Università
Ca'Foscari
Venezia

Master's Degree programme - Second Cycle
(D.M. 270/2004)
In Informatica - Computer Science

Final Thesis

–
Ca'Foscari
Dorsoduro 3246
30123 Venezia

Market Basket Analysis with Network of Products

Supervisor:
Prof. Salvatore Orlando

Candidate:
Nikhil Verma
Matriculation number - 855183

Academic Years:
2016/2017

Ca' Foscari, University of Venice

Market Basket Analysis with Network of Products

Nikhil Verma

855183

A thesis submitted in partial fulfilment for the

Degree of Master of Science

in the

Department of Environmental Sciences, Informatics, and Statistics

Computer Science

June 2017

”There were 5 Exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days.”

Eric Schmidt

ACKNOWLEDGMENTS

I feel extremely honoured for the opportunity to work under the supervision of Prof Salvatore Orlando, who guided this work from the very beginning, who was always happy to meet every time I was stuck in my work. And it is my privilege to bring on record my sincerest and heartfelt gratitude to my esteemed supervisor, for his outstanding and invaluable guidance, keen interest and tremendous cooperation.

I would like to extend my heartfelt thanks to friends from the university for their contribution to the work done in this thesis. Francesco Cagnin and Ismail Elezi for assisting me with graph theory and with whom I had productive discussions.

My thanks also go to all the members of the Dipartimento di Scienze Ambientali, Informatica e Statistica (DAIS) for providing a good learning atmosphere in the department. The same appreciation goes to the administration of Università Ca 'Foscari Venezia and the region of Veneto for their incredible support to the international student regarding accommodation, career guide, exchange programs with financial aid. Special thanks also to the local students for accommodating the international students well.

The social events organized by DAIS students body did helped mingle and make new friends. The same appreciation goes to my fellow international students, most notably the Ethiopians, Pakistanis, Iranian, Italian, etc. for their remarkable friendship and support during this journey. Life would have been difficult without you all.

I cannot end without thanking my parents, for their unconditional moral and financial support when it was most needed. Special thanks to my sister Neha Verma Dhingra for reviewing my thesis and asking dozens of critical questions about my work. I hope that with this thesis I have made them a little proud! Thanks for the trust bestowed on me.

ABSTRACT

In today's world, big data plays a crucial role in both public and private sector. Leading companies have collected a vast amount of data and information. Data mining is the process of extracting useful information from the raw data. It has become a new trend in business, to analyze the data to make more targeted business model or important decision in defining customer profiles. The facts which otherwise may go unnoticed can be now revealed by using different techniques that mines stored information. However, due to costs, privacy and data protection, only the business owned transactional data is typically available for analysis.

To find meaningful patterns in the big data is one of the active research in the data mining. Market basket analysis (MBA) is one of the most useful modeling technique in data mining. It involves the mining and analysis of Association Rules, which take the form of a famous statement such as *"people who buy diapers are likely to buy beers"*. Such information can be used as a basis for decisions about marketing activity such as promotional support, inventory control and cross sales campaign.

In this thesis, we extracted Frequent Itemsets followed by Association Rules and use this knowledge to model the data as a product network. To mine transactional data and get detailed information about the products bought together in the market basket, we performed network based approach. This gives insights into frequently bought products and products bought together defined as "Communities". Further, we analyzed these group of communities using different measures like Network density, Centrality and PageRank algorithms. From this study, we concluded that finding communities in comparison to the old traditional method of finding Association Rule is more informative and reliable in MBA.

Contents

ABBREVIATIONS

ARN Association Rules Network

CePS Centre-Piece Subgraphs

CSV Comma Separated Values

LHS Left Hand Side

MBA Market Basket Analysis

PC Potential Connections

RHS Right Hand Side

WWW World Wide Web

List of Tables

List of Figures

1

INTRODUCTION

1.1 Background

Big data has been called 'the new discovery' in today's world. But it came with a problem: 'how to interpret them'. Generating the data has now become very cheap and it has become very difficult and expensive to mine the collected data. As big data is considered valuable resource, a variety of companies are relying on it for making effective and strategic decisions. To come up with new marketing ideas, or the facts that originates from stored information which may go unnoticed can now be revealed by the data mining techniques. Therefore, different algorithms that can work with big data and generate results in time, are more needed than ever.

The data mining techniques are becoming common for many businesses. This is possible because of the high technological era we live in, which has made possible for the companies to store and gather a large amount of data. Data mining techniques gained popularity in last two decades as there was an increase in computation which motivated not only mining data but especially big data. The collection of data on the different level is considered as raw material for extracting valuable information which leads to knowledge discovery. Some of the facts and figures can be noticed directly considering the data, but we are always interested in hidden rules and patterns which reveal real and effective information.

The study and analysis of retail transaction data, known as 'Market Basket Analysis,' has become increasingly important in the past years. Market basket analysis is a data mining technique which is widely used in the consumer purchase goods. It allows us to identify 'patterns' in customer purchases by extracting co-occurrences from a store transactional datasets. It takes its name from the customers keeping all their purchases into a shopping cart (market basket) during shopping. A shopkeeper could use this information to place products frequently sold together into the same area, while e-commerce site could use it to determine the layout of their presentation. To have the right product in the right place in right time without having an overflowing warehouse is very important. Data mining techniques results in better selection and decision, less inventory cost and more revenue. To get overview of the relationships between the products, data mining algorithms can give us more information.

The main task of MBA is to discover actionable knowledge in transaction databases. Ultimately, an effective analysis should enable the dealer to draw clear and comprehensive conclusions from the data.

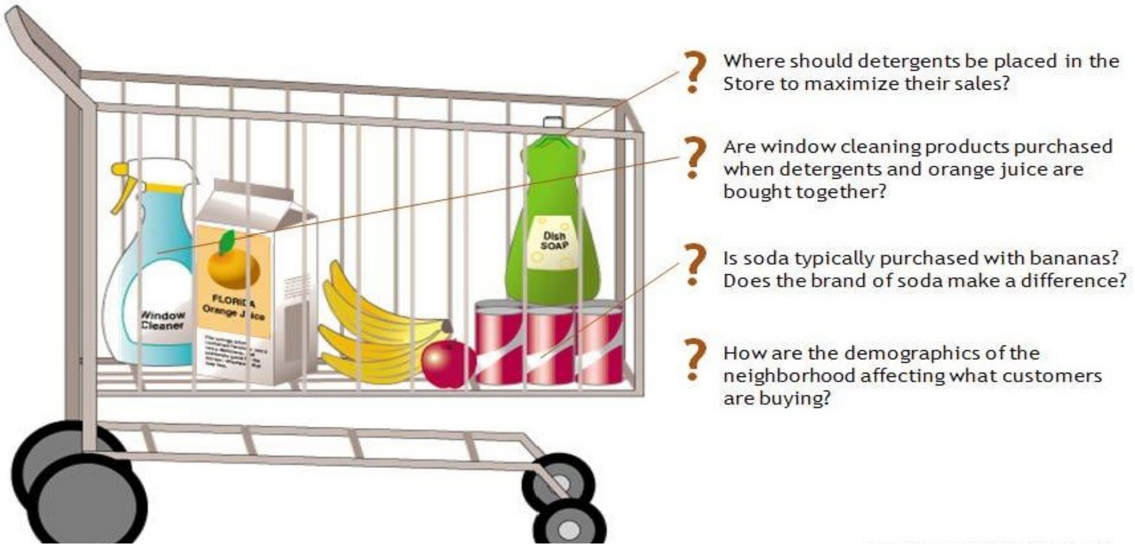

Where should detergents be placed in the Store to maximize their sales?
Are window cleaning products purchased when detergents and orange juice are bought together?
Is soda typically purchased with bananas? Does the brand of soda make a difference?
How are the demographics of the neighborhood affecting what customers are buying?
Window Cleaner
FLORIDA Orange Juice
Dish SOAP

Fig. 1.1 **Market basket (Grocery cart).**[dee14]

We can see in Figure 1.1. the number of the products bought together in a single transaction, so many questions can be raised like:

1. Which two or more products are bought together?

2. Where should be product kept in the store to increase its sale?

3. How the brand of the product affects its sale when bought together with other products?

These are the basic question raised in market basket analysis. To answer them in this thesis we use network product approach and then find the group of products called communities. Communities of products help us to determine which products are bought together frequently in the market basket. In this work, we use different measures to analyze them which provide substantial insight into customer behaviour.

1.2 Drawbacks

The best-known limitations on finding interesting rules are minimum Support and Confidence parameters. Support is the indication of how frequently the Itemset appears in the dataset, whereas confidence is an indication of how often the rule has been found true. First, there is no obvious explanation for choosing appropriate support and confidence values. If the value is chosen too high, interesting co-occurences may be lost. However, if its chosen too low, the user may be flooded with thousands of weak rules that do not represent meaningful associations. Finding patterns can be incredibly sensitive to the choice of support and confidence parameters. Similarly, to model network of products we set minimum threshold to prune items which do not share any significant information. The second issue is that transaction databases often contain hundreds or thousands of rules at reasonable levels of support and confidence, and many of those rules are simply obvious.

1.3 Motivation

Our approach is inspired by work done by Nitesh V. Chawla [RC11], who formulates the concept of "Product Network". There was not much work done on market basket analysis after Association Rules were discovered by Agarwal et al. No existing work of which we are aware has attempted to offer any procedural guidance for analysing such data. In another word, no work has addressed the question:

"Given a new market basket dataset, what method or methods should we apply in order to obtain effective insights?"

Our objective in this thesis is to find communities of products from product network built from any transactional dataset where a transaction contains set of items bought together. To have more detailed information on the products bought together was the main motivation to find communities of products and analyze them. By detecting communities of products, we can discover the strong and expressive relationships among products including relationships that are difficult to discover with Association Rules. To find communities, we built the network of products first. We tried this approach to improve the power and clarity of market basket analysis by modelling transactional dataset as a network of products. The network representation of the data allows us for the use of a diverse array of algorithms which were not available to the Association Rules community. Many items were pruned while modelling the network as we set minimum threshold.

1.4 Thesis Overview

The structure of the thesis is as follows:

-Chapter 1 starts with the introduction, motivation and drawbacks.

-Chapter 2 starts by discussing Frequent Itemsets. Followed by, briefly exploring the strengths and weakness of Association Rules analysis on the transactional dataset.

-Chapter 3 introduces the concept of product network and then we present some properties of the network, built from the transactional dataset. After that, we discover communities of products and analyze them.

-Chapter 4 starts by briefly introducing Centre-piece subgraph (CePS) and then we talk about choosing the minimum support parameter.

-Chapter 5 discusses experiments and results.

-Chapter 6 is the final chapter of the thesis. In this chapter, we review the work done in the previous chapters, make some conclusions and point out what can be done in future work.

2

ASSOCIATION RULES

2.1 Frequent Itemsets

Definition: a set of products that appears in many transactions is termed as "Frequent." We assume there is a number s, called the minimum support. If I is a set of products, the support for I is the number of transactions for which I is a subset. We say I is frequent if its support is s or more.[PBTL99]

The market basket of transactional data describes many relationships between products. Mining Frequent Itemsets is usually among the first steps to analyze a large-scale dataset containing transactions, which has been an active research topic in data mining for years [RC11]. It finds fascinating patterns such as Association Rules, correlations sequences, classifiers, clusters and many more, of which, the mining of Association Rules is one of the most popular problems.

Sets of products are typically called Itemsets and Itemsets of size k are called k-Itemsets and sets that meet the minimum support criterion are called Frequent Itemsets. Association rules are derived from the Frequent Itemsets at minimum support and confidence parameters [LCK12]. In this work, we have generated Frequent Itemset using data mining algorithms called Apriori [AS^{+}94], other algorithms which can be used are FP-growth algorithm [HPY00], Eclat [HPK11] and K-Apriori [HHR06].

