10/7/23

UNIT-1 Association Analysis

* Association Analysis can be applied to business domain. It can also used in Healthcare domain, finance.
* Dataset in Association Analysis can be represented in the form of transactions. Each transaction consists of the items that are bought in that transaction.

Difficulties with Association :-
* huge data (handling with large data)
* handling useless transaction (unintentionally, unconditionally) eg:- milk + pen
* The entire dataset can be represented in binary formate where each row represents a transaction and each column represents an item.

example :-

			milk	sugar	bread	T.p	C.p
T_1	Milk + sugar + bread	T_1	1	1	1	0	0
T_2	milk + Teapowder + sugar	T_2	1	1	0	1	0
T_3	milk + sugar + coffeepowder	T_3	1	1	0	0	1
T_4	milk + sugar + bread + jam	T_4	0	1	1	0	0

* A collection of zero or more items is known as an itemset
* k-itemset, An itemset with k items is known as a k-itemset

13/7/23

General Rule formate will be A → B (A derives B)
where A is known as Anticident and B is known as consequent
eg:- {Milk, coffee} → Sugar
for A may contains one or more items.

Support :-
An Association rule can be measured in terms of support.
→ support determines how often a rule is applicable for a given data set.

Confidence :-
confidence is a measuring parameter for Association Rule.
→ confidence determines how frequently the items in B appear in the transactions that contain A.

$$\text{Support}(A \to B) = \frac{\text{count}(A \cup B)}{N}$$

N → total no of transactions in the given dataset.

eg:-

Tid	Items
T1	I_1, I_2
T2	I_1, I_3, I_4, I_5
T3	I_2, I_3, I_4, I_6
T4	I_1, I_2, I_3, I_4
T5	I_1, I_2, I_3, I_5

Calculate the support for the Association $\{I_2 I_3\} \to I_4$

$$= \frac{count(I_2, I_3, I_4)}{N}$$

$$\Rightarrow \frac{2}{5} = 0.4$$

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

$$\Rightarrow \frac{\frac{count(A \cup B)}{N}}{\frac{count(A)}{N}} \Rightarrow \frac{count(A \cup B)}{count(A)}$$

$$\frac{count(I_2, I_3, I_4)}{count(I_2, I_3)} \Rightarrow \frac{2}{3} \quad count(A) \rightarrow count\{I_2, I_3\}$$

eg:- $\{I_1, I_3\} \rightarrow I_5$

$$\text{support}(\{I_1, I_3\} \rightarrow I_5) = \frac{count(I_1 \cup I_3 \cup I_5)}{N}$$

$$= \frac{2}{5}$$

$$\text{confidence}(\{I_1, I_3\} \rightarrow I_5) = \frac{\text{support}(\{I_1, I_3\} \rightarrow I_5)}{\text{support}(\{I_1, I_3\})}$$

$$= \frac{count(I_1 \cup I_3 \cup I_5)}{count(I_1, I_3)}$$

$$= \frac{2}{3}$$

eg:- $I_1 \rightarrow \{I_3, I_5\}$

$$\text{support}(\{I_1\} \rightarrow \{I_3, I_5\}) = \frac{count(I_1 \cup I_3 \cup I_5)}{N}$$

$$= \frac{2}{5}$$

$$\text{Confidence}(I_1 \rightarrow \{I_3, I_5\}) = \frac{count(I_1 \cup I_3 \cup I_5)}{count(I_1)}$$

$$= \frac{2}{4}$$

Note:-

* Support is used to eliminate uninteresting rules.
* confidence is used to measure the reliability of the Inference made by the rule.
* STRONG RULE:- A rule that satisfy both minimum support threshold and minimum confidence threshold is known as a strong rule.

14/7/23

Apriori Algorithm:-

Step-1:- frequent item set generation
Step-2:- Rule generation

In the step-1 we find all the item sets that satisfy minimum support threshold. These item sets are known as frequent item sets.

