“RECOMMENDATION SYSTEM USING APRIORI ALGORITHM”

A

Project Report

submitted in partial fulfillment of the

requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

by

Name	**Roll No.**
Anshika Sharma	**R610217003**
Divyanshu Singh	**R610217007**
Kaustavdeep Goswami	**R610217010**

under the guidance of

Dr. Anurag Jain
Assistant Professor, Department of Virtualization

UPES
UNIVERSITY WITH A PURPOSE

Department of Computer Science & Engineering

Centre for Information Technology

University of Petroleum & Energy Studies

Bidholi, Via Prem Nagar, Dehradun, UK

August- December, 2019

CANDIDATE'S DECLARATION

I/We hereby certify that the project work entitled **"RECOMMENDATION SYSTEM USING APRIORI ALGORITHM"** in partial fulfilment of the requirements for the award of the Degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING with specialization in **Mainframe Technology** and submitted to the Department of Computer Science & Engineering at Center for Information Technology, University of Petroleum & Energy Studies, Dehradun, is an authentic record of my/ our work carried out during a period from **August, 2019 to November, 2019** under the supervision of **Dr. Anurag Jain, Assistant Professor, Department of Virtualization.**

The matter presented in this project has not been submitted by me/ us for the award of any other degree of this or any other University.

(Anshika Sharma)
Roll No. R610217003
(Divyanshu Singh)
Roll No. R610217007
(Kaustavdeep Goswami)
Roll No. R610217010

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: _____________2019

(Dr.Anurag Jain)
Project Guide

Dr. Neelu Jyoti Ahuja
HOD, Department of Systemics
Center for Information Technology
University of Petroleum & Energy Studies
Dehradun – 248 001 (Uttarakhand)

ACKNOWLEDGEMENT

We wish to express our deep gratitude to our guide **Dr.Anurag Jain**, for all advice, encouragement and constant support he has given us throughout our project work. This work would not have been possible without his support and valuable suggestions.

We sincerely thank to our respected Program Head of the Department, **Dr.Neelu Jyoti Ahuja**, for his great support in doing our project in **Data Mining and Data Warehousing** at **CIT**.

We are also grateful to **Dr. Manish Prateek, Dean SoCS**, UPES for giving us the necessary facilities to carry out our project work successfully.

We would like to thank all our **friends** for their help and constructive criticism during our project work. Finally, we have no words to express our sincere gratitude to our **parents** who have shown us this world and for every support they have given us.

Name	**Anshika Sharma**	**Divyanshu Singh**	**Kaustavdeep Goswami**
Roll No.	**R610217003**	**R610217007**	**R610217010**

ABSTRACT

The world now stands on data. From morning till night, each and every of our work is related to data. Using this much of data, we can find some hidden patterns and relationships among these data to find out customer's marketing behavior. In this project we are focusing to get the prior knowledge of customer's behavior, that can be gained from the data assembled in common areas such as data warehouses. We are using Apriori algorithm to implement our recommendation system using different models and mining techniques. Market basket analysis is an important component of analytical system in retail organizations to determine the placement of goods, designing sales promotions for different segments of customers to improve customer satisfaction and hence the profit of the supermarket. These issues for a leading supermarket are addressed here using frequent itemset mining. The frequent itemsets are mined from the market basket database using the efficient Apriori algorithm and then the association rules are generated. As a result of this project, we will do descriptive modelling about values of data using known results found from different historical data to ultimately cut costs and increase revenue.

Keywords: Data Mining, Apriori algorithm, Market-basket analysis, Association rule

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. History

The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as "knowledge discovery in databases".

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict future outcomes. Using a broad range of techniques, we can use information to increase revenues, cut costs, improve customer relationships and reduce risks. The more complex the data sets, the more potential there is to uncover relevant insights. Data can be mined whether it is stored in flat files, spreadsheets, database tables, or some other storage format. The important criteria for the data is not the storage format, but its applicability to the problem to be solved.

Data mining is accomplished by building models. A model uses an algorithm to act on a set of data. The notion of automatic discovery refers to the execution of data mining models.

The main purpose of data mining process is to discover those records of information and summarize it in a simpler format for the purpose of increasing customer loyalty, finding hidden profitability, finding patterns and increasing customer satisfaction.

Informed decision can be made easily about product placement, pricing, promotion, profitability and also finds out, if there are any successful products that have no significant related elements. Similar products can be found so those can be placed near each other or it can be cross-sold.

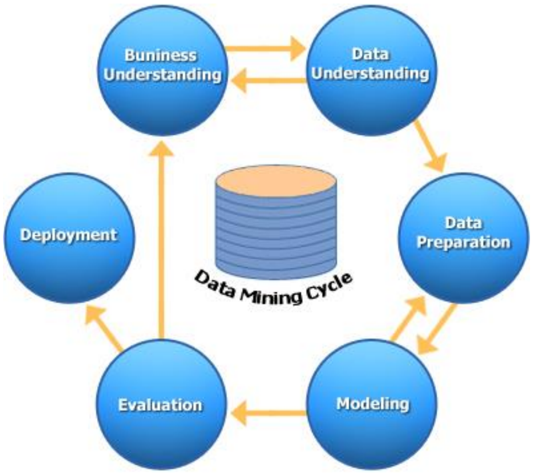


Fig 1.1: The Data Mining Cycle

1.2 System Requirements

1.2.1 Software Requirements

Programming Language	C
Operating System	Linux
Editor	Vi Editor
Compiler	gcc

Table1.1: List of software requirements

1.2.2 Hardware Requirements

Processor	Intel(R) Core(TM) i3-5200U CPU @1.79GHz Or above.
RAM	4.00 GB
External Devices	Compatible mouse and keyboard

