A MATRIX-BASED VISUAL ANALYTICS APPROACH FOR THE ANALYSIS OF ASSOCIATION RULES

by

Rakshit Varu

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
April 2021

To my family and friends (my second family) who have helped me through all the ups and downs. I wouldn't have come this far without your support.

Thank you Professor Fernando Paulovich and all my colleague for the help and guidance.

Table of Contents

List of Tables

List of Figures

Abstract

With a large amount of data, many data mining techniques have contributed to the extraction of useful information from the data. The generated information, if interpreted correctly results in a huge benefit for the user. Association rule mining is a popular data mining technique because it helps in finding hidden relationships between items in a database. The application of this data mining technique is not just limited to market data but also in health, entertainment, and census data to, name a few. With large data, the rules generated can be in thousands and needs to be assisted with proper visualization to attain the correct information that cannot be collected using textual rules. Text-based rules make it difficult for a user to analyze the results easily and get more information at the same time. Previously, a lot of efforts have been made to assist the visualization of these rules but there exists a gap in the proposed systems. In this work, we present ARMatrix, a system that provides an item-to-rule matrix-based visualization. The system comprises four modules, with the main interface providing the visualization to assist the rules with a view to compare selected rules based on different measures. The ordering options help to get a rearranged visualization based on placement or color. The freedom to find the subsets of the rules for exploration is provided using the filter section. To attain reusability and revisit a state later, the history section is provided to retrieve, delete or store the required state. The proposed system helps the general user with some knowledge of the association rules to navigate the rules and find the relevant information. The usability of the proposed technique to visualize and analyze the association rules is illustrated using two user scenarios and then confirmed from the feedback received by conducting a user test with 20 participants.

Chapter 1

Introduction

Data mining is a vast field that deals with the extraction of information from data and provide it for further use. The hidden patterns in data can provide important information that can directly or indirectly help many business domains. The applications of data mining are in multiple areas, including healthcare [25], education [4], and marketing [6] but are not just limited to them. Whether it is a classification or a recommendation problem, it is essential to provide the information in a manner that is easy to interpret.

In the past years, the increase in data has given rise to many data mining techniques that help to find hidden patterns in the data [8]. Whether it be clustering, classification, regression, or association rule learning [3], it is of utmost importance to present the results with simplicity to avoid any misinterpretation. It is done because wrongly delivered information can make a negative impact [27] in the decision-making process.

One of the data mining techniques is association rule mining which extracts hidden relations between items in a database [3]. This rule-based technique has applications varying from market data [36], intrusion detection [40], bioinformatics [5], to name a few. Despite all the benefits of these rules, some flaws need to be addressed correctly including the discovery of large rule counts [19]. The number of rules generated depends on either the threshold measures or the amount of data present in the database. The rule count can be in thousands, and it becomes difficult to look at all the rules with less cognitive load and find the relevant ones.

One possible solution to solve the problem mentioned above is visualization. Generally, the visualization process is integrated with data mining to extract information depending on the task including preprocessing of the data [29], provide explanations on working of a model [14], or representing discovered information [34]. Regardless of the number of rules, it is necessary to provide a proper visualization to understand

the results [41] and find the useful subsets so that even a general user with little knowledge about the rules can utilize this technique to gain filtered information.

Many attempts have been made to assist association rules visualization. Some frameworks approached this task using the tabular method [46, 44, 47]. Graphical assistance including node-connections level [43, 45, 46, 44, 26], scatter plots [42, 28], hierarchical tree view [30, 12], and mosaic plots [22]. Matrix view is another alternative for visualization [15, 30, 20] along with 3D visualizations [17, 37, 31, 9]. However, as the rule count increases, it becomes difficult to get important subsets and graphically manage all the coinciding items and rules. It also becomes tough to correctly interpret the rules without much cognitive burden. 3-Dimensional item-to-rule matrix [13, 32] helps solve the scalability issue if incorporated correctly. Nevertheless, these frameworks lack to provide a deeper level of freedom to analyze the rules based on different functionalities involving filtering by items or incorporating different interesting measures. Also, 3-Dimensional figures or data can confuse users while interpreting the results. This happens because 3D figures tend to distort the data creating a perceptual ambiguity [39].

In this work, we propose a technique to address the issue of visualization of a large amount of rules. The system is designed by keeping in mind a user with little knowledge about association rules so that it is easy to use and gain more information without much cognitive load.

1.1 Research Questions

Based on the problems in the visualization of association rules, our work tries to answer the following questions:

1. Is it possible to visualize a large number of rules without loosing the authenticity of the interesting measure (*support, confidence, lift, conviction, leverage, to name a few*) values required to gain insights from the rules?

2. Is there a way to perform a focused analysis on the subsets of the rules to limit the rules displayed and reduce the analysis time?

3. Is it possible to compare the rules of interest with one another for comparative analysis to gain the required results depending on the task at hand?

1.2 Proposed Solution

To overcome the issue of visualizing a large number of rules, we present ARMatrix, a framework incorporating a 2-Dimensional item-to-rule matrix to visually assist the association rules. With a large number of rules, it becomes difficult to examine multiple rules together. To solve that, we incorporate a visualization that compares multiple rules based on interesting measures. The filtering option is added to find the subsets of rules which helps to limit the number of rules based on necessity. The main features of the system are:

- A 2D item-to-rule matrix to display the rules, with detailed information about the rule provided on demand.
- Filtering option is added so that users can filter out the rules based on the items they need.
- A visualization to compare different rules by just double-clicking the item in columns.
- A history section is added to save, retrieve or delete a particular state of rules.
- Customization options for placement of rows and columns and control over each column item's color. Column placement using similarity as a measure is another feature added.

1.3 Contribution

The contributions of the thesis are the following:

- Created a technique for the visualization and analysis of the association rules and development of a framework for assistance.
- The framework and the technique cover the visualization of association rules along with the ability to do further analysis on the rules to gain insights about these rules and their subsets by focused and comparative analysis, which lack in the other works done in this field.

- Validation for the usability of the system and ease of understanding the rules is done through a user study conducted with 20 users.

1.4 Organization of Thesis

The organization of this thesis is done in 6 chapters, which are based on the paper developed during the Master's Program.

Chapter 2 deals with the background knowledge of the association rules along with the brief knowledge about the algorithms used in the calculation of association rules and the interesting measures that are associated with the rule mining.

In **chapter 3**, we review the different visualization methods that can be used for displaying association rules. We also discuss the existing system and literature plus their shortcomings.

In **chapter 4**, we present the methodology followed for achieving the research goal discussed in the previous section. We also explain the system developed in detail.

Chapter 5 contains the results of the user study conducted while the development of the system along with the example use cases for the implementation of the system.

Finally, **chapter 6** summarizes the thesis with a discussion about the assessment done. Also, the limitations, future work, and a conclusion are addressed.

Chapter 2

Background

2.1 Association Rules

Association rules are "if-then" statements that show how often things have occurred together within a database [3]. Let items $\boldsymbol{I}$ be a set of binary attributes represented by $I = i_1, i_2, ..., i_n$ and $\boldsymbol{D}$ be a set of transactions in a database, $D = d_1, d_2, ..., d_m$. Each transaction d consists set of items such that $d \subseteq I$. Association rules help to find the relations between the items in the database by finding frequent if-then associations. The rules are in the form of X $\rightarrow$ Y, both X and Y (also known as *itemsets*) belong to items $\boldsymbol{I}$, where X is antecedent (if), items in the database and Y is consequent (then), items which occurs with the antecedent. In the rule Beer $\rightarrow$ Diaper, Beer is antecedent and Diaper is consequent. The equation implies that when a customer purchases Beer, he/she also tends to purchase Diapers.

The general approach to calculating these association rules includes

1. Use minimum support threshold to find frequent itemsets.

2. Use minimum confidence threshold to extract rules from the frequent itemsets.

Depending on factors such as performance, the number of iterations, and memory, the rules are calculated using different algorithms including Apriori, FP-growth, and Eclat [38].

2.2 Interesting Measures

The association rules are calculated using different measures, with the basic measures being support and confidence. There are more than 40 different measures that exist today [23], but in this work, we consider only 7 interesting measures: *support, antecedent support, consequent support, confidence, lift, leverage and conviction.* This is done to simplify the analysis process, as too many measures might confuse the users.

To explain the measures, let's consider the equation X $\rightarrow$ Y, where X is antecedent and Y is consequent.

Formally, the support of an itemset X is the ratio of the number of transactions that contains the itemset X ($|d_X|$) to the total number of transactions ($|D|$). It is calculated by the formula,

$$support(X) = \frac{|d_X|}{|D|} \tag{2.1}$$

The support measure [3] in this paper consists of both antecedent support and consequent support. The "antecedent support" is the number of times X has appeared in the database whereas the "consequent support" deals with the frequency of Y. Support is the ratio of transactions having both X and Y ($|d_X| \cap |d_Y|$) to the total number of transactions in the database. If a rule has the support of 10% then it means 10% of the total rules contains ($X \cup Y$). It is formulated as

$$support(X \rightarrow Y) = \frac{|d_X| \cap |d_Y|}{|D|} \tag{2.2}$$

Confidence [3] is the ratio of number of transactions having both X and Y ($support($ $X \rightarrow Y)$) to the number of transactions that have X ($support(X)$). Confidence is how many times if X has appeared in a database, Y has appeared as well. The strength of the rule depends upon this measure. If a rule has confidence of 90% then it means that 90% of the transactions containing X also contains Y. The confidence can be calculated using the formula,

$$confidence(X \rightarrow Y) = \frac{support(X \rightarrow Y)}{support(X)} = \frac{|d_X| \cap |d_Y|}{|d_X|} \tag{2.3}$$

Confidence measures the probability of antecedent and consequent occurring together but does not consider the correlation between the items. For example consider a rule, $Beer \rightarrow stamps$ has the confidence of 75%. Regardless of the items Beer or stamps being related or not, the confidence of these items occurring together will still be 75%. To overcome this issue, the lift and conviction measure can be used [11].

Lift helps to find how often X and Y appear together if they are statistically independent. If X and Y are totally independent, the lift will be 1. Lift is the

correlation measure and is calculated using the formula,

$$\begin{aligned} lift(X \rightarrow Y) &= \frac{Confidence(X \rightarrow Y)}{support(Y)} \\ &= \frac{support(X \rightarrow Y)}{support(X) * support(Y)} \end{aligned} \tag{2.4}$$

If the lift is greater than 1, it indicates the dependence of the items and can be used in the prediction of consequent in unseen cases. If the lift is less than 1, it indicated that items are negatively dependent.

Conviction compares the probability that X appears without Y if they were dependent on the actual frequency of the appearance of X without Y. It can be formulated as,

$$conviction(X \rightarrow Y) = \frac{1 - support(Y)}{1 - confidence(X \rightarrow Y)} \tag{2.5}$$

The value lies between $[1, \infty]$. If the value is 1, the items are independent. If Y is highly dependent on the X, the conviction will be ∞.

Another measure that incorporates the relation between antecedent and consequent is leverage [33]. Leverage is the difference between the occurrence of items X and Y together and the occurrence of X and Y if they were independent. It can be calculated by the formula,

$$\begin{aligned} leverage(X \rightarrow Y) = support(X \rightarrow Y) - \\ (support(X) \times support(Y)) \end{aligned} \tag{2.6}$$

The value of leverage lies between -1 and 1. If the value is 0, antecedent and consequent are independent. A high positive value suggests a strong positive relationship between antecedent and consequent while a negative value suggests a negative relationship.

Chapter 3

Literature Review

The visualization process when it comes to data science or big data is divided into three categories, which helps in explaining the whole working of the model to the user with ease [29, 14, 34]. The first category is to visualize the preprocessed data to understand any patterns that can be found. The second category provides visual support for users to understand the whole model created using the data. The user involvement in the model processing is done by giving them the chance to add some interesting measures (confidence and support threshold) before getting the results. The third category, which we will also be focusing on in this work, is to visualize the result. Most of the research papers in this area provide visualization of the association rules using few common techniques: (i) Table-Based Visualization Techniques, (ii) Graph-Based Visualization Techniques, (iii) Matrix-Based Visualization Techniques, (iv) 3-Dimensional Visualization Techniques. These techniques are also summarised and compared in the work of [18].

3.1 Table-Based Visualization Techniques

The table-based technique is one of the simplest methods to display the rules. The table headers consist of the rule name and the interesting measures. The cells corresponding to the headers being the textual rule and the measure values respectively. When integrated with other visualizations like clustering, the tables can be used to explore the rules [46, 44] where they use the table as a secondary visualization. The working system proposed by them can be seen in Figure 3.1. Both works incorporate a table to show the rules having the same antecedent and consequent alongside displaying different measures.

Chunsheng and Yan [47] uses a tabular method to display rules based on natural language. The tabular format in the work of Chunsheng and Yan is used to incorporate the results generated using natural language processing and apriori algorithm.

