Visualizing Association Rules: Introduction to the R-extension Package arulesViz

Michael Hahsler
Southern Methodist University

Sudheer Chelluboina
Southern Methodist University

Abstract

Association rule mining is a popular data mining method available in R as the extension package arules. However, mining association rules often results in a very large number of found rules, leaving the analyst with the task to go through all the rules and discover interesting ones. Sifting manually through large sets of rules is time consuming and strenuous. Visualization has a long history of making large data sets better accessible using techniques like selecting and zooming. In this paper we present the R-extension package arulesViz which implements several known and novel visualization techniques to explore association rules. With examples we show how these visualization techniques can be used to analyze a data set.

Keywords: data mining, association rules, visualization.

1. Introduction

Many organizations generate a large amount of transaction data on a daily basis. For example, a department store like “Macy’s” stores customer shopping information at a large scale using check-out data. Association rule mining is one of the major techniques to detect and extract useful information from large scale transaction data. Mining association rules was first introduced by Agrawal, Imielinski, and Swami (1993) and can formally be defined as:

Let \( I = \{i_1, i_2, \ldots, i_n\} \) be a set of \( n \) binary attributes called items. Let \( D = \{t_1, t_2, \ldots, t_m\} \) be a set of transactions called the database. Each transaction in \( D \) has an unique transaction ID and contains a subset of the items in \( I \). A rule is defined as an implication of the form \( X \Rightarrow Y \) where \( X, Y \subseteq I \) and \( X \cap Y = \emptyset \). The sets of items (for short itemsets) \( X \) and \( Y \) are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule. Often rules are restricted to only a single item in the consequent.

Association rules are rules which surpass a user-specified minimum support and minimum confidence threshold. The support \( \text{supp}(X) \) of an itemset \( X \) is defined as the proportion of transactions in the data set which contain the itemset and the confidence of a rule is defined \( \text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \). Therefore, an association rule \( X \Rightarrow Y \) will satisfy:

\[ \text{supp}(X \cup Y) \geq \sigma \]

and

\[ \text{conf}(X \Rightarrow Y) \geq \delta \]

where \( \sigma \) and \( \delta \) are the minimum support and minimum confidence, respectively.
Another popular measure for association rules used throughout this paper is *lift* (Brin, Motwani, Ullman, and Tsur 1997). The lift of a rule is defined as

\[
\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{(\text{supp}(X)\text{supp}(Y))}
\]

and can be interpreted as the deviation of the support of the whole rule from the support expected under independence given the supports of both sides of the rule. Greater lift values (≫ 1) indicate stronger associations. Measures like support, confidence and lift are generally called interest measures because they help with focusing on potentially more interesting rules.

For a more detailed treatment of association rules we refer the reader to the in introduction paper for package *arules* (Hahsler, Buchta, Grün, and Hornik 2010; Hahsler, Grün, and Hornik 2005) and the literature referred to there.

Association rules are typically generated in a two-step process. First, minimum support is used to generate the set of all frequent itemsets for the data set. Frequent itemsets are itemsets which satisfy the minimum support constraint. Then, in a second step, each frequent itemsets is used to generate all possible rules from it and all rules which do not satisfy the minimum confidence constraint are removed. Analyzing this process, it is easy to see that in the worst case we will generate \(2^n - n - 1\) frequent itemsets with more than two items from a database with \(n\) distinct items. Since each frequent itemset will in the worst case generate at least two rules, we will end up with a set of rules in the order of \(O(2^n)\). Typically, increasing minimum support is used to keep the number of association rules found at a manageable size. However, this also removes potentially interesting rules with less support. Therefore, the need to deal with large sets of association rules is unavoidable when applying association rule mining in a real setting.

Visualization is successfully used to communicate both abstract and concrete ideas in many areas like education, engineering and science (Prangsmal, van Boxtel, Kanselaar, and Kirschner 2009). According to Chen, Unwin, and Hardle (2008), the application of visualization falls into two phases. First, the exploration phase where the analysts will use graphics that are mostly incompatible for presentation purposes but make it easy to find interesting and important features of the data. The amount of interaction needed during exploration is very high and includes filtering, zooming and rearranging data. After key findings are discovered in the data, these findings must be presented in a way suitable for presentation for a larger audience. In this second phase it is important that the analyst can manipulate the presentation to clearly highlight the findings.