Table 1.2: List of hardware requirements

1.3 PROBLEM STATEMENT

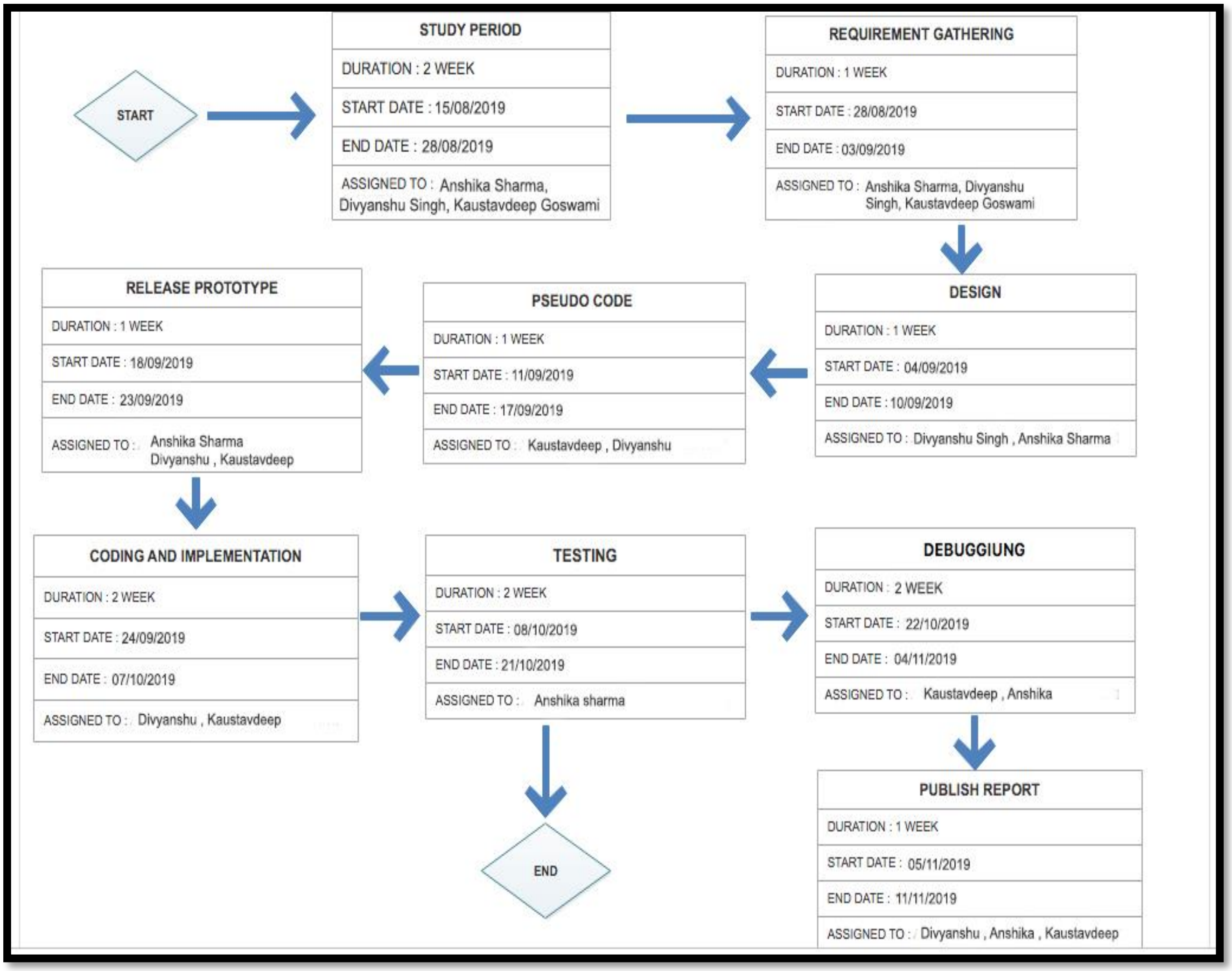
Once it is known that customers who buy one product are likely to buy another, it is possible for the company to market the products together, or to make the purchasers of one product, the target prospects for another. If customers who purchase tomatoes are already likely to purchase onions, they'll be even more likely to if there happens to be an onion crate just beside the tomato aisle. By targeting customers, the effectiveness of marketing can be significantly increased – regardless of if the marketing takes the form of in-store displays, catalog layout design, or direct offers to customers. This is the purpose of market basket analysis – to improve the effectiveness of marketing and sales tactics using customer data already available to the company.

1.4 Objectives

To create Recommendation system based on Data Mining

- To generate patterns and relationships among data elements, render relevant information, which may increase organizational revenue.
- To analyze and draw conclusions about trends in consumer's marketing behavior.

1.5 Pert Chart



STUDY PERIOD
DURATION : 2 WEEK
START DATE : 15/08/2019
END DATE : 28/08/2019
ASSIGNED TO : Anshika Sharma, Divyanshu Singh, Kaustavdeep Goswami

REQUIREMENT GATHERING
DURATION : 1 WEEK
START DATE : 28/08/2019
END DATE : 03/09/2019
ASSIGNED TO : Anshika Sharma, Divyanshu Singh, Kaustavdeep Goswami

DESIGN
DURATION : 1 WEEK
START DATE : 04/09/2019
END DATE : 10/09/2019
ASSIGNED TO : Divyanshu Singh , Anshika Sharma

PSEUDO CODE
DURATION : 1 WEEK
START DATE : 11/09/2019
END DATE : 17/09/2019
ASSIGNED TO : Kaustavdeep , Divyanshu

RELEASE PROTOTYPE
DURATION : 1 WEEK
START DATE : 18/09/2019
END DATE : 23/09/2019
ASSIGNED TO : Anshika Sharma Divyanshu , Kaustavdeep

CODING AND IMPLEMENTATION
DURATION : 2 WEEK
START DATE : 24/09/2019
END DATE : 07/10/2019
ASSIGNED TO : Divyanshu , Kaustavdeep

TESTING
DURATION : 2 WEEK
START DATE : 08/10/2019
END DATE : 21/10/2019
ASSIGNED TO : Anshika sharma

DEBUGGIUNG
DURATION : 2 WEEK
START DATE : 22/10/2019
END DATE : 04/11/2019
ASSIGNED TO : Kaustavdeep , Anshika

PUBLISH REPORT
DURATION : 1 WEEK
START DATE : 05/11/2019
END DATE : 11/11/2019
ASSIGNED TO : Divyanshu , Anshika , Kaustavdeep

Fig 1.2: Pert Chart

2. SYSTEM ANALYSIS

2.1 Motivation

A host of technological advances have resulted in generating a huge amount of electronic data, and have enabled the data to be captured, processed, analyzed, and stored rather inexpensively.

The need to understand huge, complex, information-rich data sets is important to virtually all fields in business, science and engineering. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming vital in today's increasingly competitive world. Such data (typically terabytes in size) is often stored in data warehouses and data marts.

