$

$$\text{confidence}(3 \rightarrow 2,5) = \frac{n(2,3,5)}{n(3)} = \frac{2}{3} = 66.7\%$$

$$\text{confidence}(2,5 \rightarrow 3) = \frac{n(2,3,5)}{n(2,5)} = \frac{2}{3} = 66.7\%$$

$$\text{confidence}(5 \rightarrow 2,3) = \frac{n(2,3,5)}{n(5)} = \frac{2}{3} = 66.7\%$$

$$\text{confidence}(2,3 \rightarrow 5) = \frac{n(2,3,5)}{n(2,3)} = \frac{2}{2} = 100\%$$

Example 1-Solution (Phase 2)

Rule	Confidence
2, 3→5	1.0
2→5	$S(2 \cap 5) / S(2) = 3/3 = 1.0$
3→5	$S(2 \cap 5) / S(3) = 3/3 = 1.0$
3, 5→2	1.0
3→2	$S(3 \cap 2) / S(3) = 2/3 = 0.67$
5→2	$S(5 \cap 2) / S(5) = 3/3 = 1.0$

Thus, final rules having confidence more than the threshold limit, i.e., 75% are as follows.

Selected Association rules are
2, 3→5
2→5
3→5
3, 5→2
5→2

Improvement of Apriori Algorithm- Pruning

- Apriori algorithm can be improved using Apriori property.
- *Apriori property states that all non empty subsets of a frequent itemset must also be frequent.*
- The itemset which does not satisfy Apriori property should be removed from the candidate set.
- This step is called ***pruning*** of the candidate set.
- It will help in improvement in performance of Apriori algorithm as it will reduce the search space.

Improved Apriori Algorithm- Example 2

Generate association rules according to Apriori algorithm (with pruning)

- Threshold value of Support 15% and Confidence 70%

TID	List of Items
10	I1, I2, I5
20	I2, I4
30	I2, I3
40	I1, I2, I4
50	I1, I3
60	I2, I3
70	I1, I3
80	I1, I2, I3, I5
90	I1, I2, I4

Table 9.41 C1

C1	Count
I1	6
I2	7
I3	5
I4	3
I5	2

Example 2 – Solution (Phase-1)

TID	List of Items
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

C1	Count
I1	6
I2	7
I3	5
I4	3
I5	2

L1
↙



Example 2 – Solution (Phase-1 Contd...)

Process of Generation of C_2 and L_2

L_1

C_1
I_1
I_2
I_3
I_4
I_5

Generate C_2
Candidates
from L_1

C_2
Itemset
$\{I_1, I_2\}$
$\{I_1, I_3\}$
$\{I_1, I_4\}$
$\{I_1, I_5\}$
$\{I_2, I_3\}$
$\{I_2, I_4\}$
$\{I_2, I_5\}$
$\{I_3, I_4\}$
$\{I_3, I_5\}$
$\{I_4, I_5\}$

Scan D for
count of each
candidate

C_2	
Itemset	Sup. count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_4\}$	1
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2
$\{I_3, I_4\}$	0
$\{I_3, I_5\}$	1
$\{I_4, I_5\}$	0

Compare candidate
support count with
minimum support
count

L_2	
Itemset	Sup. count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2

Example 2 – Solution (Phase-1 Contd...)

Generation of C3

L_2

Itemset	Sup. count
{I1,I2}	4
{I1,I3}	4
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2

Generation of C3

C3
I1, I2, I3
I1, I2, I5
I1, I3, I5
I2, I3, I4
I2, I3, I5
I2, I4, I5

Example 2 – Solution (Phase-1 Contd...)

