Team Members: David Lysko, Jessica Chan Vargas, Ji Ho Ahn, Kunwoo Hong

Introduction/Motivation

Over the past decade, e-commerce has grown significantly, along with the amount of transaction data generated. One of the most common analytics approaches for gaining useful insights from such data is association rule mining (ARM), which aims to discover products commonly purchased together and quantify the strength of these associations. For our project, we will build an interactive app for the visualization of rules derived from a UCI Machine Learning Repository dataset for a UK-based retailer of gifts, which includes variables such as invoice number, invoice date, and items purchased (Chen et al. 199). Our team's motivation for looking into this problem is that it is very common in business and has many practical applications. Furthermore, we as customers could benefit significantly if more organizations adapted this approach and actioned based on the findings from ARM. As we are all active consumers, we believe that sellers can provide additional discounts/promotions if they know their customers well.

Problem Definition

Retailers collect data from each purchase made, including data that includes but is not limited to the items purchased, date of purchase, and price of item purchased. With thousands of different transactions made up of multiple items per transaction, retailers' main concerns shift to understanding purchasing behavior, particularly, what items are frequently purchased together. Understanding these association rules from purchases can help to reduce overall costs and improve return on investment. Marketing teams can focus on promoting products relevant to a particular item and remove those with minor impact, leading to greater returns and more efficient marketing as fewer resources are used on irrelevant products. The advantages of the findings from association rules can assist with various levels of decision making within a retail business. However, there needs to be a clear and straightforward way to find and interpret these association rules.

Problem Statement: To help solve this problem, we will develop an interactive app which can surface the most useful association rules for a user-provided set of inputs and data filters.

Survey

The usefulness of association rules can be quantified by three metrics: support, confidence, and lift. Suppose we had a rule {flour} => {sugar}, implying sugar is commonly purchased together with flour. Then support is the proportion of all transactions containing flour and sugar, confidence is the conditional probability that sugar is purchased given that flour is purchased, and lift is the confidence of the rule divided by the support of the right-hand side. The higher the lift ratio, the greater the strength of association (Shmueli et al. 332-335).] There are numerous algorithms for ARM, but Cavique states that the most popular algorithm is the Apriori algorithm which uses confidence and support thresholds, in order to determine the various product relationships available in a market basket (Cavique 401-402). Even though the Apriori is the most popular ARM algorithm, there are two downsides of the Apriori algorithm mentioned by Ansari which are 1) heavy computation in the rule generation process 2) multiple scans of the dataset (Ansari 26). In order to generate the association rules and the appropriate visualizations, there are 2 packages in R: arules, which contains an efficient implementation of the apriori algorithm (Hornik et al. 1-25) and arules Viz (Hahsler 163-175). According to Becerikli (422), using insights from association rules allowed marketers to get a better understanding of customer behavior, find products in common, make better decision about product placement, and remove products with little impact. With the knowledge gained from the association rules, Seng and others said that the knowledge can be used to create cross-selling, upselling, targeted promotions, and inventory management strategies (Seng et al. 8054). The association rules of the online behavior data can also be used to improve the recommendation systems according to Kim and Yum (Kim and Yum 13320). With the continued uptrend of the marketing costs nowadays compared to the past as mentioned by Weber (Weber 705), association rules that are ranked/interpreted correctly can provide better information on targeted promotions as mentioned by McCormick (McCormick 2) which can increase the return on investment on the marketing costs and increase the overall revenue.

Proposed Method

Intuition: We believe our application's top two innovations are interactivity and ease of use and interpretation. This allows any user at any level of a business to understand what the rules are and what they might mean for business decisions.

Interactivity: The app allows users to specify various inputs in order to dynamically extract the association rules. These users can then explore these rules through a graph. The users can select a particular item and the rules involving the item will be highlighted in the graph.

Ease of Use/Interpretation: The graph includes details for each item and associated rule when hovered over. In addition, our app surfaces a table of the top 20 rules sorted by strength of association for a good starting point for further exploration.

Approach:

Raw data exploration

The date range for these transactions is December 2010 to June 2011. The dataset has 541,909 rows and 8 variables. The fields of interest are InvoiceNo, Description, InvoiceDate, and Country, so we are not including the other columns in our final transaction set. The dataset includes transactions from customers in 38 different countries; however, many of them (91%) are from customers in the United Kingdom. There are 4,224 distinct product names. We discovered that there are several invalid product names such as "error" or "damaged" or "???". Most of these non-products have a negative number for unit price.

