Grocery Maze

Navigating through grocery maze by using Interactive Network Visualization

Jun Haur LOK, Xingwen WANG, Shifeng XU

Abstract— In this era of information explosion, this poses challenges in continuing the current conventional way to analyze association rules. Often these conventional methods do not show the underlying complex relationships between different items. Besides that, the conventional interface of the analytical software lack of interactivity between the graphs and the underlying model. Hence, in this research paper, an alternative way to visualize the association rules is presented. Demonstrations are also included to show interactivity functions are integrated into the application to provide a more seamless process in calibrating the model and visualizing the association rules. The designed application will provide users more control and flexibility.

Index Terms— Network visualization, interactive dashboard, market basket analysis

1 Introduction

As the data becomes easier to collect and computers become more powerful, companies are constantly finding ways to extract values from the data they have. Market basket analysis is widely used by companies in the retail industry. Often these graphs used in illustrating the association rules derived from market basket analysis are static graphs. This could pose an issue when the number of items or rules increases. The traditional way of visualizing the rules also lacks interactivity, which can be hard for the readers to read the graphs when the number of rules is huge.

Besides, the convention interface of the analytical software also lacks interactions between the graphs and the underlying models. The users often need to go back and forth between the model and the relevant graphs if the users would like to change the parameters. In other words, the interface does not allow the users to calibrate the model on the fly. This can pose a challenge to the users especially when the users are required to go through the codes to make necessary changes to the parameters.

To resolve the issues mentioned above, we have explored network visualization and ShinyApp. The purpose of this research is to provide an alternative method to visualize the association rules and provide users more flexibility and control in calibrating the model.

This research paper focuses on network visualization and R Shiny can be used to support the market basket analysis. Section 1 provides an overview of this research paper. Section 2 informs the readers about the motivation behind this research. Review and past critic can be found in section 3. This will be followed by the interface and design of the application. Lastly, this paper will also include a case study to illustrate how the application could help in assisting the users to have a better understanding of the underlying relationships.

2 Motivation

In retail services, market basket analysis is often used to mine the association rules between the different items. This can be used to understand what items tend to be purchased together. The results can be used in many areas, such as how the company cross-sell or up-sell based on the insights, how to arrange the products in the stores to maximize the sales, and so on. Furthermore, to ensure the success of the strategies mentioned above, the rules are typically assessed based on a few measurements, such as confidence, support, and lift.

Below is the simple explanation of the different key measurements under market basket analysis:

	Definition
Support	Measure how frequent the item or item set appears in the transactions
Confidence	Measure the likelihood that customers would buy the products shown in the rules, given that they have the products listed on the left-hand side in their basket
Lift	Co-occurrence of products on the left-hand side and right-hand side exceeds the likelihood of products on the left-hand side and right-hand side are independent

Table 1: Definitions of key measurements under market basket analysis

After researching the conventional way of analyzing the association rules, our motivation was sparked by a general lack of interactivity in the way of analyzing association rules and lack of interactivity in model building and the relevant visualization. As such, we aim to design an application that would enable the companies to better understand the underlying relationships to guide them to make a more informed business decision.

Hence the proposed design of the application will include the following features:

- To enable the users to better visualize the association rules
- To allow for interactivity in the visualization and model building
-

3 Review and Past Critic

One of the conventional ways to analyze the association rules is through making the rules into a data table format This approach does not reveal the underlying relationships between the various items.

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[1]	{Whole Wheat Bread}	{Banana}	0.011	0.427	1.179	26,946.000
[2]	{Kiwi}	{Banana}	0.011	0.499	1.377	25,937.000
[3]	{Blackberries}	{Banana}	0.011	0.408	1.128	26,372.000
[4]	{Yam}	{Banana}	0.010	0.457	1.262	24,900.000

Figure 1: Example of association rules shown as a data table

Figure 1 shows an example of how the association rules are being converted into the data table. It is hard for the audiences to visualize how products are closely associated with bananas, but it does not show the users aside from bananas, what other products are closely associated with these products.

