Pakistan Journal of Engineering and Technology, PakJET
Multidisciplinary & Peer Reviewed
Volume: 04, Number: 01, Pages: 196- 200, Year: 2021

Discovering Beneficial Associations Among Products Through Network Analysis

Amna Iftikhar, Talha Hafeez

Department of Computer Science Bahria University, Karachi, Pakistan
Corresponding author: Talha Hafeez (e-mail: Talha.hafeez1997@gmail.com).

***Abstract*- A common problem faced by retail and online stores is to recommend and advertise the appropriate products to their customers. This problem and many others including shelf management and designing sales promotion can be solved by determining the set of products sold together. Classic approaches of market basket analysis are not of much use when a large amount of data is being processed. In this paper we use graph mining approach to determine the important associations among products that can benefit supermarkets or retail stores.**

***Index Terms*-- Networks motifs; Market basket analysis; Association rules; Centrality measures; Product network**

I. INTRODUCTION

Many stores and supermarkets maintain profiles of customers and their transactions by issuing loyalty cards. This data regarding products and customers can be used to extract important information that can help the organization to make key decisions about product placement, promotions, profitability and pricing. These cards are also useful for organizations in determining customers purchasing habits as well. Based on the products purchased by any particular customer, appropriate products can be recommended to relevant customers and products can be advertised that a customer is more likely to purchase. Furthermore, the products that are purchased together can be placed near to each other to facilitate customers.

Market basket analysis is a technique used to discover customers' purchasing habits. The aim of market basket analysis is to find out significant relationships among purchased products. A common approach used for extracting meaningful information from the stored transactions data is called association rule mining. An association rule set is defined by a set of transactions along with the required support and confidence parameters. Frequent item set discovery or association rule mining helps to determine which products are sold together and is a classical approach for market basket analysis. However, when a large amount of data is being processed, association rule mining yields a number of rules that are not of much use. Out of these rules many are obvious and redundant. In this paper we attempt to provide solutions for these challenges for increasing the influence and clearness of market basket analysis while modeling transactional data as a product network. By motifs detection, expressive relationships between products can be discovered; as well as discovery of relationships that are challenging to discover with traditional association rules is also possible. The representation of transaction data as a product network allows for the use of a number of different analysis techniques that are new and were unavailable for association rules extraction earlier.

II. LITERATURE REVIEW

Videla Cavieres et al. [1] have mentioned that when the data is of high dimension and dispersed the common techniques of market basket analysis or association rule mining are unable to discover meaningful information. Their paper presents an approach for association rule mining using graph mining techniques that makes processing of millions of transactions easier. The efficacy of the approach is demonstrated in a retail sale supermarket and wholesale supermarket chain using classical approach to extract import associations from retail stores data with the help of techniques like market basket analysis, discovery of frequent item sets and techniques for clustering like K-means clustering [3], SOM [4]. However, the results obtained were not of much use and the clusters discovered using these approaches did not show any differences and no new information can be extracted. Therefore a product network is made from the given transactional data where the nodes symbolize products and edges symbolize relationships between two nodes. Temporally Transactional Weighted Product Network is generated by applying a temporary set of filters. Then a weighted network of product–to–product relationship is obtained. This network is illustrated by an adjacency matrix that shows the weight among each pair of products. Community detection is used to determine important associations [5]. Overlapping community detection [7] is then used as an additional aspect of classic community detection. The product networks also present nodes with high degree and have spurious edges present between them. For the removal of spurious edges, a threshold s is defined after that the graph is reread for the search of edges with a weight s0 lower than s. The unconcerned edges match this criterion and are removed. The network shows expressive regions now after application of different filters. These regions describe products with an influential relationship between them.