Apriori algorithm is one of the most used algorithms for implementing Association rules because of its efficiency, it produces accurate information with huge amount of data without being complex.

2.2 Proposed System

In this project we are developing a code in C which will find hidden patterns using huge amount of data implementing data mining technique. This will act as a recommendation system of Market stores applies on market-basket analysis.

System working is mainly divided into following parts:

- Available Data.
- Admin applies algorithm on data to find the patterns.
- Admin creates association rules to be provided to store sellers.
- Sellers mark data items accordingly to increase revenue.

To implement this technique using C language and Apriori algorithm.

2.3 Modules

This project is basically divided into two parts that is Seller side and Data Admin side. Seller collects the list of items bought by customers and send it to the data admins to generate the association rules so that seller can apply them to arrange the items accordingly.

1. Store-Seller Login

    - Review Stocks
    - Edit/Update/Delete –Stock report
    - Review Recommendations
    - Release item list.

2. Administrator (Data Science)

    - Review–Stock report (to find patterns)
    - Add new items
    - Generate reports
    - Warehouses Details
    - Item Details
    - View Selected Patterns
    - Make Recommendations
    - Approves Selected Patterns
    - Send Selected-data-patterns to Sellers

3. Design

3.1 Flow Diagram

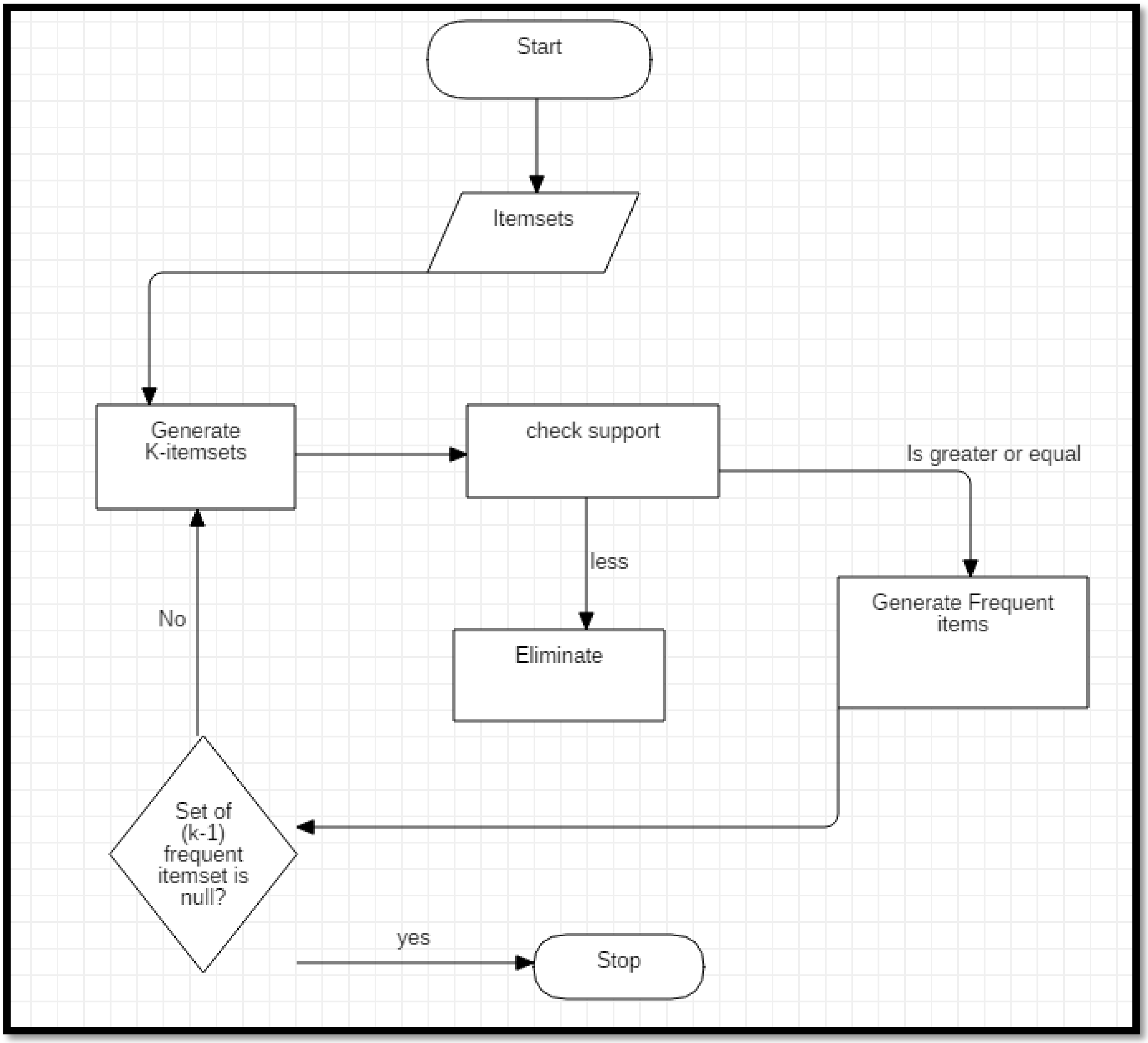


Fig. 3.1: Flow diagram

3.2 Use-Case diagram

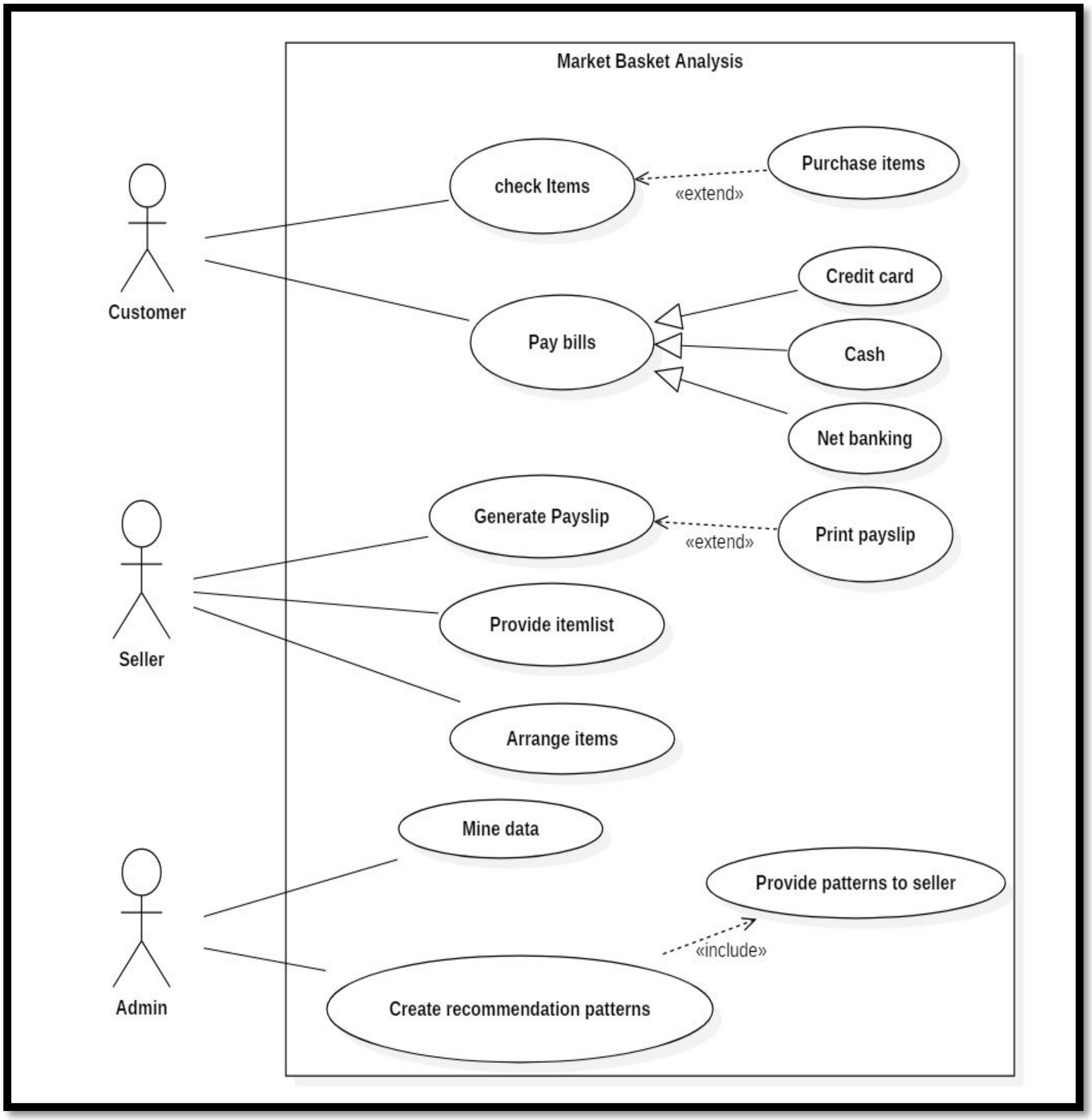


Fig. 3.2: Use case diagram

3.3 Data-Flow Diagram

3.3.1 Store-seller Side (Level 0)

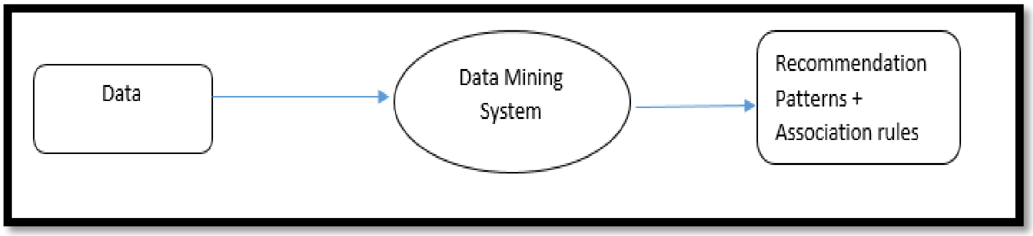


Fig. 3.3.1: Level 0 data flow diagram

3.3.2 Administrator Implementations (Level 1)

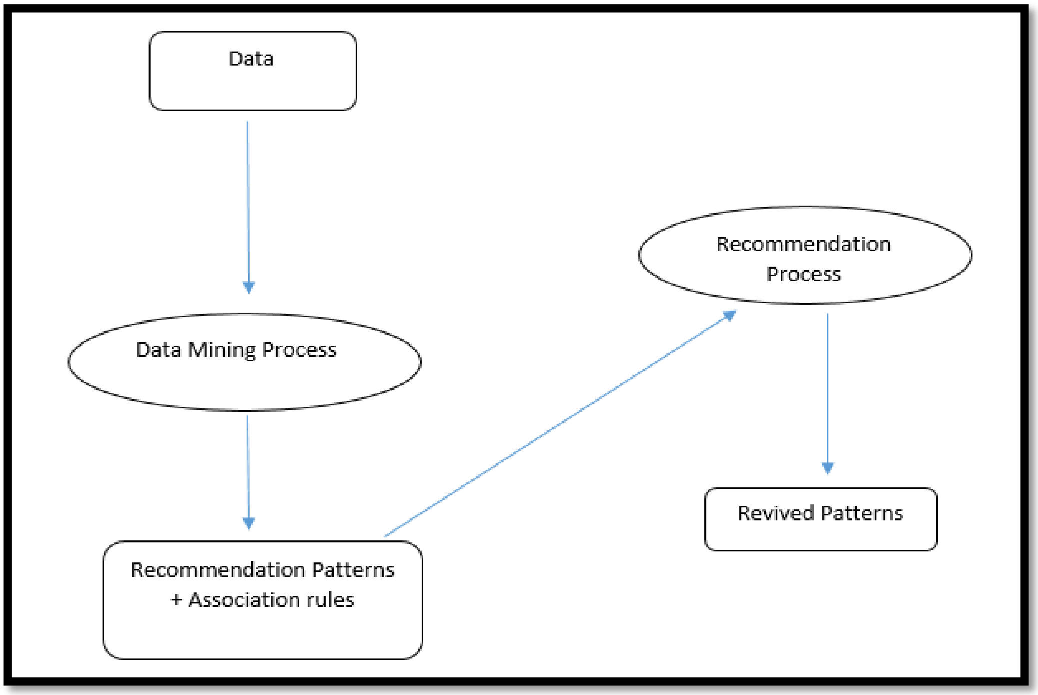


Fig. 3.3.2: Level 1 data flow diagram

4. Implementation

4.1 Basic concepts

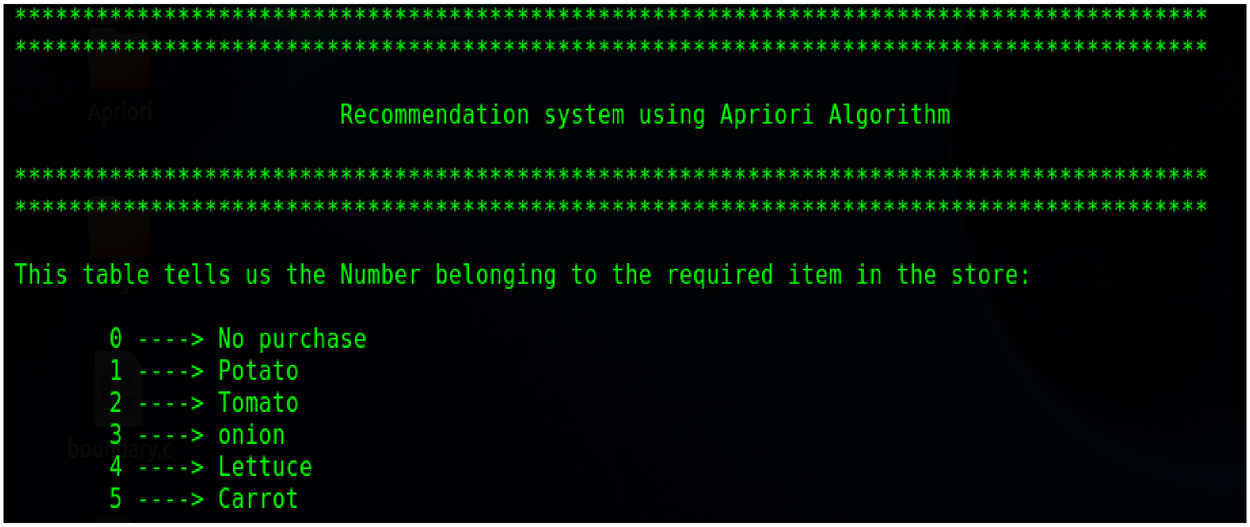
Apriori [3] is based on the Apriori property and the Association rule generation procedure of the Apriori algorithm. Initially, the binary data is transformed into real domain using linear Wiener