ASSOCIATION MODELS FOR MARKET BASKET ANALYSIS, CUSTOMER BEHAVIOUR ANALYSIS AND BUSINESS INTELLIGENCE SOLUTION EMBEDDED WITH ARIORI CONCEPT

J.M. Lakshmi Mahesh*

ABSTRACT

This paper analyzes the customer behavior by identifying the frequency set combination by applying the techniques & methods of Apriori (By R.Agarwal and R.Srikant in1994), a known Data Mining Concept .The analysis utilize the sample data collected from the customers of a specified Nationalized bank in India about its Retail services, &applying the Apriori concept for identifying frequent mining sets by developing the code in C programming language. The analysis result has been represented as a frequency between various preferences, with Findings & Recommendations The predicted Findings and the hidden patterns/information can be applied for Marketing research activities such as customer requirement analysis, Market Segmentation, Product/service segmentation and also for customer care services. The study further paves the path for further research initiatives to be taken in the field of customer behavior analysis by adopting Cluster, Classification, correlation& Market Basket analysis for prediction purposes on different input data cases.

Keywords: *Apriori, Association Analysis, Customer behavior analysis, Data Mining, Market Basket analysis, Market Segmentation, Retail Marketing*

*SCMS School of Technology and Management, Muttom, Cochin, KERALA

INTRODUCTION

The goal of any organization , especially commercial organizations is to provide products and services so as to attract and retain profitable customers and to identify the target market for developing & expanding the market in future for Potential Customer as well. To achieve this, In depth knowledge about the customers and prospects is essential to stay competitive in today's marketplace. Some of the benefits include improved targeting and product development.

Organizations began to offer multiple products and Services hoping to compete by appealing to different consumer tastes and Preferences. Consumers became discriminated which created a challenge for organizations. They wanted to get the right product to the right consumer. This created a need for target marketing- that is directing an offer to a target audience

The growth of target marketing was facilitated by two factors (1) the availability of information and (2) The increased computer power. On one side the application of target models becomes very common in the marketing industry on the otherside; Mining for associations among items in a large database is an important database mining function. Embedding both the concept the paper proceeds to identify the customer behaviour analysis, revealing market basket analysis & providing some intelligent business solutions by identifying various Association Rules embedded with the technique of the Data mining tool called Apriori algorithm.

Apriori algorithm is an influential algorithm for mining frequent Item sets for Boolean association rules. In computer science and data mining Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example collections of items bought by customers or collections of services &options preferred by customers or details of a website frequentation). As is common in Association rule mining given a set of item sets (for instance sets of retail transactions each Listing individual items purchased) the algorithm attempts to find subsets which are common to at least a minimum number of the item sets.

Apriori uses a "bottom up" approach where frequent subsets are extended one item at a time (a step known as candidate generation) and groups of candidates are tested against

the data. The algorithm terminates when no further successful extensions are found. The result may reveal the hidden patterns of relationship among data items if any.

The model can be deployed in various ways to diverse business processes or systems, depending on the business needs.

DATA COLLECTION& PREPROCESSING:

Data mining refers to extracting or 'mining" knowledge from large amounts of data. The data to be provided for Data mining applications should be preprocessed like data cleaning, integration, conversion & selection. Thus data preparation is exactly as it seems. It is finding and organizing the data for the chosen mining method. Once the data has been prepared, then the mining method can be invoked to create the mining model, which is then verified by the analyst. The most common Transactional data types used in data mining are

1. Behavioral data: Data about an individuals behaviour or relationship characteristics, such as quantities purchased, amounts spent, balances, and number or rate of interactions.

2. Demographic data

3. Production data

A sample of Behavioral type data has been analyzed for the paper .The data collection is basically primary data collected from two different Localities to get better interpretation.

CONCEPT OF APRIORI ASSOCIATION RULE MINING:.

Frequent Item Set: Frequent item set appears frequently together in a transaction data set.

Frequent pattern Mining: Frequent pattern mining searches for recurring relationships in a given data set. A typical example of frequent item set mining is a market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customer place in their shopping basket. The discovery of such associations can help retailer to develop marketing strategies by gaining insight into which products are frequently purchased together by customers.