Another approach is to plot the results in a scatterplot format. While it helps the users in identifying the product combinations with various measurements, it still does not show us the underlying structure or relationships of different products.

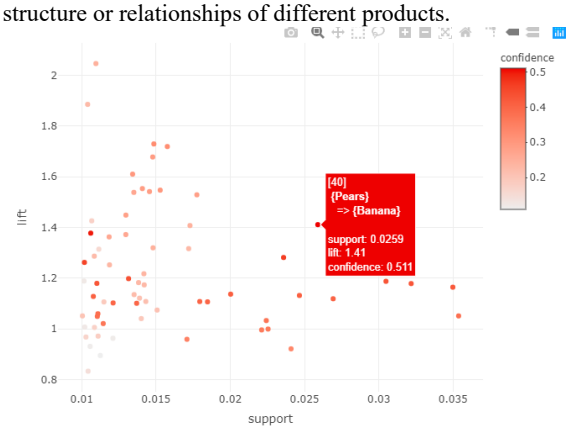

Figure 2: Example of Scatterplot on the market basket analysis results

Besides, it is impossible to visualize the overall relationships by using a data table as the number of rules increases. For example, Figure 2 shows that pears are associated with bananas. This visualization does not tell us whether there are any other items are also associated with pears.

Hence, to overcome this issue, network visualization provides a great alternative in visualizing the underlying relationships. In the article 'market basket analysis with network' by Raeder, T., Chawla, N.V., they have illustrated how they have leveraged on a static network graph in drawing insights. This approach clearly illustrated how the items are 'related' to one another. However, this static visualization method posed a challenge when the number of rules increases. It is also not possible to observe whether there are any clusters within the product groups such as there are any products that are closely associated together as a cluster and so on. Furthermore, often the items might overlap with one another as shown in Figure 3 (Kam). It is not possible to make the items to repel one another so that we could read the individual items.

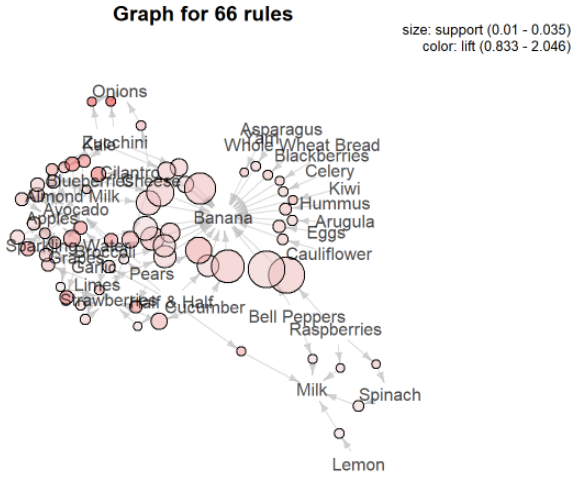

Figure 3: Static network graph

We could reduce the number of rules illustrated in the graph. However, this is not preferred as it could result in missing some important relationships between the products. Hence, to resolve this, we could further enhance the network visualization technique by including the interactivity functions such as highlighting, into the visualization.

Apart from that, there is a delink between the model building and visualizing the results conventionally, even with some on-the-shelf tools. In the visual analytics lecture note by Professor Kam Tin Seong (Figure 4), he illustrated how visualization is a part of the model building process. This does not allow the users to calibrate the model on the fly. Often this process also involves back and forth between model building and visualization. The users might want to recalibrate the model after reviewing the results.