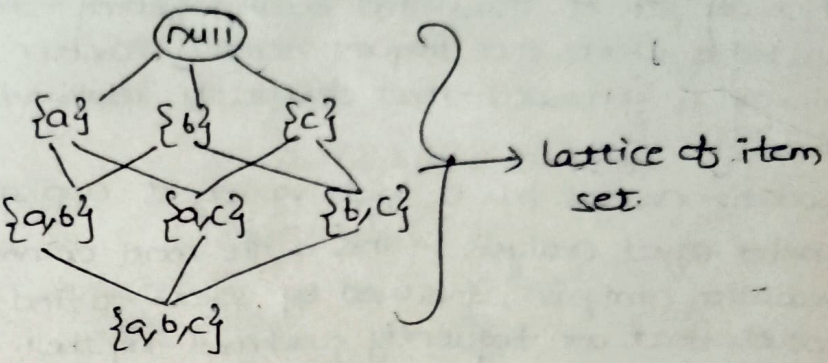
In the step2 we extract all the i confidence rules from the frequent item sets. These rules are known as strong rule

frequent item set generation :-

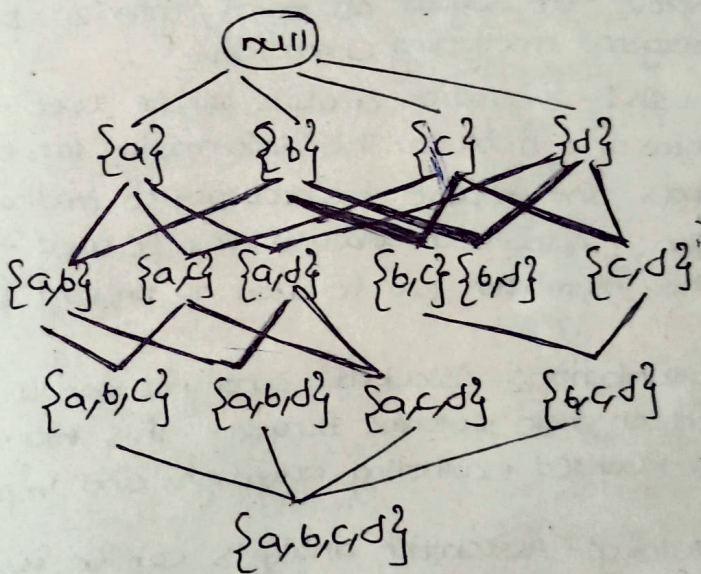
for a dataset with k items there will be $2^k - 1$ frequent item sets.

eg:- a,b,c [k=3]

we get → {a}, {b}, {c}, {a,b}, {a,c}, {b,c}, {a,b,c} → item sets



eg:- k=4



15 ie
$2^4 - 1$
15

As k might be very large the number of frequent item sets that need to be explore may become exponentially large, in order to reduce this complexity we use apriori principle.

Apriori principle I :-

All the non empty subsets of a frequent item set must also be frequent.

for eg:- for a four-item set {b,c,d} is frequent then all its subsets are frequent ({b}, {c}, {d}, {b,c}, {c,d}, {b,d})

Apriori principle 2 :-

If an item set is non frequent than all its supersets are considered to be non frequent.

for eg :- if {b} is considered to be non-frequent than all its supersets are non frequent ({a,b}, {b,c}, {b,d}, {a,b,c}, {a,b,d}, {b,c,d}, {a,b,c,d}).

for 4-item data set.

What is ~~clustering~~ Association Analysis? Explain various applications where association analysis can be applied.

Association analysis is a data mining technique that identifies relationships between variables in a data set. It is often used to find associations between products that are frequently purchased together such as milk and bread.

Association analysis works by identifying frequent itemsets, which are sets of items that occur together more often than would be expected by chance. Once frequent itemsets have been identified, association rules can be generated that describe the relationships between these itemsets.

Association analysis has a wide variety of applications, including:-

- Market Basket analysis :- This is the most common application of association analysis. It is used by sellers to find associations between products that are frequently purchased together. This information can be used to improve the layout of stores, optimize product placement, and develop targeted marketing campaigns.
- Medical diagnosis :- Association analysis can be used to find associations between symptoms and diseases. This information can be used to develop diagnostic tools and improve the accuracy of medical diagnoses.
- Fraud detection :- Association analysis can be used to detect fraudulent transactions. This information can be used to prevent fraud and protect customers.
- Customer segmentation :- Association analysis can be used to segment customers based on their purchase behaviour. This information can be used to develop targeted marketing campaigns and improve customer service.
- Web usage mining :- Association analysis can be used to find associations between web pages that are frequently visited together. This information can be used to improve the design of websites and develop targeted advertising campaigns.

