Mining Frequent Sequence from online retail data using GSP algorithm

**Harish Goud Ediga ID No: 2016A7PS0110H**

**Abstract**: -

In this project we use Generalised sequential pattern algorithm to mine the data. It’s an improvised apriori based Algorithm where the output is a sequence of patterns rather than a set, the major advantage is it reduces search space rather than searching the whole database (the out of one epoch is used as input of the other) but problem is scans the database multiple times and generates large sequences. The data analysis shows pattern which can be used for advertisements ,like if I purchased a laptop and mouse and if earlier data shows that people purchased laptop bag soon, then we can recommend user this , GSP has an edge over apriori as it takes account the order of purchase like laptop and then bag is different than bag and then laptop unlike apriori.

**Data Pre-processing:**

The data I got is online retail sales info across Europe, which contain attributes like invoice number (the invoice number for return items start with C), description of item and the no of units purchased, the price, the customerID and the country.

Each customerID might have more than one invoice numbers, each invoice has might have more than one items, there are many missing customerID and in Description column.

I deleted all the columns which have missing data as there is no way we can fill using approximation or from the existing data. All the data columns are having relation and are independent of other so there is no variable dependency, when checked for outlier there are no outliers in the data. As we are looking for patterns in the purchased data, we have to delete the rows containing cancelled data.

The description column is one hot encoded as to perform on integral data.

**GSP Algorithm (with example): -**

The starting of the algorithm is similar to apriori