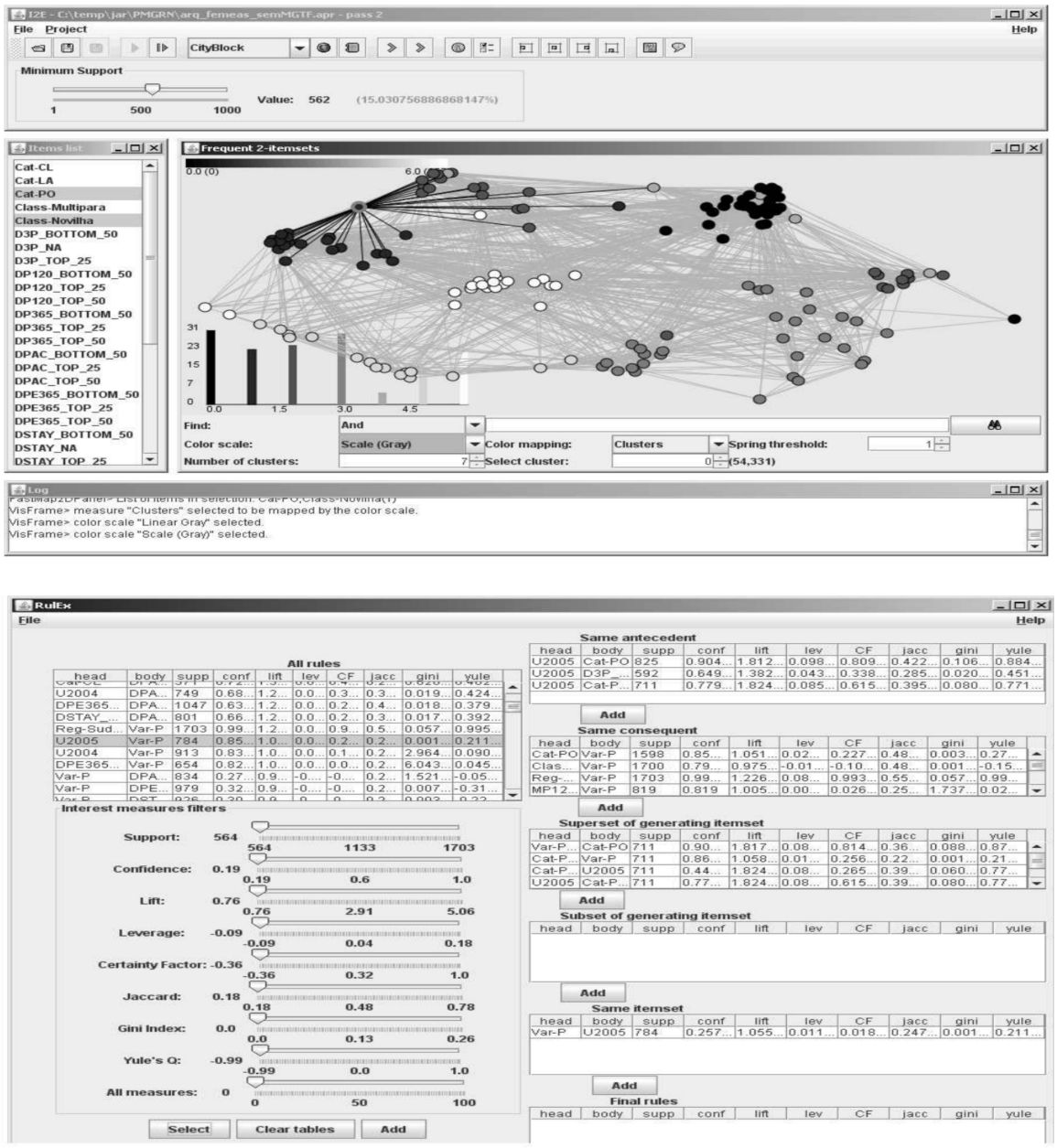

Figure 3.1: I_2E system proposed by [46]

Limitations

The Table form is easy to utilize when the dataset and the rules are less in number but gets time-consuming and difficult to interpret when the volume is high. It also affects the cognitive load and, comparison between rules becomes laborious. Being a textual-tabular technique, it can not handle different measures together. For instance, ordering the rules based on different measures or the display of more information in

a confined space becomes difficult.

3.2 Graph-Based Visualization Techniques

This technique represents the rules graphically. The items are represented using nodes of a graph, while the rules are displayed using the connections between these nodes [43]. Yamamoto and Oliveira [45] uses Hooke's Law to show the representation as springs and the force between nodes to represent the frequency of the items occurring together. The clusters generated in [46, 44] are also represented using graphs to show items as nodes and connections to show interesting measures. The work of [26] uses the method as a rule graph to display all the rules together along with the bar charts to browse the rules, while visualization for alarm association [42] uses scatterplot as a visual element to assist the alarm association rules. AssocExplorer [28] uses scatterplot (see figure 3.2) to show the items using the attributes highlighted by colors and the x-axis and y-axis being confidence and support.

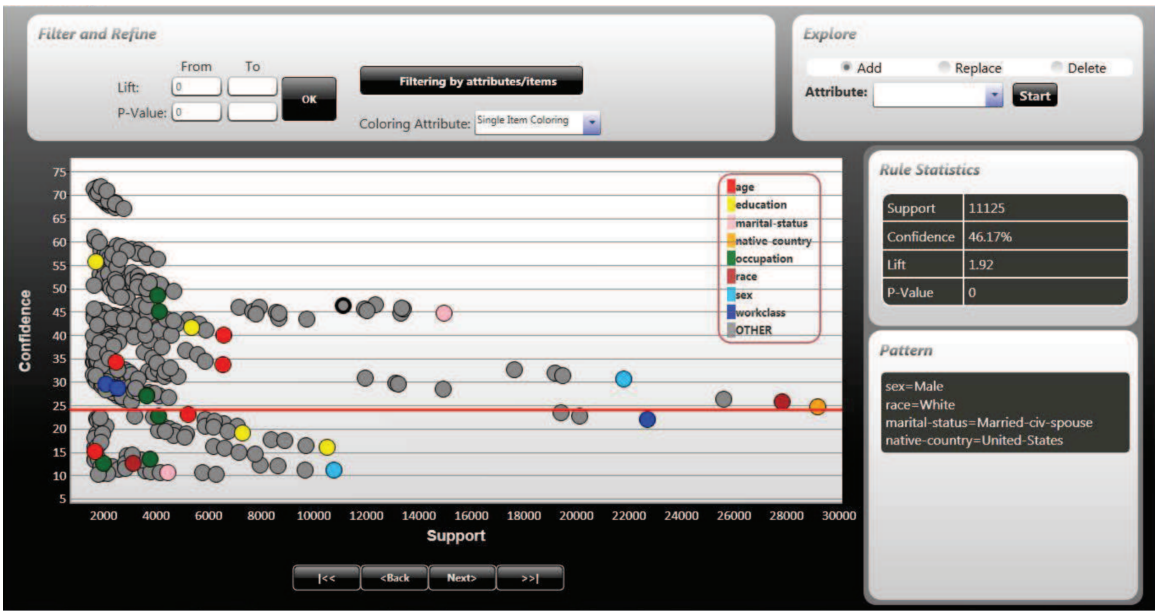

Figure 3.2: AssocExplorer by [28]

CrystalClear [30] uses the tree structure with matrix-based visualization to show a hierarchical view. Castillo-Rojas, Peralta, and Meneses [12] uses hierarchical structure as primary visualization and scatterplot as detailed visual elements. The mosaic

plot helps in the visualization of multi-item association rules using double-decker plots[22], which can be seen in figure 3.3.

Limitations

The main drawback is that when the rule number is large, some rules can coincide with one another, giving fewer rules than expected which results in loss of authenticity of the visualization. The node-link-based graph visualization can result in inter-twinned connections which gets difficult to interpret and the mosaic plots get difficult to analyze as the items increase. Our system helps to visualize all the rules without making them difficult to understand and preserving the authenticity of the rule subsets.

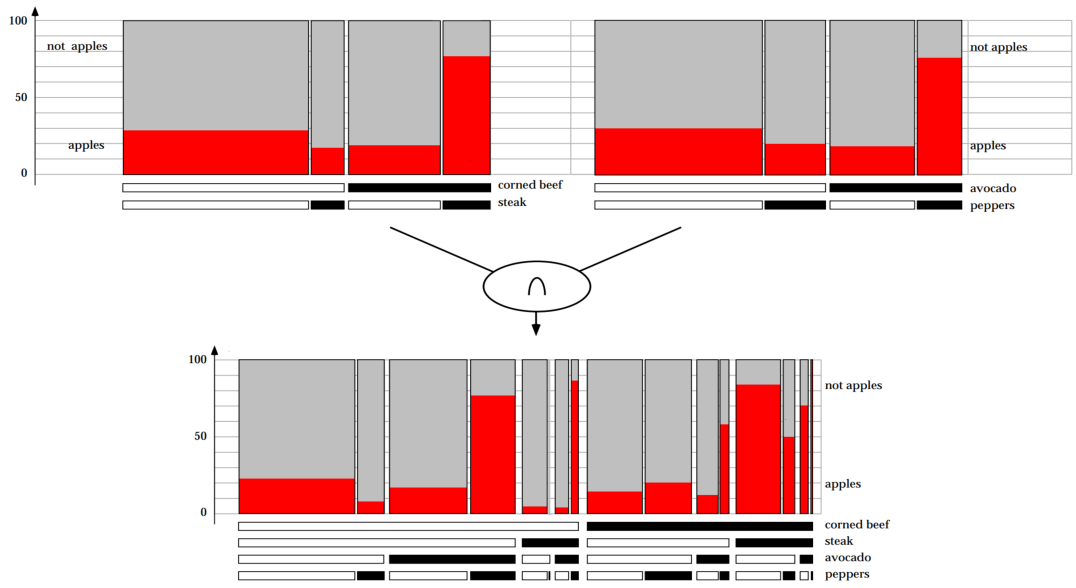

Figure 3.3: Mosaic plots to show the association rules by [22]

3.3 Matrix-Based Visualization Techniques

Matrix-based is used to display the rules to the user without much loss of the information. The method is simple to implement and to understand but also gets difficult to interpret with an increase in the number of rules. This method works for item-to-item [15] (figure 3.4) where the headers of each row and column consist of names of

the items representing the itemsets of antecedent and consequent respectively, while the cells are colored using a measured value for the rule. Item-to-rule matrix [13, 32] uses the rows to represent items and the columns to represent rules.

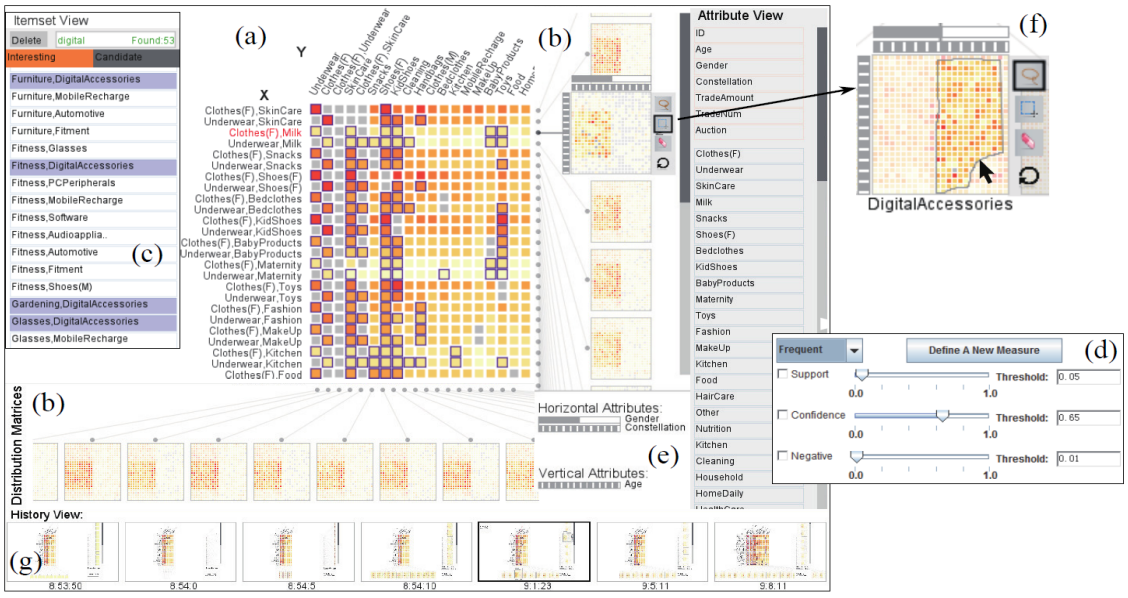

Figure 3.4: Matrix based method proposed by [15]

