Data Mining – A Review on Intelligent Prediction Model in Context of Retail Industry

Dipika S Patel [1], Dr. Hitesh Raval [2]
[1] *Shri Sarvajanik M.Sc(CA&IT)College, Mehsana, Hemchandracharya North Gujarat University, Patan, India*
2 Department of Computer Science, Sankalchand Patel University, Visnagar, India
(E-mail: dipika.pismehsana@gmail.com)

Abstract— The augmentation growth of artificial intelligence in present zestful business environs has upsurge the demand and dire of flourishing trading firm to be able to respond swiftly to the fluctuating market ultimatum both distinctly and intercontinentally, by make use of the newest data mining techniques of takeout formerly unspecified and potentially useful statistics from huge resources of primary data. Here, I will pinpoint the algorithms of Data Mining in Retail Business, and its pros and cons on consumers.

Keywords—*DataMining,RetailBusiness,Algorithm*
Introduction

I. INTRODUCTION

Database can be simply stated as the collection of records electronically, which results in fruitful information. The grouped data can be accessed, altered, managed, controlled and organized to carry out different data-processing operations. Emerging Database [2] Technologies like Centralized database, Distributed database, Personal database, End-user database, Commercial database, Non SQL database, Operational database, Relational databases are used by organizations to manage volume of day-to-day data. Data warehouse is the place where the data are stored and managed for an efficient database. [1] Data warehousing outlines the architectures and tools used by trading executives to systematically organize, understand, and use their data to make strategic decisions.

Data mining is important and helpful in trading report for hike slot of consumption, merchandise throughput, and irregular transactions, this will help in downscale the loss comeuppance to internal deceit, around 40 to 50% of the catalogue, suffered by the retail businesses around the country [3]. Data mining can also help treat uncommon patterns entailing refunds, discounts, price overrides, credit cards, store cards, debit cards, staff discounts, voids, reversals, overage and shortages due to stock listed as damaged or defective, thus making retail fraud detection much easier, accurate, timely and economically. Data mining and e-Commerce provides numerous opportunities for technical communications and confederation between professionals prepared to amass new skills and expertise.

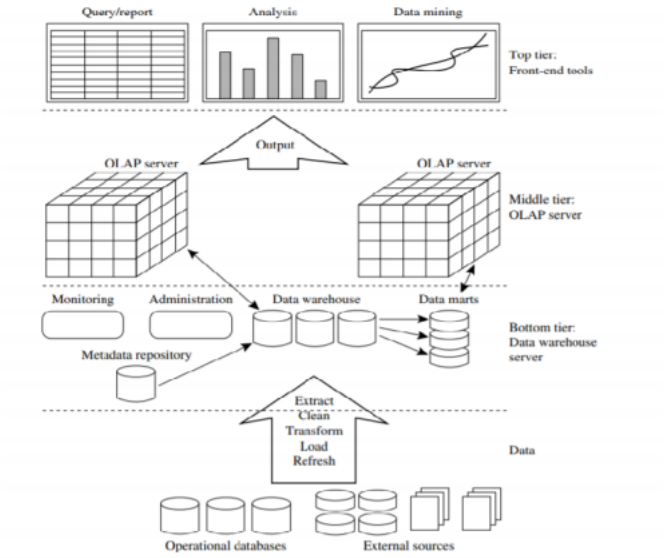

Fig.1. A three-tier data warehousing architecture

II. DATA MINING

Data mining [4] is a useful source to collect the behaviour and potential of the customer. It discovers information within the data that queries and reports can't effectively reveal. [3] Data mining is preferred to analyzing the defined data in different ways, and create it into a single useful content. This technic is very effective to optimize the customers of similar taste.

A. Data mining and its types

In data mining, association rules are created by analyzing data for frequent patterns, then using the support and confidence to locate the most important relationships within the data. [7] Data mining are mainly focused for four main purposes: (1) to enhance customer acquisition and retention; (2) to enhance fraud; (3) to analyses internal inefficiencies and then renovate operations, and (4) to outline the undetermined ground of the social network.

i. Sequential Analysis – ordered list of set of items.
ii. Classification – checks for new pattern, may vary the order also.
iii. Clustering – It groups a set of entity and clump them based on their similarity.

iv. Fostering – This will unlock the patterns in data that can leads to reasonable prophecy about the future
v. Association – This happens when, occurrence is being lined up in single event.

