Developing a Customer Model for Targeted Marketing using Association Graph Mining

Suresh. K, Pattabiraman. V

***Abstract*:**
Data mining is the procedure to find out significant information from large database by applying several mining techniques. Finding out products that are purchased together is a major issue in basket market analysis. So, developing a customer model is important for targeted marketing. The traditional dataset is taken into account because the origin of information which is available from the history of sales repository. While applying the basic techniques on transactional data analysis, it fails when the process has a greater number of transactional information. Also, it is difficult to identify suitable correlation between one product to another. In this paper Market Basket Analysis is extended towards into network level and it recommends a product network consideration it clearly states that the correlation involving products bought together by customer. This research work focuses on product to product network analysis in market basket network. The direct and indirect approach is applied in associated product network from history of retailer data. The major intention of this research work is to find the group of essential products purchase by the customer together. So, it will bring out consumer profile, product blue print, guidance from associated products and provide effective result from large number of customer wholesale outline.

***Index Terms*: *Association mining, Product Networks, Community detection, Graph mining, Market Basket Analysis*.**

I. INTRODUCTION

Data mining is procedure to find out unseen facts from huge databases. The frequent pattern mining technique plays a vital role in data mining techniques to extract buying pattern of consumers. The two major techniques are Apriori and FP-growth to find out pattern mining knowledge. The explosive increase of information provides the motivation to search out meaningful knowledge hidden within the immense database. A frequent pattern mining technique is one of important technique, to analyze large consumer product dataset, it required further additional immediate process owing to immense volume of knowledge is being updated in real time. In addition frequent pattern mining over information generates a vast range of frequent patterns and it causes a major amount of expenditure. Considering weight conditions are extremely helpful factors in reflective importance for every object in the real time, it is also necessary to use them to the mining methods to become a lot of serviceable and important designs.

Revised Manuscript Received on July 05, 2019.

Suresh.K, School of Computing Science and Engineering, VIT University, Chennai- 600127, India.
Pattabiraman. V, School of Computing Science and Engineering, VIT University, Chennai- 600127, India.

Sequence data is mostly through outcome of statistically related formats among information which is distributed in a sequence direction. It is nothing but identifying relevant pattern between series of data. A sequence can be represented by $S_i = (S_{i1}, S_{i2},\ldots S_{in})$, where S_{ij} is the 'j'th measurement value of the 'i'th sequence, assuming the length of the sequence. Data structures, definitions, attributes and domain values are considered as data characteristics which are actually suggestion to different database tables with the same key value. Different types of data dimensions are analysis considered by the high dimensional data.

Based on the customer buying pattern, marketing managers will establish good relation of association with the customers, also they can distinguish and predict the individually and all customer activities. Association algorithms, decision tree, clustering these are familiar data mining techniques, by using these techniques researcher can find out the pattern extraction to unknown dataset. Graph mining technique will overcome the existing method through traditional association algorithm method. There are two key aspects in the traditional statement. First there are only very few interested in sub graphs that are connected.

This research work continued to upgrade and value added to Market Basket Analysis (MBA) through graph mining. The suggested product networks know how to evaluate the correlation between products which are purchased together. In this paper, it takes a different method to mining transaction data, by modeling the data as product network it find out significant communities in the data which can be targeted for further analysis. The transactional data is converted into product level format. It put forward two types of product network, direct and indirect product networks in this analysis. The existing model has limited technique while analyzing co-purchased product in transactional database. To overcome this work introduced two types of pruning strategy. Direct purchase, it will visit the edges one product to another product. So co-purchased correlation is not even drawn through this technique. Interrelated network has complex and its connected large number of nodes. To rectify this complex the products are represented in product network model.

II. RELATED WORK

Russel Peard et al (2015) presented a unique way in semi supervised taxonomy and it established by weighted communication model. This work is derived by rule extraction from database, somewhere a business deal consists of number of itemset

and every product connection is allocated a weight indicating in relation to other products.

Omer M.Soysal et al (2015) proposed an approach Mostly ASsociated Patterns (MASPs) this method presented benefits of overwhelming with a reduction of computational resources to discover lengthy rule sequences. In throughout pruning procedure, MASP tree is created. Later than creating MASP tree, the rules are producing in considerably less computation time.

Russel Pears et al (2015) describes about network analytic approach to the association mining problem and this will give meaning full correlation than the existing technique. This work initially constructs the social network structure for the transactional data and implemented community detection algorithm to find out more effective relationship form data. When compare to existing algorithm in weight assignment method this work reduced the execution time.