transformation. The Wiener transformed data is partitioned using the multi-pass Kmeans algorithm. Then the Apriori procedure is executed for the K clusters in which the sets of items which are greater than minimum support (min_sup) are found iteratively. Using these frequent itemsets based on confidence, Association rules are derived. The items in the clusters are very similar, so that multiple and high informative frequent itemsets are effectively generated in the Apriori algorithm.

4.1.1 Algorithm

- Join step: is generated by joining with itself
- Prune Step: any (k-1) item set that is not frequent cannot be a subset of a frequent k-item set
- Pseudo-code:

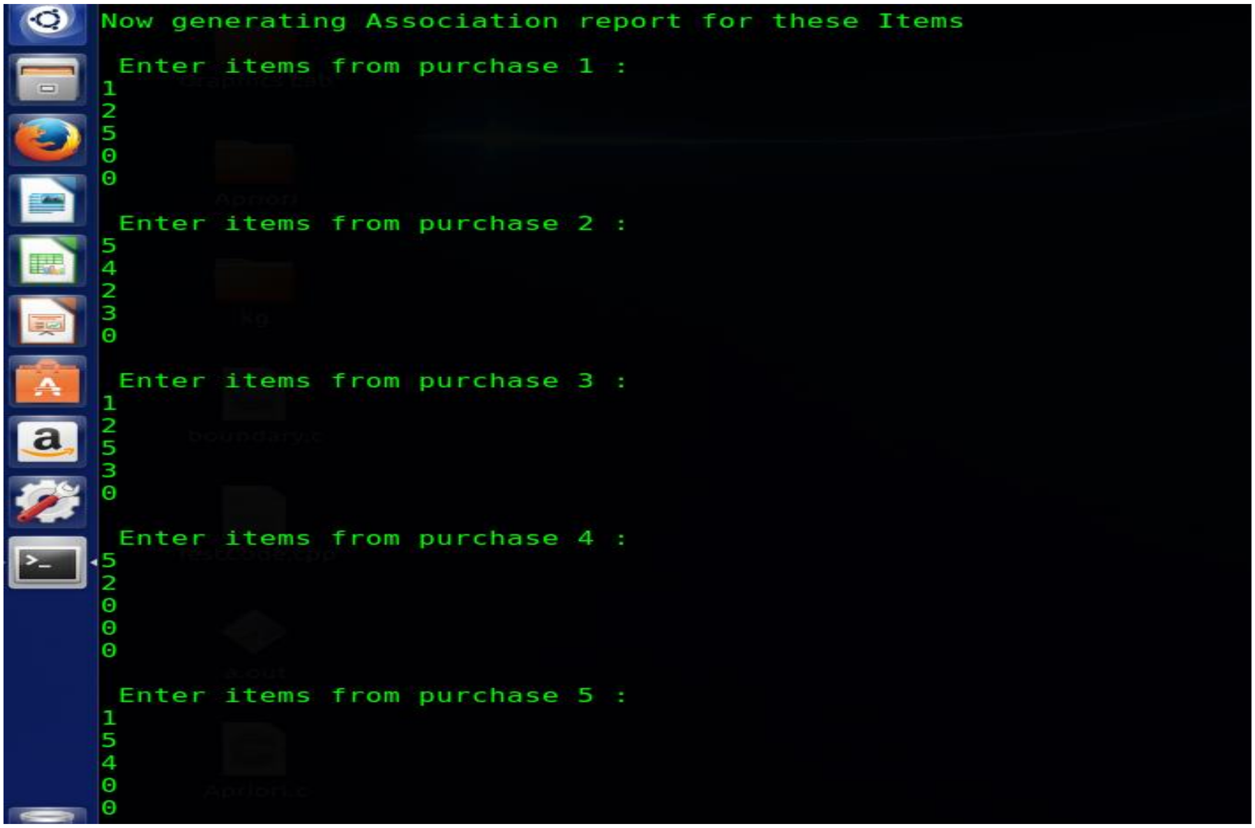
```
$C_k$: Candidate item set of size k
    $L_k$: Frequent item set of set k
    $L_1$ = {frequent items};
For (K=1; $L_k$! = $\phi$; k++) do begin
$C_{k+1}$ = candidate generated from $L_k$;
For each transaction t in database do
    Increment the count of all candidates in $C_{k+1}$
    Those are contained in t
$L_{k+1}$ = candidate in $C_{k+1}$ with min_support
  End
Return $L_{k-1}$; when $L_k$=$\phi$
```

