Churn Discovery on Large Telecom Data with Apriori Algorithm

[1]Roopa B, [2]Dr. R. Nagaraja, [3]T Shilpa

[1]Student, [2]Professor & PG Co-Ordinator, [3]Assistant Professor
Department of Information Science and Engineering,
Bangalore Institute of Technology, Bangalore, Karnataka, India.

Abstract*: Basically, customers are provided with an opportunity to switch from one service provider to another service provider in Telecommunication industries. Telecom industries are currently more concerned about the churn customers due to competition in the market with other telecom industries. In order to have a deeper understanding of customer churn, apriori algorithm is used to show the most accurate churn prediction. Telecom industries enhance their services to reduce the number of customer churn. In this project, attribute selection method based on apriori technique is used. According to this algorithm attribute selection problem can be replaced. A sample dataset is used to test the algorithm and obtain the number of customer who fall under churn based on the threshold value provided. The threshold value is calculated with mean weighted average vector. Recent, a robust churn prediction model has been applied in telecom industries.

***Index Terms*: Apriori, Seed Pattern, Data Extraction.**

I. INTRODUCTION

Dynamics in technology and rapid growth in marketplace makes retaining customers a competitive effort. Many telecommunication industries, now a day's come up with various services, schemes and offers which attract customers' service which would make them turn to the other service provider, maybe because the customer is not satisfied with the old service. The aim of those companies is to retain their old/existing Customers than to obtain new customers because the cost of operating new customer is higher than retaining the old customer. Therefore it is significant to know the customers who may switch to another service in advance.[1][2]

The customers who switch from one service provider to another, such a customer is known as churn customer. An approach which targets the long term relation with customer is called CRM - Customer Relationship Management. Due to intensive competition, most of the companies realised the importance of CM and change their product centric targeting marketing to customer centric market. The quick progress of digital system and information technology provide an opportunity to understand customer needs and construct reliable CRM system.

Customer Relationship Management framework consists of two analytical modelling.

i. Predict customers were about to churn.
ii. Provide the most effective way to react. (This can also include "do nothing ") in terms of retaining the customer.

Therefore, according to telecommunication industries, the customers switch from one company to another, would lead to great loss to the company. So as to overcome such business loss and retain existing customer and also to retain the market, the telecom industries are forced to take up alternative ways to know the churn customer in advance, find the cause for the customer churn and take immediate effort to retain the customers. This task possible by having past history of the customers which can be analysed systematically.

Telecom industry is maintaining a large volume of data, fortunately. These customers information includes billing information, call details etc. This data helps data mining techniques to predict turn customers. In turn to improve their business.

There are many data mining algorithms which has been developed to mine large dataset for providing accurate churn rate and predict churn customers, so that the caution can be taken in advance. One of the algorithms called Apriori algorithm is used in this project is to predict on customers and show the churn rate of each customer who may fall under churn. Having a deeper understanding of customer churn and showing the accurate turn prediction which is given by apriori algorithm helps the telecom industry is to become aware of a high risk customer needs and the services are enhanced to retain their customers.[1][2]

Selection of attributes is method based on apriori technique is used. According to this algorithm selection of attributes problem has been replaced by the pruning question of classifier and an indicator system has been structures for finding out the churn customers. Robust churn prediction model has been applied in the telecom industry which is based on apriori algorithm. This approach has one of the major advantages that is simpler and efficient, but its own limitation is in the form of non-availability of features selection process.

In analytical approach existing does not work well in dealing with big data in order to overcome customer churn and efficient mining technique for the unstructured data has to be implemented. In the existing system the relational database cannot process unstructured data as it can only store traditional data.

Disadvantages

i. Difficult to identify turn customers
ii. Unstructured data cannot be processed effectively
iii. In accuracy in plastering

Apriori Algorithm is termed as a computation for mining large data set and administers to learn over large database. This algorithm continues to recognize the regular individual data by extending the database bigger. These dataset are shown frequently in the database.