ASSOCIATION RULES: An association model generates combinations of items called item sets. Each item sets contains one or more items and has a relative frequency of occurrence called support in the population of transactions being analyzed. These support

values are used to construct rules that qualify relationships or affinities among items occurring together in transactions, leading to identify frequent pattern mining

Association rules are described using the following terms:

1. Rule body; one or more items in a transaction, which imply the presence of another item

2. Rule head: One item whose presence in the transaction is implied by the presence of the items in the rule body

3. Support: Percentage of all transactions containing both the body items & the head items

4. Confidence: Likelihood that the head item will be in the transaction , given the presence of body items

5. Lift: degree to which the confidence is greater (or less) than expected (Threshold values).

Example Representation of ASSOCIATION RULES: The information that customers who has own computers (P.C) will opt for Net banking (N.B) is represented in

ASSOCIATION RULE BELOW

Own PC ==>Net banking[support=2%,confidence=60%]

Rule Support support of 2% for the above means that out of complete sample data under analysis, persons having Own PC & Netbanking.

Thus support (own Pc==>Net banking) = Probability (Own PC U net banking)

Support (P.C==>N.B) = P(P.C U N.B)

Rule Confidence: A confidence of 60% means that out of total number of persons having Own Pc 60% will have Net Banking also.(Conditional probability)

Confidence (Own Pc==> Net banking) = Probability(Net banking/ Own PC)

Confidence(P.C==>N.B) = P(N.B/P.C)

Typically, association rules are considered interesting if they satisfy both a minimum support threshold value & a minimum confidence threshold. Such thresholds can be set by domain experts based on application requirements.

Association modeling is often used in conjunction with clustering

APRIORI ASSOCIATION RULE MINING: Apriori is a semial algorithm proposed By R.Agarwal and R.Srikant in1994 for mining frequent itemsets for Boolean association rules.The name of the algorithm based on the fact that the algorithm uses Prior knowledge of frequent itemset properties.

Apriori algorithm employs an iterative approach known as level wise search where k-itemsets are used to explore(k+1)-itemsets.

APRIORI PROPERTY:

"**All nonempty subsets of a frequent itemset must also be frequent**".

The apriori property is based on the following observation : .

If an itemset I , does not satisfy the minimum support (min-support then I is not frequent .If an item A is added to the itemset I, then the resulting itemset(I U A) cannot occur more frequently than I.

This property is called ANTIMONOTONE in the sense that if a set cannot pass a test , all of its supersets will fail the same test as well.(it is called antomonotone because the property is monotonic in the context of failing a test)

APRIORI ALGORITHM: The Apriori Algorithm in a Nutshell

•Find the frequent itemsets: the sets of items that have minimum support

–A subset of a frequent itemset must also be a frequent itemset

•i.e., if {AB} isa frequent itemset, both {A} and {B} should be a frequent itemset

–Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)

•Use the frequent itemsets to generate association rules.

The Apriori Algorithm: Pseudo code

•Join Step: Ckis generated by joining Lk-1with itself

•Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

Ck: Candidate itemset of size k

Lk: frequent itemset of size k

L1= {frequent items};

For (k= 1; Lk! = Lk-1; k++) do begin

Ck+1= candidates generated from Lk;

For eachtransaction tin database do

Increment the count of all candidates in Ck+1that are contained in t

Lk+1= candidates in Ck+1with min_support

End return kLk;

METHODOLOGY & IMPLEMENTATION:

Variables for association methods require a transactions table with multiple records, where each record contains a transaction identifier as one of the items in that transaction record. The record may consist of different types of data such as Interval scales (age), Binary (sex), Categorical, & ordinal variables. Many real time databases consists of Mixed data types .The sample data collected for this research paper also consists of Mixed data types, But the preprocessing technique adopted here is combine different type of variables in to categorical variables with common interval[1,m].

M-the maximum options available for each variable, which is assumed to be same for all the variables from v1,v2,v3,……vn., thus the database matrix of available attributes is of fixed length.