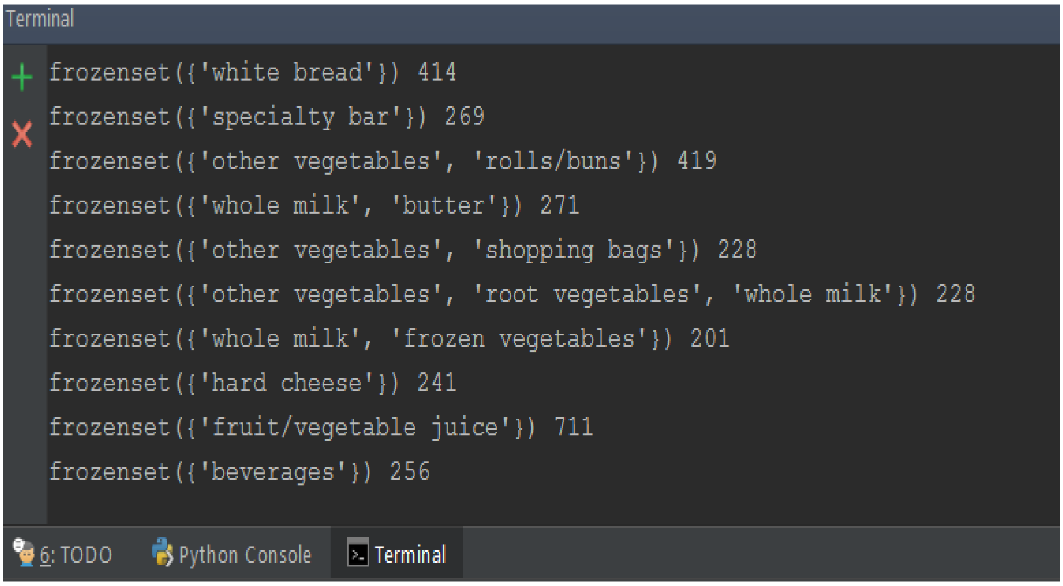
2.1.1 Apriori algorithm pseudocode

I have used "Pymining" which is a library in Python which has this data mining

Algorithm 1 Apriori algorithm

```
Require: Apriori (T, minSupport) {T is the database and minSupport is the minimum
   support}
Ensure: L1= frequent items;
   for (K = 2;$L_k$-1! = ∅; K++)
   $C_k$ = candidates generated from $L_k$-1
   {that is cartesian product $L_k$-1 x $L_k$-1 and eliminating any k-1 size itemset that is not
   frequent}
   for each transaction t in database do
   $L_k$ = candidates in $C_k$ with minSupport
   {increment the count of all candidates in $C_k$ that are contained in t}
   {end for each}
   {end for}
   return $U_k L_k$
```

algorithm [VR+07]. It uses Apriori algorithm to find Frequent Itemsets.



```
frozenset({'white bread'}) 414
frozenset({'specialty bar'}) 269
frozenset({'other vegetables', 'rolls/buns'}) 419
frozenset({'whole milk', 'butter'}) 271
frozenset({'other vegetables', 'shopping bags'}) 228
frozenset({'other vegetables', 'root vegetables', 'whole milk'}) 228
frozenset({'whole milk', 'frozen vegetables'}) 201
frozenset({'hard cheese'}) 241
frozenset({'fruit/vegetable juice'}) 711
frozenset({'beverages'}) 256
```

Fig. 2.1 **Frequent Itemset found in the dataset with minimum support = 200). We can see that there are single items like {white bread} = 414 in a transaction and similarly we have double or triple Itemsets in the dataset which are Frequent Itemsets.**

In Figure 2.1 - we can see {"{other vegetables}", "{root vegetables}", "{whole milk"}}: 228, is a Frequent Itemset where all the items in this set are brought together most of the time. 228 is the number of times these items are bought together in the dataset.

This analysis shows a association between these products as they are bought together. Finding Frequent Itemsets is very helpful because it gives us an idea mathematically that which items are important or bought frequently in the dataset. Knowing these results, we can arrange the products strategically in a store or do something to enhance the sales and production of the products. This technique is very fundamental and should be used in market basket analysis as a first step.

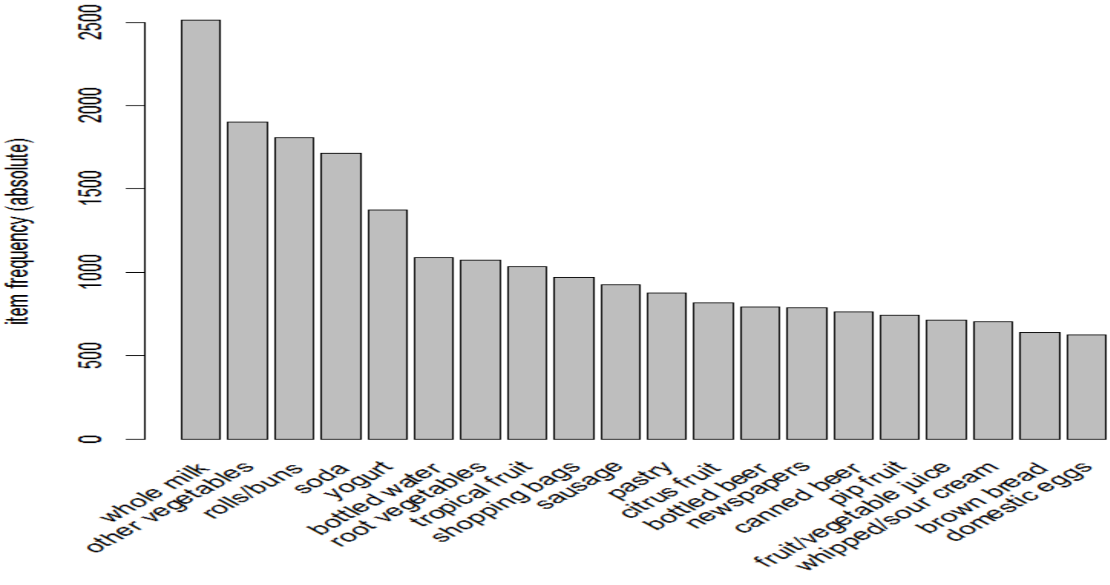


***Fig. 2.2* Item Frequency bar plot.**

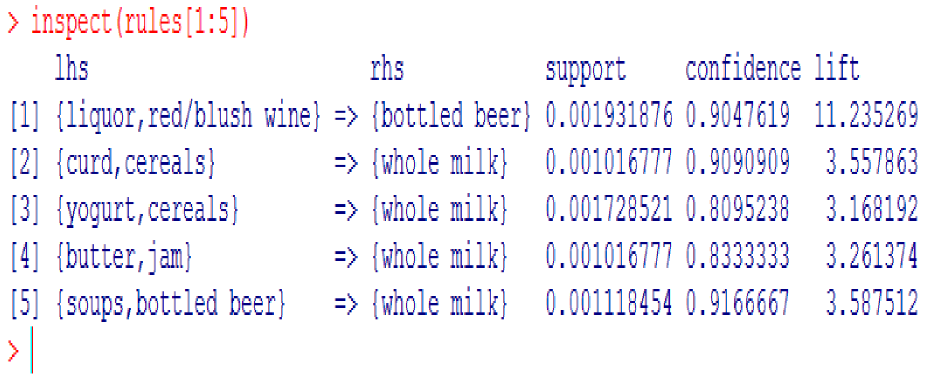
In Figure 2.2 - we can see that {whole milk} is the most frequent item in the dataset which occur with an absolute value of 2500. It shows that this item occurs in most of the transactions in the dataset. Similarly, the figure also shows other items with decreasing frequency. This information gives us insight about choosing our Support and Confidence which are key parameters for finding Association Rules.

2.2 Association Rules

Association Rule mining is a rule-based machine learning method for finding interesting relations between products in large datasets. A set of Association Rules *R (T, s, c)* is defined by a transactional database T, a minimum support parameter s and a minimum confidence parameter c. It, defines A as the set of transactions containing every product in $A(B)$. Formally, R is the set of all rules A -> B such that:

1. $\frac{(A \cap B)}{|T|} \geq S$

2. $\frac{(A \cap B)}{|A|} \geq C$



```
> inspect(rules[1:5])
    lhs                        rhs            support     confidence lift
[1] {liquor,red/blush wine} => {bottled beer} 0.001931876 0.9047619  11.235269
[2] {curd,cereals}          => {whole milk}   0.001016777 0.9090909   3.557863
[3] {yogurt,cereals}        => {whole milk}   0.001728521 0.8095238   3.168192
[4] {butter,jam}            => {whole milk}   0.001016777 0.8333333   3.261374
[5] {soups,bottled beer}    => {whole milk}   0.001118454 0.9166667   3.587512
>
```

***Fig. 2.3* Association Rules (Support 0.001 and Confidence 0.8). We can see, whenever customer buy {liquor, red/blush wine} they also tend to buy {bottled beer} with Support= 0.0019 and Confidence= 0.904, which defines a rule.**

In Figure 2.3 - we show the results obtained for Association Rules in R, where we have products in the transactions at values for support = 0.001 and confidence = 0.8. For example {whole milk} is almost bought with every Itemsets on left hand side (LHS). Likewise, whenever {soups, bottled beer} are brought together in a transaction there is an almost a chance that {whole milk} is bought with it. These kinds of rules are extracted by Association Rule mining which gives us an idea to model network approach and then discover communities for further analysis. In "Association Rule Learning", "Lift" is defined as the ratio by which the measure of performance of Association Rules exceeds the expected confidence. An Association Rule is said to be supported in a transaction database if it meets both the minimum support and minimum confidence criteria.

Association Rules have found successful application in many diverse areas and several algorithms have been developed to discover them efficiently, but they have limitations.

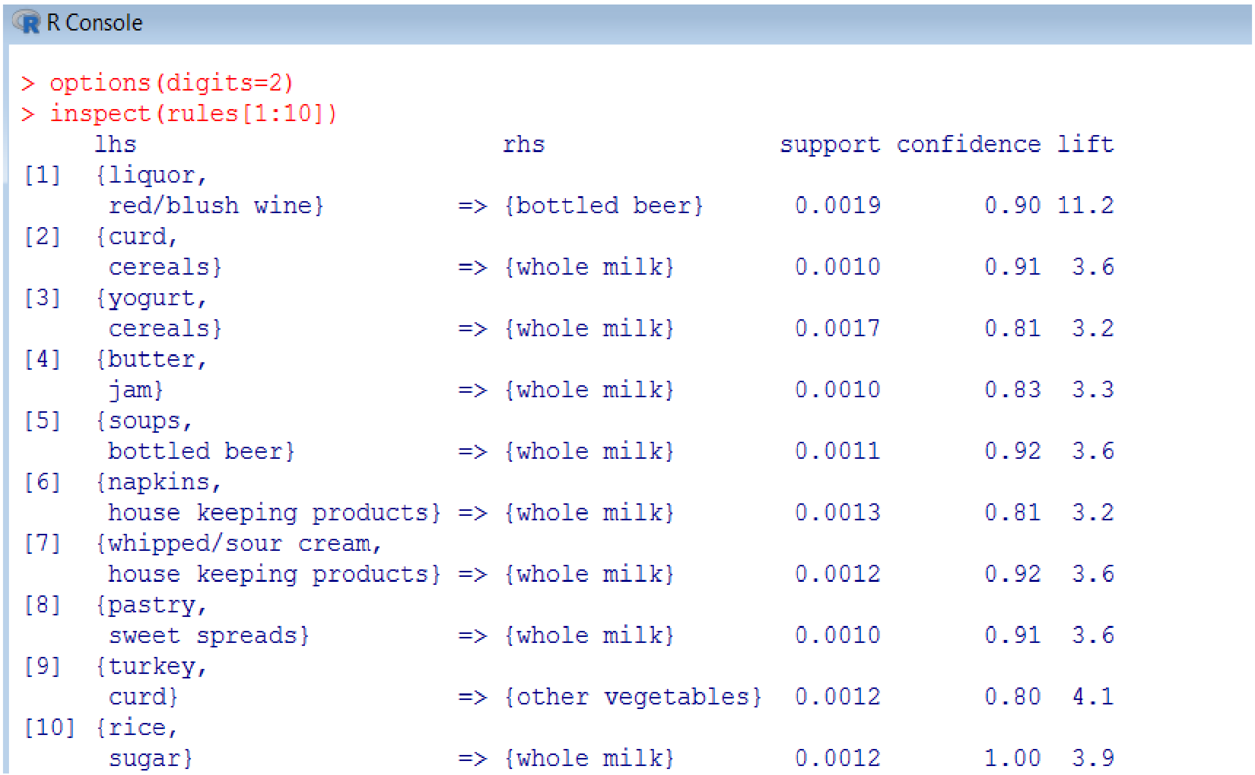
There is no obvious method for choosing appropriate support and confidence parameters. Another issue is that transaction database often contains thousands of rules at reasonable levels of support and confidence parameters, and many of them are obvious. One of the technique to address this issue is mining of Maximal or Closed Itemsets. A "Closed Itemset" is a set which is as large it can be without losing any transactions. A "Maximal Frequent" Itemset is a Frequent Itemset which is not contained in another Frequent Itemset [ZH02]. The Maximal and Closed Itemset mining are not used frequently in MBA because it depends upon the datasets. But this technique is not effective as it does not address the issue of choosing minimum support and confidence parameters.