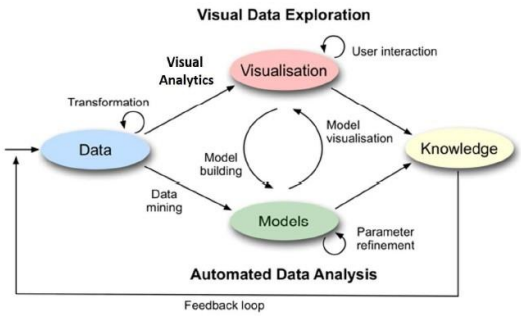

Figure 4: Model building process

4 APPLICATION (DESIGN FRAMEWORK)

Several articles and papers were carefully examined to understand how we could resolve some of the challenges mentioned above. With the features mentioned above and a few rounds of iterations, below is the final prototype:

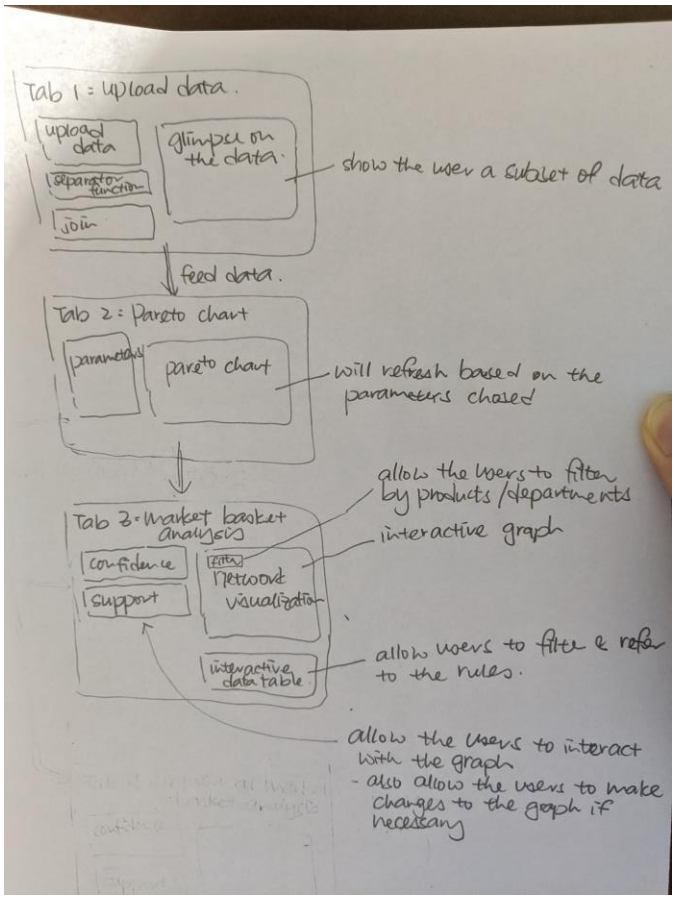

Figure 5: Final prototype on the design of the application

In each subsection, we will further discuss how various issues are being tackled in the web application and data visualization.

4.1 Web application

ShinyApp is used in this research as it is a great tool for an interactive web application. This also allows the users to calibrate the model on the fly by including appropriate parameters while building the ShinyApp. This gives the audience a deeper and better insight through our powerful, flexible, and interactive dashboards.

Apart from that, ShinyApp is using R and R with its different modern data science packages such as ggplot, tidyverse, and dplyr. This provides a very rich environment for the project team to perform exploratory data analysis, data cleansing, analysis, and visualization.

ShinyApp also allows us to build in different appropriate parameters to allow the users to interact with the model. This will allow the users to adjust/change the values of the parameters and the visualization will be updated based on the selected parameters.

Lastly, unlike the convention on-the-shelf tools, ShinyApp is easier to share across different users. Licenses are often required for convention on-the-shelf tools and it costs money to acquire more licenses to use the conventional software. This creates a hurdle for more business users to use the application if the cost is a consideration.

4.2 Data Visualization

While designing the visualization, we have used the principle of 'Overview first, zoom and filter, then details on demands' by Ben Shneiderman in his article 'The eyes have it: A task by data type taxonomy for information visualization'. This provides the users with an overview of how the relationships look like overall. A highlight button will be included to allow users to highlight the selected products. The selected R package should allow the users to be able to zoom and focus on the relationships of certain product groups.

However, these requirements would require the application to be interactive. This is another reason why R Shiny was chosen as R Shiny made this possible.