Agarwal and Srikant proposed Apriori in the year 1994.The intention of Apriori is to work on databases containing exchanges. It is viewed that all exchanges as an arrangement of the items. The limitation of Apriori calculation is that, it distinguishes the datasets which are subsets of at any rate exchanges in the database.

Apriori Algorithm to count the item set effectively make use of tree structure and bread-first search. The length of itemset is k which is generated from itemsets of length k-1. New Trends of data mining is emphasised, from the classic data mining algorithm to present techniques. Data is widely available in large amounts and turning around those data into information that are useful and knowledgeable, this algorithm plays a vital role in information industry. These knowledge and information obtained are made useful in many applications such as hospitals, analysing markets, controlling production, managing business and designing engineering works etc. It contains all frequent itemset of k-length according to lemma closure-downward.

The vast amount of data which hides the large database that consist of global patterns and relationship, for example patient data and medical diagnosis and their relationship is called data mining. The valid knowledgeable data obtained through database and the database object is represented by data mining.

Data mining deals with structured data organised in database. It does not cover many properties.. Database sorted out in a organised information is essentially managed by data mining. It reveals abnormalities, special cases, examples, anomalies or patterns that may somehow stay undetected under the gigantic volumes of information.

II. PROBLEM STATEMENT

The crucial activity in growing and competitive telecom industries is customer churn. The cost of acquiring a new customer is very high when compared to retain the old customer. So as to overcome the problem of customer churn, a telecom industry applied one of the techniques to predict the customer churn rate, i.e, Apriori algorithm.The paper adds ability to acquire dataset and fetches it into the java memory and performs the analysis using apriori algorithm and display customer churn rate. In this project threshold values for churn rate is provided dynamically by using mean weighted average vector which changes every time when the customer details are updated.

III. LITERATURE SURVEY

Roma Singh,Sonal Chaudhary In this system the description about the customers who make subscription to an enterprise based on the business and products. Here the method used to find the customer who may churn in future. Several challenges where faced in mining large dataset. A novel ranking algorithm is described in this system. Working with industrial dataset, attributes that are myriad can be grouped into customer demographic information, product details, event data, domain specific data and behavioural data. Direct feeding of attributes to the algorithm did not work in a effective manner. The elevator for the attributes can be used for selecting the attributes and ranking them. There are class of imbalance which finding out that churn customers is extremely less.

Zheng.J, Zhang. D, Stephen C. H, Zhou. X . It is a crucial and fundamental that mines frequent itemsets that are mostly investigated fields. In data mining approach, the quantitative attributes should be appropriately dealt with as well as the Boolean attributes.The structured data organized in a database are mainly dealt with data mining. It does not cover many irregularities, anomalies etc or trends that may otherwise remain undetected under the immense volumes of data. Data mining essentially manages organized information sorted out in a database.

RamaKrishna Vadakattu,Bibek Panda,Swarnim Narayan,Hashal Godhia. A novel algorithm naned Bit-Apriori which is an efficiency of apriori for frequent itemset mining.In this approach the binary string is used to describe the database and the data structure is employed. Bitwise "AND" operation is performed on binary string to support count. This approach scanj.s the database only twice. It also uses tree structure to support count which uses "&" operation.

Grahne G,J. Zhu A novel FP-array technique using FP-trees are more efficient while mining frequent itemset. Time required to traverse FP-tree is reduced.FP-growth method is incorporated using FP-array technique. New algorithms where presented by this algorithm for maximal and close frequent itemsets.MFI tree which is a variation of FP trees where introduced for maximality testing.FP close algorithm which was used for mining close frequent itemsets. Certain optimization techniques are provided further to reduce running time and memory consumption.

IV. METHODOLGY

A. SYSTEM ARCHITECTURE

System design is the way toward: illustrates how the system have to be design to make sure it meets all the requirements.