2.3 Pruning redundant rules

In section 2.2, we saw techniques which have limitations to find Association Rules. One of the methods to get rid of it is by pruning redundant rules. Pruning redundant rules in practice depends on the composition of the data. For example, if a dataset supports several rules A -> B, AC -> B, AD -> B, Maximal Itemset mining will prune the 1^{st} of these rules but leave the others [RC11]. If the first rule arises because of the others, then the pruning is useful. However, if the additional products C, D co-occur incidentally with the popular products A and B, then the remaining rules are redundant. The number of pruned rules may be very small compared to the number of rules remaining.

An alternate approach is to calculate additional interestingness measures on the rules. It can be used to either rank the rules by importance or as an additional pruning criterion. This notion of interestingness is apparently reasonable, but there are many of such measures and show that they tend to rank rules very differently. To study this phenomenon in the dataset, we found rules, at Association Rules at 0.001% support and 0.8% confidence and ranked them according to the measure of interestingness.

In Figure 2.4 - we can see that rules are ranked by their importance. Left-hand side



```
> options(digits=2)
> inspect(rules[1:10])
     lhs                         rhs                support confidence lift
[1]  {liquor,
      red/blush wine}         => {bottled beer}      0.0019       0.90 11.2
[2]  {curd,
      cereals}                => {whole milk}        0.0010       0.91  3.6
[3]  {yogurt,
      cereals}                => {whole milk}        0.0017       0.81  3.2
[4]  {butter,
      jam}                    => {whole milk}        0.0010       0.83  3.3
[5]  {soups,
      bottled beer}           => {whole milk}        0.0011       0.92  3.6
[6]  {napkins,
      house keeping products} => {whole milk}        0.0013       0.81  3.2
[7]  {whipped/sour cream,
      house keeping products} => {whole milk}        0.0012       0.92  3.6
[8]  {pastry,
      sweet spreads}          => {whole milk}        0.0010       0.91  3.6
[9]  {turkey,
      curd}                   => {other vegetables}  0.0012       0.80  4.1
[10] {rice,
      sugar}                  => {whole milk}        0.0012       1.00  3.9
```

Fig. 2.4 **Association Rules at 0.001% support and 0.8% confidence and ranked them according to the measures of interestingness of the items.**

(LHS) and right-hand side (RHS) have the combination of products bought together according to Association Rules. But if we rank them as per the importance of products bought together most of the time, it also generalizes the preference. For example, {rice, sugar} is an itemset of products which is bought together with {whole milk} as its confidence is maximum but according to ranking, this rule is positioned at 10th which gives us clear idea that pruning ranking rules is not the best way to approach "Market Basket Analysis".

2.4 Association Rules Networks

Definition – if we have set of Association Rules R and a target product z, the Association Rules Network (ARN)*(R, z)* is the unique directed hypergraph *(G)* which satisfy the following properties:

1. Any hyperedge in G corresponds to a rule in R with a one-item result or consequent.

2. There is a hyperedge corresponding to a rule whose consequent is the target product z.

3. The target product z is reachable from every vertex v in G.

4. No vertex $v \neq z$ is reachable from z.

ARN shows the limit to which rules "flows into" the target product. The resulting network can show both direct and indirect associations of the target product z. One of the challenges of data mining techniques is that they often generate many "patterns", making it very difficult for the researcher to decide which patterns are adequate and worth evaluating using different statistical techniques [PCP+09].

In ARN, the rules discovered by the Association Rules mining algorithm can be analyzed, pruned and integrated into the context of specific objectives. If there is a product of interest, then we can form a network consisting of the related products and use the network to inform a "model" that can be examined using different methods. The pruning strategy that accompanies ARN can be used to remove local inconsistencies between products, to suggest consistent statistical models. ARN's offer the following features:

1. Pruning in context: We use the ARN to prune rules in keeping our main objective because if we change the objective it will result in pruning of different rules.

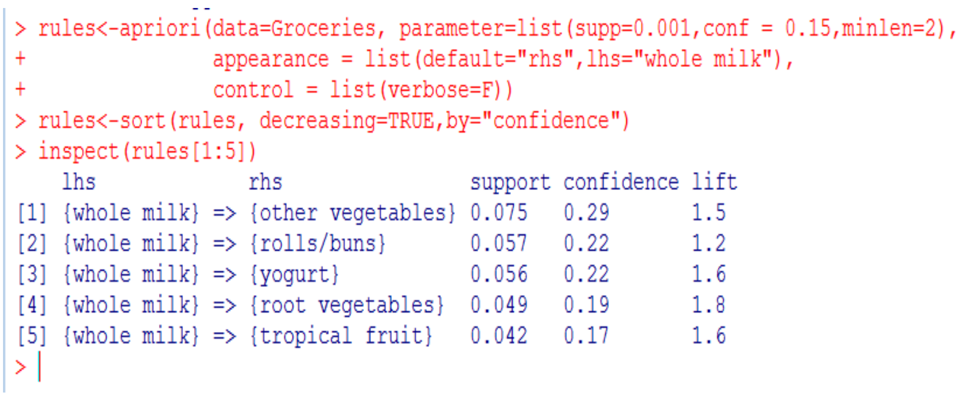
2. Network structure: It provides a mechanism for determining the network relationship between the relevant products. This can help us to draw out direct and indirect and join effects from the network.

3. Hypothesis generation for evaluation: ARN can serve as a bridge between the outputs generated by Association Rule models and their analytical evaluation.

2.5 Targeting Items

This technique is also very interesting and effective in finding rules. We get an idea of the products bought by the customers before the targeted item and after the targeted item. There are two types of targets we might be interested which we illustrated with an example of {whole milk} (from the dataset):

1. What are customers likely to buy before buying {whole milk}?
2. What are customers likely to buy if they purchase {whole milk}?



```
> rules<-apriori(data=Groceries, parameter=list(supp=0.001,conf = 0.15,minlen=2),
+                appearance = list(default="rhs",lhs="whole milk"),
+                control = list(verbose=F))
> rules<-sort(rules, decreasing=TRUE,by="confidence")
> inspect(rules[1:5])
    lhs             rhs                support confidence lift
[1] {whole milk} => {other vegetables} 0.075   0.29       1.5
[2] {whole milk} => {rolls/buns}       0.057   0.22       1.2
[3] {whole milk} => {yogurt}           0.056   0.22       1.6
[4] {whole milk} => {root vegetables}  0.049   0.19       1.8
[5] {whole milk} => {tropical fruit}   0.042   0.17       1.6
> |
```

Fig. 2.5 **Antecedents of the item {whole milk}. This technique we can set either on the LHS or RHS.**

Figure 2.5 - shows the analysis, where we target LHS, which means whenever a customer buy's {whole milk} it is quite certain that he/she will buy the set of the items on the RHS with minimum support and confidence parameters. One must have some idea of the products he/she is interested in before targeting items.

Targeting items suggest that there is no procedure currently available in the literature which addresses the problem of finding meaningful links in large transactional databases. This deficiency motivates us to formulate the network analysis for market basket. It is not obvious to solve the MBA problem, but this technique can discover expressive relationships between the products, from which we can draw direct conclusions about the nature of customer behaviour in a store.

3

NETWORK OF PRODUCTS

3.1 Modelling a Network of Products

Recently, many researchers have used network analysis to study complex systems in a wide variety of scientific, social and engineering domains[KKC12]. Examples include World Wide Web (WWW), organization network and software development process etc. Watts and Strogatez (1998) [WS98] represented the joint relation between movies and actors by [1]bipartite network containing two types of vertices.

Typically, a network is defined as a set of elements interconnected with each other. A common way to model a network is using "Graphs". A graph consists of a set of elements, called vertices with some pairs of them connected by links called edges [VCR14]. It is a model to demonstrate relationships between a set of vertices. Graphs and networks provide us structural models which help us to analyze and understand how different schemes act together.

A Product Network is a model where a vertex represents a product and an edge represents a relationship between them. In this work, we have formulated an edge between two products in such a way, it represents that both products are bought together in one or more transactions in the dataset. Each product in the transaction will represent a vertex in the network, and there will be an edge between vertices which are bought together. There can be a vertex which has no incoming and outgoing edges in the network.In that scenario, we include it in the network, only if it is bought number of

[1] A bipartite graph also called a bigraph is a set of graph vertices decomposed into two disjoint sets such that no two graph vertices within the same set are adjacent.

times greater than minimum threshold we set.

3.1.1 Limitation

If there is an edge between the products, it does not necessarily imply a confirmed relationship between products. In citation networks, two nodes linked together by an edge are necessarily related, if one paper cites another, there is a reason. Product networks are different. Simply because a person buy's {cheese} and {spaghetti sauce} in the same transaction does not provide a common motivation for the two purchases. Aditionally, a person who buys several unrelated items in a single transaction will form a group among them, regardless of the absence of any true relationships [RC11].

3.2 Discovering Communities of Products

Many real-world networks naturally contain groups of nodes that are more strongly connected to each other than they are to the rest of the network. Community detection has been applied successfully in many field of science, ranging from social network analysis to biology and molecular physics. Therefore, the remainder of the thesis will focus on the problem of community detection in product networks, and show how communities of products can be used to gain insight into the behaviour of customers in a store.

Community detection is the process of finding strong groups in a network. The problem is usually addressed as follows: given a network graph G, partition it into a series of disjoint subgraphs $G = G_1 \ldots G_n$ maximizing an objective function *f(G)* [RC11]. A network is said to have community structure if the vertices of the network can be grouped into sets of vertices, such that each set of vertices is densely connected internally. This implies that the network divides naturally into groups of vertices with inner dense connections and sparser connections between groups [Bor05]. The number of communities is generally not known beforehand. Many community detection algorithms (Blondel et a.l 2008 [BGLL08], Clauset et a.l 2004 [CNM04], Newman 2006) attempt to optimize a quantity known as "Modularity".

If a group of communities has a large fraction of its edges falling within communities, (and therefore a relatively small fraction falling between communities) then that community decomposition probably presents a strong community structure. The application to market basket analysis is clear, dividing tightly connected communities within the network of products will allow us to identify strong relationships among the products and therefore meaningful correlations in customer purchase behaviour. Furthermore, as communities can be large, they should be able to represent these connections much more expressively and with less repetition in compare to Association Rules.

We will explain briefly the concept of "Modularity" which is a measure of calculating the difference between the edges in the communities extracted from the network. It is used by "Louvain method" which is used for detecting communities of products from the network of product (Figure 3.1.).

3.3 Modularity

The formal definition of Modularity Q, of a set of communities is defined as:

$\mathrm{Q} = \Sigma\mathrm{i}(\mathrm{e}_{ii} - \mathrm{a}_i{}^2)$

where e_{ii} is the number of edges that fall within the given communities minus $a_i{}^2$ is the expected number if edges were distributed at random [New06]. Modularity measures the difference between the number of edges in-community in each set of communities discovered and the expected number of in-community edges in a random network with the same degree distribution. The range of the value of the modularity lies in between [- 1/2,1)[RC11].

3.4 The Louvain method

The Louvain method finds high Modularity partitions of large networks in short time and unfolds a complete hierarchical community structure for the network, thereby giving access to different resolutions of community detection [DMFFP11].

The algorithm is divided into two phases that are repeated iteratively. We started with a network of N vertices. First, we allocate a different community to each vertex of the network. So, in this initial partition, there are as many communities as there are vertices. Then, for each vertex i we considered the neighbour j of i and we calculated the gain of modularity that would take place by removing i from its community and by placing it in the community of j. The vertex i is then placed in the community for which this gain is maximum (in the case of a tie we used a breaking rule), but only if this gain is positive. If no positive gain is possible, i stays in its original community. This method was applied repeatedly and systematically for all vertices until no further improvement can be achieved and the first phase is then complete. This first phase stops when a local maximum of the modularity is attained.

The second phase of the algorithm focuses on building a new network whose vertices is now the communities. To do so, the frequency between the new vertices are given by the sum of the weight of the links between vertices in the corresponding two communities. Links between vertices of the same community lead to self-loops for this community in the new network. Once this second phase is completed, it is then possible to re-apply the first phase of the algorithm to the resulting weighted network and to iterate.

In Figure 3.2 - we can see that six communities of the products are discovered from the modelled network. Each community range in size and in the number of products. There are edges between vertices of each community but not between communities.