Ivan F. Videla et al (2014) Proposed graph mining techniques in transactional data. Generally graph represents significant connections between one products to another product. The set of product connections are definite as a complex networks, nodes are consider for products and edges stands for corresponds relationships between them. The proposed method is compared with circumstances of the capacity to justify its competence among different datasets through various features of uniqueness.

Shu-Ming Hsieh et al (2014) put forwarded new algorithm to execute isomorphism testing of labeled graphs and its drawback is considered as the backtracking in the middle of vertex categories and the vertex partition methodology. Considering the graph representations as complicated structure, the isomorphism testing connecting tagged graphs turn out to be one among the foremost long procedures throughout the method of graph mining. To carry out the isomorphism testing, author has been introduced CISC (Class-wise ISomorphism testing with limited baCktracking) algorithm. In multiple labeled graphs the proposed algorithms are outperforms Ullmann, FSG and VF2.

Paolo Giudici et al (2013) describes about statistical association models and graphical models on consumer behavior. The retail owners have invested large amount of money to find out the knowledge from the customer weblog data predict best product characteristic in database. Market Basket Analysis (MBA) is one among knowledge in data mining techniques to discover customer buying model by extracting efficient correlation from product one another from transactional database.

Hyea Kyeong Kim et al (2012) proposed network analysis model for market basket analysis. Generally association mining works on the basis of transaction in level by level iterations. When compare market basket analysis model, the network model focuses on more network based relationship between all products. It has two types of networks, market basket networks and co-purchased product networks. The author recommended new network model exploration be estimated to a lot of valuable with efficiency like commercial and personalized product with deep relation between products.

Olfa Nasraoui et al (2008) described about framework and finding web usage patterns from weblog data. Based on the Customer Relationship Management (CRM) they present a new method for discovering about the user profiles. The framework of web mining includes integration, preprocessing the weblog data and sessionization to continue to find out model through clustering technique.

Michihiro et al (2004) describes about graph modeling in real time which is complicated structures like geographical area, earth science, etc., and these data are basically constructed in different dimensional. Due to speed of internet the web business has been growing rapidly. Nowadays in industry, the business analysts are facing lots of difficulties to predict pattern because of large and complex data. Generally in graph objects are represented as vertices and associations between objects are correspond to edges.

III. PROPOSED WORK

If you are using *Word*, use either the Microsoft Equation Editor or the *MathType* add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation *or* MathType Equation). “Float over text” should *not* be selected.

This research work focuses on product to product network to identify the more information about the retailers and provide better service to the customers useful needs. In this section it explores how to generate these products networks and also shows how to generate frequent itemsets through community detection discovery.

A set of networks is a group element presented in a set is interrelated connecting every one. The general method to correspond toward a set of connections is a graph. A graph is a way to identify associations among sets of products. The graph contains of a established of items called nodes through a amount of sets which is joined by relations called edges. The set of products are connected as network format and it is well defined by nodes and represented as products and edges are characterizing as interaction among a combine of networks. Also it will define that what type of association is correspond through an edge.

Similarly to Ivan F. Videla-Cavieres et al (2014) this research work has the similar issues, an extremely intense products network among highly number nodes. For example the total transactions 485,136, based on the product network technique it only obtained 5,359 products among 5,362,906 edges. 31,244 have larger density of weighted edge. 4,235,093 as of those edges, it have only weight lesser than 10.

1.1. Community Detection

Community detection is procedure discovering a set of dynamic associated nodes. In several real time systems be able to be explained by difficult networks. For example social network, biological networks, mobile phone networks etc,. There are many techniques have been proposed to identify the complex structures of community structures and applied successfully to some

real complex networks. On the basis of the process results and other features, these processes can be able to categorize into graph partition technique. Everywhere each vertex fit in to only one community. There is no commonly established standard to estimate the community detection. Examples include social relationships, spreading of viruses and diseases, and the Internet and the World Wide Web, etc.

Generally **SP**eaker listener **L**abel **P**ropagation **A**lgorithm (SPLA) is a next version of the **L**abel **P**ropagation **A**lgorithm (LPA). Before give explanation to the overlap community detection process, this work explains what Community Detection is in graphical way. It is the advancement of demanding towards locates number of associated nodes. Fortunato et al (2010) according to Literature the initial issues is to come across for a characterization of community, because generally the characterization depends on the exact coordination for correlation developed. Fundamentally the major design is to be additional edges inside the group of nodes than the edges involving vertices of the community.