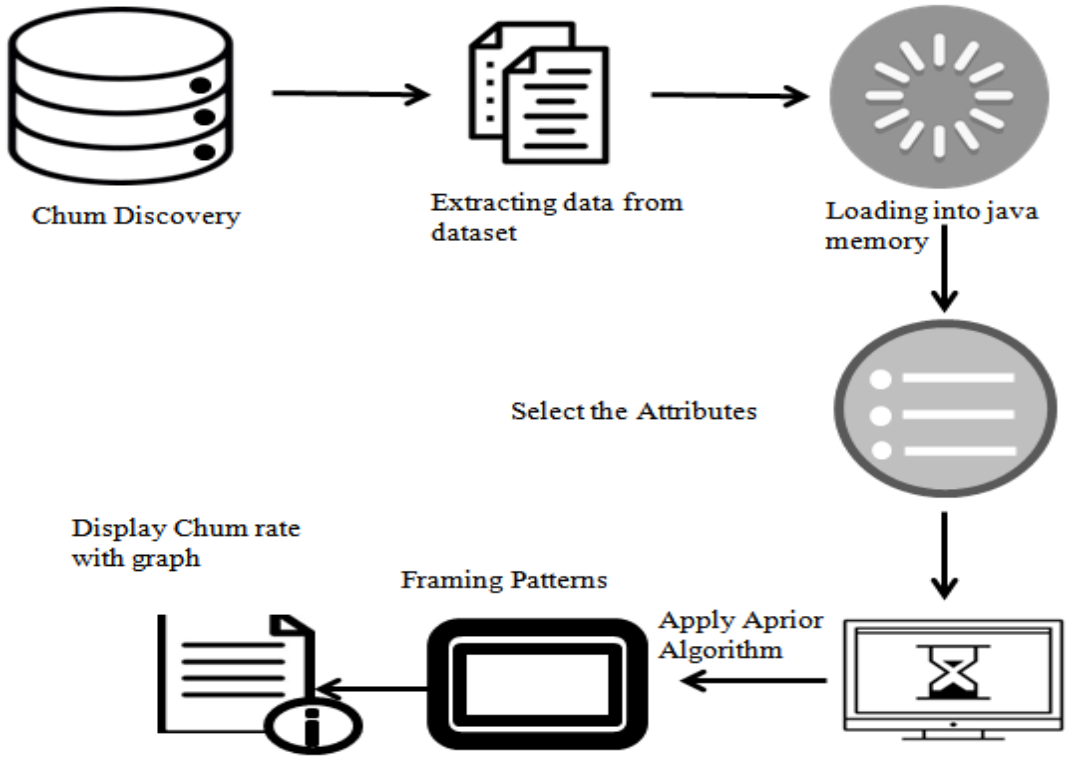

Figure 5.1: System Architecture

The Figure 5.1 shows the system architecture for churn discovery over large telecom data using apriori algorithm. In big data era, the customer churn problem and predicting the number of customer churn in telecom industry aiming to prevent customer churn, first apply Apriori algorithm improves clustering accuracy. Secondly, solve the customer churn problem with big data Apriori and big data algorithms.

A part of a program is called a module. Modules are independently developed and a program contains one or more modules. These modules are combined only when they have link with each other. There can be one more routines in a single module. There are five modules in this project. They are:

A. Admin login module
B. Acquire dataset using POI API to the data structure
C. Selection of the attributes and apply apriori and display apriori frames.
D. Display top ten phone numbers that fall under churn first through alert message. And also display all the phone numbers with churn rate. Display graph of top ten phone numbers that fall under the churn.