4.1.2 Output Screens



```
****************************************************************************************
****************************************************************************************

                        Recommendation system using Apriori Algorithm

****************************************************************************************
****************************************************************************************

This table tells us the Number belonging to the required item in the store:

       0 ----> No purchase
       1 ----> Potato
       2 ----> Tomato
       3 ----> onion
       4 ----> Lettuce
       5 ----> Carrot
```

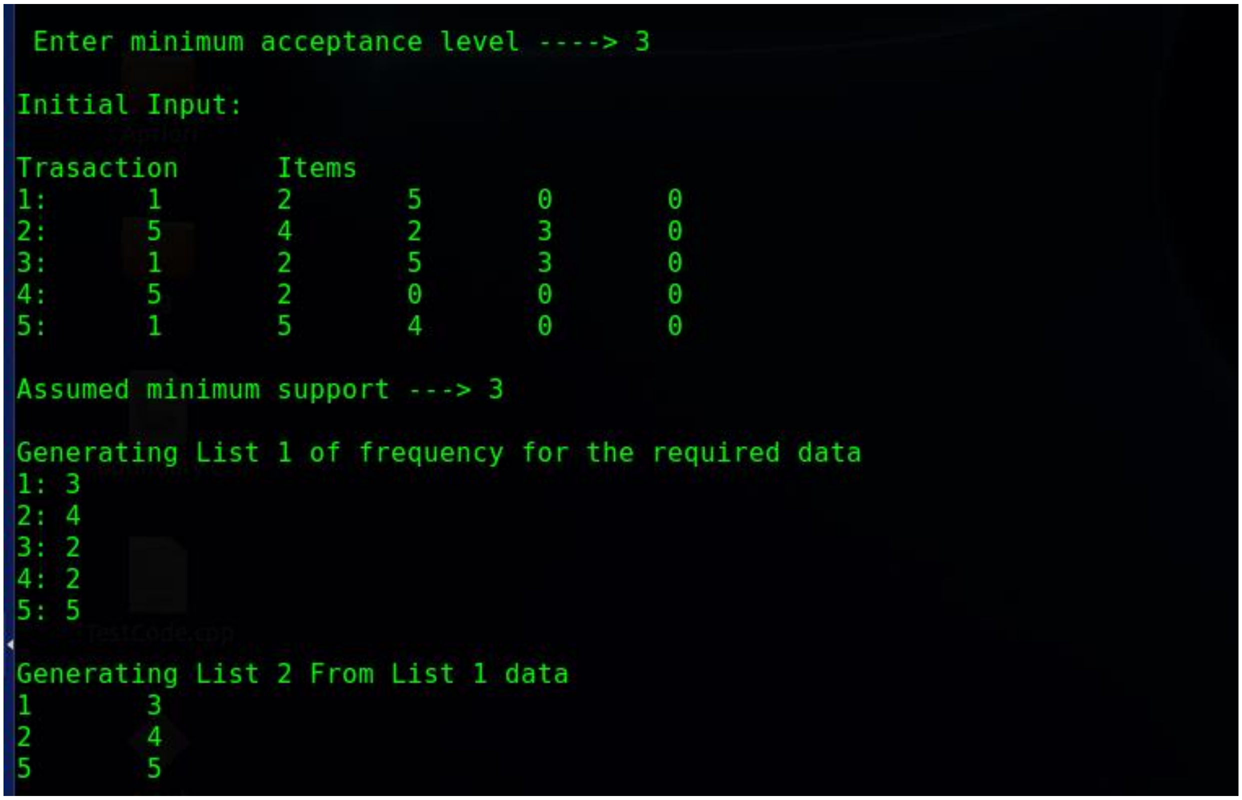
(Fig 4.1) This is the First Screen that tells you the Data item list.



```
Now generating Association report for these Items
 Enter items from purchase 1 :
1
2
5
0
0

 Enter items from purchase 2 :
5
4
2
3
0

 Enter items from purchase 3 :
1
2
5
3
0

 Enter items from purchase 4 :
5
2
0
0
0

 Enter items from purchase 5 :
1
5
4
0
0
```

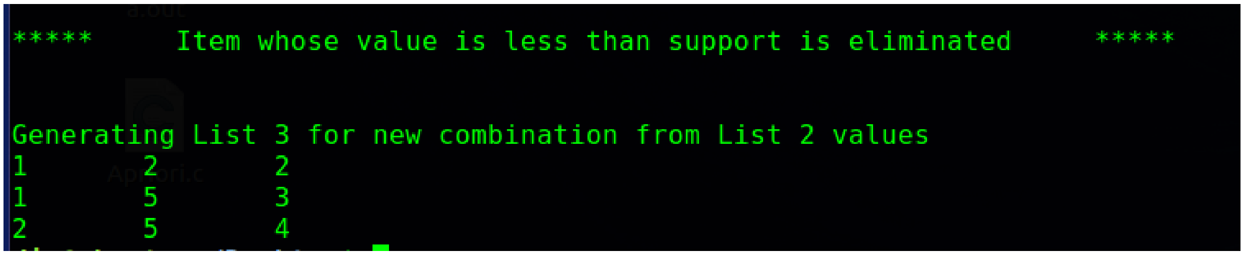
(Fig 4.2) This Screen Accepts the transactions from the admin to generate patterns.

Here we accept the Acceptance level (support value) and the initial input is displayed with the first list of item including their frequencies and list2 is also generated from list1 with the applyied rule of support value.