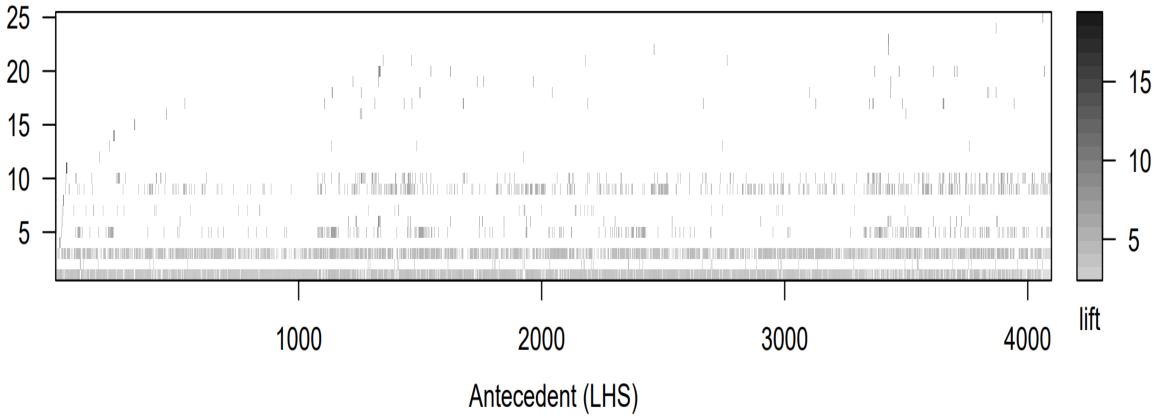

Figure 3.5: 2D matrix approach by [20]

CrystalClear [30] uses a matrix to show support as x-axis and confidence as y-axis with highlighting changes to the rules over time using color and icons. Hahsler and

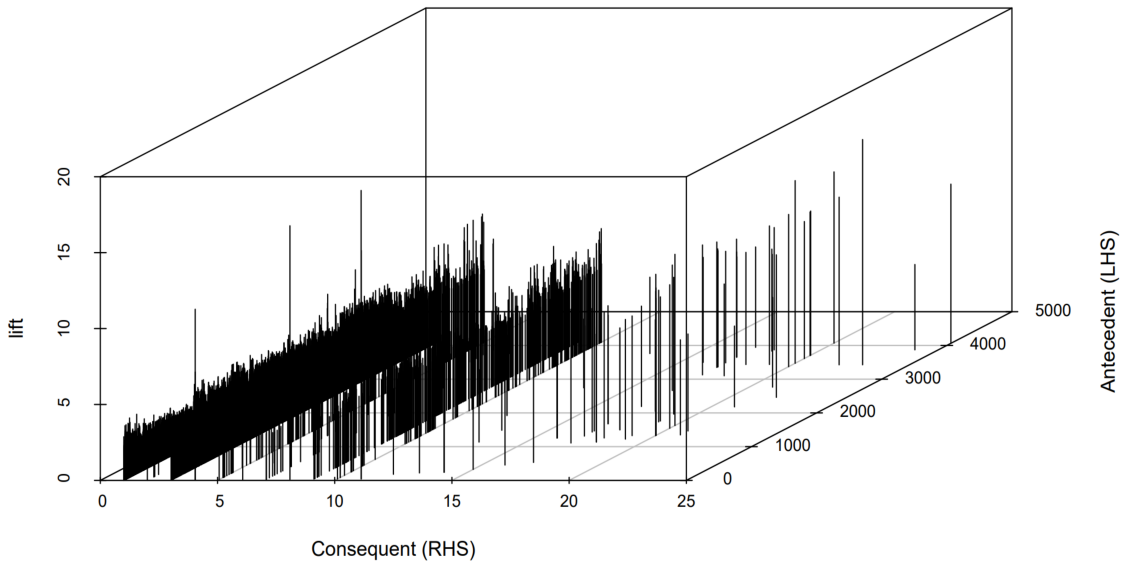

Figure 3.6: 3D matrix approach by [20]

Chelluboina [20] utilize the 2D matrix and 3D matrix to show antecedent and consequent relations using color to show different rules based on the lift as seen in figures 3.5 and 3.6. They also utilized grouped matrices created using clustering to show all the rules together.

Limitations

The item-to-item matrix visualization does not incorporate a way to deal with repeated rules and the visualization becomes too dense to handle. The item-to-rule matrix is in 3-Dimension and these matrices tend to create confusion while interpreting them. All these systems fail to incorporate more interesting measures and the grouped matrices do not always work with more than one consequent per rule.

3.4 3D Visualization Techniques

With the advancement in the visualization field, people have also tried to solve this problem using 3D plots [9, 37, 31] in which, rules are displayed with extra information using the third dimension. When it comes to user satisfaction, these plots lag as it might be tough to understand the results, or they might easily be misinterpreted [39]. The core concept of this type of visualization uses any one of the familiar techniques

mentioned above and tries to visualize the rules with better information. The work done by [17] uses a matrix to represent the clusters of rules and then provides support using 3D bar charts.

The Visualizing Association Rules for Text Mining [32] (seen in figure 3.7), and Visualization of Association Rules Over Relational DBMS [13] uses the means of matrix-based visualization along with 3D visualization to show the rules in the rule-to-item base. Even though the later paper covers the filtering of rules, it is difficult for a general user with no background in computer science to understand the system in a few try and the filters are added before generating the rules and not for the display of results. Both the paper lacks in providing the user a chance to compare the rules and also store the filters in any possible manner, whereas the former research papers fail to provide any other functionality. They fail to consider other interesting measures such as antecedent support, lift, etc.

Limitations

The visualization using these techniques does not deliver the user with the chance to do any formatting to the result, which leads to a lack of user interaction. The 3-Dimensional visualization requires intensive interaction to understand the results which can get overwhelming with time. If not provided with proper resources for visualizing 3D objects, the user might have to use their imaginations to visualize and understand the result [7].

3.5 Conclusion

In this chapter, we saw the potential drawbacks of the existing techniques and the need for a new technique. All the mentioned limitations are addressed in our system by using a 2-Dimensional matrix to represent item-to-rule which helps to display all the rules and allow the user to understand the rules, compare them, filter the rules based on items and store them for later if they wish.

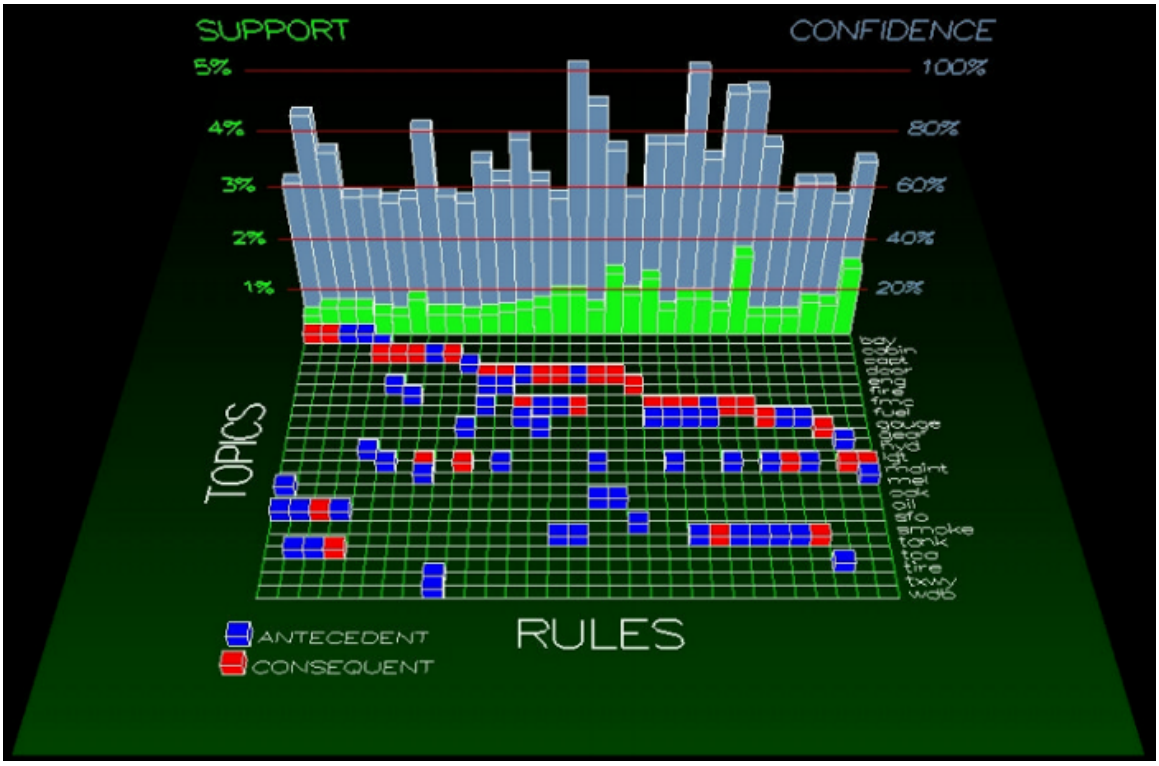

Figure 3.7: 3D item-to-rule matrix approach by [32]

Chapter 4

Methodology

In this chapter, we present the proposed system for the visualization and analysis of association rules. The sections presented cover the design guidelines followed for implementation of the technique, the overview of the *ARMatrix* system, functionalities provided to meet the design goals, and the in-depth workflow of the system.

4.1 Guidelines Considered Before Designing

This section presents design guidelines and functionalities developed by considering the techniques and the limitations explained in chapter 3. These design goals are listed below:

G1: Visual assistance for large rule counts with ease of understanding. The generated association rules can number from tens to thousands and as the number increases, it becomes difficult to analyze [19]. Based on previous literature [28, 13, 20, 46, 37], we know the problem in visualizing large rule sets and some possible solutions. Our primary design goal is to provide a better means to visualize all the rules regardless of the rule count and provide additional information about the rules. Understandably, with the increase in rule count, the analysis becomes a bit confusing to interpret [18]. The main purpose of our visualization technique is to reduce the cognitive load and provide more details [24]. For example, consider associations with a large rule count and a high number of unique items in the rules. It is necessary to provide the correct details about the rule to the user without losing the essence of the visualization.

G2: Comparative Analysis at the local level by user's interest. Based on the background about interesting measures, we know that all the rules can be differentiated based on these measures. There is a need for the rules to be compared based on interesting measures to find the rules that best suit the user's requirement. Previously, efforts have been made[37, 28, 26] to allow users to get a comparison

between the rules. Our goal is to follow the same path and provide a means for comparative analysis and find the important rule which is of user's interest. For example, Consider two rules $Orange, Milk \rightarrow Brush$ and $Orange, Milk \rightarrow Juice$ having same *confidence* value but different *lift* value. The rule with Brush as a consequent might have mostly been purchased together with Orange and Milk, but Juice is highly correlated to them as they all serve as *"breakfast items"*. By analyzing their lift values user can find the correlation between the antecedent and consequent to confirm the case, resulting in better profitability for the user.

G3: Focused Analysis of the rules-based on user's domain knowledge. Based on related work, we know that the association rules tend to get large in number. It is necessary to focus on the rules of interest from all the calculated associations, which helps in the reduction of the analysis time and lets the user extract only the needed information [15]. Providing a visualization for the analysis of these rules means allowing the user to explore the rules with freedom without much complexity. We aim to provide a way to study or analyze specific rules based on user's knowledge to get more focused insights on a domain rather than dealing with all the rules at the same time. For example, the superstore dataset generates a large number of rules which consists of many different items. But the user wants to deal with the rules that contain only the dairy items, so based on his domain knowledge, the user should be able to restrict the rules and analyze the focused group (dairy) individually.

4.2 System Overview

To implement the above design goals and functionalities, we present *ARMatrix*, a visualization framework for representing association rules. The interface includes four sections, Ⓐ Input and order section to get the rules and rearrange the visualization. Ⓑ Matrix and comparison view help to show and compare the rules to one another. Ⓒ Filter section helps to find the subsets for analysis and Ⓓ history section helps in retrieve or deleting a saved state. The steps to use or interact with the system are presented in **Figure 4.1**. Initially, *ARMatrix* returns the rules that are generated using the default values of support and confidence, set based on the dataset. The rules are displayed in the form of a 2-Dimensional item-to-rule matrix which also shows the frequency of items being antecedent and consequent, with the conviction values on the

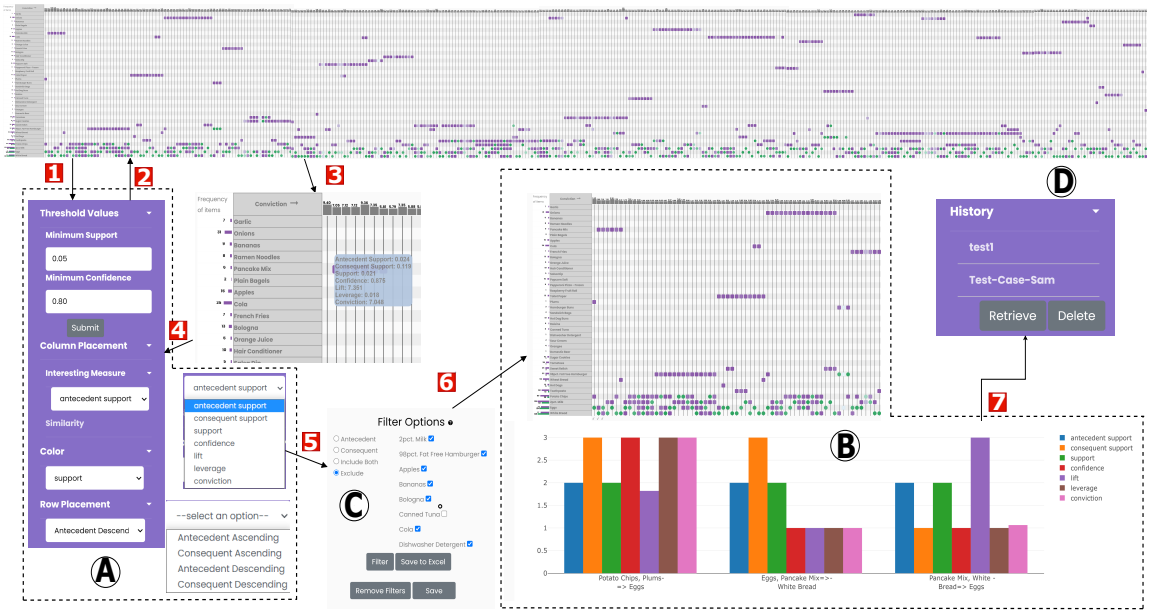

Figure 4.1: *ARMatrix* system overview and visualization. The letter indicates different modules. A) and C) are the input and order filters. B) Matrix and Comparison view, D) History tab to retrieve or delete the preserved states. The user starts with rules using default thresholds and sets the new thresholds using input section 1) and then is presented with all the rules 2). The interaction with the rules can done by simply hovering on to the items in the column to present more information. At anytime the order 4) of the visualization can be changed and the subsets can be found for analysis 5). Comparison of rules to each other can be done by double-clicking items and then selecting different measures 6). Once a desired state is reached, it can be saved and then be retrieved or deleted later.