The attributes used in this project are listed below.

a) Account Length-Account Length
b) VMail Message-voice mail
c) Day Mins-Day call minutes
d) Eve Mins- evening call minutes
e) Night Mins-night call minutes
f) Intl Mins-international call minutes
g) CustServ Calls-customer service calls
h) Churn-number of churn
i) Int'l Plan-internet plan
j) VMail Plan-voice mail plan
k) Day Calls-number of day calls
l) Day charge-day calls charge
m) Eve Calls-number of evening calls
n) Eve charge-evening call charge
o) Night calls-number of night calls
p) Night Charge-night call charges
q) Intl plan-international call plan
r) Intl charge-international call charges
s) State –customer state
t) Area code- code of area

The flow of data, modelling of process aspects are represented in a graphical representation known as data flow diagram. Data flow diagrams are the primary step that is used to create an overview of the system. Visualization of data processing can be done through data flow diagrams. The following Figures show the data flow diagrams of this project.

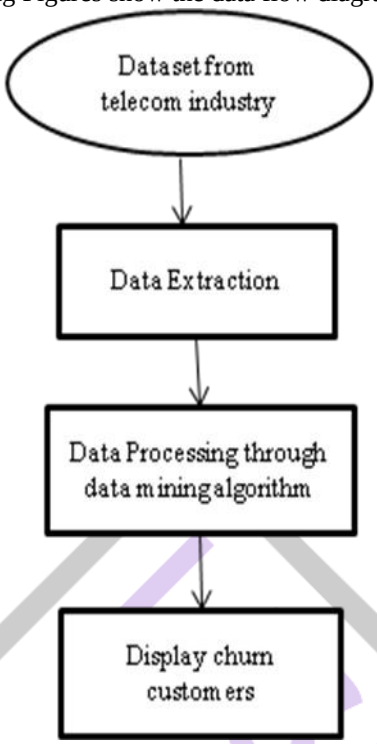

Figure 4.2: DFD Level 0:Overall Interaction of the churn discovery

V. IMPLEMENTATION

Data Extraction

Acquire dataset using POI to the data structures. In this project POI API is used to import the dataset that is acquired in MS Excel Sheet. The packages used to implement POI are org.apache.poi.ss and org.apache.poi.xssf. XSSF is XML Spreadsheet Format which reads and writes the Office Open XML (XLSX) files. XSSFWorkbook is the class which is used to read the data from excel sheet. The jar file that has to be included for this purpose is, poi.ooxml-3.6.jar-this file is used to extract data and perform data processing, apply formulae, and show the results.

Attributes Selection Module and apply Apriori and display apriori Frames

Selection of the attributes and make apriori. In this module the attributes are selected to find the customers who fall under churn. Attributes selected are taken as prioritized attributes. Prioritized attributes are those attributes which the customer use more often and the company is making more profit. After Selecting the Attributes the Apriori Frames are displayed. Firstly Seed Pattern is displayed and Next all the patterns of remaining attributes are displayed. The object used to store the selected attributes is "selatt" and the class name is allSelectedAttributes.

Display phone numbers

Calculate the churn per phone number without ignoring any attribute and display top ten phone numbers that fall under churn first through alert message. And also display all the phone numbers with churn rate.

A. APRIORI PSUEDO CODE

The following figure shows the apriori pseudo code. An influential algorithm for mining frequent itemsets for Boolean association rules is apriori algorithm.

Key Features:

i.Frequent Itemsets: These are item sets which has minimum support (denoted by $L_i$for i^{th}- Itemset).

ii.Apriori Property: These are the subset of frequent itemset must be frequent.

iii.Join Operation: To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself.

```
Apriori(T , ?)
    L1 ?  {large 1 – itemsets}
    k  ?  2
     while Lk-1 ? Ø

        Ck ?  {a ? {b} | a ? Lk-1 ? b ? a} – { c | {s | s ?  c ? | s | =k – 1}

        ? Lk-1} for transactions t ? T
            Ct ?  {c | c ? Ck ? c ?  t}
            For candidates c ? Ct
                  count[c]  ?  count[c] + 1
        Lk ?  {c | c ? Ck ? count[c] =? }
        k ?  k + 1

    Return ? Lk
```