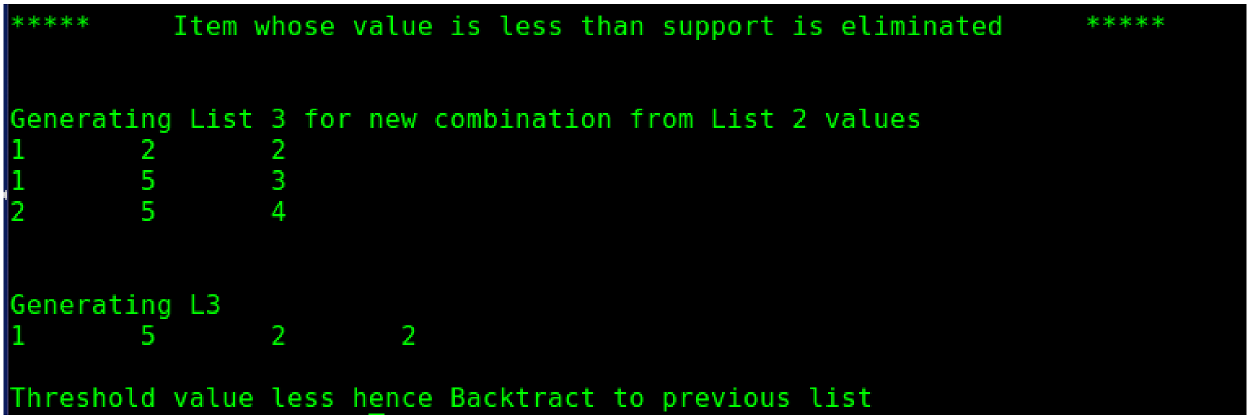
```
 Enter minimum acceptance level ----> 3

Initial Input:

Trasaction      Items
1:      1       2       5       0       0
2:      5       4       2       3       0
3:      1       2       5       3       0
4:      5       2       0       0       0
5:      1       5       4       0       0

Assumed minimum support ---> 3

Generating List 1 of frequency for the required data
1: 3
2: 4
3: 2
4: 2
5: 5

Generating List 2 From List 1 data
1       3
2       4
5       5
```

(Fig 4.3) List 2 data items displayed



```
*****     Item whose value is less than support is eliminated     *****


Generating List 3 for new combination from List 2 values
1       2       2
1       5       3
2       5       4
```

(Fig 4.4) This list shows the List 3 generated from the output of list 2.



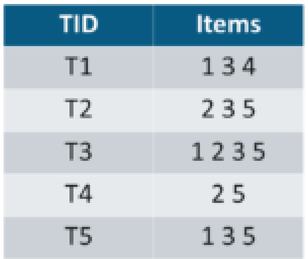
```
*****     Item whose value is less than support is eliminated     *****


Generating List 3 for new combination from List 2 values
1       2       2
1       5       3
2       5       4


Generating L3
1       5       2       2

Threshold value less hence Backtract to previous list
```

(Fig 4.5) The final list shows the List 4 generated from the output of list 3.

4.2 Methodology

4.2.1 Applying Iterations

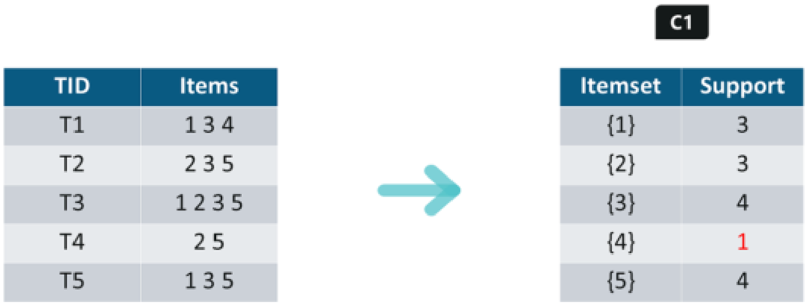
With the following data of a store.



TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

(Fig 4.6)

Iteration 1: The support value is 2 and create the item sets of the size of 1 and calculate their support values.



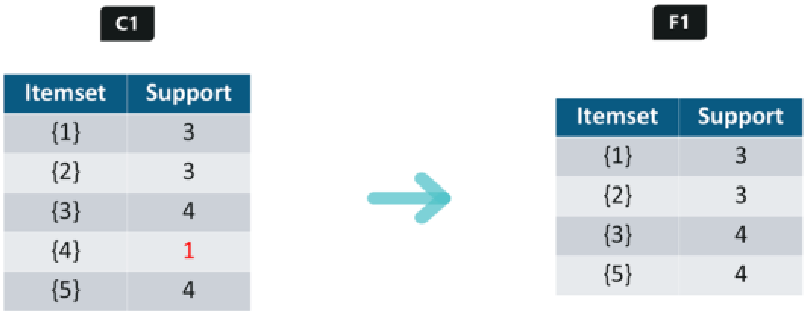
TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

C1

Itemset	Support
{1}	3
{2}	3
{3}	4
{4}	1
{5}	4

(Fig 4.7)

Item 4 has a support value of 1 which is less than the min support value. **Discard {4}** in the upcoming iterations to get the final Table F1.



C1

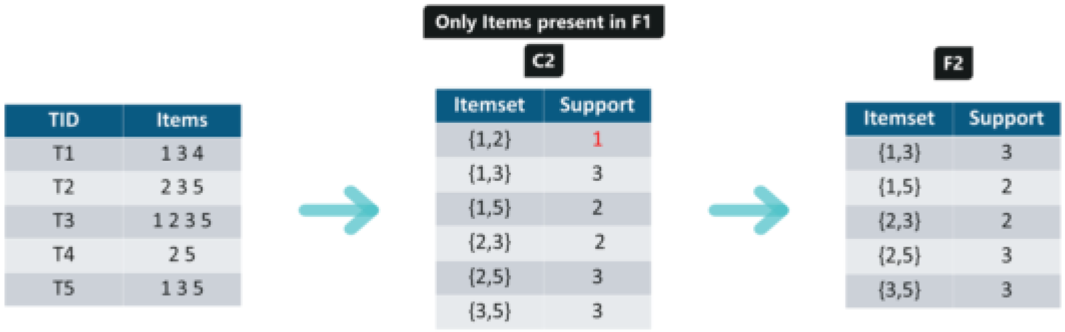
Itemset	Support
{1}	3
{2}	3
{3}	4
{4}	1
{5}	4

F1

Itemset	Support
{1}	3
{2}	3
{3}	4
{5}	4

(Fig 4.8)

Iteration 2: Create itemsets of size 2 and calculate their support values. All the combinations of items set in F1 are used in this iteration.



TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

Only Items present in F1

C2

Itemset	Support
{1,2}	1
{1,3}	3
{1,5}	2
{2,3}	2
{2,5}	3
{3,5}	3

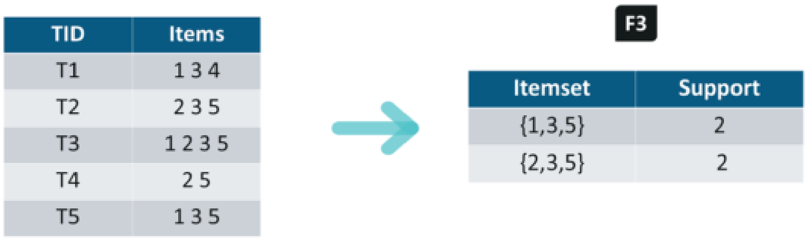
F2

Itemset	Support
{1,3}	3
{1,5}	2
{2,3}	2
{2,5}	3
{3,5}	3

(Fig 4.9)

Itemsets having Support less than 2 are eliminated again. In this case **{1,2}.**

Iteration 3: Discard **{1,2,3}** and **{1,2,5}** as they both contain **{1,2}.** This is the main highlight of the Apriori Algorithm.



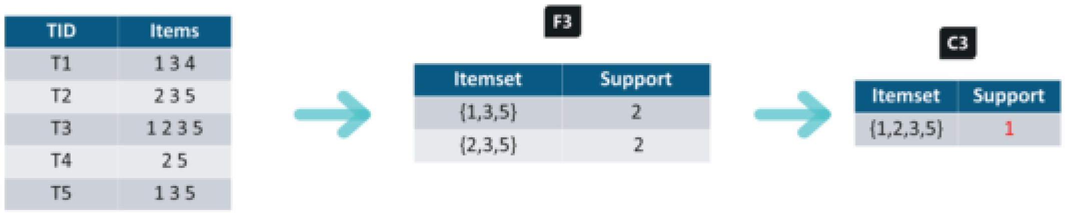
TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

F3

Itemset	Support
{1,3,5}	2
{2,3,5}	2

(Fig 4.10)

Iteration 4: Using sets of F3, create C4.



TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

F3

Itemset	Support
{1,3,5}	2
{2,3,5}	2

C3

Itemset	Support
{1,2,3,5}	1

(Fig 4.11)

4.2.2 Calculate Confidence values

Since the Support of this itemset is less than 2, stop here and the final itemset is F3. Calculate the confidence values.

With F3 we get the following itemsets:

For I = {1,3,5}, subsets are {1,3}, {1,5}, {3,5}, {1}, {3}, {5}
For I = {2,3,5}, subsets are {2,3}, {2,5}, {3,5}, {2}, {3}, {5}

Applying Rules: Create rules and apply them on itemset F3. Now let's assume a minimum confidence value is **60%.**

For every subsets S of I, you output the rule

- S –> (I-S) (means S recommends I-S)
- if **support(I) / support(S) >= min_conf value**

4.2.3 Applying association rules

{1,3,5}

Rule 1: {1,3} –> ({1,3,5} – {1,3}) means 1 & 3 –> 5

Confidence = support(1,3,5)/support(1,3) = 2/3 = **66.66% > 60%**

Hence Rule 1 is **Selected**

Rule 2: {1,5} –> ({1,3,5} – {1,5}) means 1 & 5 –> 3

Confidence = support(1,3,5)/support(1,5) = 2/2 = **100% > 60%**

5.Limitation

- Sometimes, it may need to find a large number of candidate rules which can be computationally expensive.
- Calculating support is also expensive because it has to go through the entire database.
- Apriori algorithm can be very slow and the bottleneck is candidate generation.
- For example, if the transaction DB has 104 frequent 1-itemsets, they will generate 107 candidate 2-itemsets even after employing the downward closure.
- To compute those with sup more than min sup, the database need to be scanned at every level. It needs *(n +1)* scans, where *n* is the length of the longest pattern.
-

6.Future Enhancements

From the above results it is observed that:

- Hash-based itemset counting: A k-itemset whose corresponding hashing bucket count is below the threshold cannot be frequent
- Transaction reduction: A transaction that does not contain any frequent k-itemset is useless in subsequent scans
- Partitioning: Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB.
- Dynamic itemset counting: add new candidate itemsets only when all of their subsets are estimated to be frequent

7.Conclusion

Apriori algorithm effectively generates highly informative frequent itemsets and association rules for the Market Stores. Market Stores widely used the market basket analyses to manage the placement of goods in their store layout. Related products are placed together in such a manner that customers can logically find items he/she might buy which increases the customer satisfaction and hence the profit. Customers are segmented and association rules are separately generated to satisfy their specific needs in a cost effective manner using some special promotions for the common groups. From the results it is shown that the market basket analysis using Apriori algorithm for Market stores improves its overall revenue.

8.References

[1] Sheng Chai ; Jia Yang ; Yang Cheng, The Research of Improved Apriori Algorithm for Mining Association Rules, 2007 International Conference on Service Systems and Service Management
Year: 2007 | Conference Paper | Publisher: IEEE , pp. 345-353.

[2] A.A. Raorane, R.V. Kulkarni, B.D. JitkarAssociation Rule – Extracting Knowledge Using Market Basket Analysis, Research Journal of Recent Sciences, 1 (2) (2012), pp. 19-27

[3] Patcharin Ponyiam, Somjit Arch-int, "Customer Behavior Analysis Using Data Mining Techniques", Application for Technology of Information and Communication (iSemantic) 2018 International Seminar on, pp. 549-554, 2018.

[4] Wan Faezah Abbas, Nor Diana Ahmad, Nurlina Binti Zaini, "Discovering Purchasing Pattern of Sport Items Using Market Basket Analysis", Advanced Computer Science Applications and Technologies (ACSAT) 2013 International Conference on, pp. 120-125, 2013.

[5]Md. Mahamud Hasan, Sadia Zaman Mishu, "An Adaptive Method for Mining Frequent Itemsets Based on Apriori And FP Growth Algorithm", Computer Communication Chemical Material and Electronic Engineering (IC4ME2) 2018 International Conference on, pp. 1-4, 2018.

[6] V. Saurkar Anand, V. Bhujade, P. Bhagat, A. KhapardeA Review Paper on various Data Mining Techniques
International Journal of Advanced Research in Computer Science and Software Engineering, 4 (4) (2014), pp. 98-101

Synopsis Draft verified by

Project Guide	**HOD**
Dr. Anurag Jain	**Dr. Neelu Jyoti Ahuja**
(Dept. of Virtualization)	**(Dept. of Systemics)**