Assignment 01

Web Usage Mining

Due Date: Wednesday, 10th November

1 Background

This assignment focuses on web usage mining. Given a data set that describes the behaviour of a user on a specific website, your assignment is to mine this data in order to find common surfing patterns. In other words you are to find relations of the kind: if the user starts at webpage A and then goes from webpage A to webpage B she is likely to continue to webpage C. This might seem to be similar to discovering association rules using the *a priori* algorithm, and indeed it is, but in this case the order of visiting the webpages is significant.

2 Assignment

The dataset can be found in dataset WUM.txt Each row in the data corresponds to one user session. The session is described by space separated integers where each integer corresponds to a unique webpage on the site. The order of the integers is significant as this describes the order in which the different webpages were visited during that session.

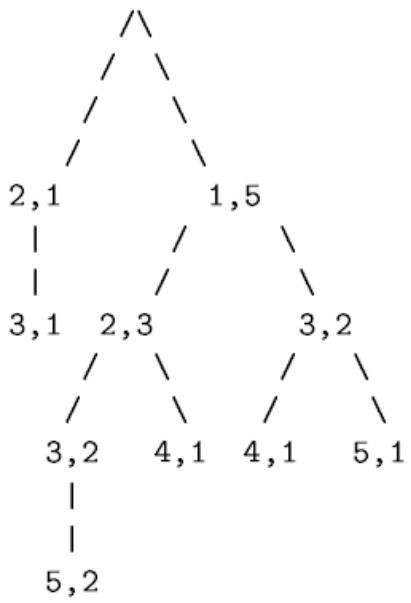
2.1 Tasks

Use your favourite programming language to turn the transaction data into a trie where each node contains an integer corresponding to a webpage *PageNumber* and another integer *NoOfTransactions* which describes how many transactions have taken this path.

For example, for the simple transaction data base shown below:

```
1 2 3 5
1 3 4
2 3
1 2 4
1 2 3 5
1 3 5
```

the trie becomes:



Your task is to then find rules of the type `1->2->3 => 5` that have confidence and support higher that some minimum values specified by the user of your program (your program should take these as inputs). The meaning of such a rule is: If a user starts a session at page 1 and continues to page 2 and then to page 3, then she is likely to continue to page 5. To find the confidence of such a rule simply take the *NoOfTransactions* in the node `1->2->3->5` and divide it by the *NoOfTransactions* in `1->2->3`. In this case the confidence of the rule is $2/2 = 100$ %. To calculate the support just take *NoOfTransactions* in node `1->2->3->5` and divide that by the total number of transactions. In this case, the support is $2/6 = 33.3$ %. Calculate these values for all nodes and return the rules which have support and confidence above the user-specified thresholds.