Figure 5.1:Shows the apriori pseudo code

In the proposed work this algorithm work as the following.
i.Collect the dataset. The dataset contains the number of attributes that are used for the proposed work.
ii. Scan the dataset and select the attributes that has to be prioritized.
iii. Create a seed pattern. And again scan the dataset to find the next prioritized attribute and create frames. Continue until all the attributes are included in the frame pattern.
iv. Obtain the threshold value for each attribute using mean weighted average vector These positive and negative variable added and stored in a variable tCount. All the negative value is divided by tCount and stored in a variable tValue. And it is multiplied by 100 and stored in a variable fValue. And the value of the variable fValue is the churn percentage of each phone number.
After calculating the churn percentage of each phone number ,top ten churn customer are displayed first and also each customer churn percentage is displayed. Top ten customer churn rate is displayed in graph.

VI. RESULTS AND DISCUSSION

In this section the outcomes of the project are reported where apriori algorithm is applied to predict accurate customer churn rate. The objective of this project is to make it simpler in finding the churn customer.

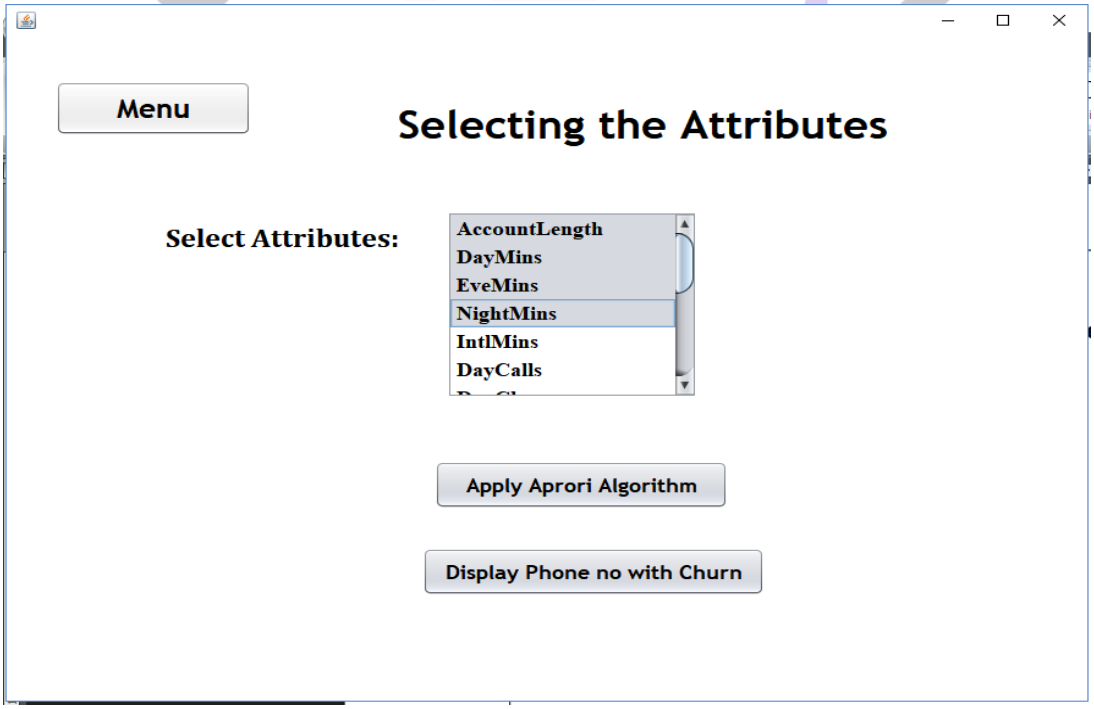

Figure6.1 :Selecting the attributes for prioritizing and perform apriori analysis

Figure 6.1 Shows the attributes selected to perform the apriori analysis. These attributes re selected by the telecom company dynamically.