Some additional examples are product recommendations, social media analysis, logistics etc...

20/1/23

*We have 2 phases in Apriori :-

1) Join
2) prune

Example-1

Transactional Data for an All Electronics Branch

TID	List of item_IDs
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

minimum support count is 2.

candidate Item set (C)

C_1

Item	support
✓ I_1	6
✓ I_2	7
✓ I_3	6
✓ I_4	2
✓ I_5	2

min^m count/support

*crisocographic order

$C_2 = L_1 \bowtie L_1$

Item	support
(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

L_3

Item	support
(I_1, I_2, I_3)	2
(I_1, I_2, I_5)	2

(selected list)

L_1

Item	support
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

let us suppose L_2

$\{(I_1, I_2)(I_1, I_3)(I_2, I_3)(I_1, I_4)\}$

$= \{(I_1, I_2, I_3)(I_1, I_2, I_4)\}$

L_2

Item	support
(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

$C_3 = L_2 \bowtie L_2$

$C_3 = \{(I_1, I_2, I_3), (I_1, I_3, I_5), (I_1, I_3, I_5), (I_2, I_3, I_4)(I_2, I_3, I_5), (I_2, I_4, I_5)\}$

$C_3 = L_2 \bowtie L_2$

Item	support
(I_1, I_2, I_3)	2
~~(I_1, I_2, I_4)~~	
(I_1, I_2, I_5)	2
~~(I_1, I_3, I_4)~~	
~~(I_1, I_3, I_5)~~	
~~(I_1, I_4, I_5)~~	
~~(I_2, I_3, I_4)~~	

* we need to eliminate certain item sets based on the apriori principle.

(frequency) for the Item set $\{I_1, I_3, I_5\}$ all its subsets are not frequent. $\{(I_1, I_3), (I_1, I_5), (I_3, I_5)\}$ ie-, they are not present in L_2. so the item $\{I_1, I_3, I_5\}$ is considered to be non frequent

L_3

Item	support
(I_1, I_2, I_3)	2
(I_1, I_2, I_5)	2

4/7/23 JOIN STEP

→ In this step we find L_k a set of candidate k item sets by joining L_{k-1} with itself. The resulting set after performing join is denoted as C_k. The join $L_{k-1} \bowtie L_{k-1}$ is performed where the members of L_{k-1} are joinable.

→ The members of L_{k-1} are said to be joinable if their first k-2 items are in common.

PRUNE STEP

→ C_k is the superset of L_k, so it may even contain non frequent item sets. All the candidates having a count greater than the minimum support count are included in L_k.

→ The item sets in C_k are included if all its subsets are frequent.

$C_4 = L_3 \bowtie L_3$

$= \{(I_1, I_2, I_3, I_5)\}$ X ⟶ because its subset $\{I_1, I_3, I_5\}$ is not frequent.

When only the $C_k = \emptyset$ ie {empty} the algorithm terminates.

Ex:- T100 {M, O, N, K, E, Y} minimum support is 3.

T200 {D, O, N, K, E, Y}

T300 {M, A, K, E}

T400 {M, U, C, K, Y}

T500 {C, O, O, K, I, E}

Candidate Itemset

C_1

Item	support
M	3
O	4
X N	2
K	5
E	4
Y	3
X D	1
X A	1
X U	1
X C	2
X I	1

L_1

Item	support
M	3
O	4
K	5
E	4
Y	3

$C_2 = L_1 \bowtie L_1$

Item	support
X(M, O)	1
(M, K)	3
X(M, E)	2
X(M, Y)	2
(O, K)	4
(O, E)	3
X(O, Y)	2
(K, E)	4
✓(K, Y)	3
X(E, Y)	2