Pareto chart is also included as one of the tabs under the ShinyApp as shown in Figure 6. This is to help the users to understand the distribution of products sold. With that, the users would be able to know what the top-selling products are. This would be very handy when the companies design the strategies based on the association rules. This is because companies might encounter the scenarios that the top products that are strongly associated might not be the most popular products. With the Pareto chart, the company can pick product combinations that do not have too low sales and with good measurements such as lift.

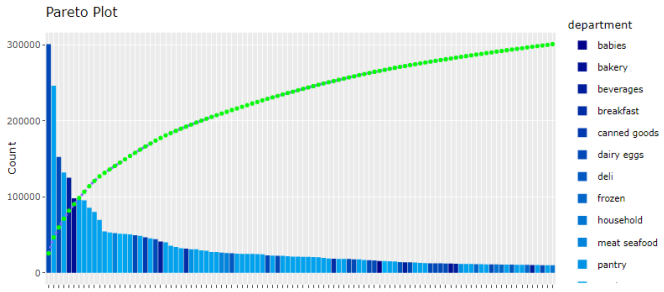

Figure 6: Example of the Pareto chart

We will also use network visualization to better visualize the underlying relationships. In the book 'Finding Beautiful Insights in the Chaos of Social Network Visualizations' by Adam Peter, he explained how network analysts focus on relationships instead of just the individual elements that can explain social, cultural, or economic phenomena. He further argued that how the elements are connected is just as important as the elements themselves. This is true for our case. Often, one of the main applications of market basket analysis is to understand how the items are related to one another and how strong are the relationships. This would guide the companies in coming up with appropriate strategies.

Igraph and visNetwork were compared to weight the pros and cons of each package. visNetwork is chosen as it allows us to plot an interactive graph. This is particularly important especially when the number of rules is big. The package allows us to highlight certain rules. On the other hand, the network graph produces by igraph are static.

visNetwork is also compatible with R Shiny even though it is using vis.js, which is a java library. This package also allows us to more customization if necessary.

Finally, the design of visualizations in this research also referred to the recommended taxonomies stated in the article of 'Interactive Dynamics for Visual Analysis' by Heer J. and Shneiderman B. Functions such as highlight, zoom and so on are included to the interface to allow the users to interact with the applications.

5 CASE STUDY: INSTACART GROCERY DATA

In this case study, we will illustrate the advantage of using a network visualization technique in illustrating the association rules.

The dataset from Instacart is used in this research. This data is accessed from the Instacart website. The full dataset consists of 33.8 million from over 200,000 unique customers. The transactions contain the info of the products purchased made by the customers.

arules is used in finding the underlying association rules. However, arules package ignores the time effect while finding the association rules. This could pose an issue as the underlying customer preference could have changed over time. Furthermore, this dataset also contains the different transactions made by the customers over time and the sequence of the transaction, this allows us to explore the sequential market basket analysis. Hence, to account for the time effect, the arulesSequence package is also used to discover the underlying relationships.

Other than that, it also contains info such as the departments the purchased products belong to and the aisle where the products are being placed.

One issue faced during the analysis is the large number transactions contained in the dataset. There are close to 30 million transactions in the dataset. This poses a challenge in terms of the runtime of the analysis. Also, in practice, we will have a customised customer strategy for different segments of customers. So, in this research, we will focus more time on understanding customer buying behavior for more loyal customers. Top 5% of our customers in terms of sales will be used in this analysis.

We also noted that there are many variations to the product names although they seem to be the same type of products (Figure 7). Hence, to resolve this, the product names are being recoded to group similar products together.

	product_id	product_name	aisle_id	department_id
1	143	Organic Lemons	24	4
2	5876	Organic Lemon	24	4

Figure 7: Example of different product names for the same product

The original data is being divided into different datasets. Instead of performing joining so many data joining over the web application, we will merge some of the datasets before uploading them into ShinyApp.

Another issue with the dataset is there seems to have some mismatch between the departments and food. For example, during the analysis, we noted that strawberry appeared in multiple departments and one of the departments it appeared under is 'Dairy eggs' as shown under Figure 8. Due to the lack of information to verify this, we will not change the data.