Many researchers introduced visualization techniques like scatter plots, matrix visualizations, graphs, mosaic plots and parallel coordinates plots to analyze association rules (see Bruzzese and Davino (2008) and Jentner and Keim (2017) for a recent overview). This paper discusses existing techniques and demonstrates how their implementation in *arulesViz* can be used via a simple unified interface. We extend most plots using techniques of color shading and reordering to improve their interpretability. Finally, this paper also introduces a completely new method called “grouped matrix-based visualization” which is based on a novel way of clustering rules. Clustering rules is usually based on distances defined on the items included in the rules or on shared transactions covered by the rules. However, here we cluster rules especially for visualization using similarities between sets of values of a selected interest measure.

The rest of the paper is organized as follows. In Section 2 we give a very short example of how to prepare data using package *arules* and then introduce the unified interface provided.
by **arulesViz** for association rule visualization. In Sections 3 to 8 we describe the different visualization techniques and give examples. Most of the techniques are enhanced using color shading and reordering. Grouped matrix-based visualization in Section 5 is a novel visualization technique. In Section 9 compares the presented visualization techniques. Section 11 concludes the paper.

2. Data preparation and unified interface of arulesViz

Before we start, we set the number of displayed significant digits to two to make the output easier to read, and we set the seed for the random number generator for predictability.

```r
> options(digits = 2)
> set.seed(1234)
```

To use **arulesViz** we first have to load the package. The package automatically loads other needed packages like **arules** (Hahsler et al. 2010) for handling and mining association rules. For the examples in this paper we load the “Groceries” data set which is included in **arules**.

```r
> library("arulesViz")
> data("Groceries")
```

Groceries contains sales data from a local grocery store with 9835 transactions and 169 items (product groups). The summary shows some basic statistics of the data set. For example, that the data set is rather sparse with a density just above 2.6%, that “whole milk” is the most popular item and that the average transaction contains less than 5 items.

```r
> summary(Groceries)
```

transactions as itemMatrix in sparse format with
9835 rows (elements/itemsets/transactions) and
169 columns (items) and a density of 0.026

most frequent items:

<table>
<thead>
<tr>
<th></th>
<th>whole milk</th>
<th>other vegetables</th>
<th>rolls/buns</th>
<th>soda</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2513</td>
<td>1903</td>
<td>1809</td>
<td>1715</td>
</tr>
<tr>
<td>yogurt (Other)</td>
<td>1372</td>
<td>34055</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

element (itemset/transaction) length distribution:

<table>
<thead>
<tr>
<th>sizes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2159</td>
<td>1643</td>
<td>1299</td>
<td>1005</td>
<td>855</td>
<td>645</td>
<td>545</td>
<td>438</td>
<td>350</td>
<td>246</td>
<td>182</td>
<td>117</td>
<td>78</td>
<td>77</td>
<td>55</td>
</tr>
<tr>
<td>Min.</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>1st Qu.</td>
<td>46</td>
<td>29</td>
<td>14</td>
<td>14</td>
<td>9</td>
<td>11</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Qu.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
includes extended item information - examples:

1  frankfurter sausage meat and sausage
2  sausage sausage meat and sausage
3  liver loaf sausage meat and sausage

Next we mine association rules using the Apriori algorithm implemented in \texttt{arules}.

\begin{verbatim}
> rules <- apriori(Groceries, parameter=list(support=0.001, confidence=0.5))

Apriori

Parameter specification:
  confidence minval smax arem aval originalSupport maxtime support minlen
  0.5 0.1 1 none FALSE TRUE 5 0.001 1

Algorithmic control:
  filter tree heap memopt load sort verbose
  0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [157 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.01s].
writing ... [5668 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].