top of each column. By setting the threshold values ❶, new rules can be generated and visualized at any point ❷. Hovering on the item highlights the row and column associated with that item alongside showing the interesting measures related to the column (rule) ❸. Then the user re-orders the rules based on different criteria and by selecting the appropriate options ❹. Once the user attains the correct order, the user can find subsets of the rule for specific analysis by checking all needed items and selecting the filter options (antecedent, consequent, include both, and exclude) ❺. By double-clicking the rules, they can be compared to one another based on different interesting measures ❻. The whole process can be iterated multiple times depending on the analysis requirements regardless of the sequence. Once the desired state is attained, the rules can be preserved for later use or access by saving the rule. The state can then be deleted or retrieved at any point by selections in the history section ❼.

ARMatrix is implemented as a website using python in the back-end [35] and D3 [10], plotly Javascript and Jquery as front-end. All the sections are explained further in the next sections.

4.3 Functionalities

To cover all the design goals and provide more useful features we created a system that helps in the visualization of association rules by incorporating the following functionalities:

F1: Attribute Selection. Similar to all the related work, the user is given a chance to set the minimum support and minimum confidence for the threshold to extract the rules and to change them interactively.

F2: Rule and Item Positioning and Coloring. The main focus of our system is to provide the user with the ability to analyze the rules based on different measures. For that, we use 2D matrix visualization with sorting of rules (columns) based on interesting measures using position and color. Another placement factor provided is to sort the rules by the similarity of items in the rules. Similar to the rule positioning, the items (rows) are sorted based on the number of times the item has occurred as antecedent or consequent in either increasing or decreasing order.

F3: Finding and Comparing the Rules of Interest. The rules are filtered

according to the items being present only in Antecedent, Only Consequent, Both, or Excluding items. The selected rules can be compared to one another based on interesting measures using a bar chart. The filtered rules can be stored and revisited at any point by the user to make the comparison of different filters easier. Also, the current rules can be extracted as a CSV file if necessary.

F4: Rule Details and Information. Detailed information about the rule can be provided by hovering onto the cell. And the rule can be displayed in the text format on click. The number of times an item has been antecedent and consequent is displayed on the left side of the item using a bar graph with antecedent as blue and consequent as orange. The conviction value for the rules is visualized using the histogram at the top of the table matrix.

4.4 Data Preprocessing

To derive the rules from the association rules model, the data should be in a transactional format. If not, we convert the data in the right format of transactions $d_1, d_2, ..., d_m$. This is done by transforming the data into transactions of binary items from the item set $I = \{i_1, i_2, ..., i_n\}$ so that each transaction $d \subseteq I$. Each transaction consists of values 1 or 0 depending upon the item being present or absent, respectively. For this conversion, we primarily get all the unique values from the database and create transactional data with each transaction having items as 1, if that row in the database has that element, and 0 if the element is absent. The data before and after processing can be seen in figure 4.2 and 4.3 respectively.

	age	gender	Chest pain	RBP	serum cholestoral	Blood Sugar>120mg/dl	ECG results	Max heart rate	Exercise	Oldpeak	ST segment	Vessels	Thal	Disease
0	70.0	1.0	4.0	130.0	322.0	0.0	2.0	109.0	0.0	2.4	2.0	3.0	3.0	2
1	67.0	0.0	3.0	115.0	564.0	0.0	2.0	160.0	0.0	1.6	2.0	0.0	7.0	1
2	57.0	1.0	2.0	124.0	261.0	0.0	0.0	141.0	0.0	0.3	1.0	0.0	7.0	2
3	64.0	1.0	4.0	128.0	263.0	0.0	0.0	105.0	1.0	0.2	2.0	1.0	7.0	1
4	74.0	0.0	2.0	120.0	269.0	0.0	2.0	121.0	1.0	0.2	1.0	1.0	3.0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
265	52.0	1.0	3.0	172.0	199.0	1.0	0.0	162.0	0.0	0.5	1.0	0.0	7.0	1
266	44.0	1.0	2.0	120.0	263.0	0.0	0.0	173.0	0.0	0.0	1.0	0.0	7.0	1
267	56.0	0.0	2.0	140.0	294.0	0.0	2.0	153.0	0.0	1.3	2.0	0.0	3.0	1
268	57.0	1.0	4.0	140.0	192.0	0.0	0.0	148.0	0.0	0.4	2.0	0.0	6.0	1
269	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2

270 rows × 14 columns

Figure 4.2: Data before preprocessing

	age=70.0	gender=Male	Chest pain=4.0	RBP=130.0	serum cholestoral=322.0	Blood Sugar>120mg/dl=False	ECG results=2.0	Max heart rate=109.0	Exercise=No	Oldpeak=2.4	...	Max heart rate=123.0
0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	0.0
1	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	...	0.0
2	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	...	0.0
3	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0
4	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	...	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
265	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0
266	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	...	0.0
267	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	...	0.0
268	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	...	0.0
269	0.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	...	0.0

270 rows × 386 columns

Figure 4.3: Data after preprocessing

4.5 Rule Extraction

The main aim of this paper is to visualize the results of association rules. To do that, we are generating the rules using the fpgrowth algorithm in the mlxtend library in python [35]. The model uses minimum support and minimum confidence as threshold measures. FPgrowth algorithm is selected as it is a time-efficient process and tends to work faster than the apriori algorithm as it scans the database only twice to extract and confirm the rules. The fpgrowth algorithm [21] works on the divide-and-conquer principle and for us, it extracts the same number of rules as apriori in less time. The threshold values are either set by default for each dataset or taken by the user to calculate the rules to be displayed (**F1**). The model works in a two-part process, a) the frequent itemsets are generated using the data and minimum support as a measure. b) Providing the frequent itemsets as a parameter and setting *confidence* as a metric measure to get the rules for visualization. The process filters out the frequent rules (*using support*) with a probability of the rule actually being true (*using confidence*).

4.6 Visual Assistance

4.6.1 Matrix and Comparison View

The visualization of the rules (**G1**) is presented using a 2-Dimensional item-to-rule matrix. Each row in the matrix represents an item $i_1, i_2, ..., i_n$, and the column represents a rule $X \rightarrow Y, (X, Y) \subseteq I$ which consists of both antecedent X and consequent

Y (left and right side of the rules, respectively). The items in a single rule are connected with a black line at the center of the column to reduce the cognitive load while examining a rule. The placement and the color of the cell or the rules are based on the interesting measures selected in the Ⓐ section from Figure 4.1(**F2, G1**). For each rule, the antecedent is represented by the rounded square and the consequent by the circle that can be seen in A from Figure 4.4.

The conviction measure for each rule is displayed as a grey-colored histogram on top. The conviction measure consists of values from $[1, \infty]$. It becomes difficult to scale it using color as the highest value that can be attained by a rule is ∞. While mounting the values to colors and bars, we found that the bar representation for these values helps to preserve the authenticity of this measure value and display the correct information to the user. The height of each bar is calculated by scaling the values of the measure between fixed numbers, where infinity values are considered 0 but are represented using the ∞ symbol to show the bars for only for the numeric value of the measure.

The frequency of each item is an antecedent and the consequent is calculated and scaled between 0 and a fixed number. This is done to make a scalable visualization while preserving the information it carries. These values are then displayed as bar graphs on the left side of each item using the scaled values as height and real frequency is displayed as text(**F4**), which can be seen in B of Figure 4.4. The purple color bars represent antecedent frequency for that item and the green color represents consequent frequency. This item-to-rule matrix helps to display a large number of rules on the screen without much cognitive load.

Each rule that needs to be compared to one another(**G2**) can be selected by double-clicking the items in the column that represents the rule.

Once selected, a tick appears on the bottom of the selected column for indication. The chosen rules are then compared to one another using bar graphs. Each bar represents an interesting measure while the group of bars represents a rule. The data for the height of the bars is normalized to prevent a large gap between the heights. This is done as different interesting measures have a different range of supportable values and non-normalized data is difficult to interpret. The original value of the measures is attained upon hovering onto the object. When dealing with thousands of

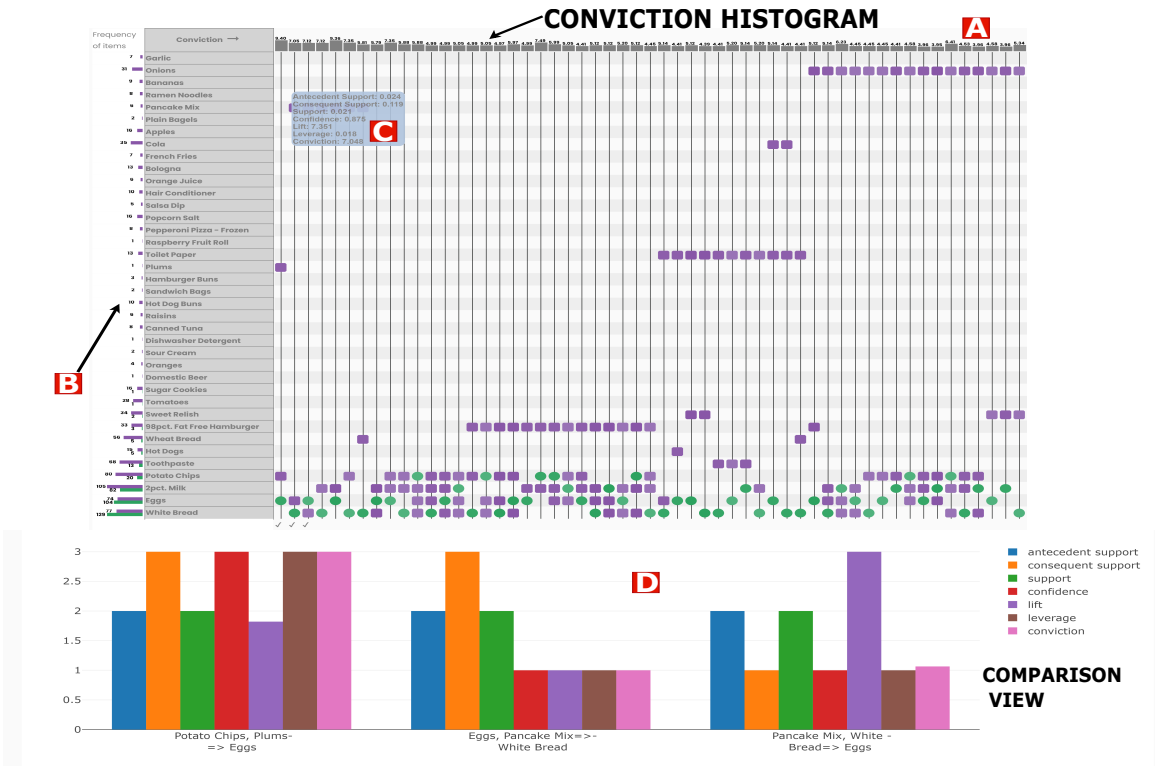

Figure 4.4: A) Matrix view; B) Bar graphs showing the frequency of each items being antecedent (purple) and consequent (green); C) Tool tip on hover on items showing all the interesting measures for that rule; D) Comparison view

rules, it becomes difficult to keep the track of interesting measures of each rule. Some rules can only be differentiated based on a few interesting measures meaning, consider 10 rules from a set of 1000 rules having the same support and confidence value but different values for other interesting measures (*e.g., lift, conviction, leverage*). To avoid any confusion or to provide a means to examine them together, the user can select all of these 10 rules together for analysis. These features help in examining and comparing multiple rules together based on their measure values which can be seen in the comparison view section of Figure 4.4(**F3**).

4.6.2 Rule Ordering and Filtering

At first, all the generated rules are placed based on the increasing value of their 'antecedent support' from left to right and the color is scaled to fit the 'consequent support' values of each rule (**G2**). The antecedent items of each rule are represented using rounded squares filled with variations of purple color whereas circles with green

color variations represent consequent items. The user is asked to select the measure value for ordering the rules using placement and color(**F2**). Upon selecting the interesting measures for column placement, a selection option is provided to set the measure. The selected measure is then used to sort the rules based on the increasing order of that rule's measure and then rearranged in the matrix.