Although this approach is commonly used in the graph-partitioning literature, the only condition is the size of the communities are not normally known in advance. If

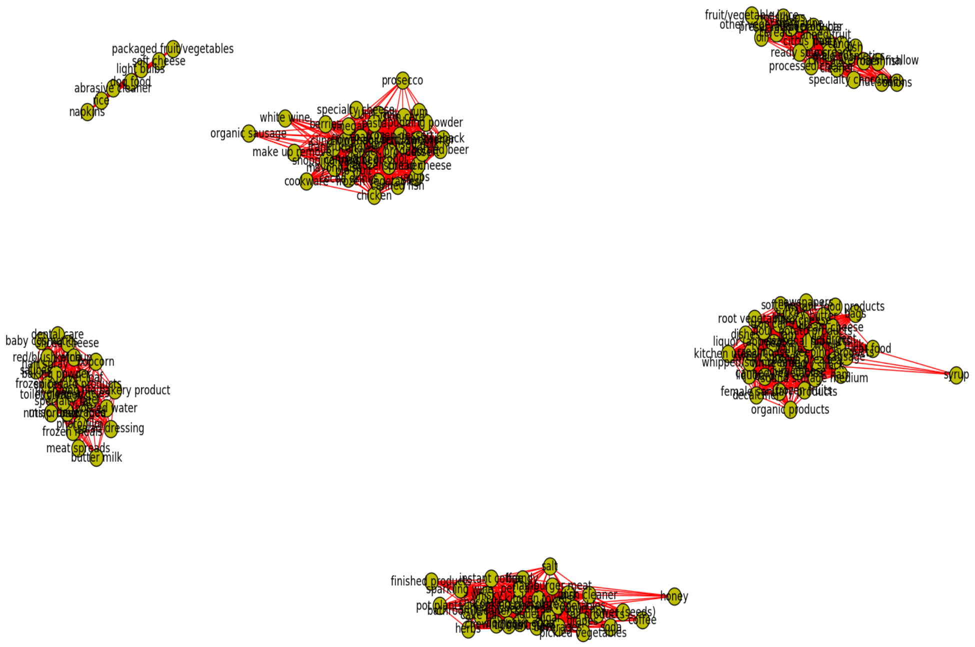


Fig. 3.1 **Group of communities found after applying Louvain method on network of products (threshold = 50). Total 6 communities are discovered, all the communities are compact and isolated from each other. All the items in the communities show the strong connection with each other in the transactions from the dataset.**

there is no limit on community size, then we are, free to select the network in a way that puts all the vertices in one of our two groups and none in the other, which promises we will have zero intergroup edges. In the following section, we will implement different measures to analyze these communities and visualize them.

3.5 Measuring the utilities of Communities

In this section, we will measure the utilities of the communities. Specifically, we wish to answer the question:

after finding sets of the communities in a network of product, which communities are most useful to a human analysis?

3.5.1 Network Density

Our objective is to find communities which has high number of edges between the vertices because that will be beneficial for drawing significant information between the products and customer purchasing behavior. Network density is the measure of finding communities with actual connection between vertices. In a network, there are vertices and edges. A vertex represents a product and the connections between the vertices are called 'edges', also, it represents the relationship between them.

Network density describes the part of the potential connections in a network that are actual connections. A "Potential Connection" or PC is a connection that could potentially exist between two 'vertices'. This person could know that person; this computer could connect to that one. Whether they do connect is irrelevant when you're talking about a potential connection. By contrast, an "Actual Connection" is one that exists. This person does know that person; this computer is connected to that one. We used this technique to find the community which has more potential connections. [Bor05].

3.5.2 Calculating Network Density

Network density is the part of actual connections divided by potential connections. In the below chart, *'PC'* is 'potential connection' and *'n'* is the number of the vertices in the network.

In Figure 3.3 - examples *'A'* and *'B'* demonstrate items where the number of actual connections between vertices is the same as the number of potential connections. You can't draw any new edge to connect these vertices, they are already connected and perfectly 'dense'.

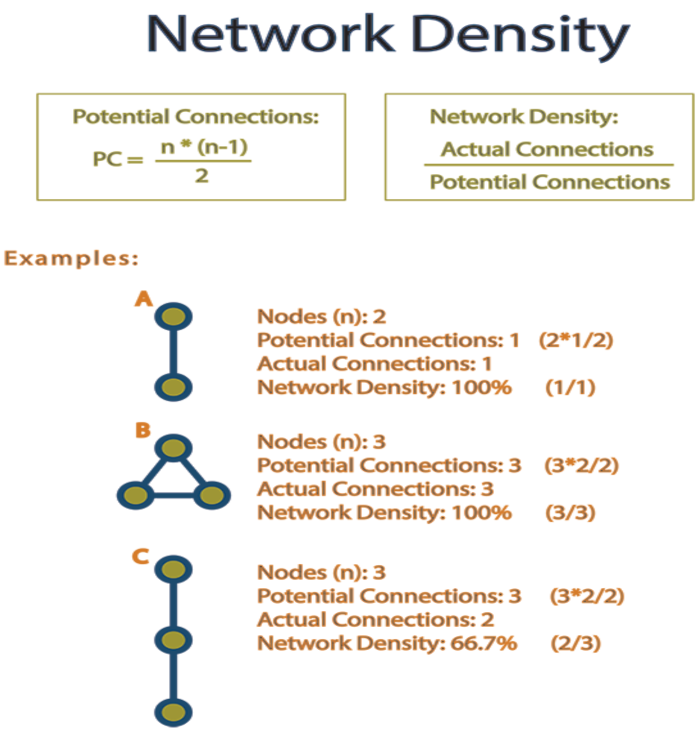

Network Density
Potential Connections:
PC = n * (n-1) / 2
Network Density:
Actual Connections / Potential Connections
Examples:
A
Nodes (n): 2
Potential Connections: 1 (2*1/2)
Actual Connections: 1
Network Density: 100% (1/1)
B
Nodes (n): 3
Potential Connections: 3 (3*2/2)
Actual Connections: 3
Network Density: 100% (3/3)
C
Nodes (n): 3
Potential Connections: 3 (3*2/2)
Actual Connections: 2
Network Density: 66.7% (2/3)

Fig. 3.2 **Shows the formula of calculating Network Density and a small example we explain below how it is calculated [Ros13].**

Now look at example *C*. Like example *B*, there are three vertices. But in this case, two of the vertices (the top and bottom ones) aren't connected to each other. This little network is missing one of its potential connections, and as a result, its network density drops to two-out-of-three, or 66.7%.

3.6 Centrality

Now we want to consider most important products in the communities. In network analysis, indicators of centrality identify the most important vertices within a graph. Centrality application includes determining the most influential person in a network or in our case most important product in the network [Fri91]. Centrality has been applied to understand employment opportunities, differential growth rates among medieval cities, political integration in the context of the diversity of Indian social life and many more [YD09].

3.6.1 Betweenness Centrality

In product network analysis, to identify important vertices is a crucial task. We will see how this measure is computed and how to use the library Networkx to create a visualization of the network where the vertices with the highest betweenness value are highlighted. The betweenness focuses on the number of visits through the shortest paths. If a walker moves from one vertex to another vertex via the shortest path, then the vertices with many visits have a higher centrality. The betweenness centrality is defined as:

$$B(v) = \Sigma(s \neq t \in v)\frac{sv(s,t)}{s(s,t)}$$

where *s(s,t)* is a total number of shortest paths from node s to node t and $s_v(s,t)$ is the number of those paths that pass through v [wik17].

It is essential to find important vertices in the communities to have a better understanding of the items in the network. Betweenness centrality highlights the vertices which are very influential on the way the information spreads over the network. Looking at the network, we can point out which are the most important items in the group of communities. It returns centrality value which ranks each vertex in the network.

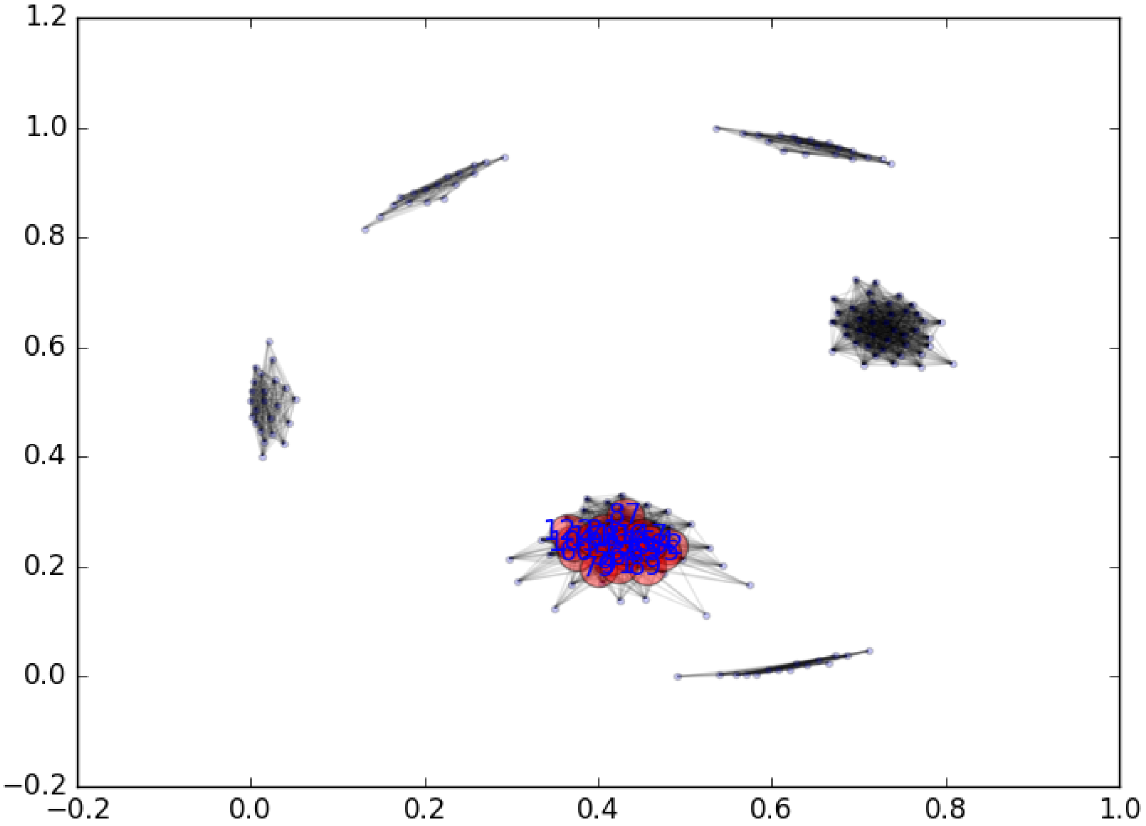


Fig. 3.3 **Betweenness centrality for each community discovered from the product network. We can see the highlighted vertices in each community, these vertices have high betweenness centrality value and are bought most of the time together.**