B. Time-oriented data/Temporal data

Temporal data refers to data, which defines the state in time. Process in Data Mining is carrying out by a common set of steps in process. The hardware requirement is based on database system and mining tools. Based on the type of data sampling and intuition is applied. Next analyzing is processed, where the significance and the trends are determined. The next immediate step is interpretation. It includes, business cycles, seasonality and the population the pattern applies to. Finally, exploitation is done, where both business and technical activity is processed. Prediction is more recommended way to exploit.

III. TYPES OF OUTPUT

Datamining technique generally produce an output like

- Buying the patterns of the client; consortium among client enumeration characteristics; predicting on client's response.
- Patterns of crooked credit card utilization; defining the "loyal" customers; credit card investment by group of clients; foretell of customers who are willing to alter their credit card connection.
- Predicting who will purchase the new policies in insurance; behavior patterns of high-risk customers; supposition of crooked deportment.
- Attribute of patient behavior and estimating how frequent they visit the office.

The datamining process includes the following tools:

- Rapid Miner
- Mahout
- Orange
- Weka

IV. ALGORITHEM

It is a set of heuristic data which is molded for calculation which defines the model from data. Below is the frequent pattern mining algorithm.

A. Apriori algorithm:

This algorithm is a very first algorithm which has been published to defined frequent item set mining. In later years it was reconstructed by R Agarwal and R Srikath and called it as Apriori algorithm. To reduce the search space, this algorithm utilizes 'Join' and 'Prune'.

The probability that item I is not frequent is if:

• P (I) is less than the minimum support threshold.

• P (I+A) is less than the minimum support threshold, then I+A is not frequent, where A also belongs to itemset.

• If an itemset set has value less than minimum support then all of its supersets will also fall below min support, and thus can be ignored. This property is called the Antimonotone property

a. Join: This step generates (K+1) itemset from K-item sets by joining each item with itself.
b. Prune Step: This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate item sets.

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

TABLE 1: Transactional data sample

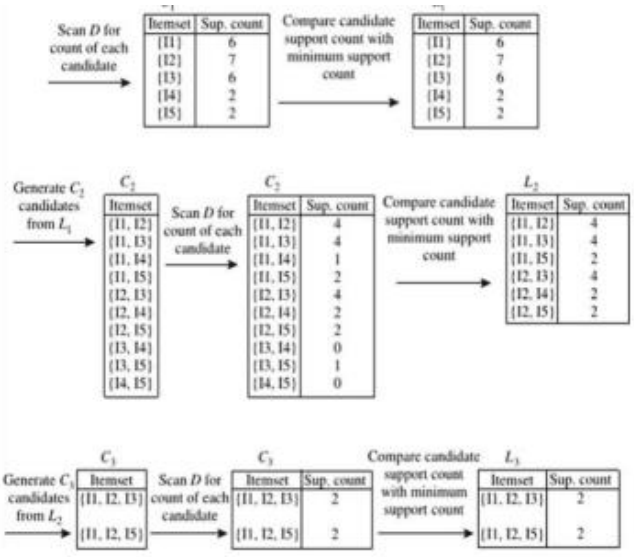

Fig.2. Generation of candidate itemset and frequent itemset.

1. Advantages: It reduces the size of candidate item sets significantly. It provides a good performance gain.
2. Disadvantages: A large number of candidate itemset may still need to be generated if the total count of a frequent k-item sets increases. The entire database is required to be scanned repeatedly and a huge set of candidate items are required to be verified using the technique of pattern matching.

B. FP-Growth algorithm:

Frequent Pattern Growth (FP-Growth) algorithm data are being represented in the form of tree, hence it is called as frequent pattern tree. The structure of tree will maintain the associate in common to item set.

Following are the steps considered in frequent pattern:

a. Step 1 – Scanning of database.
b. Step 2 – Construct the FP tree.
c. Step 3 – Scan and examine the data transaction.
d. Step 4 – The obtained data base is examined.
e. Step 5 – Incrementing the count of item set.
f. Step 6 – Mining of created FP tree.
g. Step 7 – Based on item set path, FP tree is constructed.
h. Step 8 – Generating frequent pattern.