> rules

set of 5668 rules

The result is a set of 5668 association rules. The top three rules with respect to the lift measure, a popular measure of rule strength, are:

\begin{verbatim}
> inspect(head(rules, n=3, by="lift"))

   lhs                        rhs                    support  confidence
[1] {Instant food products,soda} => {hamburger meat} 0.0012    0.63
[2] {soda,popcorn} => {salty snack}          0.0012    0.63
\end{verbatim}
However, it is clear that going through all the 5668 rules manually is not a viable option. We therefore will introduce different visualization techniques implemented in arulesViz. All implemented visualization techniques share the following interface:

```r
> args(getS3method("plot", "rules"))

function (x, method = NULL, measure = "support", shading = "lift",
       interactive = NULL, engine = "default", data = NULL, control = NULL,
       ...) NULL
```

where `x` is the set of rules to be visualized, `method` is the visualization method, and `measure` and `shading` contain the interest measures used by the plot. Further arguments are described in the manual page.

Using `engine`, different plotting engines can be specified to render the plot. The default engine typically uses `grid`, many plots can also be rendered using the engine "htmlwidget" resulting in an interactive HTML widget.

In the following sections we will introduce the different visualization methods implemented in arulesViz and demonstrate how easy it is to use them.

### 3. Scatter plot

A straight-forward visualization of association rules is to use a scatter plot with two interest measures on the axes. Such a presentation can be found already in an early paper by Bayardo, Jr. and Agrawal (1999) when they discuss sc-optimal rules.

The default method for `plot()` for association rules in arulesViz is a scatter plot using support and confidence on the axes. In addition a third measure (default: lift) is used as the color (gray level) of the points. A color key is provided to the right of the plot.

```r
> plot(rules)
```

This plot for the rules mined in the previous section is shown in Figure 1. We can see that rules with high lift have typically a relatively low support. Bayardo, Jr. and Agrawal (1999) argue that the most interesting rules (sc-optimal rules) reside on the support/confidence border, which can be clearly seen in this plot. We will show later how the interactive features of this plot can be used to explore these rules.

Any measure stored in the quality slot of the set of rules can be used for the axes (vector of length 2 for parameter `measure`) or for color shading (`shading`). The following measures are available for our set of rules.
These are the default measures generated by Apriori. To add other measures we refer the reader to the function \texttt{interestMeasure()} included in \texttt{arules}. For example we can customize the plot by switching lift and confidence:

\begin{verbatim}
> plot(rules, measure = c("support", "lift"), shading = "confidence")
\end{verbatim}

Figure 2 shows this plot with lift on the y-axis. Here it is easy to identify all rules with high lift.

\textbf{Unwin, Hofmann, and Bernt (2001)} introduced a special version of a scatter plot called \textit{Two-key plot}. Here support and confidence are used for the x and y-axes and the color of the points is used to indicate “order,” i.e., the number of items contained in the rule. Two-key plots can be produced using the unified interface by:

\begin{verbatim}
> plot(rules, method = "two-key plot")
\end{verbatim}
Figure 2: Scatter plot with lift on the y-axis.

Figure 3: Two-key plot.
The resulting Two-key plot is shown in Figure 3. From the plot it is clear that order and support have a very strong inverse relationship, which is a known fact for association rules (Seno and Karypis 2005).

In addition to using order for shading, we also give the plot a different title (main). Other control options including pch (best with filled symbols: 20–25), cex, xlim and ylim are available and work in the usual way expected by R-users.

For exploration, the scatter plot method offers interactive features for selecting and zooming. Interaction is activated using interactive=TRUE.

```r
> sel <- plot(rules, measure=c("support", "lift"), shading = "confidence", + interactive = TRUE)
```

Interactive features include:

- Inspecting individual rules by selecting them and clicking the inspect button.
- Inspecting sets of rules by selecting a rectangular region of the plot and clicking the inspect button.
- Zooming into a selected region (zoom in/zoom out buttons).
- Filtering rules using the measure used for shading by clicking the filter button and selecting a cut-off point in the color key. All rules with a measure lower than the cut-off point will be filtered.
- Returning the last selection for further analysis (end button).

The result of an example interaction is shown in Figure 4. Using a box selection the rules with the highest lift are selected. Using the inspect button, the rules are displayed in the terminal below the plotting device.

**4. Matrix-based visualizations**

Matrix-based visualization techniques organize the antecedent and consequent itemsets on the x and y-axes, respectively. A selected interest measure is displayed at the intersection of the antecedent and consequent of a given rule. If no rule is available for a antecedent/consequent combination the intersection area is left blank.