Figure 8: Screenshot of strawberry that falls under 'Dairy eggs' department

However, after examining the results, we noted that this dataset is not appropriate to use sequential market basket analysis. The result shows that people tend to buy the same items over time. This also suggests that user preference has not changed significantly over time. Hence, the remaining of the paper will focus on how the network visualization on the market basket analysis.

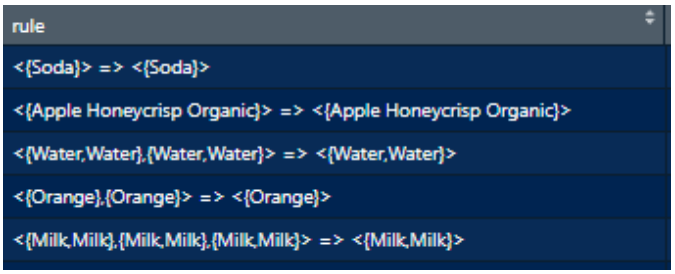

Figure 9: Top 5 association rules under sequential market basket analysis

The detailed steps on how the association rules are run can be found under the wiki page for this research project (https://wiki.smu.edu.sg/1920t2isss608/Group01_proposal).

Figure 10 is the network visualization produced by the application built.

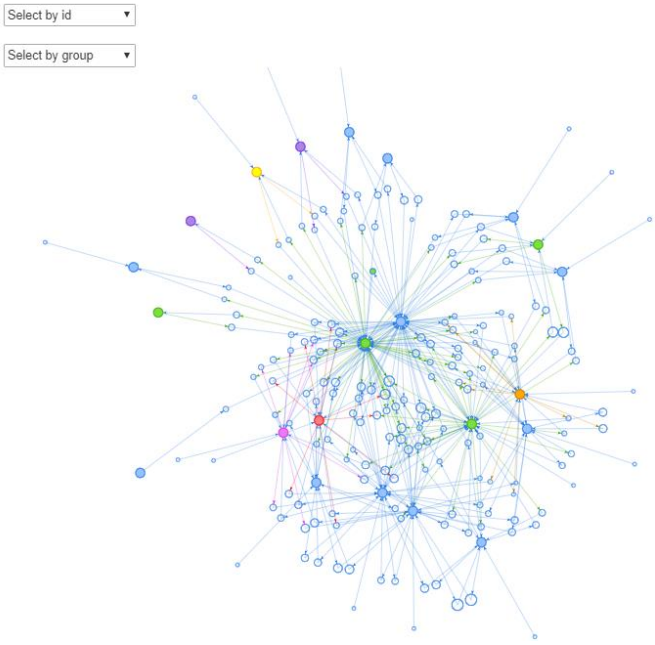

Figure 10: Final network visualization on the association rules

In the following example, we will contrast how the interactivity of the graph has helped the users to quickly identify the interested product, compared to a static graph. For example, if the user would like to know the products that are closely associated with chocolate. It would be much easier to tell from the associated products by highlighting the chocolate and associated products from the graph (Figure 12).

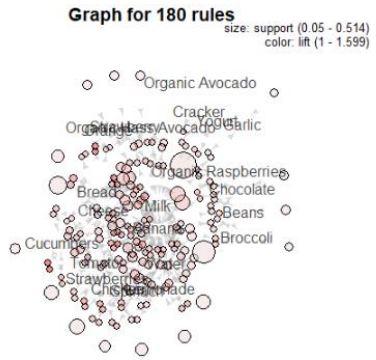

Figure 11: Static network graph

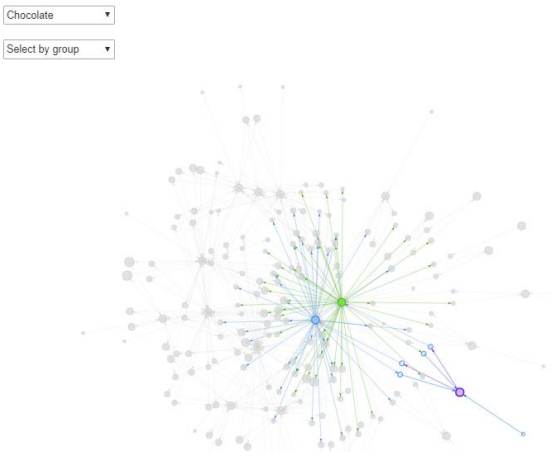

Figure 12: Network graph with chocolate & the relevant associated products being highlighted