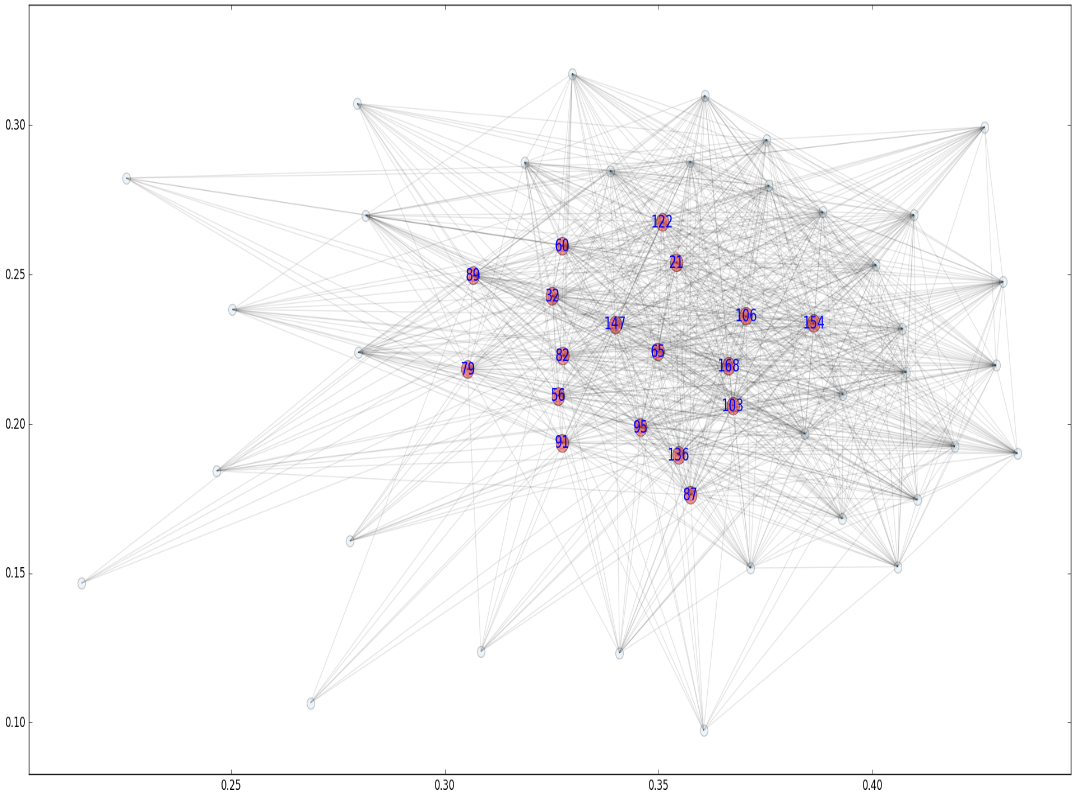
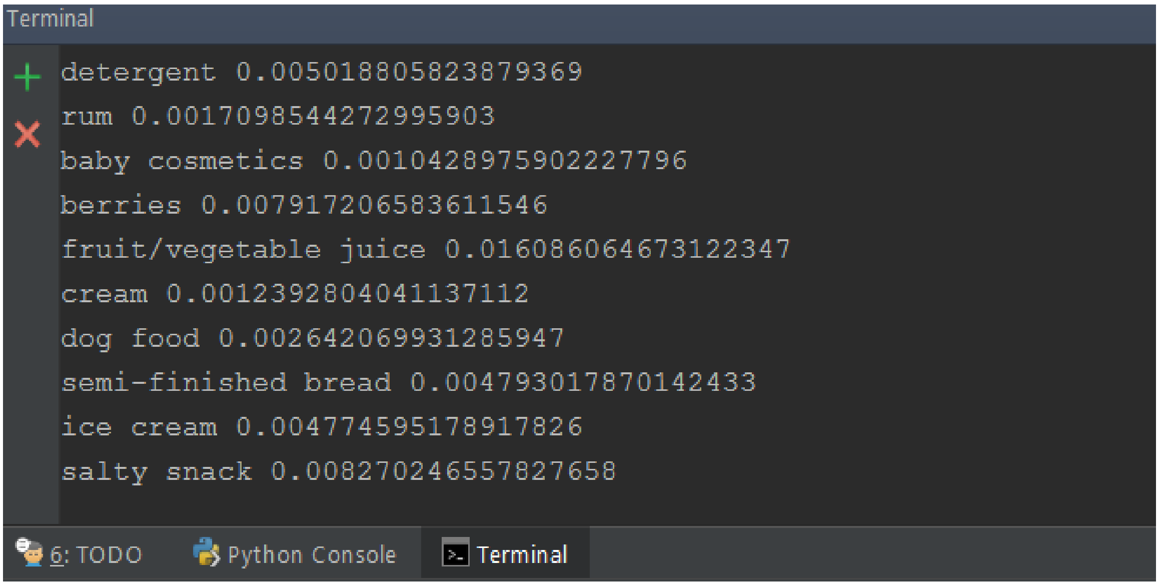


Fig. 3.4 **Magnified view of one of the community discovered highlighting the vertices which are important in the group. Betweenness centrality points out the important products which are more influential in the whole network. This technique analyses more into communities and break down the fundamental question of which products are bought most of the time together in the transactions in market basket analysis.**

3.7 PageRank

PageRank is implemented by counting the number of links to a vertex to determine how important that vertex is. Important vertices are likely to receive more edges from other vertices. One of the limitation of Centrality method is that, if a vertex with a high value of centrality has edges going out from it then all those other vertices will have high centrality. This basic problem is resolved by using PageRank algorithm. The aspects which determine the PageRank of the vertex are the number of edges it receives, the centrality of the vertex from where edges are coming and the link propensity of the linkers [CN06].



```
detergent 0.005018805823879369
rum 0.0017098544272995903
baby cosmetics 0.0010428975902227796
berries 0.007917206583611546
fruit/vegetable juice 0.016086064673122347
cream 0.0012392804041137112
dog food 0.002642069931285947
semi-finished bread 0.004793017870142433
ice cream 0.004774595178917826
salty snack 0.008270246557827658
```

Fig. 3.5 **Products and their values ranked by PageRank algorithm. These values associated with the items shows the importance of the items which are frequent in the dataset.**

Figure 3.6 - shows the PageRank value for the products in the discovered communities. What makes this algorithm important is that the graph could equally well represent the relationship between vertices but as there might be more than one edge coming from a vertex which increases the PageRank value for the vertex which receives that edge.

3.8 NetworkX Library

We discuss, briefly the Networkx package for building Network of Products. NetworkX is a Python language software package and an open-source tool for the creation, manipulation and study of the structure, dynamics and functions of complex networks. It is a computational network modelling tool. It can load, store and analyze networks, generate new networks, build network models and draw networks. The first public release of the library was all based on python, and the library can engage with other programming languages too such as *C*, *C++* and *FORTRAN* [SS08]. Therefore, for constructing a network of products we used this package which has all inbuilt functions which we require to draw a graph. In a graph, vertex represents the products or items and edges between them correspond that both the products are bought together. We checked the frequency of items bought together most of the time in the transactions by keeping track of the number of edges between them. The dataset can be in any format and networkx package contains all the functions to read and write.

4

CENTRE-PIECE SUBGRAPHS

4.1 Introduction to Centre-Piece Subgraphs (CePS) on Market Basket Data

Now we know, which are important vertices in the communities, in this section, we briefly introduce the concept of Centre- Piece Subgraphs on market basket analysis. Graph mining has become very popular for finding out communities, partitioning the network and finding out significant relationship between vertices in the network [TF06]. All the methodologies we have discussed so far has a limitation that we need to specify a minimum support and confidence parameter or minimum threshold in case of modelling network of products to extract valuable information. Strong relationships which does not support minimum threshold, remain undiscovered or are missed out. Centre-piece subgraph (CePS) is limited in size by the budget parameter b, it is not necessary to further limit them with minimum support and confidence parameter. Therefore, CePS can discover relationships between all the products that create the network. This technique is useful for either verifying the results suggested by other techniques or explaining the relationships that are undiscovered by other methods.

CePS describe the relationship with the neighbourhood of a vertex or set of vertices but the difference comes how they describe or defines this neighbourhood. The *CePS Cp (G, b, Q, k)* is a subgraph H of the graph G, where H contain all query vertices in the set Q and at most b other vertices. The selection of H must maximize an objective function *g(H)*. The parameter k is the number of query vertices to which a vertex must be strongly related to being considered a candidate for the subgraph. Association Rules Network (ARN)[TF06] defines the neighbourhood of the target product, for example Z as the set of sets of products that are either direct or indirect cause of Z within ruleset R. A CePS defines the neighbourhood of the query vertices in the set$Q$ as the set of b products that are closely related to the member of Q according to the objective function *g()*.

The advantage of *CePS* in relation to the market basket analysis is that they allow analyses for all the vertices in the network. While community detection can find links in the network, with the requirement of minimum support to find useful relationships, the same thing is with Association Rules. Thus, in both cases, the number of products about which we can learn useful information is significantly limited. We conclude that the *CePS* are primarily useful for either verification of hypotheses suggested by other techniques or for explaining unexpected results arrived at by another method. We have highlighted *CePS* just to give an insight into the unexpected results, which can be obtained by implementing this technique and included it in the future work section.

5

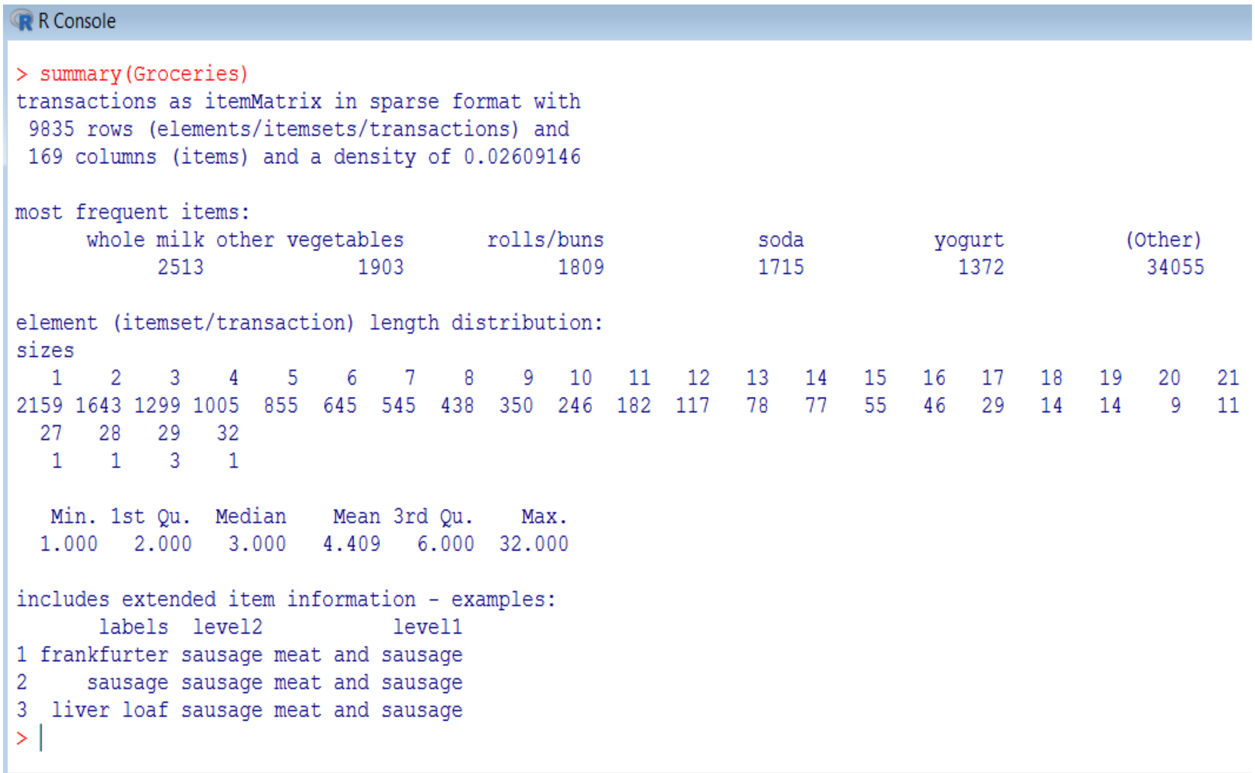
DATA ANALYSIS AND RESULTS

5.1 Experimental

5.1.1 Dataset

To evaluate our methods, we experimented our approach on publicly available dataset called Groceries. The Groceries dataset contains real-world transactions data for certain period from a typical local grocery outlet. The dataset contains 9835 transactions and 169 items [TF06]. It is a CSV file where column represents a list of the items and row represents the number of the transactions in the dataset.

In figure 5.1 - we can see that we have 9835 rows which represents transactions in the dataset and 169 columns which represents products. Each transaction has set of products bought together. $\{$Whole milk$\}$ = 2513, $\{$other vegetables$\}$ = 1903, $\{$rolls/buns$\}$ = 1809, $\{$soda$\}$ = 1715, $\{$yoghurt$\}$ = 1372, others = 34055 are the most Frequent Itemsets in our dataset.



```
> summary(Groceries)
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:
      whole milk other vegetables       rolls/buns             soda           yogurt          (Other)
            2513             1903             1809             1715             1372            34055

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21
2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46   29   14   14    9   11
  27   28   29   32
   1    1    3    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   4.409   6.000  32.000

includes extended item information - examples:
       labels  level2           level1
1 frankfurter sausage meat and sausage
2     sausage sausage meat and sausage
3  liver loaf sausage meat and sausage
> |
```

Fig. 5.1 **The summary of the Groceries dataset in R.**

5.1.2 A strategy for Market Basket Analysis

The work we did has allowed us to make significant observations about market basket analysis on our transactional dataset. We would like to re-state our main observations:

1. First, we discovered Frequent Itemset from our transactional dataset and then we removed the items which were not frequent by setting the minimum support and confidence parameters. Then, the items left in the transaction dataset gave us the motivation to find Association Rules, which was important because the association between the products bought together in the market basket gives us knowledge about the customer behavior.

2. In the next step, we derived Association Rules which is an old and traditional way of finding relationships between the products, with a minimum support and confidence parameters. As an initial step, this technique was effective and gave us a general idea about the items bought together in the dataset.

3. Then we mined Maximal or Closed Itemsets because there was redundancy in the rules. But these techniques fail to prune away many redundant rules.

4. After that, we ranked rules which we found in step 2, based on how many times they were bought together in the transactions. We found that there are many such measures to choose and the ranking done by them is not consistent. Therefore, we decided that it might be difficult to choose an appropriate measure in the absence of prior knowledge.

5. Now, we had set of items in the transactions in our dataset which are more frequent and actionable. Next, we modeled a network of products which is a graph where vertices represent the items in the dataset, and the edges showed the relationship between the items bought together. We noticed the difference between the number of edges between the items in the dataset. There were vertices which didn't have any incoming or outgoing edges but still present in the network as they passed minimum threshold criteria.