L_2

Item	support
(M, K)	3
(O, K)	4
(O, E)	3
(K, E)	4
(K, Y)	3

$C_3 = L_2 \bowtie L_2$

$\{(O, K, E), (K, E, Y)\}$

Subset is not frequent (K, E, Y) → (E, Y)

Item	support
(O, K, E)	3

L_3

Item	support
(O, K, E)	3

24/7/23

*Consider the minimum confidency as 60% and compute the association rules. (Example-1)

Let us suppose the frequent item set as L_4 where L_4 can be written as (S; L) i.e.,

S is the starting element of L_4 and L is list of items such that $L = L_1 - S$. We can frame Association rules for L_4 such that $S \rightarrow (L - S)$ or any other combination which satisfies the minimum confidence.

L_1

(I_1, I_2, I_5)

$I_1 \rightarrow (I_2, I_5)$ $(I_1, I_2) \rightarrow I_5$

$I_2 \rightarrow (I_1, I_5)$ $(I_1, I_5) \rightarrow I_2$

$I_5 \rightarrow (I_1, I_2)$ $(I_2, I_5) \rightarrow I_1$

$$\text{confidence}(\{I_1\} \rightarrow \{I_2, I_5\}) = \frac{\text{count}(I_1 \cup I_2 \cup I_5)}{\text{count}(I_1)}$$

$$\Rightarrow \frac{2}{6} = \frac{1}{3} \times 100 = 33.66\% \quad \times$$

$$\text{Confidence}(I_2 \rightarrow \{I_1, I_5\}) = \frac{\text{count}(I_2, I_1, I_5)}{\text{Count}(I_2)}$$

$$= \frac{2}{7} \times 109 = 28\% \quad \times$$

$$\text{Confidence}(I_5 \rightarrow \{I_1, I_3\}) = \frac{\text{count}(I_5, I_1, I_3)}{\text{count}(I_5)}$$

$$\Rightarrow \frac{2}{2} = 1 = 100\% \quad \checkmark$$

$$\text{confidence}(\{I_1, I_2\} \rightarrow I_5) = \frac{\text{Count}(I_1, I_2, I_5)}{\text{Count}(I_1, I_2)} = \frac{2}{4} = \frac{1}{2} = 50\%$$

$$\text{confidence}(\{I_1, I_5\} \rightarrow I_2) = \frac{\text{count}(I_1, I_5, I_2)}{\text{Count}(I_1, I_5)} = \frac{2}{2} = 1 = 100\%$$

$$\text{confidence}(\{I_2, I_5\} \rightarrow I_1) = \frac{\text{count}(I_2, I_5, I_1)}{\text{count}(I_2, I_5)} = \frac{2}{2} = 1 = 100\%$$

L_2 (I_1, I_2, I_3)

$I_1 \rightarrow (I_2, I_3)$	confidence $= \frac{2}{6} \times 100 = 33.66\%$ ✗
$I_2 \rightarrow (I_1, I_3)$	confidence $= \frac{2}{7} \times 100 = 28\%$ ✗
$I_3 \rightarrow (I_1, I_2)$	confidence $= \frac{2}{6} \times 100 = 33.66\%$ ✗
$(I_2, I_3) \rightarrow I_1$	confidence $= \frac{2}{4} \times 100 = 50\%$ ✓
$(I_1, I_3) \rightarrow I_2$	confidence $= \frac{2}{4} \times 100 = 50\%$ ✗
$(I_1, I_2) \rightarrow I_3$	confidence $= \frac{2}{4} \times 100 = 50\%$ ✗

$L_4 \rightarrow (I_5 \rightarrow \{I_1, I_3\}), (\{I_1, I_5\} \rightarrow I_2), (\{I_2, I_5\} \rightarrow I_1)$

TID	I_1 Rice	I_2 pulses	I_3 Milk	I_4 Apple	I_5 oil
t1	1	1	1	0	0
t2	0	1	1	1	0
t3	0	0	0	1	1
t4	1	1	0	1	0
t5	1	1	1	0	1
t6	1	1	1	1	1

Support = 50%
Confidence = 60%.