Pruning of C3

Generation of C3

C3
I1, I2, I3
I1, I2, I5
I1, I3, I5
I2, I3, I4
I2, I3, I5
I2, I4, I5

Pruned C3

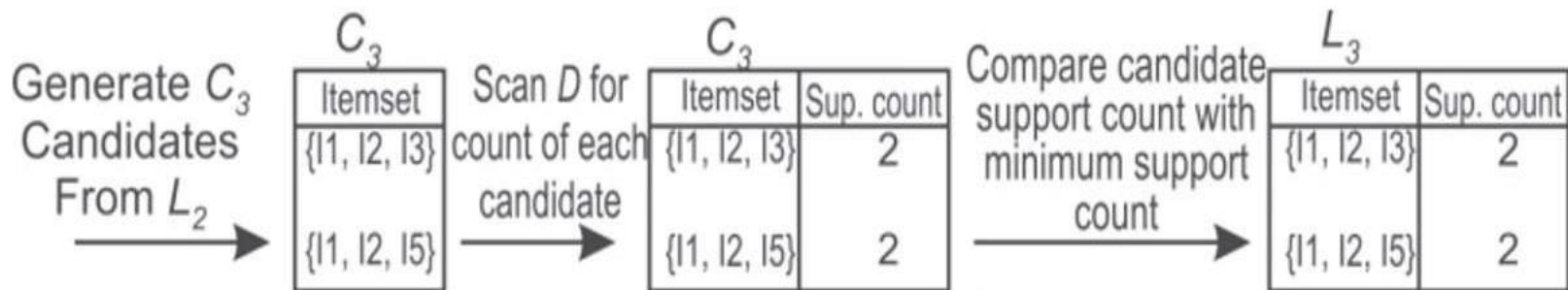
Pruned C3
I1, I2, I3
I1, I2, I5

L_2

Itemset	Sup. count
{I1,I2}	4
{I1,I3}	4
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2

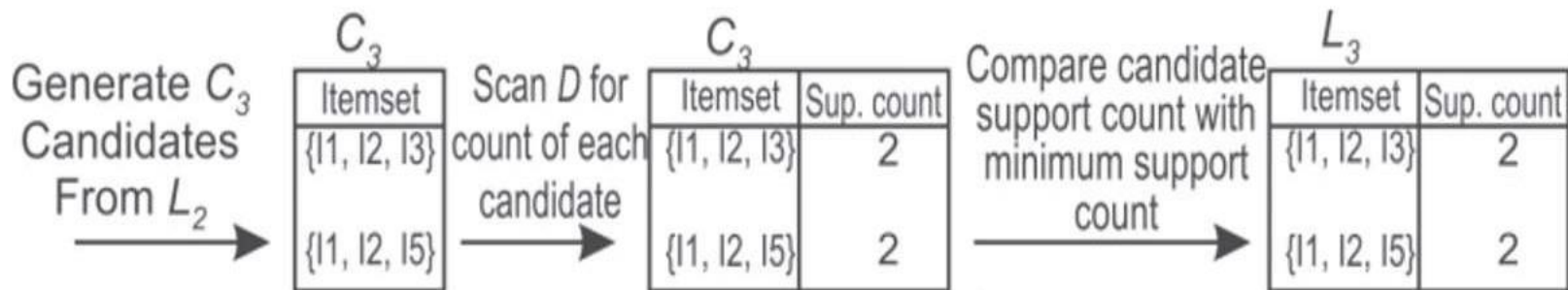
Example 2 – Solution (Phase-1 Contd...)

Generation of L_3



Example 2 – Solution (Phase-1 Contd...)

Generation of L_3



Example 2 – Solution (Phase-1 Contd...)

Generation and Pruning of C4

C4

C4
I1, I2, I3, I5

But this itemset is pruned by the Apriori property because its subset (I2, I3, I5) is not frequent as it is not present in L3. Thus, C4 is null and the algorithm terminates at this point, having found all of frequent itemsets as shown in Table :

Pruned C4

C4
NULL

Example 2 – Solution (Phase-2 Contd...)

Let us apply this rule to frequent 3-itemsets (I1, I2, I3) and (I1, I2, I5) found in case of example

For first frequent itemset (I1, I2, I3), non-empty subsets are {{I1},{I2},{I3},{(I1, I2)}, {(I1, I3)},{(I2, I3)}}.

For every non-empty set, the rule will be generated as follows:

I1→I2, I3 [Here, (I1) is *s* and I2 and I3 are *l-s*]

I2→I1, I3

I3→I1, I2

I1, I2→I3

I1, I3→I2

I2, I3→I1

The next will be to calculate the confidence for each rule as shown below.