Through the interactive network visualization, we discovered a few interesting product combinations. For example, Figure 13 shows that garlic is associated with milk. This implies that there is a chance we might find garlic in the grocery basket, provided milk is already in the basket.

This visualization can be used to verify our initial hypothesis. In this study, our initial hypothesis is garlic would be associated with some of the vegetables since garlic is sometimes used in cooking. This graph proved the hypothesis is wrong as garlic is not associated with any of the vegetables. It could be that the customers rarely use garlic in their cooking. Alternatively, it could be the customers have used some other ingredients which are a substitute for garlic in their cooking.

Nevertheless, this graph also indicates that it may not be a good idea to run target promotion on garlic. The potential benefit the company could extract from the promotion on garlic is likely to be minimal due to its low association with other items.

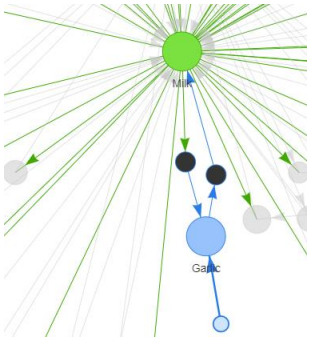

Figure 13: Screenshot of garlic is associated with milk

If we were to look at the top 100 rules by sorting the support descending, Garlic, as shown under, is not associated with the rest of the product. It would be hard to identify this if we were to use a data table to analyze the association rules.

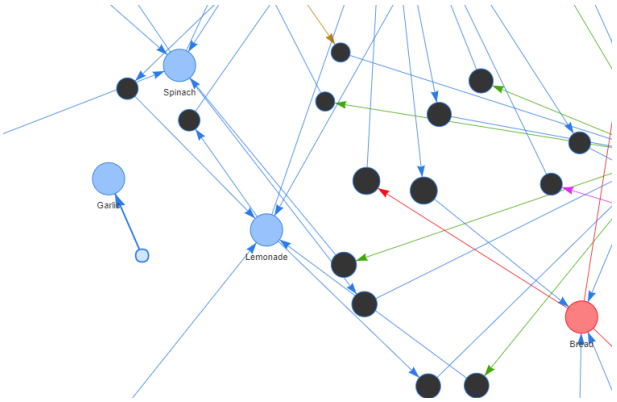

Figure 14: Screenshot of garlic not associated with the rest

6 Conclusion

In this paper, we have demonstrated how we have used both ShinyApp and network visualization to draw insights from the association rules. This would help the companies to have a better understanding of how the different items are 'related' to one another. The companies would know how they would like to cross-sell or up-sell the products to boost product sales.

Nevertheless, the network graph used in this research paper is not the only type of network graph. The design could further improve by trying different network graphs so we can compare the different types of the network graph. Crosstalk function could be added to the application to link the graphs together. This would enhance the user experience as well.

Acknowledgments

The authors wish to thank Professor Tin Seong KAM for the guidance provided along with this research.

7 References

Heer J., S. B. (n.d.). Interactive Dynamics for Visual Analysis.

Kam, T. S. (n.d.). Week 1 Lecture Note of Visual Analytics and Applications.

Perer, A. (n.d.). *Finding Beautiful Insights in the Chaos of Social Network Visualizations.*

Raeder, T. C. (2010). Market basket Analysis. *Social Network Analysis and Mining*, 97–113(2011).

Shneiderman, B. (2005). The eyes have it: A task by data type. *IEEE Conference on Visual*, 336-343.