Data cleaning and transformation

The dataset is formatted such that each transaction is split into a row for each item purchased, resulting in multiple rows for the same transaction. The dates are first formatted so that they are guaranteed to be uniform for proper date filtering later on. We then filtered the data so that only valid products and purchases are included. This process involves filtering out rows where the quantity field is less than or equal to zero, as well as prices that are less than or equal to zero. Additionally, invalid product descriptions such as "Sale Error" or "AMAZON FEE" are removed. After cleaning, we filter out the four relevant columns: InvoiceNo, Description, InvoiceDate, and Country. The Apriori algorithm requires the dataset to be in a binary format. The dataset is transformed such that transactions form the rows and the items form the columns. If an item was purchased then the matrix position shows as 1, if not then 0.

Transaction exploratory data analysis

After data cleansing and converting the data into a transaction format, we end up 19,847 unique transactions or "baskets" with 4,008 different products. The density of this item matrix is 0.006518492 which indicates that the data is sparse. Most customers are purchasing only a handful of items frequently, other purchases seem to be one off during this time period**.** This observation confirms Ansari's claim that the distribution of supermarket like data tend to have a long-tailed distribution (18)**.** The median number of items in each transaction is 8 with a mean of 26.3. There is an outlier with 1,107 items in a single transaction. Most customers have a small number of items in each transaction.

R Shiny App

The Shiny app consists of three key pieces of code: 1) User Interface 2) Server 3) ShinyApp function which takes UI and Server as inputs and generates the application

User Interface: Within this part of the code, sliders for support, confidence, and lift are defined. For each slider, min, max, and default values are specified. An InvoiceDate range filter is also defined.

The UI code also defines the visual layout of the app. This project uses a sidebar layout, which displays user inputs on the left and outputs on the right. There are two outputs: an interactive network graph of association rules and a tabular display below it. The interactive network graph supports the selection of a specific item of interest (which highlights the rule(s) to which it belongs), zooming in or out, and hovering over a particular node to display the relevant association rule details.

Server: This part of the code is used to dynamically filter the data based on inputs provided, run the apriori algorithm, and return the resulting association rules.

A variable called rules is assigned to a reactive container which listens for changes to inputs. When any of support, confidence, lift, or InvoiceDate inputs are changed, the following actions take place:

- Clean dataset is filtered based on InvoiceDate range
- A new transactions object is created
- The apriori algorithm (from the arules library) is run on the transactions object subject to support and confidence input thresholds
- The resulting rules are filtered based on lift threshold

The resulting rules are then used to generate two outputs, which are linked to those defined in the UI part of the code: 1) a network graph of association rules (using arulesViz plot function) and
a table of association rules (limited to a maximum of 20 rules, based on descending lift and confidence).

Experiments/Evaluation

Testbed

In exploring the association rules visualization application of the online retail dataset, we are focusing on evaluating several key points: what filters are applicable to the dataset and will add to the design of the application, does the application function as intended, what rules are being found and what patterns emerge from those rules, and is it intuitive and easily navigable?

Application Design Experiment

The application includes different filters to provide an interactive model of the association rules. One of the filters initially planned was for Country but was not included in the final model. We noticed that the Country field is highly skewed towards purchases from the United Kingdom as 91% of transactions were from the UK. When filtering by other countries, there is too little data to create reliable association rules, if any rules are made at all. Therefore, the final application includes only the date and threshold filters.

Application Evaluation

Deployment

The application needs to be tested on quality of performance when deployed through shinyapps.io. Upon uploading our app code and clean dataset to the Free version of shinyapps.io, we were unable to run the app due to insufficient memory (Free version is limited to 1GB). Several modifications to the app were tried to see if the memory issue could be resolved:

- Removing the InvoiceDate range input
- Removing the association rule table output
- Including only dates between Jan to Jun 2011

Despite these attempts, the memory remained insufficient. In a real-world business setting, a company could easily spend $39/month to upgrade to the Basic tier of shinyapps.io, which would provide additional memory capacity. For our project, we have instead found that we could successfully run the app on shinyapps.io if we limited the uploaded data to the first 100,000 rows of the cleaned dataset. A link to this interactive test app is here: https://dlysko3.shinyapps.io/Arules_App_Test

This is meant only to demonstrate how this app would look and work if deployed in a business-setting, whereby users could simply access the app via a URL. Other than the dataset, the functionality and appearance is identical to running it locally on one's computer. **Please note, however, that the observations and conclusions in our report are based on running the app locally with the full clean dataset.**