Attributes/options	option1	option 2	option3 … … … option m
Attribute1(A)	A1	A2	A3 … … … Am
Attribute2 (B)	B1	B2	B3 … … … Bm
…….			
Attribute n (…)	……		

Extracting and transforming the Data:The assumptions for extracting & transferring data for this paper includes

- .M size assumes to be fixed as 5 (ie) Maximum options for each attribute is fixed as 5

- N size is 24 as the value for the following 24 attributes , mostly related to Banking sector are collected from the customers.

1. Gender 2. Location 3. Age 4. Marital Status 5 .Occupation 6.Income 7. Family size 8.Education 9. knowledge of Internet Browsing 10. Types of accounts in Banks 11 Account holders period 12. Retail facilities availed 13. loan details 14. Frequency of visit to Bank 15. Grading Banks customer relationship 16.Preference of Technology Banking over Traditional Banking 17. Personal Internet connection details 18. Banking Electronics Services availed 19. Usage of ATM 20. Awareness of Net banking 21.Reasons for preferring Net banking 22. problems with NET BANKING 23. Rating the products of Bank 24. other services like Insurance and Mutual fund

- Using Reduced Minimum support levels: Each level of selection list has its own threshold values. The deeper (more levels of combination) the level the smaller the corresponding threshold is.
- The customer data base consist of

Customer /attributes	A	B	…….N
Customer (C1)	A1/A2/A3/A4/A5	B1/B2/B3/B4/B5 ……..	
……			
Customer (CS)	……	……	…….

PROCESS DESIGN (ALGORITHM FOR THE STUDY):

STEP1: The system accepts number of questions and question code from the administrator. For every question five codes starting from 1 till 5 will be created automatically by the program and stored as “OPTION DATA FILE”(example:. Code for question “Location of the customer” is B then program automatically generates B1, B2, B3, B4, and B5).Thus the option file creates the possible option for all the questions

STEP 2: Based on code the corresponding actual option will be stored in the “ACTUAL DATA FILE”.(ex. For B1-it will store Location1)

STEP 3: After creating the initial 2 data files, by retrieving questions and question code from ‘OPTION FILE” & its corresponding actual from “ACTUAL DATA FILE” data will be collected from customers through questionnaire and creates “CUST_DATA”.

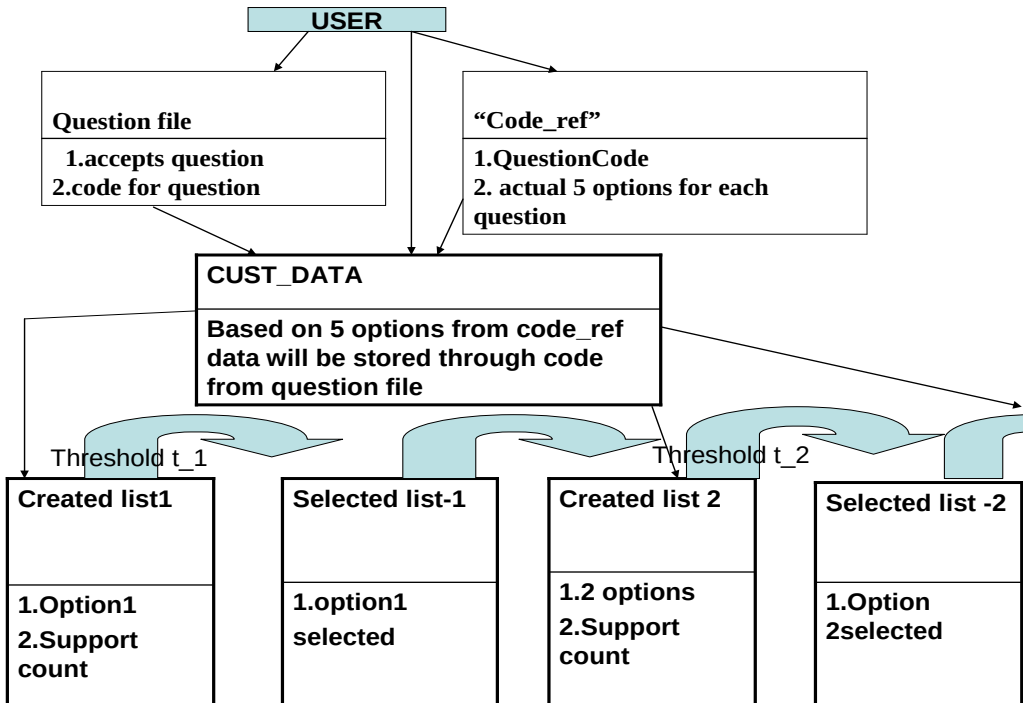
STEP 4: Main process starts by first creating list1 combination and computes its support count-c_list

STEP5: From the previous list 1, based on the support count >= threshold value the selected list 1 will be created.