6. Once we build a network of products, our next step was to discover communities of products. Community detection was not easy to find within the network, as it was hard to find a group of products which are closely related. To achieve that, we used Louvain technique which use modularity to find complete hierarchal community structures.

7. We discovered 6 communities of products. To analyze discovered communities of products, we have used network density, centrality (betweenness centrality) and PageRank algorithms to extract useful and valuable information from these communities.

8. We also discussed ARN because this technique is useful to explore the core of the networks - the only condition was that the chosen target product should appear in many Association Rules, for example, {whole milk}. ARN are highly sensitive to the choice of the target product. But even for very popular products, this technique sometimes does not share much information that is the reason we have included this method in future work section.

5.1.3 Methodology Used

To do that, we have used R software because it has all the in-built functionalities. Further we modelled a network of products. First, we set minimum threshold equal to 50 and then used Python programming language. Python has Networkx library which has all the functions to model a graph. So, we configured vertices as products and assigned an edge between two vertices if they are brought together in one or more transactions in the dataset. The only condition in this was the minimum threshold. Once we modelled the network of products our next step was to extract communities. As we have discussed in section 3.2, we used Modularity which is a measure used by Louvain method to extract communities of products from the network. We used Python language because we found it more convenient and reliable and easy to use. Finally, we discovered 6 communities and analysed them using different measure discussed in section 3.5.

5.2 Results and Discussion

5.2.1 Frequent Itemsets

By extracting Frequent Itemsets, a shop keeper can get idea what is commonly bought together. What is more important is larger sets of items that occur much more frequently in the entire dataset. We estimated that most of the customers buy {vegetables} and {whole milk} together. But that is of little interest, because these products are popular products individually in the dataset. There is no surprise to people who like {chewing gums}, but it offers the retailers an opportunity to do some smart marketing.

***Tab. 5.1* shows the Frequent Itemsets at minimum threshold = 200**

Itemsets	Frequency
{‘whole milk’, ‘soda’}	394
{‘other vegetables’, ‘yoghurt’}	427
{‘margarine’}	576
{‘rolls/buns’, ‘yoghurt’}	338
{‘root vegetables’, ‘other vegetables’, ‘whole milk’}	228
{‘whipped/sour cream’, ‘other vegetables’}	284
{‘hygiene articles’}	324
{‘yoghurt’, ‘other vegetables’, ‘whole milk’}	219
{‘other vegetables’, domestic eggs’}	219
{‘brown bread’, ‘whole milk’}	284
{‘pastry’, ‘whole milk’}	327
{‘butter milk’}	275
{‘newspaper’}	785
{‘beef’, ‘whole milk’}	209
{‘tropical fruit’, ‘soda’}	205
{‘chewing gum’}	207
{‘bottled beer’, ‘whole milk’}	201
{‘whole milk’}	2513
{‘other vegetables’, ‘soda’}	322

5.2.2 Association Rules

After we looked for Association Rules at minimum support and confidence parameters. As know that, there is no obvious method for choosing appropriate values for support and confidence. If the value is chosen too high, we lose interesting associations. However, if the value is chosen too low, we will be flooded with thousands of weak rules, which do not imply meaningful associations. Therefore, Association Rules are incredibly sensitive to the choice of support and confidence parameters.

In Table 5.2 - we can see that **80%** of customers who buy {yoghurt} and {cereals} also buy {whole milk} and almost **2%** of customers buy all these products together. A

Tab. 5.2 **shows first 10 Association Rules at support = 0.001 and confidence = 0.8.**

LHS	RHS	Support	Confidence
{liquor, red/blush wine}	{bottled beer}	0.0019	0.90
{curd, cereals}	{whole milk}	0.0010	0.91
{yogurt, cereals}	{whole milk}	0.0017	0.81
{butter, jam}	{whole milk}	0.0010	0.83
{soups, bottled beer}	{whole milk}	0.0011	0.92
{napkins, housekeeping products}	{whole milk}	0.0013	0.81
{whipped/sour cream, housekeeping products}	{whole milk}	0.0012	0.92
{pastry, sweet spreads}	{whole milk}	0.0010	0.91
{turkey, curd}	{other vegetables}	0.0012	0.80
{rice, sugar}	{whole milk}	0.0012	1.00

shop keeper can use this type of information to help them find new favourable circumstances for cross selling their products to the customers. Choosing minimum support and confidence parameter is very crucial step in determining Association Rules.

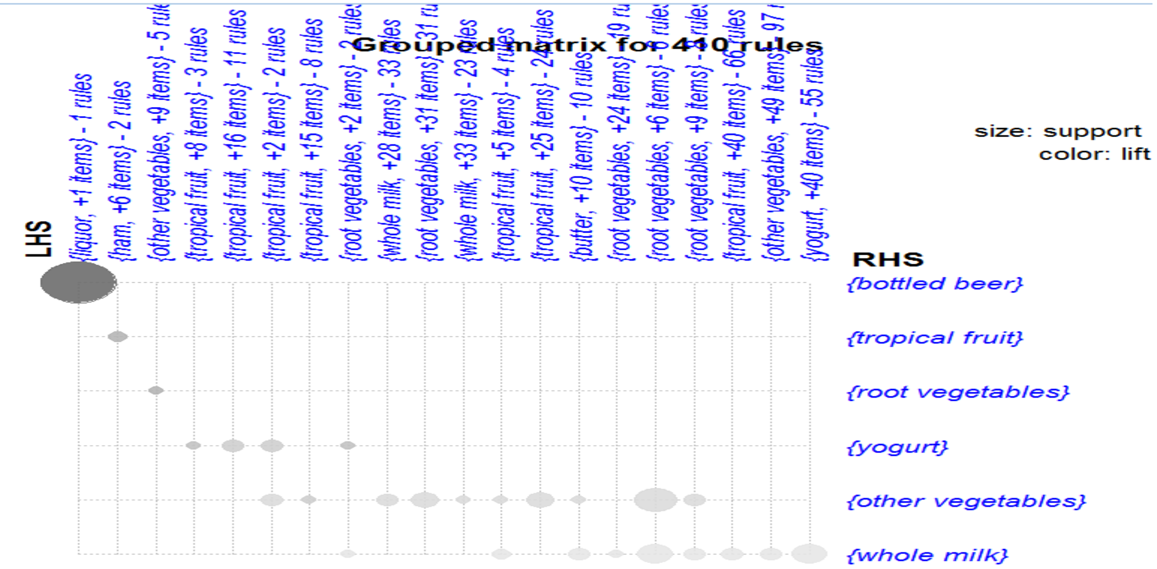


Fig. 5.2 **Shows the grouped matrix for 410 rules. The LHS shows the group of the products which are bought with the RHS with the number of the rules generated.**

5.2.3 Network of Products

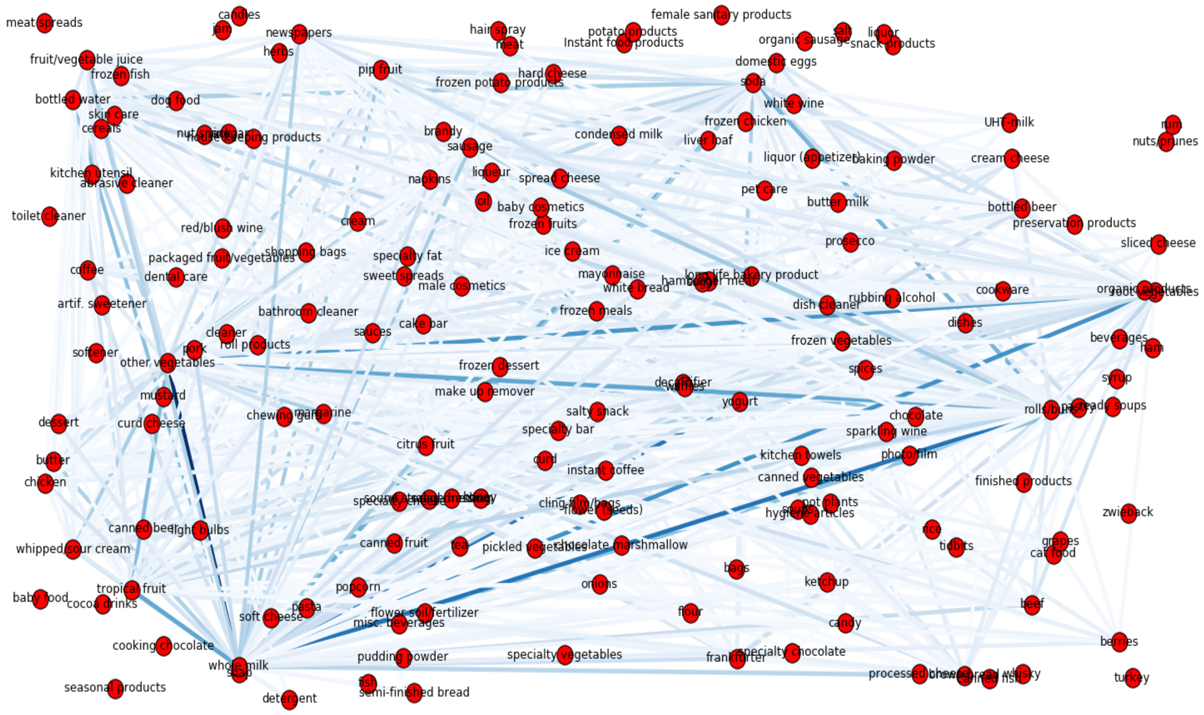


Fig. 5.3 **The graphical representation of Network of products built from the transactional dataset [RC11]. We can see {meat spreads}, {season products}, {turkey} etc. are the products which have no incoming and outgoing edges, which suggests that these items are bought alone in a transaction most of the time. Also, edges which are denser implies that the products which are related to it are bought more than once in the dataset.**

Figure 5.3 - shows the network of products where vertex represents the product and an edge between vertices represent that both are bought together. To remove the noisy edges created by coincidental purchases and to improve the quality of our subsequent analysis, we establish a minimum threshold σ times. Note that, in the pruned network, the weight of any remaining edge is unchanged. Pruning will help us to eliminate the products which are less frequent and are still bought together with other products. Product networks we build is very dense, with many connections per node, but many of these edges might be meaningless showing false associations generated by chance. The

network we modelled contains **169** products and almost **9636** edges between them. We have formulated the number of edges between vertices through the brightness of color. For example, if the color of the edges is light in the network, it means it is less frequent and if it is dark that means the number of edges between the products are more frequent.

In our analysis, we have set the threshold = 50. Also, we considered the products which are bought alone in the transactions more than 50 times. For example, products like {meat spreads}, {candles} etc. Finally, all the products in the pruned network are of importance and we analyze them further to extract valuable information.

5.2.4 Discovered Communities

Next, we discovered communities from the pruned network. To do that, we have used Louvain method which uses the concept of Modularity to find these communities. Overall, we have discovered **6** communities from the pruned network. Each community range in size from 7 products to over 50. To extract valuable information, we analysed these communities using the different measures defined in section 3.7. The group of products which are highly rated, generally are well-connected with a clear idea.

***Tab. 5.3* shows all the products in Community 1.**

No.	**Product**
1	{rice}
2	{packaged fruit/vegetables}
3	{soft cheese}
4	{dog food}
5	{light bulbs}
6	{abrasive cleaner}
7	{napkins}

Table 5.4 - shows the PageRank values for the products bought together in this community. It counts the number of edges to vertex and determine how important that vertex is in the community. We can see all the products are important in this community

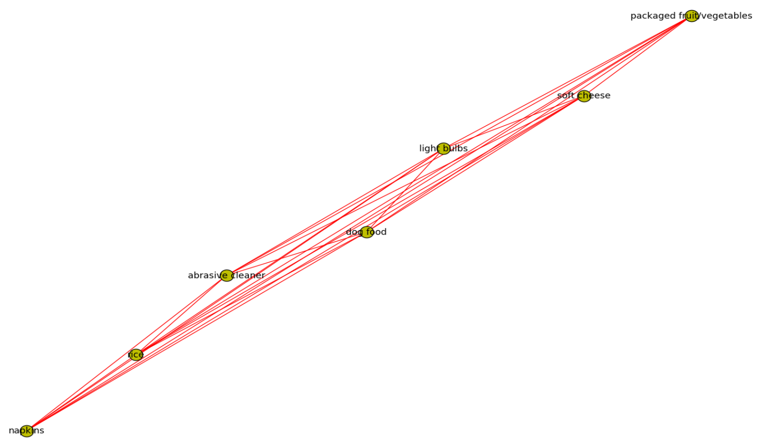


Fig. 5.4 **Shows the first community. The community is densely connected, as all the vertices are connected to each other with a clear message, that customers often buy these products most of the times together.**

Tab. 5.4 **PageRank values for the products in the first Community 1.**

Product_Name	**PageRank**
{packaged fruit/vegetables}	0.003
{soft cheese}	0.005
{light bulbs}	0.001
{dog food}	0.002
{abrasive cleaner}	0.001
{rice}	0.003
{napkins}	0.012

and form the group of products bought together frequently most of the time. Centrality is another method we have used which points out the nodes which are very influential on the way the information spreads over the community. In this community, there are 7 products and all the products are visited significantly.