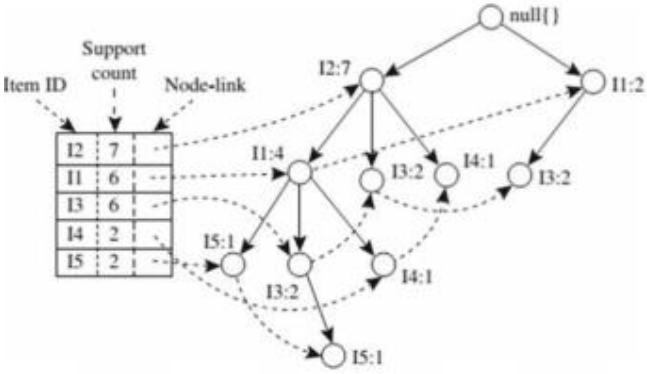

Fig. 3. Frequent pattern tree (FP-Tree).

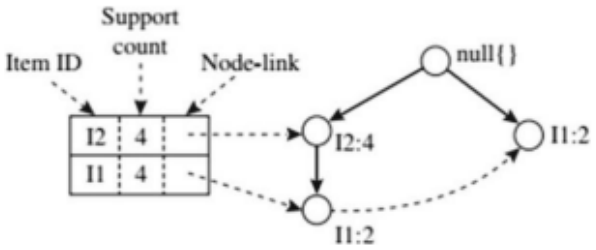

Fig.4. Conditional FP-Tree associated with Node I3.

Item	Conditional pattern base	Conditional FP-tree	Frequent patterns generated
I5	{{I2, I1: 1}, {I2, I1, I3: 1}}	{I2: 2, I1: 2}	{I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}
I4	{{I2, I1: 1}, {I2: 1}}	{I2: 2}	{I2, I4: 2}
I3	{{I2, I1: 2}, {I2: 2}, {I1: 2}}	{I2: 4, I1: 2}, {I1: 2}	{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}
I1	{{I2: 4}}	{I2: 4}	{I2, I1: 4}

TABLE 2: Conditional Pattern Base and conditional FP-Tree.

1. Advantage: The cost of search is low.
2. Disadvantage: For a larger number of data set, this process seems to be time consuming.

C. Éclat algorithm:

1. Need to receive tidlist for every item (DB scan)
2. Tidlist of {a} is same the list of proceeding holding {a}
3. The Intersect tidlist of {a} with the tidlists of other items, results in tidlists of {a,b}, {a,c}, {a,d},… = {a}-conditional database (if {a} removed)
4. Repeat the process from 1 on {a}-conditional database
5. Repeat for all other items.

Following holds the pros and cons of the discussed algorithm:

1. Advantages: The depth – first will reduce memory requirements. It is faster than Apriori. There is no any need to scan the given database for the support of (K+1) itemset.
2. Disadvantage: The TID-set is long and expensive.

.

itemset	TID_set
I1	{T100, T400, T500, T700, T800, T900}
I2	{T100, T200, T300, T400, T600, T800, T900}
I3	{T300, T500, T600, T700, T800, T900}
I4	{T200, T400}
I5	{T100, T800}

TABLE 3: Transactional data in vertical data format.

itemset	TID_set
{I1, I2}	{T100, T400, T800, T900}
{I1, I3}	{T500, T700, T800, T900}
{I1, I4}	{T400}
{I1, I5}	{T100, T800}
{I2, I3}	{T300, T600, T800, T900}
{I2, I4)	{T200, T400)
{I2, I5}	{T100, T800}
{I3, I5}	{T800}

TABLE 4: Item sets in vertical data format.

itemset	TID_set
{I1, I2, I3}	{T800, T900}
{I1, I2, I5}	{T100, T800}

TABLE 5: 3-itemsets in vertical Data format.

D. Tree Projection algorithm:

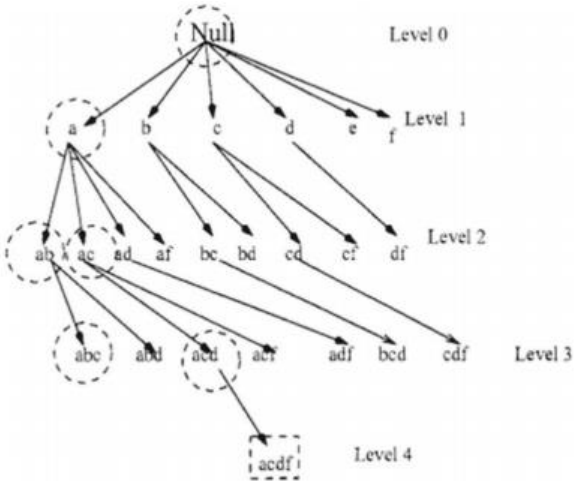

Fig.5. Lexicographic Tree

This constructs a lexicographic tree. In this algorithm, the hold up of each chronic itemset in every transaction is summed up and estimated onto the lexicographic tree as a node. This will enhance the production of averaging the total transactions that has a particular chronic itemset [2]. In the hierarchical structure of a lexicographic tree, only the subset of transactions that can probably has the chronic itemset will be forage by the algorithm. This search is carried by cross the lexicographic tree with a top-down approach.