Raeder et al. [6], present a different approach for mining product transaction data. First data is modeled as a product network and expressive communities (clusters) are discovered and then they

are analyzed further. Each node in the product network symbolizes a product and an edge between two products signifies that those two products were bought together in the same transaction operation. Communities are detected using strongly connected nodes. These communities were then analyzed using utility measures like association rule networks and center-piece subgraphs. The researchers aim to provide a solution for the question. "Given an unseen market

Approach	Pros	Cons
K-Means Clustering	Easy to implement and interpret	High values of clustering co-efficient on vertices of small degree
Frequent Itemset discovery	Identification of important trends.	Does not work for large datasets
Community Detection	Uncover meaningful relations among products.	Reveal very little about the crux of the network.
Motif Analysis	Identify recurring patterns in the network	Computationally expensive for large data sets.

TABLE I
COMPARISON OF DIFFERENT APPROACHES USED FOR MARKET BASKET ANALYSIS

basket dataset, what set of steps should I follow to conduct a thorough, complete analysis?" by providing a list of steps that comprehensively concludes their research.

Srivastava et al. [9] have built an association rules network by using association rules. They have shown that data repositories can be categorized as experimental or observational. Observational data is collected without any particular task in mind and is undirected. Therefore, observational data is high dimensional data whereas experimental data is low dimensional or small. Experimental data is collected with a specific task in mind and is directed. Therefore, in experimental data feature selection that is required for some target variable is not necessary whereas for observational data feature selection is important to extract the meaningful relationships that exist among the data. Association rule network (ARN) can be used for selecting relevant features by using the following four steps: First the data is prepared for association rule mining (This is done by converting the data into transactions and then continuous valued variables are converted into discrete values). Then required support and confidence are selected to apply an association rule mining algorithm to determine association rules. After that Association Rule Network ARN is built and for feature selection they applied a clustering algorithm on ARN. The paper is similar to the work done by Srivastava et al. [9] that use motif analysis to discover significant associations. In this paper we attempt to discover meaningful information from Amazon Co-Purchased network using motif analysis, community detection as well as by using centrality measures.

III. METHODOLOGY

This section describes the basic process model proposed in this paper as shown in Fig. 1 and is divided into two main parts. Part one defines how the transactions data is collected, the nature of the data and how this data is used to build a products network. Second part defines the proposed network analysis methodology. This section is further divided into four sub-sections. First, degree distribution is used to understand the nature of the network, then centrality measures are used to detect significant products in this network. Furthermore, community detection and motif analysis are used to discover important associations among products.