Another column ordering functionality provided is based on the similarity. On selecting 'Similarity', the rules are primarily converted into a list of lists filled with the one-hot encoded value of the items. Then they are utilized to generate a sparse matrix $A_{N\times N}$ where N is the number of rules. Each element A_{ij} consisting of distance between the rules *i and j*. Having binary data, we focus on finding a distance measure that can accommodate binary elements and extract the proximity distance. After reviewing the literature about the distance and similarity measures for binary data, we selected *jaccard distance* as a measure for similarity [16]. The rules are clustered together using hierarchical clustering which uses a sparse matrix as the distance measure and the cluster numbers are then assigned to each rule. The rules are then rearranged based on the increasing values of the cluster number to put similar rules together.

To provide an additional layer of details to differentiate the rules based on their measure value, the rules are displayed with different shades of color depending on the value of the interesting measure which is selected by the user(**G3**). On selecting the interesting factor, the chosen measure of each rule is scaled to fit the color scheme of purple for antecedent and green for consequent. The darkest shade of the color represents the maximum value of the measure, and the lightest shade represents the minimum value. The *"conviction"* measure contains a number ranging $[1, \infty]$, value which makes the color scaling difficult. The scaling method treats ∞ as the lowest number and assigns the lightest shade of color, which results in providing wrong information. Therefore, *"conviction"* is not provided as a measure to color the columns.

Each item contains a bar graph representing the frequency of it being an antecedent and consequent on the left side. The rows or items can also be ordered based on their frequency by selecting from the options. *'Antecedent Ascending'* and *'Antecedent Descending'* rearrange the rows by increasing and decreasing the order of

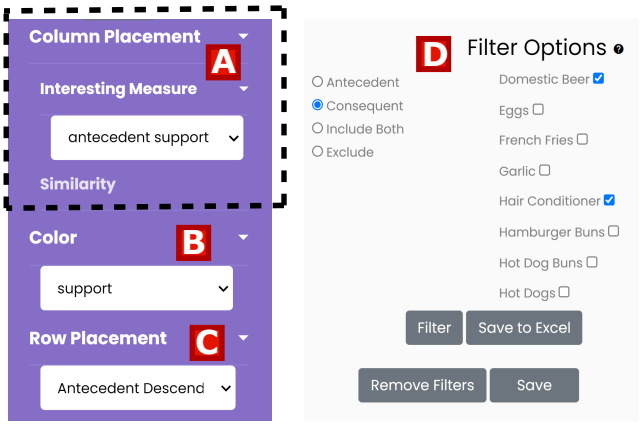

Figure 4.5: A) Columns Placement options of Interesing Measure and Similarity allows the columns to be rearranged according to selection; B) Column Color options to change the shade of color for each rule based on selected measure; C) Rows can be rearranged by selecting Row Placement options; D) Filter options to get a subset of rules based on selected items and condition

the frequency of the item being an antecedent respectively. *'Consequent Ascending'* orders the rows by increasing frequency of items being consequent and *'Consequent Descending'* order by decreasing frequency(**F4**).

The filtering of the rules(**G3**) is done by selecting all the items that are needed to be filtered using the checkbox, then the measure to find the subset is selected. On selecting *'Antecedent'*, all the rules that contain selected items as only antecedent items will be displayed. If *'Consequent'* is selected, the rules that contain items as the only consequent will be displayed. *'Include Both'* returns the rules where the checked items are present. For getting the rules which do not contain the selected items, *'Exclude'* filter is used. The generated rules can have all the checked items or their subsets. These filters help to visualize the necessary subsets of the rules for analysis without losing the authenticity of the rule measures(**F3**). This helps to limit the number of rules and also to gain relevant information about the subsets. For example, consider a supermarket dataset, the analyst wants to check the rules containing only vegetables. So the necessary subsets related to vegetables can be obtained from total rules by applying filters. All the items are considered while calculating rules so that no item is excluded for rule calculation, and filters are applied only for visualization of the rules.

All the ordering and filtering options can be removed by simply clicking the "Remove Filters" button which brings the rules back to its original state.

4.6.3 History Management

In the whole system, the freedom of filtering is of immense priority, but it is easy to forget the rules or filters. So this section helps the user to store the selected filters and review them when needed(**F3**). The filters are preserved by clicking on the save button provided on the bottom of the filter options that as seen in section D of Figure 4.5.

Once a state is saved, it is displayed in the row-wise list format, and it can either be retrieved or deleted by clicking on the name first then on the respective buttons as shown in Figure 4.6. Another small functionality added to this section is for the user to get the current rules in a CSV format by clicking on the "Save to Excel" button which is seen in Figure 4.5. This functionality helps to generate an offline file with the necessary subsets or rules that can be useful when the system might not be accessible.

Figure 4.6: History section helps to choose from previously saved selection state and then either retrieve the rules or delete them.

4.6.4 Detailed Information

The rules can be sort using any two interesting measures, but the user can attain more information regarding the particular rule by hovering on the items in that column. Upon hovering, a tooltip is generated for that column. The tooltip consists of all the interesting measures values associated with that rule.

By clicking any item on the column, the entire rule is seen in textual form to avoid any confusion(**F4**) and acknowledge the correct rule. This functionality is helpful

when the rules contain too many items. Also upon hovering, that row and column is highlighted showing all the items in that column and all the columns having that item.

Chapter 5

Results

In this chapter, we evaluate the effectiveness of the tool using two user scenarios that illustrate the usability of our system in understanding and analyzing Association rules. We also present a User study to measure the efficiency of the tool in real-time practice.

5.1 Use Case 1: Market Data set

The most popular example to explain the use of association rules has been the market basket dataset. The database used for this user scenario [1] consists of 1361 transactions and 255 different items. When set to default threshold of support and confidence, 0.02 and 0.7(**F1**) respectively, which can be seen in figure 5.1 (A). We found around 364 rules with 38 unique items. The user of the system can be an analyst as well as a shop owner with a little knowledge about the interesting measures. All the rules generated can be seen in (Figure 5.1 (B)).

For this use case, we have Sam, a shopkeeper who is dealing with a shortage of some items in his store. He plans to use our system to re-shelf the aisles of his store along with analyzing to determine the implications of out-of-stock items on the other items sold together with them.

So he starts his analysis by looking at the rules which are arranged in each column and the items in each row (**G1**). He then hovers upon each rule to get all the interesting measure values for that particular rule and also checks the frequency of each item being an antecedent and consequent (**F4**). For the ease of understanding the rules and analysis, the user rearranges the rows of the matrix based on decreasing frequency of items being an antecedent among the total number of rules. He does this by clicking the row placement dropdown list on the left and selecting *"Antecedent Descending"* as a measure, which places items having a higher frequency as antecedent on top and low-frequency ones on the bottom (**F2, G2**).

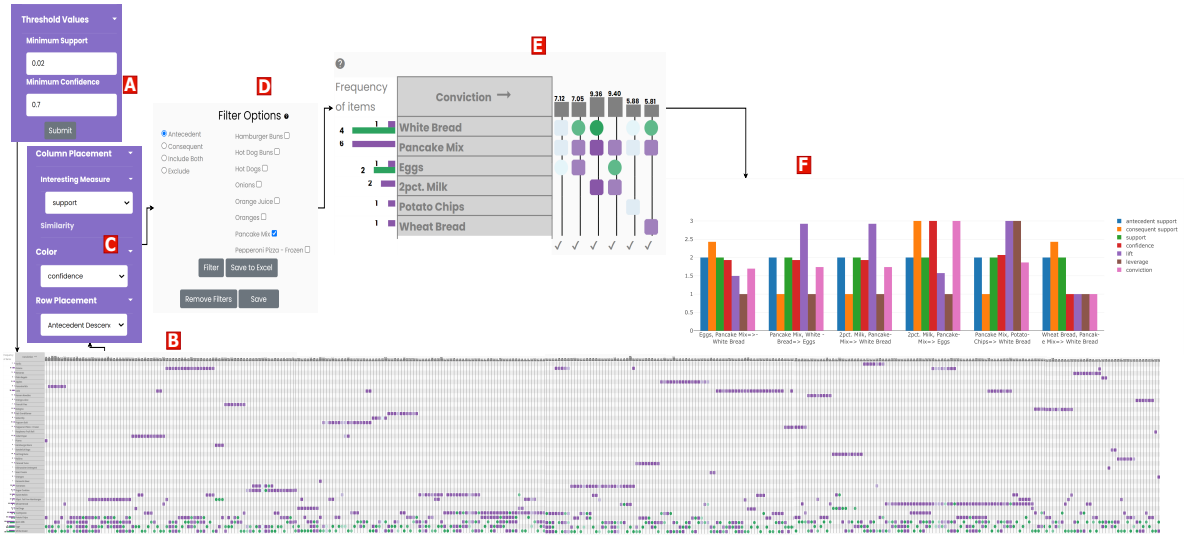

Figure 5.1: ARMatrix visualization with *market-basket* dataset. A) The user sets the threshold values and clicks submit to generate the rules; B)The 364 rules are displayed to the user; C) The user selects the column placement as support, color measure as confidence and row placement measure as Antecedent Descending; D) User filters the rules to get the desired rule and can also save the rules to csv file or save in memory for later use; E) User is presented with the filtered rule; F) The user can select the items in columns to compare based on interesting measures and is then presented with the bar graph

Further, Sam sets the placement of the columns as an interesting measure and selects support to rearrange the rules. He finds that *Orange, Eggs → 2pct. Milk* has the highest support by scrolling to the right end of the matrix. To further analyze the rules by adding another layer of detail, he sets the color of the matrix to be adjusted based on the confidence measure as can be seen in Figure 5.1C. While analyzing the results, he found that even though the rule had high support, it does not have high confidence, meaning even though the items have occurred together frequently, the rule does not always turn out to be true(**F2**).

Sam saves the rules and filters for now as *"Sam-Store-Data"* to revisit after he has done his inventory. After getting a list of unavailable items, he retrieves the state from the history section and continues his analysis. Sam decides to fill up the breakfast items first and only has the place to put 4 different items. So he looks for the rules that contain breakfast items and notices that there are only six rules having pancake mix that can be seen in section E of Figure 5.1. He then adds antecedent as a filter with pancake as an item, to display only those 6 rules and conduct a focused analysis (Figure 5.1 D). Just for ease, to know all the elements in a particular rule, he clicks on the item in the column and gets the rule in 'Selected Rule' section on the right in a textual format(**F4**). Now he has 6 items in the rules but only 4 slots available, so double clicks on all the rules for comparing them using different interesting measures(**G2, F3**). In the graph filters section, he selects confidence as a measure and notices that *Pancake Mix, Eggs → White Bread* has the highest confidence and the rules having potato chips and wheat bread has somewhat low confidence. So he selects lift as the measure of comparison to find the correlation between items (Figure 5.1F). He observes that Pancake Mix, White Bread, and Eggs have higher lift values, so he decides to put these three items along with 2pct. Milk on the shelf as the rules containing them have higher confidence and lift than the chips and wheat bread. By doing so, Sam can attain maximum profit as these items are often sold together and have a higher correlation between them.

Once done with the re-shelving task, he removes all the filters and checks the unavailable item-list that includes
[Raisin, Toilet Paper, Plums, Bologna, Dishwasher Detergent, Hair Conditioner, and Cola]. To analyze the repercussion of absent items on other products, he selects all

Figure 5.2: Raisin, Toilet Paper, Plums, Bologna, Dishwasher Detergent, Hair Conditioner, and Cola are selected as items with Include Both as a filter to get the rules containing these items as both antecedent and consequent.

these items in the filter options using checkbox and sets *Include Both* as the filter which can be seen in Figure 5.2. After that, he clicks on the filter button to get the new rules that contain all the unavailable elements (**F3, G3**).

Sam finds out that the number of rules having Dishwasher Detergent, Plums and Raisins is less which results in less effect to the sales in the store. While Bologna, Toilet Paper, Hair Conditioner, and Cola (being one of the highly repeated items in the rules) are more frequently appeared in the rules, which means that these items being absent affects the sales of other products more greatly. To increase the sales margin and the profitability for his store, he orders more quantity of the highly frequent items.