Application Functionality

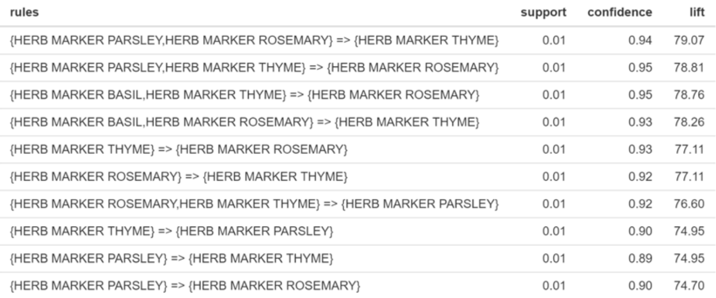
Upon the initial evaluation of application, we made a few high-level observations. The maximum support in the rules generated from dataset is 0.04. This means that at maximum 4% of transactions contain the products in a particular association rule. The date range filter can also be too restrictive if your date range is too small and will result in 0 rules. It is important to note that the network graph located at the top right corner of the application displays the 100 top generated rules by lift and the table at the bottom displays the top 20 rules sorted by lift, confidence.

Test Parameters	What was tested?	Observations
No date filters, different thresholds	Support filter	• Max value without error 0.04 • 1 rule generated with support set at 0.04 and showed in both network graph and table
No date filters, different thresholds	Confidence	• Max value with error 0.98 • 1 rule generated at max confidence and displayed both on table and network graph
No date filters, default thresholds	Network graph & dropdown	• Displays rule clusters generated some interconnected based on a single common purchased

		• Network graph dropdown correctly highlights selected rule or rules with selected product • Able to zoom in and out of graph and drag graph • Hovering over rule pill shows rule details
Date filters: (12-01-2010-01-01-2011) and (11-01-2011-12-31-2011), default thresholds	Date filter	• 2010 – 0 rules generated • 2011- 1 rule generated, as expected Christmas related items
No date filters, confidence 0.09, support 0.01, lift def	Date filters, table, network graph	• 5 rule clusters in graph
Different thresholds & date filters	Performance	• Network graph and table rules rendered in acceptable time, If no rules were found, it generated an error

Association Rules

The results generated by the algorithm vary depending on the threshold limits. Setting confidence threshold at 0.50, support threshold 0.01and lift > 1 generate rules related to herb purchases. Each of the top 20 rules contain different combinations of herb purchases and most rules contain 2 or 3 itemsets. The top rule is {HERB MARKER PARSLEY,HERB MARKER ROSEMARY} => {HERB MARKER THYME} with support 0.01, confidence 0.94, and lift of 79.07. According to this rule, customers who purchase parsley and rosemary are also likely to purchase thyme. About 1% of transactions contain these 3 items. The confidence is high as is the lift value. The lift value that is greater than 1 indicates a positive association between the antecedent and consequent. The high lift value means there is a greater chance for there to be a preference to buy thyme if the customer has already purchased parsley and rosemary.



rules	support	confidence	lift
{HERB MARKER PARSLEY,HERB MARKER ROSEMARY} => {HERB MARKER THYME}	0.01	0.94	79.07
{HERB MARKER PARSLEY,HERB MARKER THYME} => {HERB MARKER ROSEMARY}	0.01	0.95	78.81
{HERB MARKER BASIL,HERB MARKER THYME} => {HERB MARKER ROSEMARY}	0.01	0.95	78.76
{HERB MARKER BASIL,HERB MARKER ROSEMARY} => {HERB MARKER THYME}	0.01	0.93	78.26
{HERB MARKER THYME} => {HERB MARKER ROSEMARY}	0.01	0.93	77.11
{HERB MARKER ROSEMARY} => {HERB MARKER THYME}	0.01	0.92	77.11
{HERB MARKER ROSEMARY,HERB MARKER THYME} => {HERB MARKER PARSLEY}	0.01	0.92	76.60
{HERB MARKER THYME} => {HERB MARKER PARSLEY}	0.01	0.90	74.95
{HERB MARKER PARSLEY} => {HERB MARKER THYME}	0.01	0.89	74.95
{HERB MARKER PARSLEY} => {HERB MARKER ROSEMARY}	0.01	0.90	74.70

Figure A. Table of top 20 rules

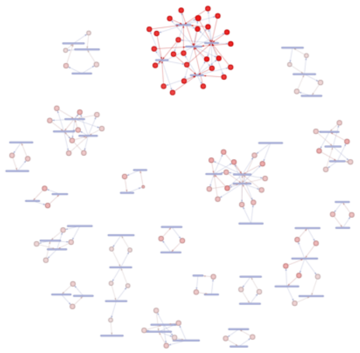


Figure B. Network graph

Using the same test parameters as mentioned above, the network graph displayed several smaller networks that are not all interconnected. Herb products form the largest network which contains about 30 rules. The next largest network contains regency tea set products is made up of 13 rules. There are only a handful of these tea products: plates in green, pink or rose pattern, and green or rose teacups and saucers. All of these items seem to be influencing the purchase of each other. For example, when looking at rules involving herbs, a purchase of thyme indicates a strong association of purchasing rosemary. However, when purchasing thyme and parsley, a separate rule shows a stronger association to purchasing rosemary. This would indicate that purchasing multiple items increases the likelihood of purchasing a target item than having purchased just one.