1. Advantage: In this matrix structure is used to provide a more effective scheme for calculating the frequent itemset that have very low level of support count.
2. Disadvantage: The major issue faced is that different representations of the lexicographic tree present different limitations in terms of efficiency at memory consumption.

E. COFI algorithm:

Co-Occurrence Frequent Item set (COFI) uses a pruning method that lessen the usage of memory space convincingly. Its intelligent pruning method defines relatively small trees from the FP-Tree on the fly, and it is based on a special property that is originated from the top-down approach mining algorithm. Some examples of the COFI-Trees are shown below [6].

T1	A	G	D	C	B
T2	B	C	H	E	D
T3	B	D	E	A	M
T4	C	E	F	A	N
T5	A	B	N	O	P
T6	A	C	Q	R	G
T7	A	C	H	I	G
T8	L	E	F	K	B
T9	A	F	M	N	O
T10	C	F	P	G	R
T11	A	D	B	H	I
T12	D	E	B	K	L
T13	M	D	C	G	O
T14	C	F	P	Q	J
T15	B	D	E	F	I
T16	J	E	B	A	D
T17	A	K	E	F	C
T18	C	D	L	B	A

TABLE 6: Transactional Database

STEP 1

Item	Counter	Item	Counter
A	11	N	3
B	10	O	3
C	10	P	3
D	9	Q	2
G	4	R	2
E	8	I	3
H	3	K	3
F	7	L	3
M	3	J	3

STEP 2

Items	Counter
A	11
B	10
C	10
D	9
E	8
F	7

STEP 3

Items	Counter
F	7
E	8
D	9
C	10
B	10
A	11

1. Advantages: It provides a good execution run time and has a better memory consumption. This is because of the following two implementations: (1) A non-recursive method is used during the process of mining to traverse through the COFI-Trees in order

to generate the entire set of frequent patterns. (2) The pruning method implemented in the algorithm has removed all the non-frequent patterns, so only frequent patterns are left in the COFI-Trees.

2. Disadvantages: The threshold value of the minimum support is low.

F.TM algorithm:

This represents repeated itemset using the vertical data representation like the Eclat algorithm. The transaction IDs of all the itemset are transfigure and charted into a catalog of transaction interim at another location. Then, interim will be performed between the transaction gaps in a depth-first search order all over the lexicographic tree to sum up the item sets.

Item	Mapped transaction interval list
1	[1,500]
2	[1,200], [501,800]
3	[1,300], [501,600]
4	[601,800]

Fig.6. Transaction Mapping

1. Advantage: This algorithm shows a good performance over the FP-Growth and EClat algorithms on data sets that contain short frequent patterns.
2. Disadvantage: The TM algorithm is deliberate in processing speed compared to the FP-Growth algorithm.

G.P-Mine algorithm:

It mines recurrent itemset using a parallel disk-based approach on a multi-core processor. VLDBMine will issues a jam-packed version of data set on link. A Hybrid-Tree (HY-Tree) are extremely used for storing the required data. The efficiency is improved by pre-fetching process, where multiple projections has been loaded. Finally, the overall frequent set is being constructed [1]. By gathering the information. The architecture of the P-Mine algorithm is shown in Figure 8.

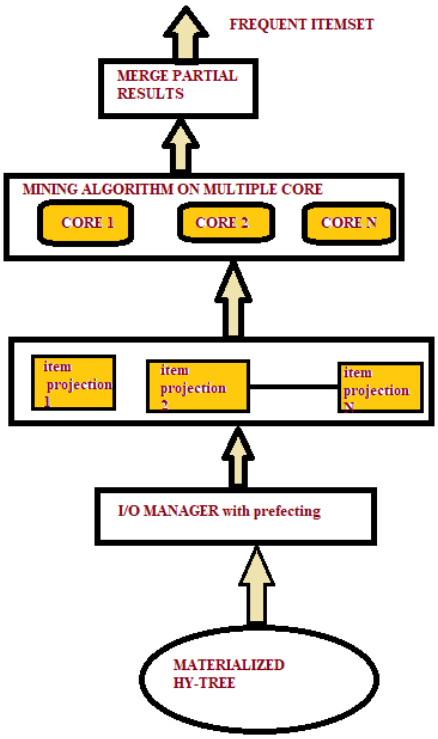

FIG.8. Architecture of the P-Mine Algorithm.