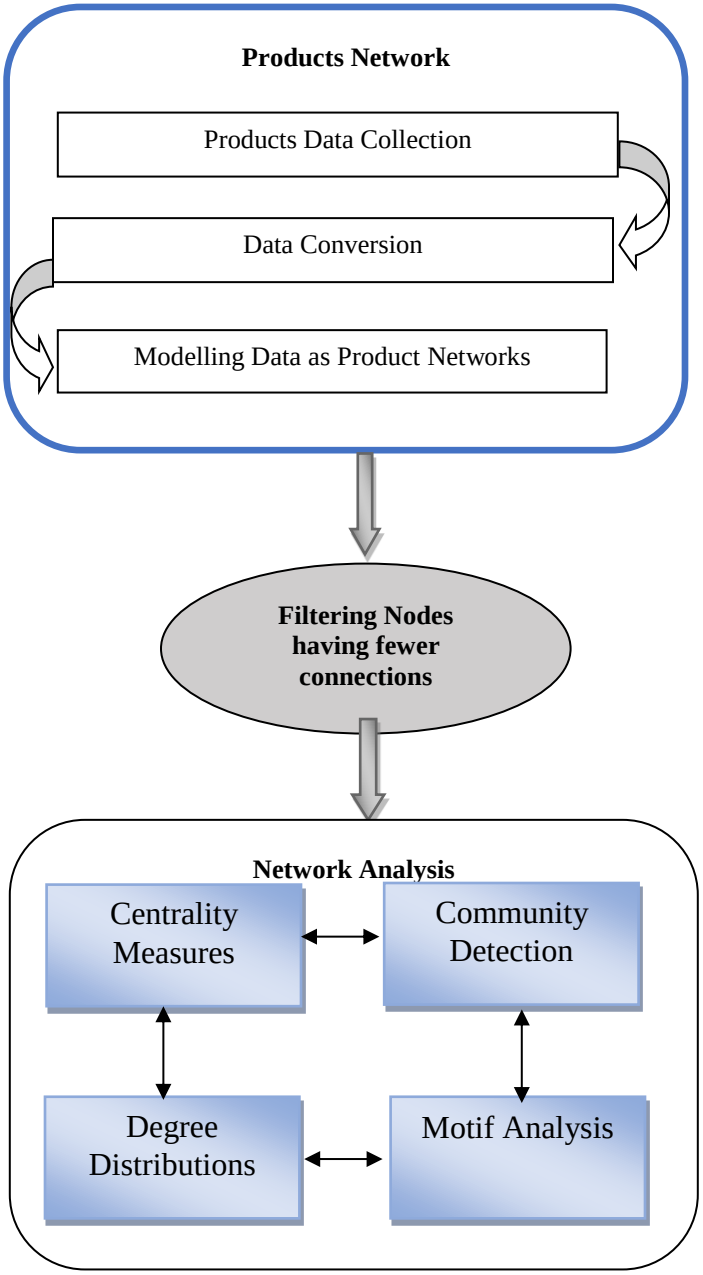

FIGURE 1. Process Model

A. NATURE AND DETECTION OF DATA SET

The dataset used in this paper is Amazon product co-purchasing network data from Stanford Network Analysis Project. The data was collected by crawling Amazon websites. It is based on

"customers who bought this item also bought" feature. If a product i is frequently co-purchased with product j, the graph contains a directed edge from i to j. The data consists of 262111 nodes and 1234877 edges. The data is in a text file, from which is converted into a CSV file to be used in Gephi software for network analysis.

Id	15
ASIN	1559362022
Title	Wake Up and Smell the Coffee
Group	Book
salesrank	518927
Similar	5 1559360968 1559361247 1559360828 1559361018 0743214552
Categories	3-Books[283155]-Subjects[1000]→Literature & Fiction[17] →Drama[2159]-United States[2160] →Authors, A-Z[70021]-(B)[70023]-Bogosian, Eric[70116] →Arts & Photography[1]-Performing Arts[521000]-Theater[2154]-General[2218]
Reviews	total: 4 downloaded: 4 avg rating: 4 2003-6-27 cutomer: A39QMV9ZKRJXO5 rating: 4 votes: 1 helpful: 1 2004-2-17 cutomer: AUUVMSTQ1TXDI rating: 1 votes: 2 helpful: 0 2004-2-24 cutomer: A2C5K0QTLL9UAT rating: 5 votes: 2 helpful: 2 2004-10-13 cutomer: A5XYF0Z3UH4HB rating: 5 votes: 1 helpful: 1

TABLE II: SAMPLE AMAZON PRODUCT METADATA

The data was then imported for further analysis in Snap and NetworkX libraries of Python. Along with the edge-list representation of co-purchased products, the metadata is also present for each product that contains all the information about each product including product group categories, reviews etc. as shown in Table II.

B. NETWORK MODEL

Product networks tend to be very dense in nature. Figure 2 shows the visualization of Amazon products network used in this paper. We strive to examine different properties of product networks. Each product is represented by a node in the network, and any two products were bought together in a transaction have an edge between them. Product networks differ from other interaction networks because of the fact that the presence of an edge does not necessarily indicate an established relationship between products. For example, citations or cell phone calls networks, do not suffer this problem to almost the same degree. Citation networks have edges that link two nodes that are related due to some reason (one paper cites another). Similarly, cell phone networks have a very few incidental links, (wrong numbers, random personal business or telemarketing), but mostly there is a connection between people who call one another. On the contrary, product networks might contain an edge between two products that were bought together but do not have any connection between them. For example there is no common motivation for purchase of two products like cheese and towels present just because a person buys them in the same transaction. Also, if a person buys several unrelated items in a single transaction, a clique would be formed among them, despite having any true relationship. All of this results in product networks being very dense and heavy tailed, with each node containing a huge number of connections, even though many of these connections don't have any meaning.