In existing method, most of the proposed work on community detection algorithm mentioned without priority explanation. Mathematically the community based nodes and edges are explained as follows:

Graph G = {V,E} set of 'n' nodes is represented by 'V' and set of 'M' edges is represented by 'E'. Finding the number j of sub graphs is not known previously and it is determined by the algorithm, based on the maximization of a function f(k). Newman et al (2004) defined the modularity function as:

$$\text{Modularity} = \sum_j (ejj - aj)^2 \quad (1)$$

Modularity is one kind of method which is applied in the structure of complex networks. This was intended to determine the potency of splitting up a complex network into different components. Modularity is frequently used in development techniques for community detection problem. 'ejj' is represented as part of edges with the purpose of join vertices to the rest of vertices. 'A_j' is referred as part of edge endpoints which is fall in community 'j'.

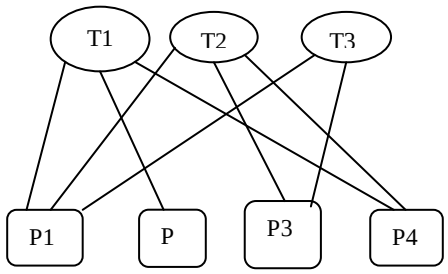

Fig-1 Transaction data in Network Structure

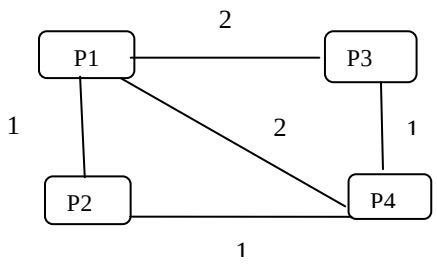

Fig-2 Product to Product Undirected Graph

Table-1 Product to Product Matrix

	P 1	P 2	P 3	P 4
P 1	0	1	1	1
P 2	1	0	0	1
P 3	1	0	0	1
P 4	1	1	1	0

Based on the customer purchase data, product matrix is constructed. Here the $\{P_1, p_2 \ldots P_n\}$ are referred as product value for both rows and column respectively. If the product was purchase, then it is referred as '1' in the matrix format otherwise it is '0', fig-1 shows corresponding product network format in table-1. In fig-2 co-purchased network construction shown it generally indirect product construction is represents the total number of customers who are all purchased the product in number of times.

1.2. Algorithms for Overlapping Community Detection

In today's world, due to large amount of data the numbers of techniques were proposed to find overlapped communities. This technique is very different from one another with respect to different types of complex network. Xie, J et al (2011) described new method established on the overlapping community detection technique. In this paper author proposed algorithm it has been classified each one into five types of classes based on the concept of community detection algorithm. However, in COPRA, each node updates its label in a synchronous way and does not consider the old knowledge in the previous iteration. In addition, the maximum number of communities each node cannot to be defined.

Table -2 Direct associations between product transaction data

	P 1	P 2	P 3	P 4	P 5	P 6
P 1	-	-	-	1	-	1
P 2	-	-	1	-	1	-
P 3	-	1	-	-	1	-
P 4	1	-	-	-	-	1
P 5	-	1	1	-	-	-
P 6	1	-	-	1	-	-

Table -3 Indirect association between product transaction data

	P 1	P 2	P 3	P 4	P 5	P 6
P 1	-	-	-	1	-	1
P 2	-	-	2	-	2	-
P 3	-	2	-	-	2	-
P 4	1	-	-	-	-	1
P 5	-	2	2	-	-	-
P 6	1	-	-	1	-	-

In this work, there are two different ways by which the product are being associated (a) direct and (b) indirect way.

Table 2 and Table-3 represent an example of direct and indirect product to product matrix. Indirect product to product network is indicating network matrix. Direct associated nodes are nothing but travel through more than one node from product to product transaction data. In order to show the products are being associated, this work implements the following algorithm. Consider each product (Pi) being associated with one another by the way the transaction (Ti) are occurred. For a direct association (DTi), find the total transactions (Ti) between the products being occurred. For a indirect association, find the inter transaction between the products being related occurred

The algorithm shown, it has to remain two vectors of vertex labels: A(x,1) and B(x,1); A(x,1) represents the preceding label in favor of vertex x. Every vertex label has set of pairs (C, B(x,1)), where C symbolize community identifier and B(x,1) correspond to the coefficient.

1. Every highest point is connected among a label, which is to be an identifier such as a numerical.
2. From the beginning, each vertex is considered as a single tag.
3. The repetitively, every vertex 'x' bring up to date its label by substitute to the label which utilize via maximum amount of associations.
4. The associated nodes are used one label for number frequency, then only one label has been used at random.
5. Subsequent towards a number of iterations, the similar tag have a tendency turn out to be connected by means of all part of a neighborhood

Algorithm for Overlapping Community detection.