$\frac{50}{20} = \frac{\frac{50}{100}}{\frac{2}{3}}$

Item	support
I_1	4 67%
I_2	5 83%
I_3	4 = 67%
I_4	4 = 67%
I_5	3 = 50%

L_1

Item	support
I_1	4
I_2	5
I_3	4
I_4	4
I_5	3

$C_2 = L_4 \bowtie L_4$

Item	support
(I_1, I_2)	4 67%
(I_1, I_3)	3 60%
(I_1, I_4)	2 33%
(I_1, I_5)	2 33%
(I_2, I_3)	4 67%
(I_2, I_4)	3 50%
(I_2, I_5)	2 33%
(I_3, I_4)	2 33%
(I_3, I_5)	2 33%
(I_4, I_5)	2 33%

L_2

Item	support
(I_1, I_2)	4
(I_1, I_3)	3
(I_2, I_3)	4
(I_2, I_4)	3

$C_3 = L_2 \bowtie L_2$

Item	support
(I_1, I_2, I_3)	3 = 50%

$\times (I_2, I_3, I_4)$

↓ Subsets are not frequent

L_3

Item	support
(I_1, I_2, I_3)	3

(I_1, I_2, I_3)

$I_1 \rightarrow \{I_2, I_3\} = \frac{\text{Count}(I_1 \cup I_2 \cup I_3)}{\text{count}(I_1)} = \frac{3}{4} \times 100 = 75\%$ ✓

$I_2 \rightarrow \{I_1, I_3\} = \frac{3}{5} \times 100 = 60\%$ ✓

$I_3 \rightarrow \{I_1, I_2\} = \frac{3}{4} \times 100 = 75\%$ ✓

$\{I_2, I_3\} \rightarrow I_1 = \frac{3}{4} \times 100 = 75\%$ ✓

$\{I_1, I_3\} \rightarrow I_2 = \frac{3}{3} \times 100 = 100\%$ ✓

$\{I_1, I_2\} \rightarrow I_3 = \frac{3}{4} \times 100 = 75\%$ ✓

27/7/23 FP-Growth Algorithm:- (frequency-pattern)

Step-1:- Condense the dataset by constructing a FP-tree

Step-2:- To extract association rules from the FP-Tree.