The selected list 1 now happens to be the input code for the next iteration to create the combination list-2. Thus the apriority property all nonempty subsets of a frequent item set must also be frequent". Will be preserved.

STEP 6: the process 5 continues till it reaches required combination selected list given by the user.

PROCESS DIAGRAM:



RESULTS & INTERPRETATIONS: Few hidden associations & interesting information/findings of the research study has been given below

Query: list out some of the frequencies in 9 items/services combination

G2	I1	O2	R1	S1	T1	V3	W2	X2	19
G2	J1	Q1	R1	S1	T1	V3	W2	X2	18
I1	J1	Q1	R1	S1	T1	V3	W2	X2	19

FINDINGS: From the above 9 items combination list it is very evident that the subitems combination whose count value is high has contributed for super set(ex.G2,I1,J1,R1,S1,T1,V3,W2,X2)

CODE	EXPLANATION	% OF PREFERENCE
G2:	family size3-5	80%
I1:	internet-brow –know	86%
J1:	savings a/c	83%
R1	ATM	77%
S1	Use ATM-always	60%
T1	net bank aware –few	65%
V3	never-face-problem-Net-Bank	56%
W2	rate-service-good	86%
X2	insurance-–no	78%

HENCE APRIORI PROPERTY HAS BEEN PROVED ie the All nonempty subsets of a frequent itemset must also be frequent".

Query: For the question gender and the option female and for the question tech_trad and the option always has preferred by 27 customers for the question gender and the option female and for the question tech_trad and the option sometimes has preferred by 18 customer's total number of female customers: 49

FINDINGS: P (tech _always U tech_sometimes)/P (female-customers)

➔(27+18)/49

➔45/49 =91.8%

Query: How many female, Post graduates having Basic internet knowledge registered for Netbanking

P(a2∩h3∩i1∩r3)

Result:NOT GREATER THAN THRESHOLD VALUE

For the question gender and the option male and for the question tech_trad and the option always has preferred but 24 customers

For the question gender and the option male and for the question tech_trad and the option sometimes has preferred but 23 customers

Total number of male customers: 51

FINDINGS:

P (tech _always U tech_sometimes)/P (Male-customers)

➔(24+23)/51

➔47/51 =92%

Thus, the preference for technology banking then traditional banking for Male & female has been analyzed & an interesting concept that Both Male &Female customers prefer technology banking than Traditional banking has been revealed ,but still female post graduates registered for netbanking is very less (it is below the threshold value) the Reasons may be - Opportunities are not provided for them ,or due to lack of motivation ?.The findings recommend that the market segment consisting of female graduates for NetBanking services can be wide up, by providing appropriate SPECIAL services like WOMEN CARDS.

Query: Revealing hidden associations

For the question electronics service 17 customers have given their option as net_banking

P (Netbanking) =17

For the question reasons_n_bank 11 customers have given their option as reserve-air,rail

P (reserve)/P (Netbanking)

➔11/17 =64.7%

FINDINGS: Majority customers, Registerd for NetBanking gave the most preferred service as for booking tickets

Query: Revealing hidden associations:

For the question visit_freq 51 customers have given their option as very_rare

FINDINGS: The result portrays that nearly 51% of the customers visit bank rarely may be considered as the mark of technology improvement but at the same time it may leads to some of the CYBER CRIMES related to e-banking as the existence of personal touch between the customers & bank people could not be established &developed.

Query: indirect queries:

For the question awareness 18 customers have given their option as all

FINDINGS: P (Net banking-awareness)/n

➔ 18/100 =18%.