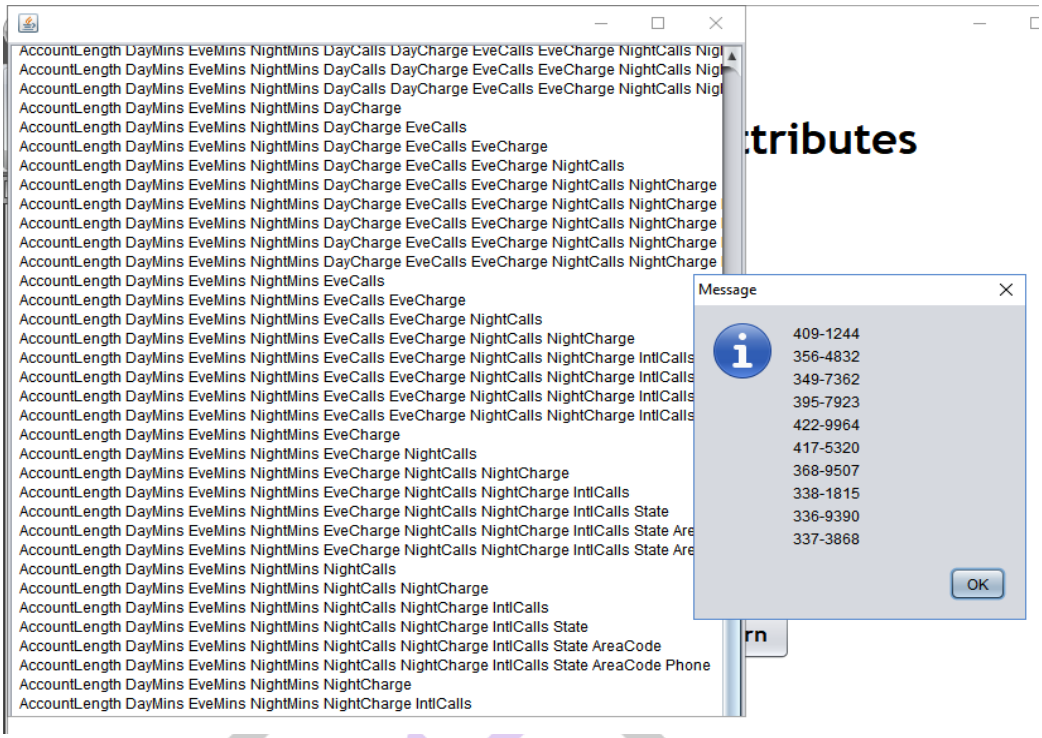

Figure 6.2: Display top ten phone numbers of the churn customers

Figure 6.2 shows the frames patterns. It displays the seed patterns of the selected attributes and continues to display the patterns including remaining attributes. It also shows the all possible apriori frame patterns.the selected attributes will be the seed pattern.seed pattern comparing with remaining attributes display all possible apriori patterns.