Figure 5.4 - shows all the products in the community are strongly connected to each other, which means it is composed of many of the store's most popular items. The strong relationships between the vertices suggest that, whenever the customer buys items from this community, it is much likely that he/she buys other related items too, which are

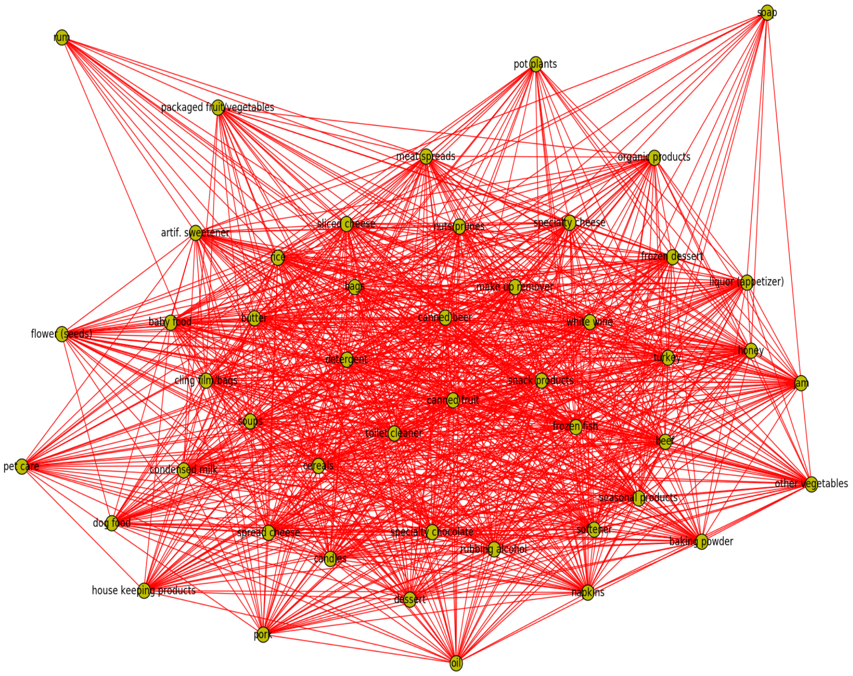


Fig. 5.5 **Shows the second community which is densely-connected. In this community, all the vertices are linked to each other with more than one edge between them, which means these products are bought frequently in one or more transactions in the dataset.**

present in the community.

Table 5.6 - shows the products with high betweenness centrality values for community 2. The betweenness centrality focuses on the number of visits through the shortest path. If we have edges from one vertex to another vertex via the shortest path, then the products have higher centrality. High values for Betweenness centrality tells us that associated products are significantly important in the community. These products plays important link between all the products in the community when they are bought together in market basket.

Table 5.10 – shows the products which are important and are more influential in the community 4. Betweenness centrality points out them with high values in compare to other products in the community. For example, product {fozen chicken} has the highest

***Tab. 5.5* shows all the products in Community 2.**

No.	Product
1	{pot plants}
2	{soap}
3	{organic products}
4	{meat spreads}
5	{packed fruit/vegetables}
6	{rum}
7	{artif. Sweetner}
8	{sliced cheese}
9	{nuts prune}
10	{specialty cheese}
11	{frozen dessert}
12	{liquor(appetizer)}
13	{rice}
14	{bags}
15	{makeup remover}
16	{white wines}
17	{turkey}
18	{honey}
19	{jam}
20	{flower(seeds)}
21	{baby food}
22	{butter}
23	{detergent}
24	{canned beer}
25	{snack products}
26	{cling film/bags}
27	{soups}
28	{cereals}
29	{toilet cleaner}
30	{canned fruit}
31	{frozen fish}
32	{beer}
33	{other vegetables}
34	{seasonal products}
35	{baking powder}
36	{napkins}
37	{oil}
38	{dessert}
39	{pork}
40	{candles}
41	{rubbing alcohol}
42	{spread cheese}
43	{dog food}
44	{pet care}
45	{condensed milk}
46	{housekeeping products}
47	{speciality chocolate}

Tab. 5.6 **Betweenness Centrality values for the important products in Community 2.**

Product_Name	Betweenness centrality
{frozen fruit}	2.788
{frozen chicken}	6.789
{bags}	1.033
{toilet paper}	3.483

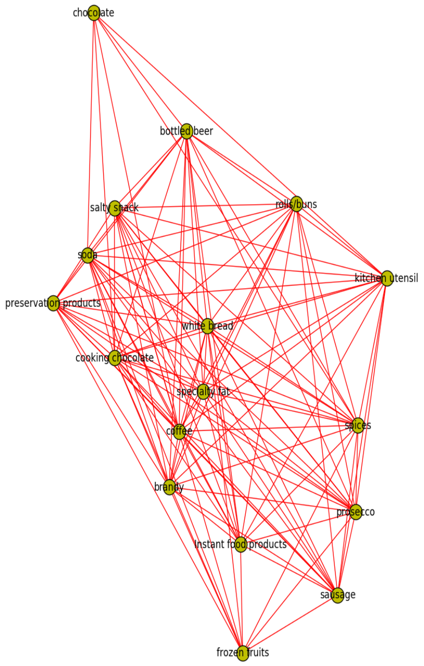


Fig. 5.6 **Shows the third discovered community with one or two products which are not linked to all the other products in the community. Vertex {chocolate} is at distance from other vertices in the community which tells that it is not bought with other products frequently most of the time in the community. But it implies that it is bought either alone or with some of the products in community because it passes minimum threshold criteria which we have set. Rest, all the items in the community are closely related to each other.**

value and is most important in the community.

Tab. 5.7 **shows all the products in Community 3.**

No.	Product
1	{chocolate}
2	{preservation products}
3	{soda}
4	{salty snack}
5	{bottled beer}
6	{cooking chocolate}
7	{white bread}
8	{rolls/buns}
9	{brandy}
10	{coffee}
11	{specialty fat}
12	{kitchen utensil}
13	{spices}
14	{instant food products}
15	{frozen fruits}
16	{sausage}
17	{prosecco}

Tab. 5.8 **PageRank values for the products in the first Community 3.**

Product_Name	PageRank
{chocolate}	0.011
{preservation products}	0.001
{soda}	0.026
{salty snack}	0.008
{bottled beer}	0.011
{cooking chocolate}	0.001
{white bread}	0.010
{rolls/buns}	0.028
{brandy}	0.001
{coffee}	0.011
{specialty fat}	0.001
{kitchen utensil}	0.001
{spices}	0.002
{instant food products}	0.002
{frozen fruits}	0.001
{sausage}	0.018
{prosecco}	0.001

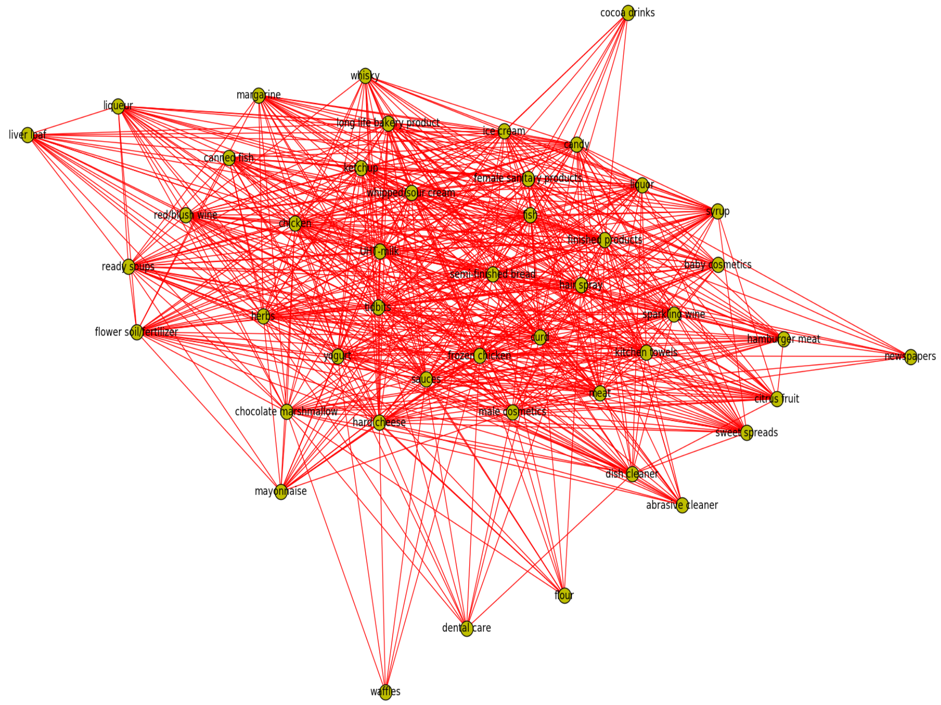


Fig. 5.7 **Shows the fourth community where, products like {cocoa drinks} and {waffles} are not completely linked to other products in the community. Rest all the products are heavenly dense and concentrated, and show strong relationship in the community.**

***Tab. 5.9* shows all the products in Community 4.**

No.	**Product**
1	{cocoa drinks}
2	{liver loaf}
3	{liqueur}
4	{margarine}
5	{whisky}
6	{ice cream}
7	{candy}
8	{liquor}
9	{syrup}
10	{baby cosmetics}
11	{hamburger meat}
12	{newspaper}
13	{citrus fruits}
14	{sweet spreads}
15	{abrasive cleaner}
16	{dish cleaner}
17	{flour}
18	{dental care}
19	{waffles}
20	{mayonnaise}
21	{flower soil/fertilizer}
22	{ready syrups}
23	{red blush/wine}
24	{canned fish}
25	{long life bakery products}
26	{ketchup}
27	{chicken}
28	{whipped sour cream}
29	{UHT-milk}
30	{yoghurt}
31	{herbs}
32	{chocolate marshmallow}
33	{hard cheese}
34	{sauces}
35	{frozen chicken}
36	{semi-finished bread}
37	{female sanitary products}
38	{fish}
39	{finished products}
40	{kitchen towels}
41	{meat}
42	{sparkling wine}
43	{male cosmetics}
44	{hair spray}

Tab. 5.10 **Betweenness Centrality values for the important products in Community 4.**

Product_Name	**Betweenness centrality**
{curd}	0.342
{fish}	1.256
{hair spray}	2.477
{frozen chicken}	6.789

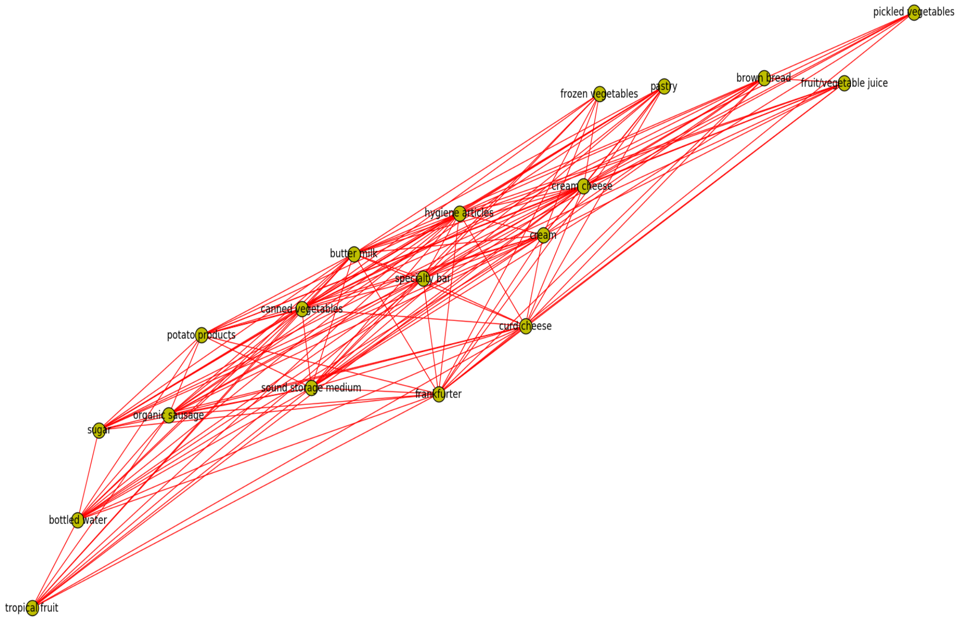


Fig. 5.8 **Shows the fifth community where almost all the products are linked to each other. Shape of the community does not make the difference in determining the strong relationships in the community. This community gives us clear idea about the products bought together most of the time from the dataset.**

