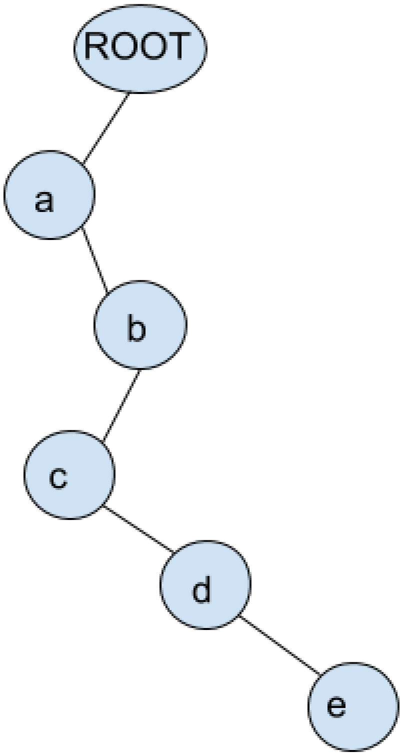
Possible Improvement(s)

Consider the path from the root to one of the leaf in the below **OIMASP** tree (only one path is shown)



According to **OIMASP** algorithm

1. For path (a → b → c → d → e)
2. Rules generated will be (a → b), (ab → c), (abc → d) & (abcd → e)

OIMASP does not say anything about the rules in which more than one items are present on the right hand side i.e (a → bc), (a → bcd), (a → bcde), (ab → cd), (ab → cde) & (abc → de). Let us refer these rules as **excluded-rules** from now onwards.

The **excluded-rules** may or may not be an association rule. It is true that **excluded-rules** will satisfy the **threshold support** (because of the way **OIMASP** tree is generated) but we cannot conclude the same about **threshold confidence.** So we need to check whether **excluded-rules** are satisfying **threshold confidence** or not.

One way to check is to scan transaction database once for each **excluded-rules**. The complexity of modified **OIMASP** algorithm will be $\Theta(|D| * |D| * log|D|)$ where $|D| = M * N$ and $M = \#rows\ \&\ N = \#columns$. D is the transaction database associated with root node.

How $\Theta(|D| * |D| * log|D|)$?

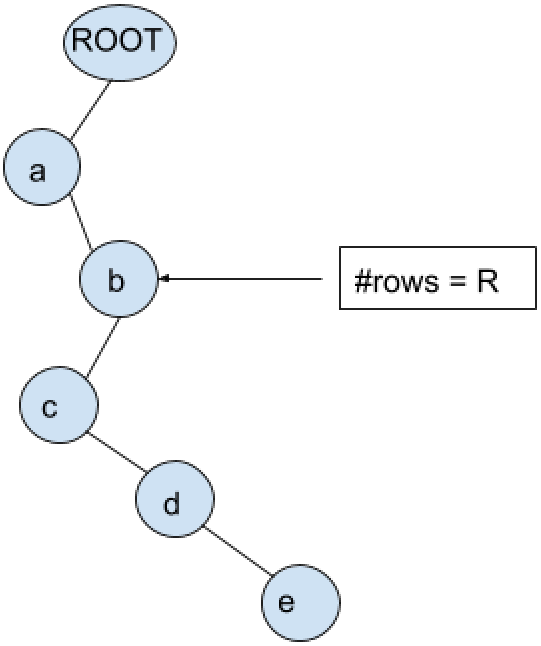
Let height of root node is 0 and height of the tree is H.

$$\begin{aligned} Total\ computation &= \sum_{h=0}^{H} (2^{H+1} - 2^{h+1})|D| \\ &= \Theta(|D| * |D| * log|D|) \end{aligned}$$

We know that data is associated with each node of the **OIMASP** tree.

Let us see how to minimize the search space by exploiting the data stored at each node instead of scanning the complete transaction database.

We will see for one of the excluded rules say (ab → cd)



For excluded-rule **(ab → cd)** we will go to the node **b** (in general, last element on the left hand side of the rule). Then, we will find number of rows (data at node **b**) in which **cd** is present (in general, right hand side of the rule). Say it is r.

Then,

$$confidence(ab \ \rightarrow cd) = \frac{r}{R}$$

If confidence is calculated by exploiting the data stored at node then complexity of modified **OIMASP** algorithm will be $\Theta(|D| * |D|)$ where $|D| = M * N$ and $M = \#rows \;\&\; N = \#columns$.

How $\Theta(|D| * |D|)$?

Let height of root node is 0 and height of the tree is H.

$$
\begin{aligned}
Total\ computation &= \sum_{h=0}^{H} (2^{H+1} - 2^{h+1}) \tfrac{|D|}{2^h} \\
&= \Theta(|D| * |D|)
\end{aligned}
$$

The space complexity of modified **OIMASP** algorithm ($\Theta(|D| * |D|)$).

$$Space\ complexity = O(|D| log |D|)$$

How to improve the space complexity? i.e. $O(|D|)$

Previously, first **OIMASP** tree was generated and then rules were obtained from the tree.

Now, we will generate rules while generating **OIMASP** tree

- At some instant say we are at node **d**
- The path from root to **d** will look like root → **a** → **b** → **c** → **d**
- At this moment data is stored only at node **a**, **b**, **c**, and **d**. No other nodes in the partially generated tree will have data.
- Check the confidence of excluded-rules (**a** → **bcd**) and (**ab** → **cd**)

While generating the tree in Depth-First-Search (**DFS**) fashion, for every new node we visit, the above steps will be followed to find acceptable **excluded-rules**. The amount of data in the stack will be atmax the summation of data size at each node in the path from root to leaf.

Therefore, $Space\ complexity = \sum_{h=0}^{H} \frac{|D|}{2^h} \; = \; O(|D|)$