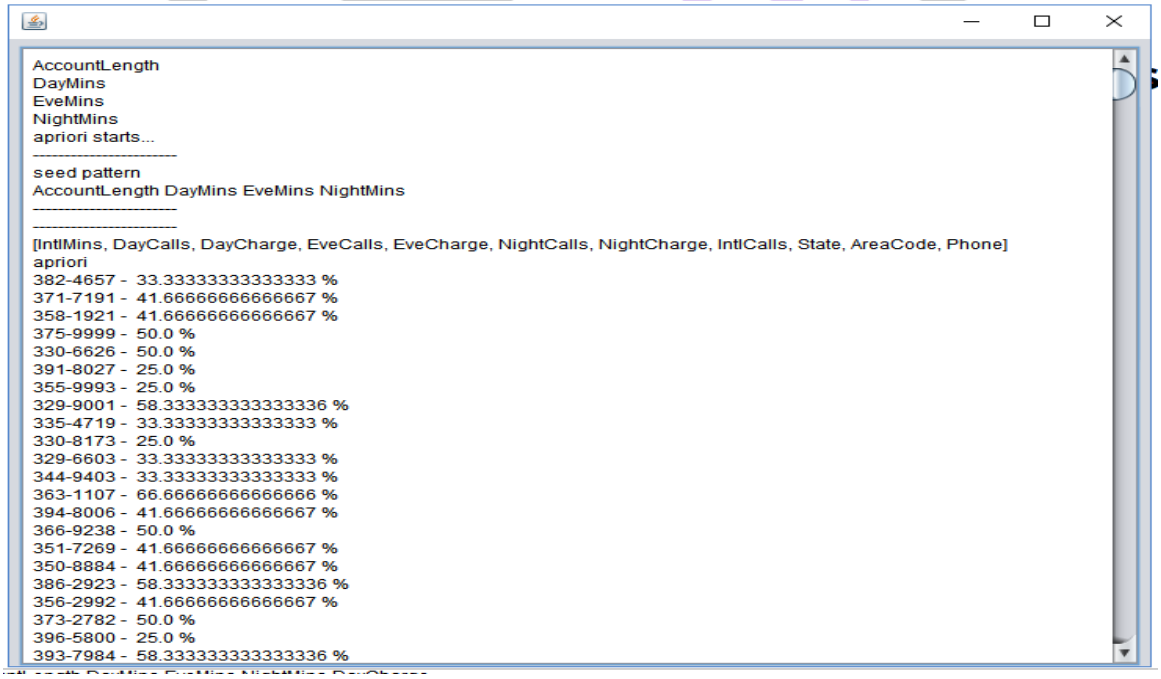

Figure 6.3: Display all the phone numbers with churn rate in percentage

Figure 6.3 shows the seed pattern, selected attributes and all the attributes. It also displays the churn percentage of each customer. Threshold set was 50% for this scenario. Phones number that cross the threshold value is considered as the churn customers

VII. CONCLUSION AND FUTURE WORK

Apriori Algorithm uses full resources in a unified form. It is a simple algorithm when compared to other algorithm mentioned in literature survey. This algorithm is implemented in this project to find the churn customers in a telecom industry. The performance is improved in finding out the churn customers through this technique. A robust churn prediction model can be built in telecom industries experiment with the dataset show the accurate result

The performance of the proposed system is affected on failure, it should scan dataset from start to obtain the result. It requires many number of dataset scans. Proposed method can also be extended to add the functionalities like set theory of redundancy which improves the frequent mining patterns which is very useful in future data analysis apriori algorithm can be applied to other real time data such as crime department, hospital etc with association rule mining .

.

REFERENCES

[1] Mr. Reubeu Sam, Ms. Shakila Shaik , Mr. ArjunKumar Tiwari "Analysis And Prediction Of Churn Customers for Telecommunication Indusry" International Conference I-SMAC 2017.

[2] Dr. M. BalaSubramanian, M. Selvirani "Churn Prediction in Mobile Telecom System Using Data Mining Techniques"International journal Of Scientific and Research Publication-ijsrp,vol-4,Issue 5,page no.5,2014

[3] Ramakrishna Vadakattu, Bibek Panda, Swarnim Narayan, Harshal Godhia. "Enterprise Subscription Churn Prediction" IEEE International Conference on Big data 2015.

[4] Roma Singh, Sonal Chaudhary "Data Mining Approach Using Apriori Algorithm: The Review"IOSR, Vol-4, issue 2, 4, 2012.

[5] Zheng . J, Zhang. D, Stephen C. H, Zhou. X, "An efficient algorithm for frequent itemsets in data mining ", IEEE – 2010.

[6] Grahne G., J. Zhu, "Fast algorithms for frequent itemset mining using FP-Trees", IEEE Transaction on Knowledge and Data Engineering, 17 (10), pp. 1347-1362, 2005.

[7] Review of Apriori Based Algorithms on MapReduce Framework. Accessible from: https://www.researchgate.net/distribution/268512540_Review_of_Apriori_Based_ Algorithms_on_MapReduce_Framework .