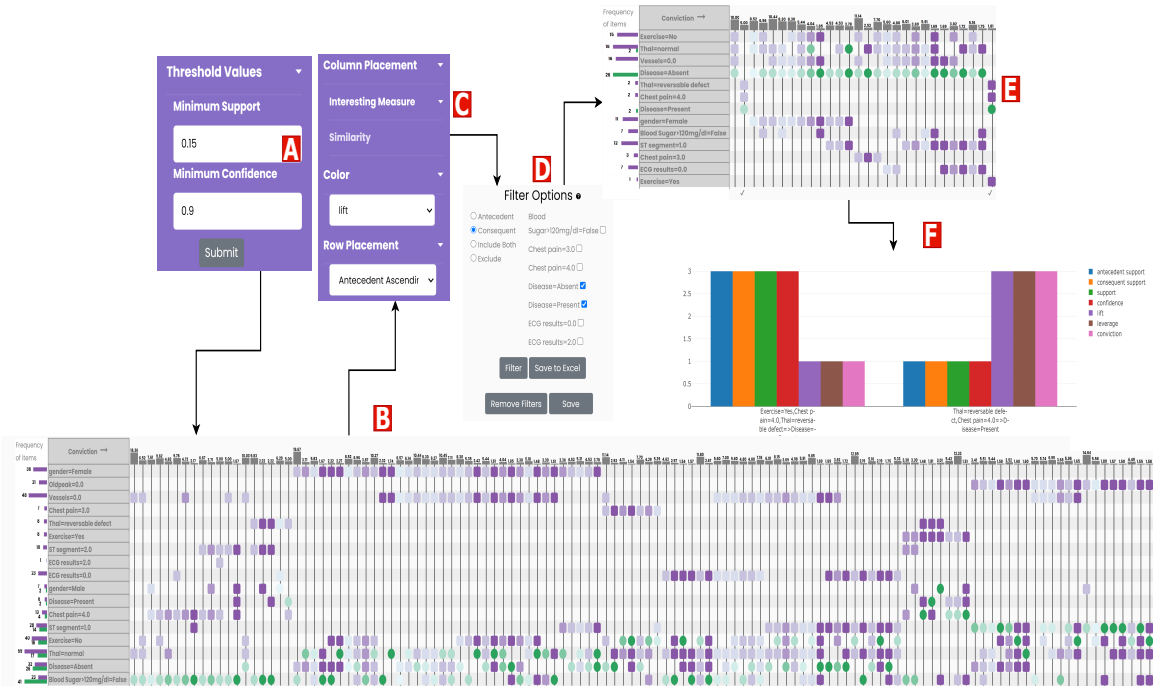

Figure 5.3: ARMatrix visualization with *Statlog-Heart* dataset. A) The user sets the threshold values and clicks submit to generate the rules; B) The 119 rules are displayed to the user; C) User filters the rules to get the desired rule and can also save the rules to csv file or save in memory for later use; D) The user selects the column placement as Similarity, color measure as lift and row placement measure as Antecedent Ascending; E) User is presented with the filtered rule; F) The user can select the items in columns to compare based on interesting measures and is then presented with the bar graph

5.2 Use Case 2: Heart/Medical Data set

Another application of association rules is in the health sector where relations between symptoms can help to identify the disease. For example, recently due to a pandemic, if the patients have symptoms such as fever, dry cough, tiredness, difficulty in breathing, to name a few, that might indicate a possible COVID case.

In this use case, the database used is Statlog Heart Data Set from UCI Machine Learning Repository [2] which is converted into a transactional dataset consisting of 270 transactions with items in the transaction as 1 or 0 based on the item being present or absent in each transaction respectively. Dr. Rae, a Cardiologist, wants to investigate the inter-association of the symptoms and their connection to the disease being present or absent as rules, so she sets the threshold values of support as 0.15

and confidence as 0.9 and gets 119 rules with 17 different items(**F1**) including both symptoms and disease, as can be seen, is Figure 5.3(A,B)(**G1**).

Dr. Rae decides to analyze only the disease being present or absent(**G3**) as a consequent, so he selects "Disease=Absent" and "Disease=Present" as items in the filter options using checkbox and sets consequent as a filter, which can be seen in Figure 5.3C. After clicking filter button the matrix gets updated with 28 rules(**F3**), 26 rules with disease absent and 2 rules with disease present, which can be checked by adding the consequent frequency of the consequent, *Disease=Absent* and *Disease=Present*.

To further analyze, Dr. Rae sets similarity as column placement and lift value as color(**F2**). She notices that there is one more item "Thal=normal" which acts as a consequent, so distinguish it from the disease items, row placement is set to Antecedent Ascending as can be seen in Figure 5.3D, to get the rules rearranged(**F3**).

Dr. Rae checks the rule Thal=normal, Chest pain=3.0, Exercise=No → Disease = Absent by clicking on the items in that column. While clicking, Dr. Rae hovers on to the column items revealing all the interesting measure values(**F4**). She notice that two rules with disease being present has similar symptoms but different lift values, indicating that the symptom *Exercise=Yes* results in negative correlation with the disease (Figure 5.3E).

To confirm the analysis, Dr, double clicks both the rules for comparison(**G2**), then checks all the interesting measures by selecting the graph filters (Figure 5.3F) and finds that the rule without exercise value has better interesting measure than that having exercise in it. So she concludes that it is necessary to examine exercise-induced angina to make the judgment for the disease. She then saves the filters for further analysis using the 'Save' button with the name "Disease-consequent"(**F3**).

5.3 User Study

5.3.1 Study Population and Tasks

A study was conducted to evaluate the effectiveness of the *ARMatrix* technique for the visualization of association rules. The main aim of the study was to analyze the ease-of-use and to know if the users were able to get meaningful insights about the

rules while using the system. The application for the conduction of the study was sent to the Research Ethics Board for review. The user study was remotely supervised and conducted online.

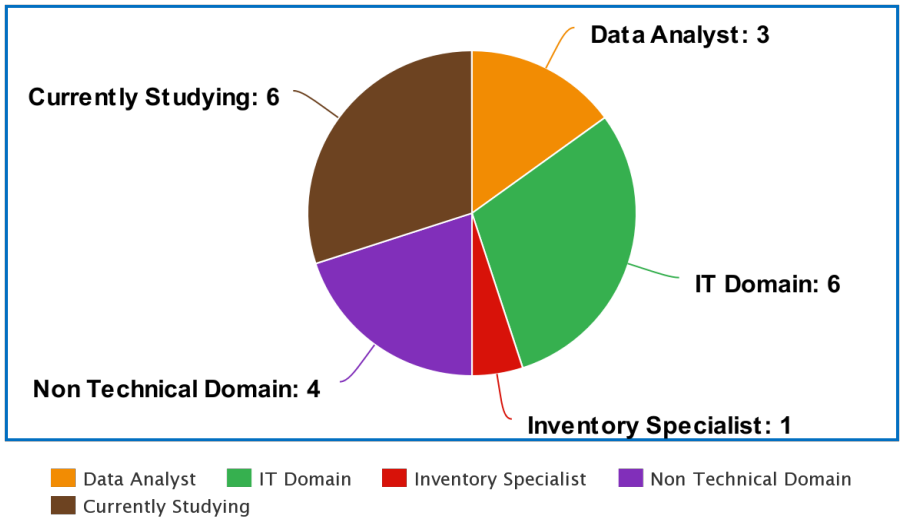

Figure 5.4: Job profiles of the participants

Twenty participants were selected for the study, out of which 3 people were working as a Data Analyst, 1 was an Inventory Specialist, 6 were working in an IT domain (Software Developer, Data Engineer, etc.), 4 were working in Non-IT Domain (restaurants or shops) as seen in figure 5.4. The rest of the participants involved graduate and doctoral students with good knowledge of association rules. Twelve of the total participants had a good understanding of Association rules and the remaining had very little knowledge. Seventeen people had an extensive grasp of the data analytics tools while the other 3 had neutral exposure. The participants were interacting with our matrix system for the first time but 14 of them had an understanding of matrix-based visualization (well to very well familiarity), the other 6 had neutral or not well familiarity. The summary of these questions can be seen in table 5.1.

Firstly, the users had to answer the demographic questionnaire which also covered their educational background, job profile, and their usage of visual representations and interactive user interfaces in daily life. The area of study for 65% participants was Computer Science, whereas the rest of the participants had their degrees in Information Technology and Internetworking (15% and 20% respectively). From the total population, around 55% participants had a university-level education. One person

Table 5.1: User results for familiarity based questions

Questions	Very Well	Well	Neutral	Not Well	Not well at all	I prefer not to answer
How familiar are you with Association Rules?	6	6	6	1	1	0
How familiar are you with Data Analytics Tools, such as Microsoft Excel or Tableau?	8	9	3	0	0	0
How familiar are you with matrix representation of data?	9	5	5	1	0	0

completed the Post-Doctoral and the rest 40% were post-graduates. The bar charts in figure 5.5 represent the numerical counts for the area of study and the highest level of completed education for the participants.

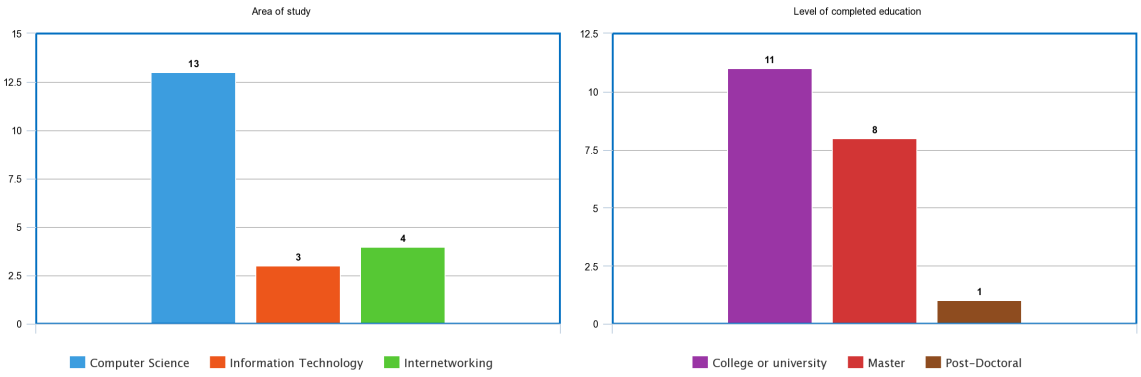

Figure 5.5: Area of study and level of education results from demographic questions

When it comes to interactive user interfaces, 95% of participants used them frequently where one participant didn't use them often. Eight of the participants extremely often used visual tools like ours, whereas 4 participants and 7 participants used them very often and often respectively. The detailed distribution of the participants based on usage of tools and interfaces can be seen in figure 5.6.

The participants were provided with a video tutorial about the overview of association rules `https://youtu.be/b0camSVo010`. The tutorial was to help the users brush up on the fundamentals of association rules and interesting measures. Another video was presented to explain the working of our *ARMatrix* system `https://youtu.be/7wvWPZYZF7I`. All the users were supervised while they explored our

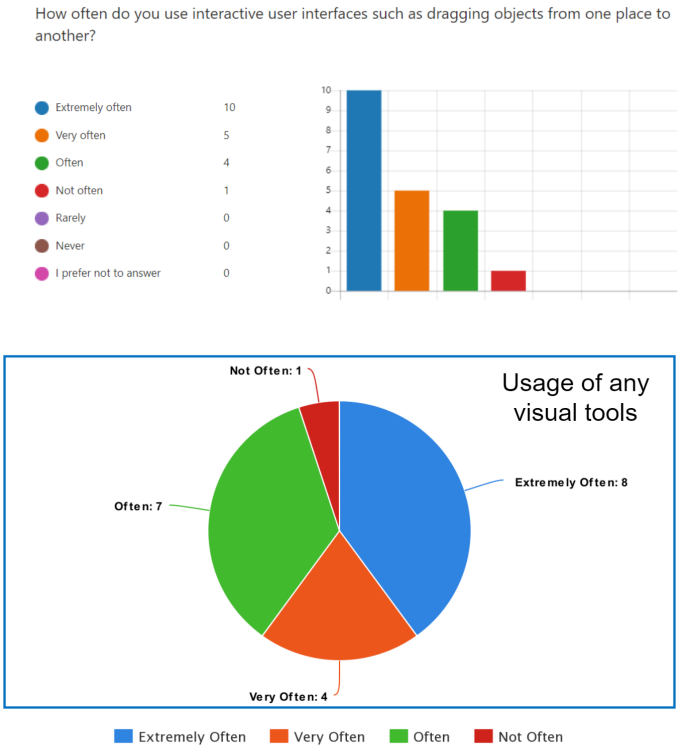

Figure 5.6: Results for the usage of interactive systems and visualization tools asked in demographic questionnaire

system freely and understood the working before answering questions in the study.

They were asked a set of quantitative and qualitative questions followed by two subjective questions to get overall feedback on the system and if they wanted to have any additional features. For the study, we used the market basket data and the user was asked to interact with the rules using the *ARMatrix* visual system and the textual rules in the form of an excel sheet.

The user study was designed for the participants to perform several tasks using the different functionalities of the system such as finding important subsets, compare different rules, differentiating them based on interesting measures, and answer the questions based on the tasks. After that, the users were presented with the sets of 5-point Likert scale questions to understand their perspective of the system's complexity and experience.

5.3.2 Study Results and Feedback

The average time taken by the participants to complete all the tasks along with qualitative and subjective questions was 32'10". The study has 7 task-based questions that cover the use of textual rules in the form of excel sheets (3 questions) and the use of the visual system to answer the questions (4 questions). The generalized type of questions asked to users is formalized in table 5.2.

	Goals	Questions
Task 1	Q2	Filter the rules with item i as antecedent and find the frequency of item j being a consequent
Task 2	Q2, Q3	Filter and compare the rules to find the rule with highest support and confidence
Task 3	Q1	What is the frequency of item i being an antecedent?

Table 5.2: The tasks, goals and the questions asked in the user study.