from example 1; support = 2

$L_1 = L$

I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

⇒

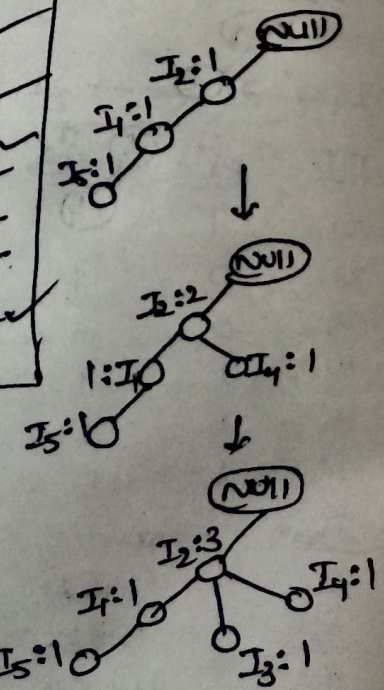
I_2	7
I_1	6
I_3	6
I_4	2
I_5	2

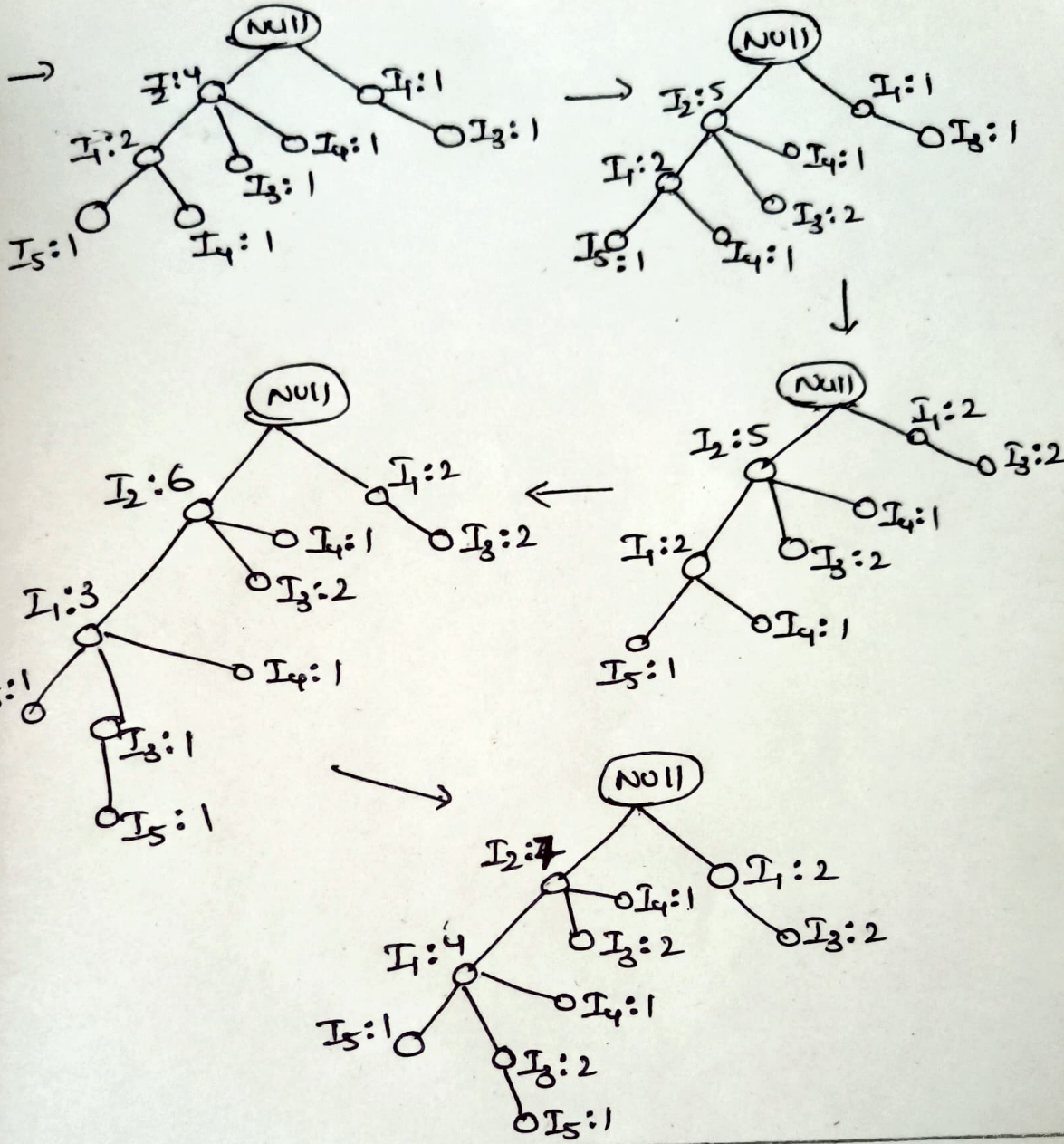
descending order of support count

⇒

T_{100}	I_2, I_1, I_5 ✓
T_{200}	I_2, I_4 ✓
T_{300}	I_2, I_3 ✓
T_{400}	$I_2 I_1 I_4$ ✓
T_{500}	$I_1 I_3$ ✓
T_{600}	$I_2 I_3$ ✓
T_{700}	I_1, I_3 ✓
T_{800}	$I_2 I_1 I_3 I_5$ ✓
T_{900}	$I_2 I_1 I_3$

FP Tree




Null
I2:4
I1:1
I1:2
I4:1
I3:1
I3:1
I5:1
I4:1
Null
I2:5
I1:1
I4:1
I3:1
I1:2
I3:2
I5:1
I4:1
Null
I1:2
I3:2
I2:5
I4:1
I1:2
I3:2
I4:1
I5:1
Null
I2:6
I1:2
I4:1
I3:2
I3:2
I1:3
I5:1
I4:1
I3:1
I5:1
Null
I2:4
I4:1
I1:2
I3:2
I1:4
I3:2
I4:1
I5:1
I3:2
I5:1