Interface Analysis and Navigation

Lastly, the application should be intuitive and easily navigated. This requires evaluating the user interface and making design choices such that any person can easily interact with the filters and analyze the rules that are created. In testing, the built-in functions from the arulesviz package show distinct rule nodes and item nodes: item nodes are portrayed as bars and rule nodes are circles of varying red hue based on the strength of association. This coincides with the table of rules below the graph. This table contains the items and the values of the three thresholds for each rule.

In the graph itself, all of the details of the rules are easily accessible by hovering over a rule node, showing all the details concerning the association rule. The graph is supported further through the visual connections between the item and rule nodes. Items on the left-hand side of a rule will point to a rule node with a blue-gray arrow; the node will use a red arrow to point to the right-hand side item. When selecting a particular item in the drop down, all rules and connections associated with the item are highlighted. This not only makes all of the information for each rule readily accessible, but users can get an overview of what kinds of associations exist for that item.

The graph is easily navigated by using the scroll wheel to zoom in and out, as well as clicking and dragging around the empty space to move around the graph. The networks within the graph can be adjusted to the user by dragging the different nodes. This allows the user to customize their view and focus on different rules or network of rules, making the rules clearer.

Conclusion

By leveraging RShiny and RStudio, we were able to successfully mine the online retail data set and create an interactive online application that allows the user to easily navigate and visualize the important association rules. These unique capabilities provide business owners with a view into customer purchasing patterns which allow them to focus on a particular product or specific time period. This analysis could even be done on a weekly basis with more data.

Many of the association rules consist of related products but vary in color or type. These purchasing patterns are expected as this retailer sells unique all occasion gifts.

- Rules consist of complementary and supplementary products
  - Complementary: teacups, tea plates, and tea saucers
  - Supplementary: paper cups, napkins, plates
- Cross-selling-promotions

With this application, we have shown a successful proof of concept for any business to better understand their customers' purchasing behaviour at any level of the business.

Distribution of work

All team members have contributed equally.

Works Cited

Ansari, Sohaib Zafar. *Market basket analysis: trend analysis of association rules in different time periods*. Diss. 2019.

Cavique, Luís. "A scalable algorithm for the market basket analysis." Journal of Retailing and Consumer Services 14.6 (2007): 400-407.

Chen, Daqing, Sai Laing Sain, and Kun Guo. "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining." *Journal of Database Marketing & Customer Strategy Management* 19.3 (2012): 197-208. UCI Machine Learning Repository. Online Retail Data Set. archive.ics.uci.edu/ml/datasets/online+retail.

Hahsler, Michael. "arulesViz: Interactive Visualization of Association Rules with R." The R Journal. 9.2 (2017): 163-175.

Hornik, Kurt, Bettina Grün, and Michael Hahsler. "arules-A computational environment for mining association rules and frequent item sets." Journal of statistical software 14.15 (2005): 1-25.

Kim, Yong Soo, and Bong-Jin Yum. "Recommender system based on click stream data using association mining." Expert Systems with Applications 38.10 (2011): 13320-13327.

Kurniawan, Fachrul, et al. "Market Basket Analysis to identify customer behaviors by way of transaction data." *Knowledge Engineering and Data Science* 1.1 (2018): 20.

McCormick, Tyler, Cynthia Rudin, and David Madigan. "A hierarchical model for association rule mining of sequential events: An approach to automated medical symptom prediction." (2011).

Seng, Jia-Lang, and T. C. Chen. "An analytic approach to select data mining for business decision." *Expert Systems with Applications* 37.12 (2010): 8042-8057.

Shmueli, Galit, et al. "Association Rules and Collaborative Filtering."Data Mining for Business Analytics: Concepts, Techniques, and Applications in R. Hoboken: John Wiley & Sons, Inc., 2018. 329-351.

Svetina, Marko, and Jože Zupančič. "How to increase sales in retail with market basket analysis." Systems Integration (2005): 418-428.

Weber, John A. "Managing the marketing budget in a cost-constrained environment." Industrial Marketing Management 31.8 (2002): 705-717.