The participants were asked to execute the tasks mentioned above using both textual rules and the ARMatrix system and comparative analysis (seen in figure 5.7) was done for measuring the efficiency of the visual system in comparison to the textual rules. For Task 1, 65% of the participants selected the correct answer using textual rules, while 95% were able to answer correctly to the same task using our visual system. When it came to Task 2, only 55% of the participants selected the right answer in contrast to 95% correct answers using the ARMatrix system. Finally, in Task 3, 85% and 95% participants gave the correct answer using textual rules and the visual system respectively.

We observed that although more than 50% of the participants were able to answer correctly using both the textual rules and visual system, 100% of them said they were confident about the answers they got using the visual system. Also, all of the participants(100%) agreed that the approach of using a visual system to analyze and explore association rules was more efficient, which can be seen in figure 5.8. Also, they acknowledged that they were more confident about the answers they gave using the visual system.

The results for the 5-point Likert-scale-based evaluation to test the effectiveness of the functionalities and the usability of the visual system are summarized in the figure 5.9 & 5.10.

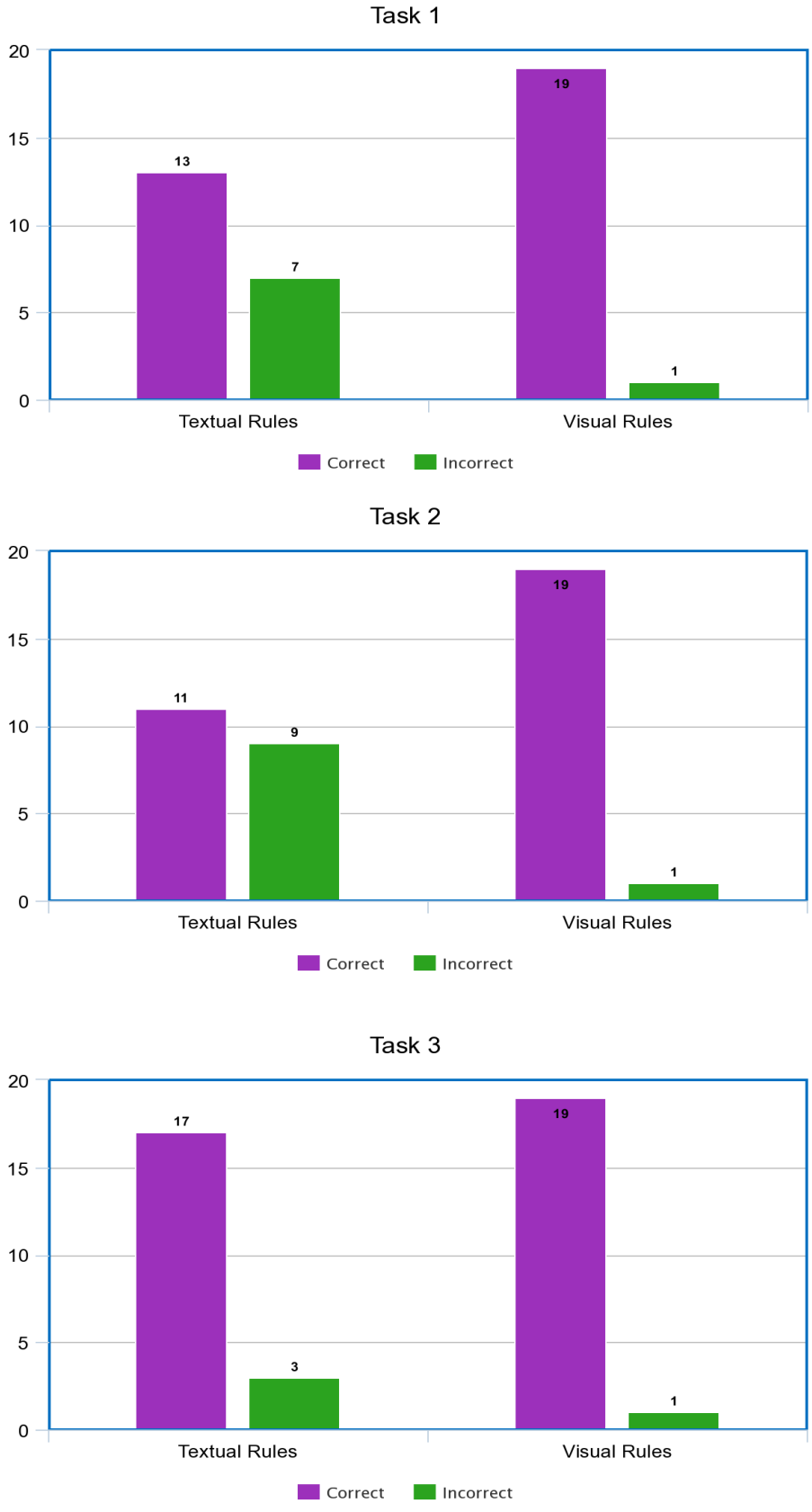

Figure 5.7: Results given by users for task 1 using both textual rules and visual system

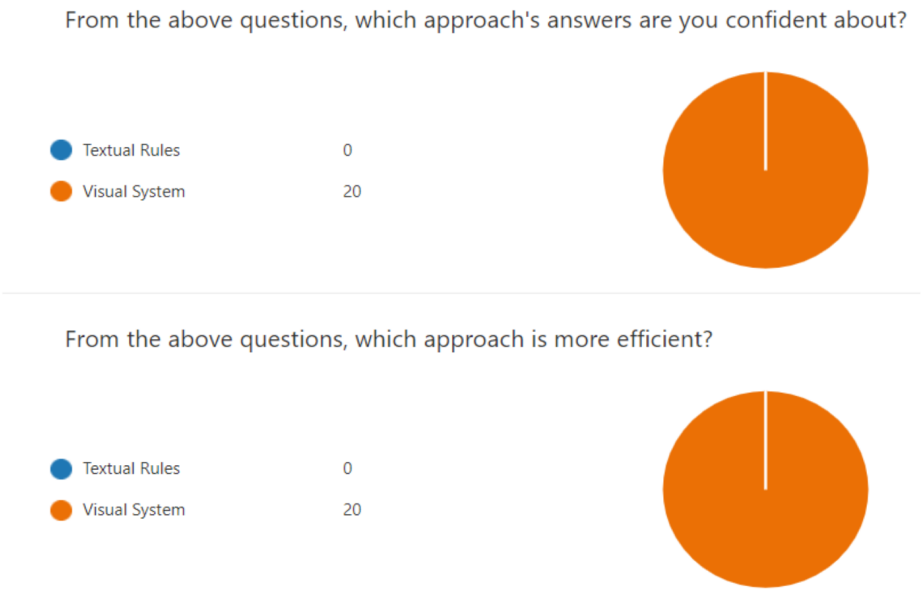

Figure 5.8: Results given by participants regarding which approach is more efficient and which approach's answer they were confident about

The feedback gathered using the subjective questions leads us to understand the people's reaction towards our system. Most of the participants mentioned that the system is "Easy to use" and a "Good Interactive Platform". One of the data analysts said that the "System was easy to use and understand. The system can have many applications with the rule-based models." One participant liked the idea and stated that "If generalized for data for every superstore, I would definitely use it and recommend it for inventory". Among positive responses, some participants also suggested that a way to show the different subsets together can fasten the decision-making process based on these rules.

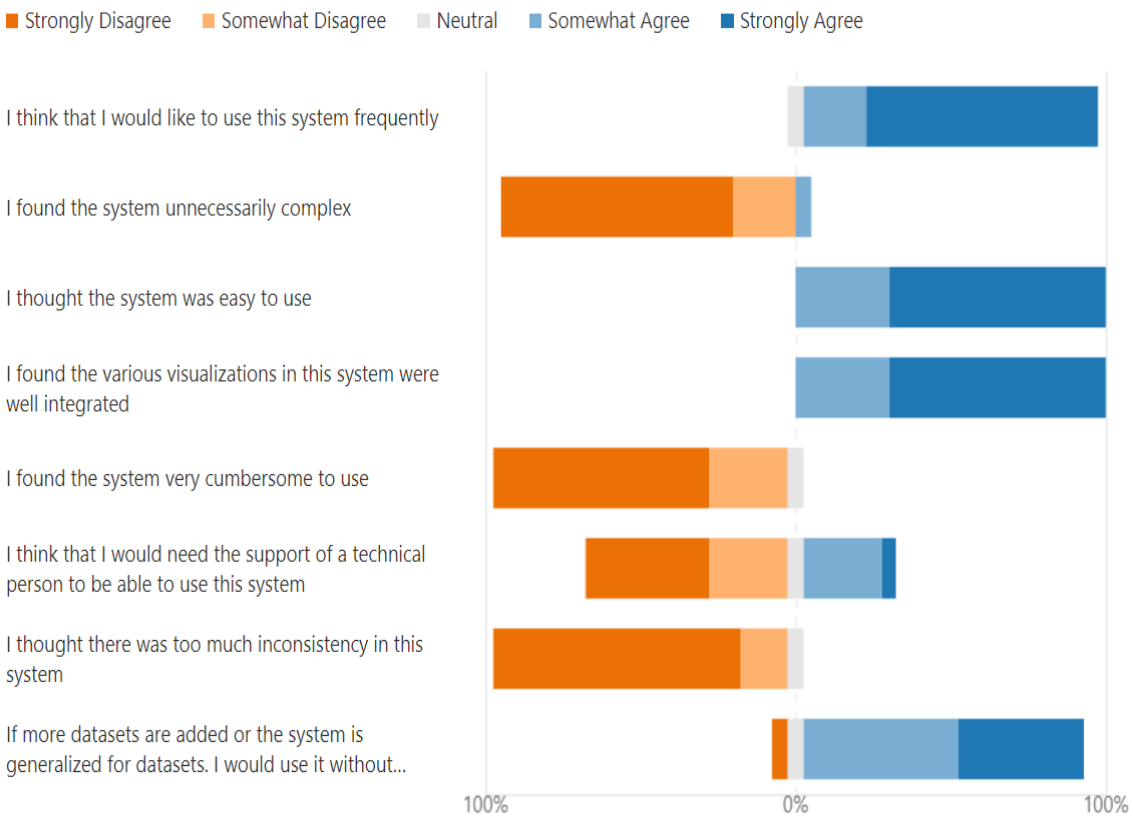

Figure 5.9: User feedback for the software usability

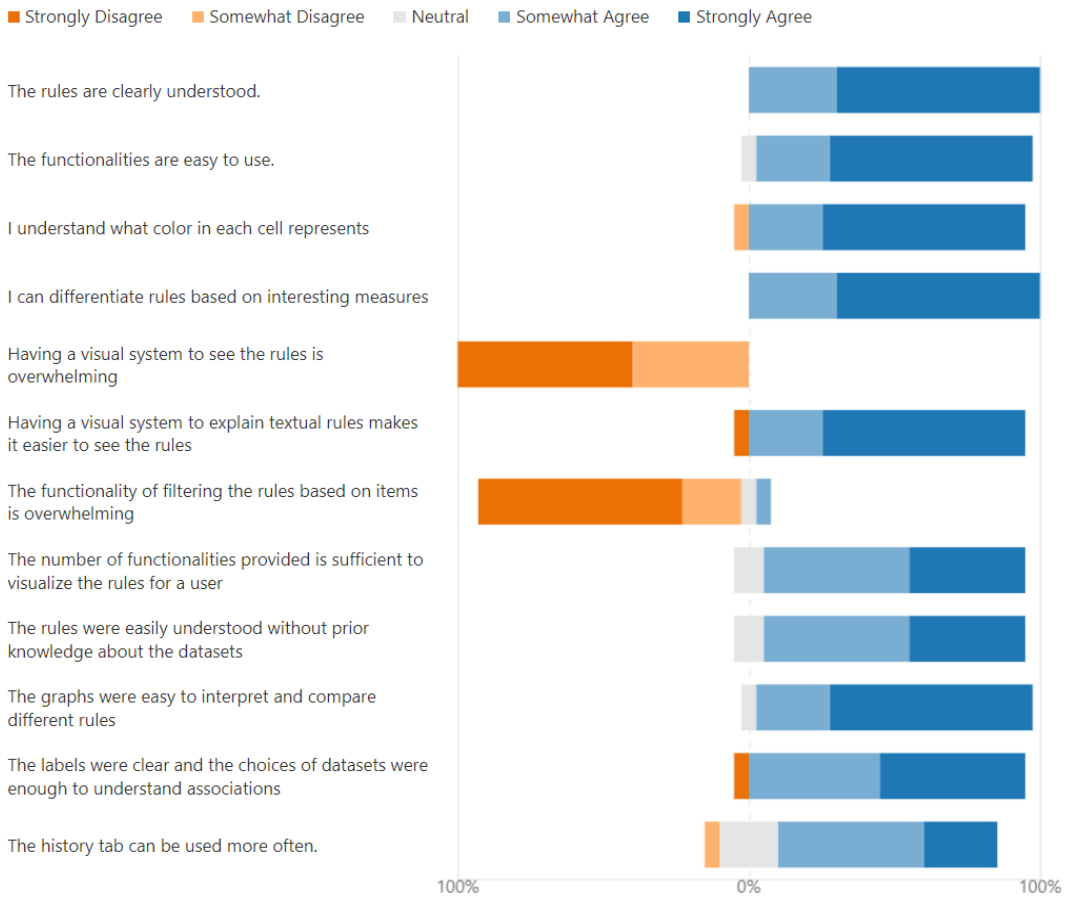

Figure 5.10: User feedback for the interactivity with the ARMatrix system

Chapter 6

Conclusion

A method to visualize the association rules has presented in this thesis. The visualization consists of a 2D matrix to show the association rules where columns represent each rule and rows represent items. The user with little knowledge about the association rules has been kept in mind while designing the system that any person with little background knowledge of association rules and interesting measures can use this system to visualize the rules and do further analysis. The usability of the proposed technique is validated by 20 users from different work and educational domains. The results and feedback from the participants help to infer that the proposed *ARMatrix* system can be used for the visualization and analysis of association rules. Despite the positive reviews, we also found some limitations that are discussed in the next section.