I1→I2, I3; Confidence = $S(I1 \cap I2 \cap I3) / S(I1) = 2/6 = 0.3$

I2→I1, I3; Confidence = $S(I1 \cap I2 \cap I3) / S(I2) = 2/7 = 0.28$

I3→I1, I2; Confidence = $S(I1 \cap I2 \cap I3) / S(I3) = 2/5 = 0.4$

I1, I2→I3; Confidence = $S(I1 \cap I2 \cap I3) / S(I1 \cap I2) = 2/4 = 0.5$

I1, I3→I2; Confidence = $S(I1 \cap I2 \cap I3) / S(I1 \cap I3) = 2/4 = 0.5$

I2, I3→I1; Confidence = $S(I1 \cap I2 \cap I3) / S(I2 \cap I3) = 2/4 = 0.5$

Example 2 – Solution (Phase-2 Contd...)

Now, let us apply this rule to second frequent 3-itemset (I1, I2, I5). For this frequent itemset, non-empty subsets are {I1}, {I2}, {I5}, {(I1, I2)}, {(I1, I5)}, {(I2, I5)}.

For every non-empty set the rule will be generated as follows:

I1→I2, I5 (Here, (I1) is s and I2 and I5 are $I-s$)

I2→I1, I5

I5→I1, I2

I1, I2→I5

I1, I5→I2

I2, I5→I1

The next step will be to calculate the confidence for each rule as shown below.

I1→I2, I5; Confidence = $S(I1 \cap I2 \cap I5) / S(I1) = 2/6 = 0.3$

I2→I1, I5; Confidence = $S(I1 \cap I2 \cap I5) / S(I2) = 2/7 = 0.28$

I5→I1, I2; Confidence = $S(I1 \cap I2 \cap I5) / S(I5) = 2/2 = 1$

I1, I2→I5; Confidence = $S(I1 \cap I2 \cap I5) / S(I1 \cap I2) = 2/4 = 0.5$

I1, I5→I2; Confidence = $S(I1 \cap I2 \cap I5) / S(I1 \cap I5) = 2/2 = 1$

I2, I5→I1; Confidence = $S(I1 \cap I2 \cap I5) / S(I2 \cap I5) = 2/2 = 1$

Now, there are three rules whose confidence is more than minimum threshold value of 70%, and these rules are as follows:

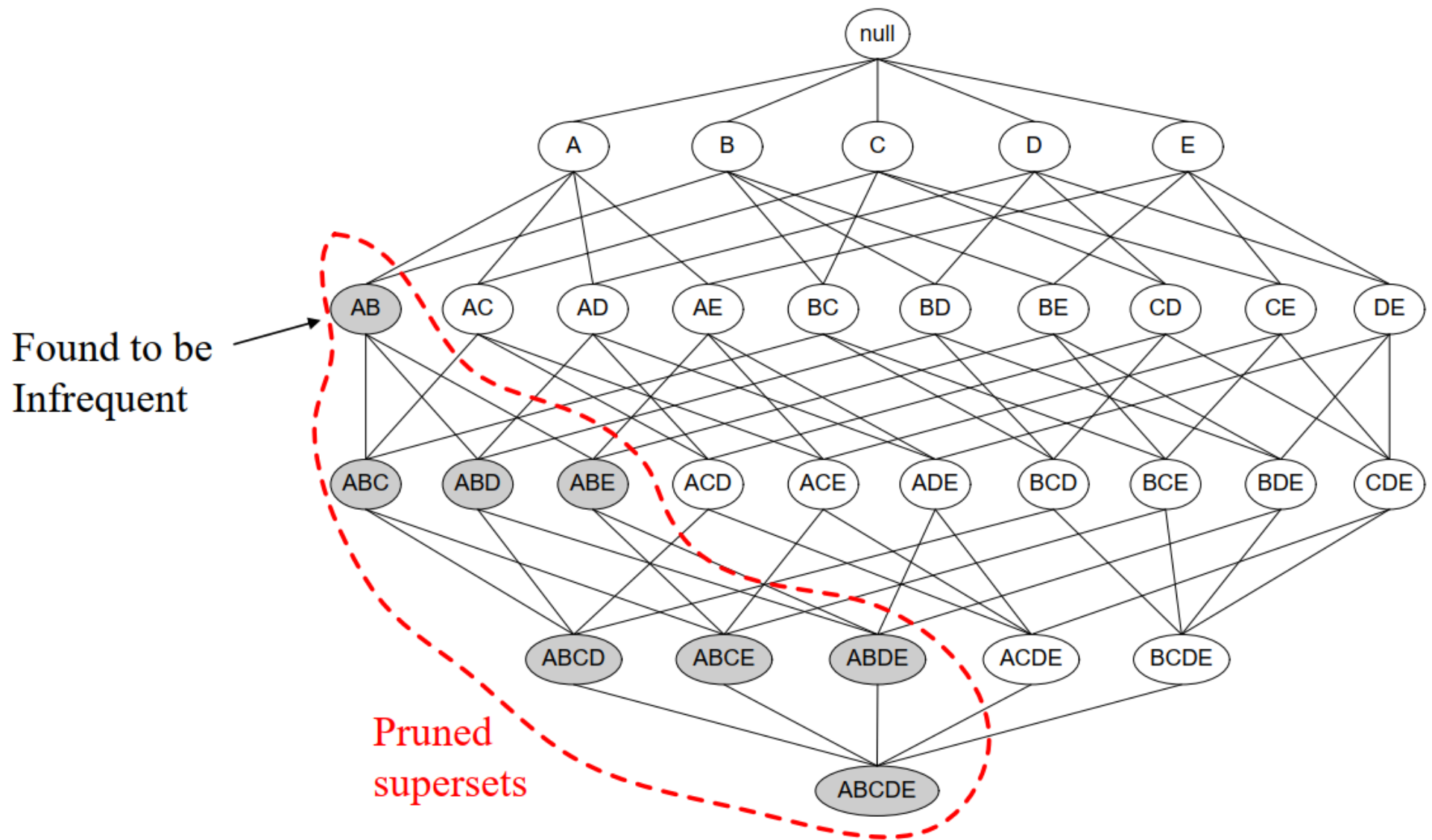
I5→I1, I2

I1, I5→I2

I2, I5→I1

Example 2 – Solution (Phase-2 Contd...)

Association Rule	Discussion	Confidence	More than threshold limit Or Not
$I5 \rightarrow I1, I2$	Already identified	1.0	Yes
$I5 \rightarrow I1$	Implicit	No need to calculate it will be more than or equal to $I5 \rightarrow I1, I2$	Yes
$I5 \rightarrow I2$	Implicit	No need to calculate it will be more than or equal to $I5 \rightarrow I1, I2$	Yes
$I2, I5 \rightarrow I1$	Already identified	1.0	Yes
$I2 \rightarrow I1$	Not found earlier, confidence need to be calculated	Confidence of $I2 \rightarrow I1$ $= S(I2 \cap I1) / S(I1) = 4/6 = 67\%$	No
$I5 \rightarrow I1$	Already found from $I5 \rightarrow I1, I2$ given in row 2	No need to calculate it will be more than or equal to $I5 \rightarrow I1, I2$	Yes (Already listed)
$I1, I5 \rightarrow I2$	Already identified	1.0	Yes
$I1 \rightarrow I2$	Not found earlier, so confidence needs to be calculated	Confidence of $I1 \rightarrow I2$ $= S(I2 \cap I1) / S(I2) = 4/7 = 57\%$	No
$I5 \rightarrow I2$	Already found from $I5 \rightarrow I1, I2$ given in row 3		Yes (Already listed)



-
- Maximal Frequent Itemsets
 - Closed Frequent Itemsets

If the frequent itemsets are:

$\{A\}$, $\{B\}$, $\{A,B\}$, $\{A,B,C\}$

and $\{A,B,C\}$ is frequent but no 4-itemset containing A is frequent, then:

A frequent itemset whose all immediate supersets have strictly lower support. \hookrightarrow $\{A,B,C\}$ is a maximal frequent itemset
It is frequent, But no larger itemset has the same support.

$A,B\}$ 3

$\{A,B,C\}$ 3

$\{A,B,C,D\}$ 2

Here:

$\{A,B\}$ has support 3

But its superset $\{A,B,C\}$ also has support 3 \rightarrow same support

$\{A,B\}$ is NOT closed (because a larger set has same support)

But: $\{A,B,C\}$ IS a closed frequent itemset (because all its supersets have lower support)