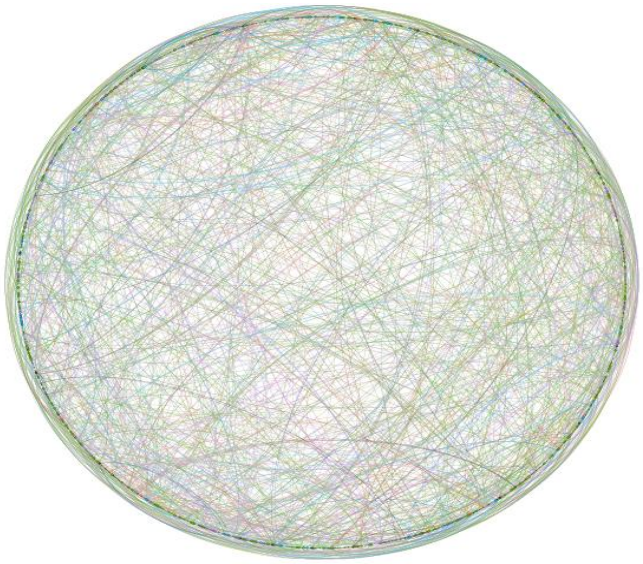

FIGURE 2: Amazon Network Product Visualized in Gephi

IV. NETWORK ANALYSIS

Since the network consists of a large number of nodes, there are many nodes which occur spontaneously and do not represent a true relationship among products. These unintentional associations should be pruned from the network to simplify the network and also to improve the quality of our analysis. This can be done by defining the minimum in-degree a node can have. Removing unwanted edges from the network corresponds to specifying of minimum support and confidence threshold in association rule mining.

A. DEGREE DISTRIBUTION

The degree distribution of the modeled network shows a heavy tailed distribution with some nodes having a very high degree while others having low degree and therefore depict the scale free nature of the network.

B. COMMUNITY DETECTION

within a network represent a group of nodes that are more strongly linked to each other than they are linked to the rest of the network. These communities are helpful in analyzing important parts of a network. For example, in a telephone network a group of people usually family or friends form a strong community. Similarly, a group of products that are mutually purchased can form a community that can help retails gain insight into the customers' purchasing habits. After removing unwanted edges there are 2141 nodes and 2657 edges in the network. There are a total of 688 connected components in this network containing from a maximum of 880 to a minimum of 1 product in their network. The first connected component consists of 880 products and is further analyzed using motif analysis. In second strongly connected component there are three products as shown in Fig. 3. This component consists of two books and a video Strange Kids, Safe Kids. This video is targeted for both parents and children regarding children safety. These books and video seem to be recommended for adult to middle aged married women who are interested in home repairing, beauty, and safety of kids. The strong relationship among these products depict that they are usually purchased together.