6.1 Discussions

In this section, we will discuss some of the limitations we found in our work and how we plan to address them in the future.

We developed the *ARMatrix* for the assistance of the visualization of association rules. This technique can be accommodated for different domains by converting the database to transactional data sets. The matrix and comparison view together in our approach assist the user for the analysis of the rules which can be generated by setting the minimum threshold values of support and confidence. We verified the usability and applications of our technique by incorporating two use cases and a user study. Our approach makes it easy to incorporate the visualization of a large number of rules but, as the number of rules increases, the matrix visualization becomes difficult to display without the scrolling option. We have tried to address the scalability issue by displaying the rule as text on clicking the rule.

The approach towards focused analysis helps to get insights on the subsets of the total rules based on the user's knowledge. It makes it easy for the user to gain an advantage on the subset level while preserving the details of the overall rules. The problem with this approach is that it is expected of the user to have deeper domain knowledge about the data set. The lack of such detail leads to get falsified information and jeopardizes the purpose of focused analysis in the first place.

Another problem that we observed is that different information can be deduced from our technique based on the user's interest. Different interesting measures lead to getting a different kind of information and hence lead to observing the different context of the rules. To assist the issue related to the user's interest, we incorporated 7 different types of commonly used interesting measures to satisfy the context-based observations.

For future work, we plan to incorporate a module to assist the visualization for the inner working of association rules to explain the methodology behind them. We also plan to add a top-down approach by clustering the similar rules and display rules in a cluster on demand, which will allow the user to only look at a particular subset of interest at a time, reducing the total number of rules to be displayed. Additionally, we tend to provide a way to visualize the change occurring in the rule over time when new data is added.

Bibliography

[1] Market Basket Dataset. http://csci.viu.ca/~barskym/teaching/DM2012/labs/LAB7/PartII.html.

[2] UCI machine learning repository. http://archive.ics.uci.edu/ml/datasets/statlog+(heart), institution = "University of California, Irvine, School of Information and Computer Sciences".

[3] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993.

[4] Abdulmohsen Algarni. Data mining in education. *International Journal of Advanced Computer Science and Applications*, 7, 06 2016.

[5] Gowtham Atluri, Rohit Gupta, Gang Fang, Gaurav Pandey, Michael Steinbach, and Vipin Kumar. Association analysis techniques for bioinformatics problems. pages 1–13, 01 2009.

[6] Radhakrishnan B, Shineraj G, and Anver Muhammed K. M. Application of data mining in marketing, 2013.

[7] Kim Bartke. 2d, 3d and high-dimensional data and information visualization. In *Seminar on Data and Information management*, pages 1–22. Citeseer, 2005.

[8] M. Bharati and Bharati Ramageri. Data mining techniques and applications. *Indian Journal of Computer Science and Engineering*, 1, 12 2010.

[9] Julien Blanchard, Fabrice Guillet, and Henri Briand. Exploratory visualization for association rule rummaging. 05 2004.

[10] Heer Jeffrey Bostock Michael, Ogievetsky Vadim. D3: Data-Driven Documents. *IEEE Transaction on Visualization and Computer Graphics*, 7(12):2–3, 1997.

[11] Sergey Brin, Rajeev Motwani, Jeffrey Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. *ACM SIGMOD Record*, 26, 12 2001.

[12] Wilson Castillo, Alexis Peralta Rojas, and Claudio Meneses. Augmented visualization of association rules for data mining.

[13] Sharma Chakravarthy and Hongen Zhang. Visualization of association rules over relational dbmss. In *Proceedings of the 2003 ACM Symposium on Applied Computing*, SAC '03, page 922–926, New York, NY, USA, 2003. Association for Computing Machinery.

[14] Angelos Chatzimparmpas, Rafael Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, page (Open Access), 03 2020.

[15] Wei Chen, Cong Xie, Pingping Shang, and Qunsheng Peng. Visual analysis of user-driven association rule mining. *Journal of Visual Languages Computing*, 42:76 – 85, 2017.

[16] SHC Choi, Sung-Hyuk Cha, and Charles Tappert. A survey of binary similarity and distance measures. *J. Syst. Cybern. Inf.*, 8, 11 2009.

[17] O. Couturier, T. Hamrouni, S. B. Yahia, and E. M. Nguifo. A scalable association rule visualization towards displaying large amounts of knowledge. In *2007 11th International Conference Information Visualization (IV '07)*, pages 657–663, 2007.

[18] Carlos Fernandez-Basso, M. Ruiz, Miguel Calvo-Flores, and Maria Martin-Bautista. A comparative analysis of tools for visualizing association rules: A proposal for visualising fuzzy association rules. 01 2019.

[19] Enrique García, Cristóbal Romero, Sebastian Ventura, and T. Calders. Drawbacks and solutions of applying association rule mining in learning management systems. *CEUR Workshop Proceedings*, 305:13–22, 01 2007.

[20] Michael Hahsler and Sudheer Chelluboina. Visualizing association rules in hierarchical groups. In *In 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms*, 2011.

[21] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1):53–87, Jan 2004.

[22] Hofmann Heike, Arno Siebes, and Adalbert Wilhelm. Visualizing association rules with interactive mosaic plots. pages 227–235, 01 2000.

[23] Mojdeh Heravi and Osmar Zaïane. A study on interestingness measures for associative classifiers. pages 1039–1046, 01 2010.

[24] B. Jiang, C. Han, and X. Hu. A finite ranked poset and its application in visualization of association rules. In *2008 IEEE International Conference on Granular Computing*, pages 322–325, 2008.

[25] Neesha Jothi, Nur'Aini Abdul Rashid, and Wahidah Husain. Data mining in healthcare – a review. *Procedia Computer Science*, 72:306–313, 2015. The Third Information Systems International Conference 2015.

[26] Mika Klemettinen, Heikki Mannila, Pirjo Moen, Hannu Toivonen, and A. Verkamo. Finding interesting rules from large sets of discovered association rules. *Proceedings of the Third International Conference on Information and Knowledge Management*, 02 1995.

[27] Hans-Peter Kriegel, Karsten M. Borgwardt, Peer Kröger, Alexey Pryakhin, Matthias Schubert, and Arthur Zimek. Future trends in data mining. *Data Mining and Knowledge Discovery*, 15(1):87–97, Aug 2007.

[28] Guimei Liu, Andre Suchitra, Haojun Zhang, Mengling Feng, See-Kiong Ng, and Limsoon Wong. Assocexplorer: An association rule visualization system for exploratory data analysis. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 08 2012.

[29] Alessandra Maciel Paz Milani, Fernando V. Paulovich, and Isabel Harb Manssour. Visualization in the preprocessing phase: Getting insights from enterprise professionals. *Information Visualization*, 19(4):273–287, 2020.

[30] Kian-huat Ong, Kok-leong Ong, Wee Keong Ng, and Ee-Peng Lim. Crystalclear: Active visualization of association rules. 02 2003.

[31] M. S. Ounifi, H. Amdouni, R. B. Elhoussine, and H. Slimane. New 3d visualization and validation tool for displaying association rules and their associated classifiers. In *2016 20th International Conference Information Visualisation (IV)*, pages 152–158, 2016.

[32] Pak Chung Wong, P. Whitney, and J. Thomas. Visualizing association rules for text mining. In *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99)*, pages 120–123, 1999.

[33] Gregory Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In Gregory Piatetsky-Shapiro and William J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

[34] E. Rabelo, M. Dias, C. Franco, and R. C. S. Pacheco. Information visualization: Which is the most appropriate technique to represent data mining results? In *2008 International Conference on Computational Intelligence for Modelling Control Automation*, pages 1228–1233, 2008.

[35] Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018.

[36] Ayse Sagin and Berk Ayvaz. Determination of association rules with market basket analysis: Application in the retail sector. *Southeast Europe Journal of Soft Computing*, 7, 05 2018.

[37] Z. B. Said, F. Guillet, P. Richard, F. Picarougne, and J. Blanchard. Visualisation of association rules based on a molecular representation. In *2013 17th International Conference on Information Visualisation*, pages 577–581, 2013.

[38] Thabet Slimani and Amor Lazzez. Efficient analysis of pattern and association rule mining approaches. *International Journal of Information Technology and Computer Science*, 6(3):70–81, Feb 2014.

[39] Danielle Albers Szafir. The good, the bad, and the biased: Five ways visualizations can mislead (and how to fix them). *Interactions*, 25(4):26–33, June 2018.

[40] Arman Tajbakhsh, Mohammad Rahmati, and Abdolreza Mirzaei. Intrusion detection using fuzzy association rules. *Applied Soft Computing*, 9(2):462–469, 2009.

[41] Alfredo Vellido, José Martín-Guerrero, Fabrice Rossi, and Paulo Lisboa. Seeing is believing: The importance of visualization in real-world machine learning applications. 01 2011.

[42] Q. Xu, C. Li, B. Xiao, and J. Guo. A visualization algorithm for alarm association mining. In *2009 IEEE International Conference on Network Infrastructure and Digital Content*, pages 326–330, 2009.

[43] S. Yamada, T. Funayama, and Y. Yamamoto. Visualization of relations of stores by using association rule mining. In *2015 13th International Conference on ICT and Knowledge Engineering (ICT Knowledge Engineering 2015)*, pages 11–14, 2015.

[44] C. H. Yamamoto, M. C. F. de Oliveira, M. L. Fujimoto, and S. O. Rezende. An itemset-driven cluster-oriented approach to extract compact and meaningful sets of association rules. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 87–92, 2007.

[45] Claudio Yamamoto, Maria Cristina Oliveira, and De Oliveira. Visualization to assist users in association rule mining tasks. 01 2005.

[46] Claudio Yamamoto, Maria Cristina Oliveira, and Solange Rezende. Including the user in the knowledge discovery loop: interactive itemset-driven rule extraction. pages 1212–1217, 01 2008.

[47] Zhang Chunsheng and Li Yan. The visual mining method of apriori association rule based on natural language. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 572–575, 2016.

Appendix A

Consent Form

DALHOUSIE UNIVERSITY

CONSENT FORM

A Matrix-based Visual Analytics approach for the Analysis of Association Rules

You are invited to take part in a research study being conducted by, Rakshit Varu, an MCS graduate student in the Faculty of Computer Science at Dalhousie University. The purpose of this research is to analyze and verify the ease-of-use and usefulness of our proposed system for visualization and exploration of association rules.

If you choose to participate in this research, you will be asked to perform pre-set operations and analysis through our Webapp and anonymously answer questions regarding its usability, which are listed below. The survey should take approximately 45-60 minutes.

- You will complete a general questionnaire regarding your knowledge or level of understanding for such systems.
- You will be given a tutorial on how to use the software.
- You will be given a practice session to use the software.
- You will be given an evaluation questionnaire.
- You will perform four tasks of searching answers the system.
- You will submit the post-study questionnaire and comment.

Your participation in this research is entirely your choice. You do not have to answer questions that you do not want to answer (by selecting prefer not to answer), and you are welcome to stop the survey at any time if you no longer want to participate. All you need to do is close your browser or browser window. I will not include any incomplete surveys in my analyses. If you do complete your survey and you change your mind later, I will not be able to remove the information you provided as I will not know which response is yours.

Your responses to the survey will be anonymous. This means that there are no questions in the survey that ask for identifying details such as your name or email address. All responses will be saved on a secure Dalhousie computer. Only Rakshit Varu and Prof. Fernando Paulovich will have access to the survey results.

I will describe and share general findings of this research in a journal and/or conference publication including the usability and functionality of the system. All the data collected will be destroyed 5 years after reporting the results.

The risks associated with this study are no greater than those you encounter in your everyday life.

To thank you for your time for completing the evaluation you will automatically be entered for a draw to win a $50 gift card for participating in the survey. Your contact information will not be linked in any way to your survey responses.
You should discuss any questions you have about this study with Rakshit Varu or Prof. Fernando Paulovich. Please ask as many questions as you like before or after participating. My contact information is rk996644@dal.ca.

If you have any ethical concerns about your participation in this research, you may contact Research Ethics, Dalhousie University at (902) 494-3423, or email ethics@dal.ca (and reference REB file # 20XX-XXXX)."

If you agree to complete the survey, please answer this email with "I accept the consent agreement", and the link to the survey will be sent to you.