In the VLDBMine data structure, the data are being represented in dataset thus the performance and scalability are improved. [5] This is because the HY-Tree of the VLDBMine data structure facilitate the data to be fussily impressed to provide a massive support to the data-intensive stuffed process with reduced cost. The performance id optimised at each node, when the mining is executed at different processor. Optimization is enhanced to a maximum level when the processor holds the multiple level.

H. EXTRACT algorithm:

EXTRACT utilises a Galois lattice, which is a mathematical concept. The structure of the EXTRACT algorithm is shown in Figure 9. It is segmented into four functions for scheming the support count, merge the item sets, discard the item sets that are rerun, and uprooting the association rules from the recurrent item sets [5].

First step involves the calculation of the support count of each frequent 1-itemset that satisfied the minimum support. The item set which is not supported will be removed. Next step involves the combining process. The redundant item set will be neglected. The association rules that satisfied the minimum confidence will be generated after mining. That which doesn't satisfy will be removed.

1. Advantages: EXTRACT outperforms the Apriori algorithm for mining more than 300 objects and 10 attributes with an execution time that does not exceed 1200 ms.

2. Disadvantage: The mined data will not be stored in a disk. Repeated use of algorithm is required for the execution of the latest set of data.

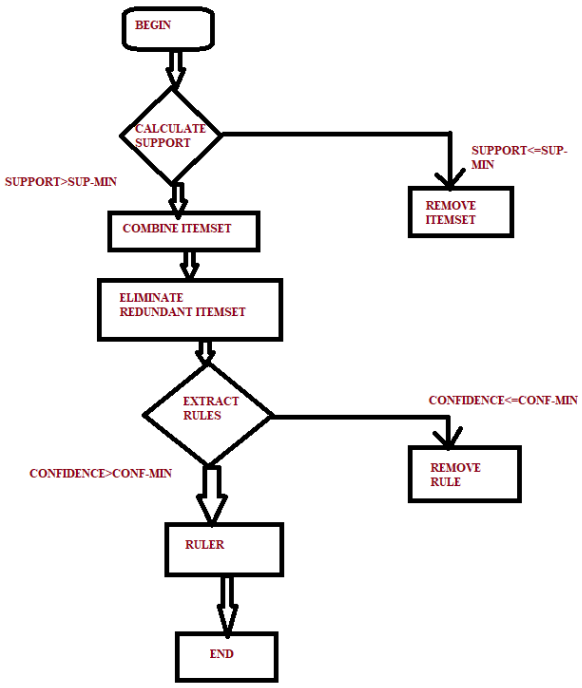

Fig.9. Architecture of the EXTRACT algorithm.

V. CONCLUSION AND FUTURE SCOPE

The purpose of this research is to examine the most important information which is hidden insight the data warehouse to analyze customer buying behaviour or to find business potential customers while what have been identified in and studied in previous research there is a lack of mining complex knowledge from complex data mining across multiple heterogeneous data sources. [3] Data mining is an influential tool that should be used with extreme care for raising customer contentedness, dispensing good, efficacious products at sensible and low-budget rate. This will result in business to perform large aggressive and advisability. This should not be utilized in any way that may cause intemperate pressure, financial backdrop or emotional stress.

REFERENCES

1. Interactive Media in Retail Group (IMRG). (2012) Press archive, http://www.imrg.com, accessed January 2012..
2. Kumar, V. and Reinartz, W.J. (2006) Customer Relationship Management: A Databased Approach, Hoboken, NJ: John Wiley & Sons.
3. Hughes, A.M. (2012) Strategic Database Marketing 4e: The Masterplan for Starting and Managing a Profitable, Customer-based Marketing Program, McGraw-Hill Professional, USA.
4. Davenport, T.H. (2009) Realizing the Potential of Retail Analytics: Plenty of Food for Those with the Appetite. Working Knowledge Report, Babson Executive Education.
5. Fuloria, S . (2011) How Advanced Analytics Will Inform and Transform U.S. Retail. Cognizant Reports, July, http://www.cognizant.com/InsightsWhitepapers/How-Advanced-Analytics-Will-Inform-and-Transform-US-Retail.pdf, accessed January 2012.
6. Jonathan Wu, Business Intelligence: The Value in Mining Data, DM Review online, February, 2002