Tab. 5.11 **shows all the products in Community 5.**

No.	Product
1	{pickled vegetables}
2	{brown bread}
3	{fruit vegetable/juice}
4	{pastry}
5	{frozen vegetables}
6	{cream cheese}
7	{cream}
8	{curd cheese}
9	{frankfurter}
10	{specialty bar}
11	{hygiene articles}
12	{butter milk}
13	{canned vegetables}
14	{sound storage medium}
15	{potato products}
16	{organic sausage}
17	{bottled water}
18	{tropical fruit}

Tab. 5.12 **PageRank values for the products in the first Community 5.**

Product_Name	PageRank
{pickled vegetables}	0.005
{brown bread}	0.012
{fruit vegetable/juice}	0.016
{frozen vegetables}	0.011
{cream cheese}	0.009
{cream}	0.001
{curd cheese}	0.001
{frankfurter}	0.012
{specialty bar}	0.005
{hygiene articles}	0.008
{butter milk}	0.006
{canned vegetables}	0.003
{sound storage medium}	0.009
{potato products}	0.001
{organic sausage}	0.001
{bottled water}	0.019
{tropical fruit}	0.022

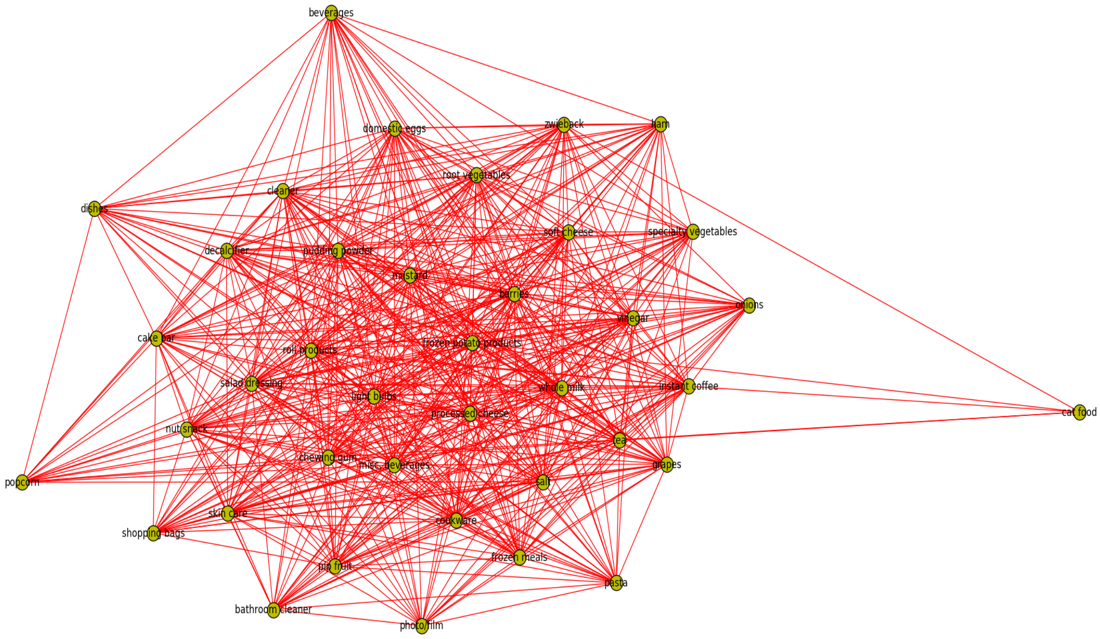


Fig. 5.9 **Shows the sixth community where the influence of product {cat food} is not much. We can see the frequency of edges between this product and other products in the community is not much. The reason for this is the Association Rules involving {cat food} and the other products, bought together are less frequent, which make this product loosely related to other products in the community. Another reason is that it is bought most of the time in a transaction but alone. Apart from it this community also has strong linkage between the vertices and gives us clear view about the products bought together most of the time.**

***Tab. 5.13* shows all the products in Community 6.**

No.	**Product**
1	{dishes}
2	{beverages}
3	{popcorn}
4	{domestic eggs}
5	{clearer}
6	{decalcifier}
7	{cake bar}
8	{nut snack}
9	{shopping bags}
10	{skin care}
11	{salad dressing}
12	{roll products}
13	{pudding powder}
14	{bathroom cleaner}
15	{pip fruit}
16	{chewing gum}
17	{light bulbs}
18	{mustards}
19	{root vegetables}
20	{zwieback}
21	{ham}
22	{soft cheese}
23	{berries}
24	{frozen potato products}
25	{processed cheese}
26	{misc. beverages}
27	{photo film}
28	{cookware}
29	{frozen meals}
30	{pasta}
31	{grapes}
32	{tea}
33	{whole milk}
34	{salt}
35	{vinegar}
36	{instant coffee}
37	{onions}
38	{cat food}
39	{specialty vegetables}

Tab. 5.14 **Betweenness Centrality values for the important products in Community 6.**

Product_Name	**Betweenness centrality**
{mustard}	0.021
{decalcifier}	8.450
{salad dressing}	3.176
{whole milk}	0.071

5.2.5 Comparative Analysis

We have compared results from Association Rules and discovered Communities and we noticed that both approaches show the variations. For example, the rule {curd, cereals} -> {whole milk} have high confidence of **0.91** and depicts one of the rule which states that whenever a customer purchase {curd, cereals}, he/she also purchase {whole milk} in the same transaction. Whereas communities of products give us a complete overview of products brought together most of the time. This information is more detailed because now we know all set of products which are brought in one or more transactions in the dataset. Also, we have analyzed these communities to find most important products. For example, the product {frozen chicken} in one of the community has centrality value of **6.789** which means that this product is brought most of the time with other products in the group. This variation in the results tells us that community detection can be used as an explanatory step in MBA.

5.2.6 Discussion

These results, suggests that finding communities of the products can play a useful role in market basket analysis. Later, we analyzed these communities by using different methodologies to extract valuable relationships between the products. All the communities provide a lot of information about the purchasing behavior of customers and give us an idea that how we can arrange items in the store to increase the sales and productivity. Therefore, community detection can be used as a explanatory step in the analysis of market basket.

Regardless of that, sometimes extracting a group of communities has drawn no significant results. The reason behind this is widely available transactional datasets in the market. Also, there is no structured way of choosing different techniques for analysis of communities. Because to find a group of products and even Association

Rules we should set a minimum threshold. This is what motivated us to briefly discuss and introduce the concept of Centre-Piece Subgraph problem in the section 4.1.

6

CONCLUSIONS AND FUTURE WORK

As the final chapter of this thesis, we are going to review the work done in the previous chapters, make conclusions, and discuss what can be done in the future work.

6.1 Conclusions

In this thesis, we have used the application of network techniques to the problem of market basket analysis:

Given an unseen market basket data set, what steps we can take to carry out a thorough, complete analysis?

There has been a lot of prior work done on market basket analysis, and the use of Association Rules, in particular, has been extensively reported[RC11]. In this work, we elaborated the different methods to analyze the discovered communities. We have used graphs to build the network of products, where we eliminated the less frequent items from our dataset by setting the minimum threshold. The results which we got after detecting communities are used in analyzing the customer purchasing behavior. We have implemented Louvain method in Python language to identify groups of communities[DMFFP11]. The discovered results in this study can be used especially in a cross-sale recommendation and planning in marketing strategy.

Further, we implemented graph theory to find a network of products. A network of products includes nodes which represented the products and edges which showed the frequency between the products bought together in the transaction dataset. We visually represented the level of links between every two products in different communities. This information gave us an idea about the products brought together most of the time in the dataset.

After studying the properties of network and then detecting communities of products within the networks, one can extract more strong relationships between the products which were not easy to find with traditional Association Rules. Next, we implemented different techniques to analyze discovered communities. To find, which products plays a major role in a community, we used network density. It calculates a value for each node and then we ranked all the nodes in the network. Analyzing communities, not only express the strong relationship between products than Association Rules - but also gives the structural information. This knowledge helps in making financial decisions like where to place items in a store to have more profit or making promotions, about the products which are bought most of the time together. Similarly, we calculated PageRank and Centrality values for items which are frequent and form most of the rules in the dataset.

Recently, more and more companies have started to focus on understanding customer purchasing behavior to increase sales. Market basket analysis is not only restricted to find rules, but it is growing dramatically in multiple fields. MBA with a network of products helps companies to know more about their customer purchase behavior. Based on above discussion, we have proposed a general structure, which detects community of products from a network of products, and discovers the precious information regarding relationships between the products.

6.2 Future Work

If we set a minimum threshold for detecting sets of communities, we may lose information, as most of the products are pruned from the network. This leads to loss of the information about the products which are pruned. Secondly, it has a direct effect on the sales and production of products in a store, because we cannot discover valuable information about those pruned products. Since MBA is all based on data. If we are not able to choose the related data, the results will not provide enough valuable information.

Therefore, in future work, instead of detecting communities we can use Centre-piece subgraph. The major difference between the two approaches is that latter- finds relationships between all the products in the network so that all the products can be analyzed and ranked. Same with ARN where you choose a target product of your interest and centralize it, and discover all the valuable information, provided that the chosen target product should appear in a large number of association rules. The framework discusses detection of communities as an initial step, in the future, using Association Rules Network we can explore relationships within the dense network and Centre – Piece Subgraphs to explore special relationships and to validate hypotheses.

Bibliography

[AS$^+$94] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994. 6

[BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. 16

[Bor05] Giuseppe Borruso. Network density estimation: analysis of point patterns over a network. *Computational Science and Its Applications–ICCSA 2005*, pages 126–132, 2005. 15, 19

[CN06] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006. 24

[CNM04] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004. 16

[dee14] deepclimate.org. Market basket example, 2014. xi, 2

[DMFFP11] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. Generalized louvain method for community detection in large networks. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pages 88–93. IEEE, 2011. 17, 51

[Fri91] Noah E Friedkin. Theoretical foundations for centrality measures. *American journal of Sociology*, 96(6):1478–1504, 1991. 20

[HHR06] Michael Hahsler, Kurt Hornik, and Thomas Reutterer. Implications of probabilistic data modeling for mining association rules. In *From Data and Information Analysis to Knowledge Engineering*, pages 598–605. Springer, 2006. 6

[HPK11] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011. 6

[HPY00] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM Sigmod Record*, volume 29, pages 1–12. ACM, 2000. 6

[KKC12] Hyea Kyeong Kim, Jae Kyeong Kim, and Qiu Yi Chen. A product network analysis for extending the market basket analysis. *Expert Systems with Applications*, 39(8):7403–7410, 2012. 14

[LCK12] Annie Loraine Charlet and Ashok Kumar. Market basket analysis for a supermarket based on frequent itemset mining. 2012. 6

[New06] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006. 16

[PBTL99] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *International Conference on Database Theory*, pages 398–416. Springer, 1999. 6

[PCP$^+$09] Gaurav Pandey, Sanjay Chawla, Simon Poon, Bavani Arunasalam, and Joseph G Davis. Association rules network: Definition and applications. *Statistical analysis and data mining*, 1(4):260–279, 2009. 12

[RC11] Troy Raeder and Nitesh V Chawla. Market basket analysis with networks. *Social network analysis and mining*, 1(2):97–113, 2011. xii, 4, 6, 10, 15, 16, 34, 51

[Ros13] Gideon Rosenblatt. Network density, 2013. xi, 20

[SS08] Daniel A Schult and P Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, volume 2008, pages 11–16, 2008. 25

[TF06] Hanghang Tong and Christos Faloutsos. Center-piece subgraphs: problem definition and fast solutions. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 404–413. ACM, 2006. 26, 27, 28

[TKS02] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41. ACM, 2002.

[VCR14] Ivan F Videla-Cavieres and Sebastián A Ríos. Extending market basket analysis with graph mining techniques: A real case. *Expert Systems with Applications*, 41(4):1928–1936, 2014. 14

[VR$^+$07] Guido Van Rossum et al. Python programming language. In *USENIX Annual Technical Conference*, volume 41, page 36, 2007. 7

[wik17] wikipedia. Betweenness centrality, 2017. 21

[WS98] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’networks. *nature*, 393(6684):440–442, 1998. 14

[YD09] Erjia Yan and Ying Ding. Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the Association for Information Science and Technology*, 60(10):2107–2118, 2009. 21

[ZH02] Mohammed J Zaki and Ching-Jui Hsiao. Charm: An efficient algorithm for closed itemset mining. In *Proceedings of the 2002 SIAM international conference on data mining*, pages 457–473. SIAM, 2002. 10