1. Assign the color for the old vertex as A(x,1).
2. If A(x,1)=B(x,1), assign the minimum value for the vertex by color B(x,1).
   Else, assign the vertex of color B(x,1).
3. If the minimum value for the vertex B(x,1) ≠ A(x,1).
4. Repeat step2.
5. For each vertex x, C = C(B(x,1),1),
   If, C ← C(B(x,1),1) – {(C,A(x,1))} ∪ {(C,A(x,1) ∪ {x})}.
   S ← S(x,i) – {(C,1)} ∪ {(C ∩ B(x,1)}.
   Else, C ← C(B(x,1),1) ∪ {(C,{x})}.
   S ← S(x,1) ∪ {(C, B(x,1))}.
6. For each (C,1) in S(x,1),
   If i ≠ empty: C ← C(B(x,1),1) – (C,A(x,1)).
7. Disconnect the split vertices.

IV. RESULT AND DISCUSSIONS

The arrangement procedure of the temporally transactional customer buying network is classified into three phases. In the primary phases it used to construct a set of sequential data as like as similar to the Table 1. In second phase, after getting phase one information it tries create transactional itemset into bipartite network where every product is associate with other products with the aim of obtained in that precise transaction as in fig. 1. Here, the set of transactions 'T' and products 'P' represents the incoherent sets necessary to construct a bipartite network. In third phase, it shifts beginning to product matrix corresponding to the association network which is shown in Fig. 2. Subsequent towards this procedure it acquired a weighted association for product matrix. These set of connections are able to signifying by product matrix performance the association between couple of network products. The weight is referred as different types of transaction, in which similar transactions are involved concurrently. The networks present high degree nodes with meaningless edges between them. To remove meaningless edges, a threshold 'S' has to be defined, then the graph is fully revised in the search of edges with a weight S_0 lower than S(S< S1). The edges that match with these criteria are removed.

The weight is the number of transactions, in which a couple of products are present simultaneously. The networks present high degree nodes with meaningless edges between them. To remove meaningless edges, a threshold 'S' has to be defined, then the graph is fully revised in the search of edges with a weight S_0 lower than S(S< S1).

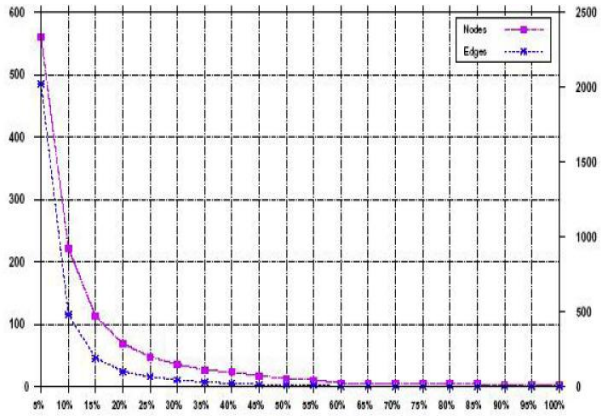

Fig-3 Graph depicting the number of nodes and edges.

When a threshold is affect, the set of connections explores essential sectors (i.e.) density of network as can see in Figs. 4 and 5 these results are achieved later than implement a threshold equivalent to 5%

and 10% respectively. These sectors communicate itemsets through a significant correlation among them. In the direction of identify these correlations and providing them an important understanding implemented by a community detection technique. These statistics were acquired later than involve in a threshold which is equivalent to the 5% and 10% considered as important edges threshold. These sectors illustrate efficient itemsets through a foremost association involving them. In the direction of identify these relationships and significant analysis is based on the community detection technique.

Graph illustrates the number of nodes and edges from transactional data. It is understandable to facilitate while the number of gains is high, the quantity of nodes and edges go downwards which is shown in Fig-3. The major improvement of deciding like 5% or 10% of threshold is clearly mentioning the individual data. In existing method threshold is fixing based on support and confidence value and it has only fallen on some percentage value. But here threshold is not static and it will categorize from 5% to 100%. These express thresholds are self-adaptable to the number of nodes and examine extremely strong with complex networks.

V. CONCLUSIONS

This research work has uncovered a distinct procedure for market basket analysis by using graph mining method. It is moreover one of the rapid and simplest ways to extract useful data from a lots of product sales transactions. The difference between indirect product network and direct network, the stock within the updated network is associated after they obtain particular time. This works shown that it will be one of the suitable techniques to find frequent itemsets in transactional information. It is very important to say that the information given is very useful for retail, representing the associations that don't seem to be comprehensible used for the market analyst of the retail. If a business can understand customers and transfer this knowledge as a resource to select out probable partners for alliances, it can be way to induce edges from business in retail market.