Number of customers having basic knowledge about Internet browsing & not registered for Net banking:

For the question browsing knowledge 86 customers have given their option as yes P (Basic Internet knowledge) ➔ 86

For the question electronics_service 17 customers have given their option as net_banking : P(net_banking)➔17

If the assumption made such that all the NetBanking customers posses basic knowledge of Net browsing then

P (Basic Inernet knowledge∩not registered for Netbanking)>=1-P (17/86)

➔ >= 69/86

=80% of the Account holders posses knowledge about Internet, do not access Net Banking

RECOMMENDATIONS: It shows that banks may organize camps & other methods to improve the knowledge of their customers about NETBANKING.

Query: Revealing hidden information:

For the question education and the option postgraduateand for the question awareness and the option all has preferred but 10 customers

➔P(graduate ∩aware –Net bank)/P(Post_graduates)

10/42 =23.8%

FINDINGS: Thus the awareness about the NetBanking features among the graduates & post graduates is also very less.& moreover, The customers having Own Internet connection & not registered for Netbanking Is 80%,yet another market segment for netbanking.

CONCLUSION

Organizations are increasingly interested in retaining existing customers as well as targeting non customers. Measuring customer satisfaction provides an indication of how successful the organization is at providing products and /or services to the market place. In the present scenario customer satisfaction is the key stone for the success of the every organization .Therefore it is necessary to evaluate the changes happening in the tastes and desires of the customers.

Thus Apriori , being a valuable Data mining technique , can be used to discover knowledge for the purpose of explaining current behavior, predicting future outcomes and to provide support for Bank's Decision making Processes and also for some other Business Intelligence Purposes.

Scope/limitations: The study is restricted to only limited number of branches of the Bank

1. The survey includes all the limitations inherent in the questionnaire.
2. The response from the respondents may be not being sufficient, since the samples are from the small group of people.

3. The study is just to analyze only certain data variables known to be common in analyzing the consumer behavior and there may be many other hidden factors contributing to the association.

FUTURE DIRECTIONS FOR FUTURE RESEARCH:

The program can extended to n number of questions by just changing the value of the variable data defined as macros in the program coded in C language.

The support threshold values may be changed in the same manner by adopting increasing, decreasing or constant values for the threshold values at each iteration

The same code can be applied to other MARKET BASKET ANALYSIS by modifying the input details.

The association rules may be applied in analyzing FINANCIAL PORTFOLIOS.

The proposed study can be extended by collecting samples from vast area and the mining concept can be applied to large database to reveal more associations and target market.

Instead of database, Data Ware house can also be applied. After data cleaning and filtering .Program code can also be enhanced to accept more than 1 option for the questions.

To identify the target market other Advanced Data mining tools Like Neural Networks and Genetic algorithms can also be applied which is the emerging trend in the research area. It is hoped that the study will open up further Areas of Research

REFERENCES

1. Jiawei Han and Micheline Kamber -Data Mining Concepts and Techniques ,Second Edition, Elsevier, 2006
2. Daniel T.Larose -Data mining methods and models
3. Chuck Ballard et al -Dynamic Ware housing:Data Mining Made Easy-.
4. Philip Kotler -Marketing – Management
5. J S Chandan -Management Theory & practice
6. Ellis horowitz,sartaj sahni,sanguthevar rajeshekharan.-Fundamentals of computer algorithms
7. Gilles brassard, Paul Bratley- Fundamentals of algorithms
8. Alex Berson, Stephen J Smith -Data Warehousing , Data Mining, & OLAP –Tata McGraw Hill, 2004

9. Sinha, Thomson Learning -Data Warehousing, 1Edn.
10. Anny levitin. - Introduction to the design and analysis of algorithms
11. Gupta .S.C (2004)-. Fundamentals of Statistics
12. Sara baase,Allen Van Gelder computer algorithms - Introduction to design and analysis
13. Gupta.S.C & Kapoor.V.K -. Fundamental of Mathematical Statistics
14. Walpole.Myers.myers.ye .-Probability & statistics for Engineers & Scientists
15. Elhance D.n ,Veena Elhance & Aggarwal .B.M. -Fundamentals of Statistics.
16. Aho, hopcroft, & ullman. -The design and analysis of computer algorithms
17. .R.Panneerselvam.- Design and Analysis of Algorithms
18. Michael T.Goodrich. Roberto Tamassia.- Algorithm Design