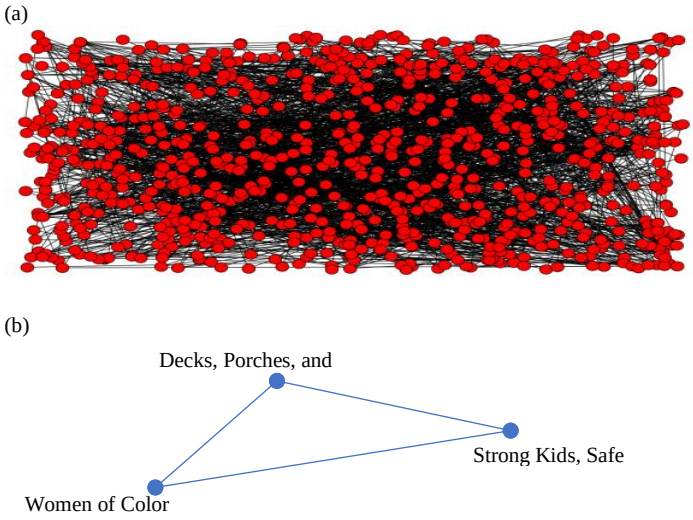

FIGURE 3: The first two a and b connected model of our network

C. CENTRALITY MEASURES

Centrality measures are used to indicate the most significant nodes in a network. To identify the most central nodes in our network we use betweenness and closeness centralities. Closeness centrality identifies those nodes that have shortest path from that node to all others nodes in the network. Whereas, betweenness centrality depicts those nodes that act as a bridge to all other nodes in the network. Table III shows the formulae of closeness centrality and betweenness centrality. Table IV shows the top ten products that are significant due to high values of betweenness and closeness measures. Some nodes appear in both columns. For example a DVD named The Time Machine. This shows that these products are frequently bought or are popular products.

Closeness	Betweenness
$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$	$C(x) = \frac{1}{\sum_y d(y,x)}.$

TABLE III
FORMULAE OF CLOSENESS AND BETWEENNESS CENTRALITIY

Closeness	Betweenness
The Time Machine(DVD)	Losing Matt Shepard(Book)
Committed(Music)	The Time Machine(DVD)
The Narcissistic Family(Book)	BizPricer Business Valuation Manual w/Software(Book)
Ultimate Sniper(Video)	Committed(Music)
Harley-Davidson Panheads, 1948-1965/M418(Book)	Ultimate Sniper(Video)
Made Again(Music)	Harley-Davidson Panheads, 1948-1965/M418(Book)
The Cheyenne Social Club(Video)	Stories for a Teacher's Heart (Stories For the Heart)(Book)
Stories for a Teacher's Heart (Stories For the Heart)(Book)	More Than You Think You Are(Music)
More Than You Think You Are(Music)	A Touch Of Tranquility(Music)
Implementing E-Learning (Book)	Implementing E-Learning (Book)

TABLE IV
TOP TEN PRODUCTS WITH HIGH BETWEENNESS AND CLOSENESS

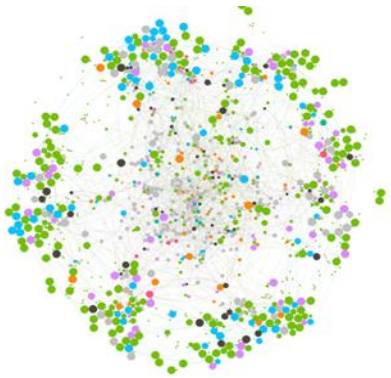

FIGURE 4: Central Nodes Having High Value Nodes Are Shown By Their Size

Subgraphs that incur repeatedly in a network are called network motifs. They can be used to indicate prominent associations among products and can help retailers gain insight into which products are bought together. In our analysis all 3-node motifs are searched through the pruned network. In our dataset 764 nodes were found to represent all 3-node motifs in Amazon Co-Purchasing network. This network of motifs is shown in Fig. 4. Out of these motifs only those are considered for further analyses that have high frequency of occurrence as shown in Fig. 5 and 6. This network shows important associations among products ranked according to the most occurring pair of products in the motifs network. These motifs are a

representation of the fact that when a customer purchased some item, he/she went ahead and purchased both or either of the two products. For example Motif ID 1 shown in Table represents the most prominent relationship that is between products 2353, 2501 and 4429. Product 2353 is music item named Committed produced by a group of six male students. The group also won the musical competition named "The Sing Off". The second product in this motif is a book named The Narcissistic Family in which the author presents a therapeutic model for treating and understanding adults that come from abusive or negligent families. The book was very popular among teenagers who feel worthless, have emotional problems and are bullied or abused at home. Product 4429 is also a book named "Harley-Davidson Panheads" which provides information, repair and maintenance instructions for Harley Davidson Panheads motorcycle. Therefore, it can be incurred from the above product details that young teenagers who listen Committed, having Harley Davidson Panheads are also interested in reading The Narcissistic Family