REFERENCES

1. Hyea Kyeong Kim, Jae Kyeong Kim, Qiu Yi Chen, “A product network analysis for extending the market basket analysis”, Expert Systems with Applications 39(2012) 7403 – 7410
2. Corrado Loglisci, Michelangelo Ceci, Donato Malerba, “Relational mining for discovering changes in evolving networks”, Neuro computing 150 (2015) 265-288.
3. Shu-Ming Hsieh, Chiun-Chieh Hsu, Yen-Wu Ti, Chi-Jung Kuo, “Reducing the bottleneck of graph-based data mining by improving the efficiency of labeled graph isomorphism testing”, Data and knowledge Engineering 91(2014) 17-33.
4. Russel Pears, Songwut Pisalpanus, Yun Sing Koh, “A graph based approach to inferring item weights for pattern mining”, Expert Systems with Applications 42 (2015) 451-461.
5. Omer M. Soysal, “Association rule mining with mostly associated sequential patterns”, Expert Syatems with Applications 42 (2015) 2582-2592.
6. Michihiro Kuramochi and George Karypis, “An efficient algorithm for discovery frequent subgraphs”, VoL.16 No.9 September (2004).
7. Ivan F. Videla-Cavieres, Sebastian A. Rios, “Extending market basket analysis with graph mining techniques: A real case”, Expert Systems with Applications, 41(2014) 1928 – 1936.
8. Bose, I., & Mahapatra, R. K. (2001). Business data mining – A machine learning perspective. Information and Management, 39, 211–225.
9. Giudici, P., & Passerone, G. (2002). Data mining of association structures to model consumer behavior. Computational Statistics and Data Analysis, 38, 533–541.
10. Nasraoui, O., Soliman, M., Saka, E., Badia, A., & Germain, R. (2008). A web usage mining framework for mining evolving user profiles in dynamic web sites. IEEE Transactions on Knowledge Data Engineering, 20(2), 202–215.
11. Borges, J., & Levene, M. (2007). Evaluating variable-length markov chain models for analysis of user web navigation sessions. IEEE Transactions on Knowledge Data Engineering, 19(4), 441–452.
12. Nasraroui, O., Spiliopoulou, M., Srivastava, J., Mobasher, B., & Masand, B. M. (2007). Advances in web mining and web usage analysis. WebKDD (Vol. 4811). PA, USA: Springer.
13. Jiang, T., & Tuzhilin, A. (2006). Segmenting customers from population to individuals: Does 1-to-1 keep your customers forever. IEEE Transactions on Knowledge Data Engineering, 18(10), 1297–1311.
14. Yeh, M.-Y., Dai, B.-R., & Chen, M.-S. (2007). Clustering over multiple evolving streams by events and correlations. IEEE Transactions on Knowledge Data Engineering, 19(10), 1349–1362.
15. Hu, M. Y., Shanker, M., Zhang, G. P., & Hung, M. S. (2008). Modeling consumer situational choice of long distance communication with neural networks. Decision Support System, 44(4), 899–908.
16. Xin Guan, Guidong Sun, Xiao Yi and Qiang Guo, “A Novel Data Association Algorithm for Unequal Length Fluctutant Sequence”, Procedia Engineering 99 (2015) 1190 – 1202.
17. Reader, T., & Chawla, N. V. (2009) Modeling a store’s product space as a social network. In 2009 International conference on advances in social network analysis and mining (PP. 164-169). IEEE.
18. Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3), 75 – 174
19. Newman. M and Girvan. M (2004). Finding and evaluating community structure in networks, Physical Review E, 69(2), 026113.
20. Xie, J., Szymanski, B., & Liu, X. (2011). SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In ICDMW 2011, 11th IEEE international conference on data mining.
21. Xie, J., Kelley, S., and Szymanski, B.(2013). Overlapping community detection in networks: The state of the art and comparative study. ACM Computing Surveys, 45(4), 1-37.
22. Gregory, S. (2010). Finding overlapping communities in networks by label propagation. New journal of Physics, 12(10),103018
23. Raghavan, U., Albert.R and Kumara. S(2007). Near Linear time algorithm to detect community structures in large-scale networks, Physical Review E, 1-12.
24. Agrawal, R., & Srikant, R. (1994). Fast algorithm for mining association rules in large database. Research Report RJ 9839, IBM Almaden Research Center, Santiago, Chile
25. Corrado Loglisci, Michelangelo Ceci, Donato Malerba, “Relational mining for discovering changes in evolving networks”, Neurocomputing 150(2015) 265 – 288.