Motif ID	Frequency Of Occurrence
1	15
2	13
3	11
4	11
5	10
6	9
7	8

Table V:
DETECTED 3-NODE MOTIFS IN NETWORK AND SORTED WITH THEIR FREQUENCY OF OCCURRENCE

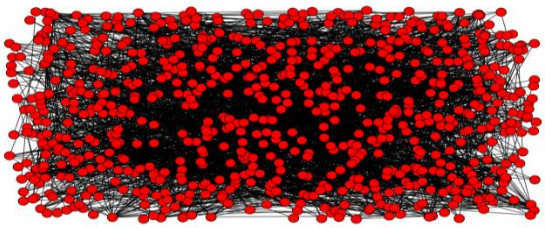

FIGURE 5: Network Of All Three Motifs Identified

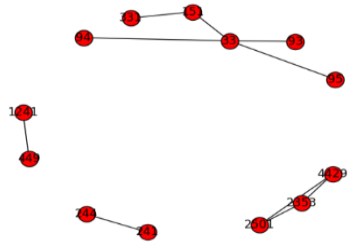

FIGURE 6: Significant motifs ranked according to the most occurring pairs of products

V. CONCLUSION

Numerous researches have been presented over products in the context of market basket analysis. In this paper we strive to present a framework for effective market basket analysis with the help of centrality measures, motif analysis and community detection. First basic network analysis techniques were applied and it was observed that many nodes and edges don't provide any significant information. Filtering techniques were applied and nodes having fewer connections were discarded. Then different degree distributions and centrality measures for products network were observed in order to understand the nature of the network and to detect significant products in the network respectively. Since product networks are very dense, there are a large number of communities that don't give much insight into the true relationships among products. Therefore, 3 node motifs are discovered from the network to identify significant associations between products which is much difficult to find using classical approaches of association rule mining.

REFERENCES

[1] Videla-Cavieres, Ivan F., and Sebastián A. Ríos. "Extending market basket analysis with graph mining techniques: A real case." Expert Systems with Applications vol. 41, no. 4, 2014: 1928-1936.

[2] Telgarsky, Matus, and Andrea Vattani. "Hartigan's Method: k-means Clustering without Voronoi." AISTATS. 2010.

[3] Kohonen, Teuvo. "The self-organizing map." Proceedings of the IEEE vol. 78, no. 9, 1990: 1464-1480.

[4] Dorso, Claudio Oscar, and A. D. Medus. "Community detection in networks." International Journal of Bifurcation and Chaos vol. 20, no. 2, 2010: 361-367.

[5] Xie, Jierui, Stephen Kelley, and Boleslaw K. Szymanski. "Overlapping community detection in networks: The state-of-the-art and comparative study." ACM Computing Surveys (csur), vol. 45, no.4, 2013: 43.

[6] Raeder, Troy, and Nitesh V. Chawla. "Market basket analysis with networks." Social network analysis and mining, vol. 1, no.2, 2011: 97-113.

[7] Loraine Charlet, Annie, Ashok Kumar "Market Basket Analysis for a Supermarket based on Frequent Itemset Mining" IJCSI International Journal of Computer Science Issues, vol. 9, 2012.

[8] Chawla, Sanjay. "Feature Selection, Association Rules Network and Theory Building." FSDM. 2010.

[9] Srivastava, Abhishek. "Motif analysis in the amazon product co-purchasing network." arXiv preprint arXiv:1012.4050, 2010.

[10] Hahsler, Michael, and Radoslaw Karpienko. "Visualizing association rules in hierarchical groups." Journal of Business Economics, 2011: 1-19.

[11] Norulhidayah Isa, Nur Syuhuba Mohd Yusof, and Muhammad Atif, "The Implementation of Data Mining Techniques for Sales Analysis using Daily Sales Data" IJATCSE, 2019.