Formally, the visualized matrix is constructed as follows. We start with the set of association rules

$\mathcal{R} = \{\langle a_1, c_1, m_1 \rangle, \ldots, \langle a_i, c_i, m_i \rangle, \ldots, \langle a_n, c_n, m_n \rangle\}$

where $a_i$ is the antecedent, $c_i$ is the consequent and $m_i$ is the selected interest measure for the $i$-th rule for $i = 1, \ldots, n$. In $\mathcal{R}$ we identify the set of $K$ unique antecedents and $L$ unique consequent. We create a $L \times K$ matrix $\mathbf{M}$ with one column for each unique antecedent and one row for each unique consequent. Finally, we populate the matrix by setting $\mathbf{M}_{lk} = m_i$ for $i = 1, \ldots, n$ and $l$ and $k$ corresponding to the position of $a_i$ and $c_i$ in the matrix. Note that $\mathbf{M}$ will contain many empty cells since many potential association rules will not meet the required minimum thresholds on support and confidence.
Figure 4: Interactive mode for scatter plot (inspecting rules with high lift).
Ong, leong Ong, Ng, and Lim (2002) presented a version of the matrix-based visualization technique where a 2-dimensional matrix is used and the interest measure is represented by color shading of squares at the intersection. An alternative visualization option is to use 3D bars at the intersection (Wong, Whitney, and Thomas 1999; Ong et al. 2002).

For this type of visualization the number of rows/columns depends on the number of unique itemsets in the consequent/antecedent in the set of rules. Since large sets of rules typically have a large number of different itemsets as antecedents (often not much smaller than the number of rules themselves), the size of the colored squares or the 3D bars gets very small and hard to see. We reduce the number of rules here by filtering out all rules with a low confidence score.

```r
> subrules <- rules[quality(rules)$confidence > 0.8]
> subrules

set of 371 rules
```

```r
> plot(subrules, method = "matrix", measure = "lift")
```

The resulting plot is shown in Figure 5. Since there is not much space for long labels in the plot, we only show numbers as labels for rows and columns and the complete itemsets are printed to the terminal for look-up. We omit the complete output here, since this plot and the next few plots print several hundred labels to the screen. The output looks like:

Itemsets in Antecedent (lhs)

```
[1] "{liquor,red/blush wine}"
[2] "{curd,cereals}"
[3] "{yogurt,cereals}"
[4] "{butter,jam}"
[5] "{soups,bottled beer}"
(lines omitted)
[343] "{tropical fruit,root vegetables,rolls/buns,bottled water}"
[344] "{tropical fruit,root vegetables,yogurt,rolls/buns}"
```

Itemsets in Consequent (rhs)

```
[1] "{bottled beer}"    "{whole milk}"    "{other vegetables}"
[4] "{tropical fruit}"  "{yogurt}"    "{root vegetables}"
```

An alternative representation is to use 3D bars (method “matrix” with engine “3d”) instead of colored rectangles.

```r
> plot(subrules, method = "matrix", engine = "3d", measure = "lift")
```

The 3D visualization is shown in Figure 6.

5. Grouped matrix-based visualization
Figure 5: Matrix-based visualization with colored squares.

Figure 6: Matrix-based visualization with 3D bars.
Matrix-based visualization is limited in the number of rules it can visualize effectively since large sets of rules typically also have large sets of unique antecedents/consequents. Here we introduce a new visualization technique (Hahsler and Chelluboina 2011; Hahsler and Karpienko 2016) that enhances matrix-based visualization using grouping of rules via clustering to handle a larger number of rules. Grouped rules are presented as an aggregate in the matrix and can be explored interactively by zooming into and out of groups.

A direct approach to cluster itemsets is to define a distance metric between two itemsets $X_i$ and $X_j$. A good choice is the Jaccard distance defined as

$$d_{\text{Jaccard}}(X_i, X_j) = 1 - \frac{|X_i \cap X_j|}{|X_i \cup X_j|}.$$ 

The distance simply is the number of items that $X_i$ and $X_j$ have in common divided by the number of unique items in both sets. For a set of $m$ rules we can calculate the $m(m-1)/2$ distances and use them as the input for clustering. However, using clustering on the itemsets directly has several problems. First of all, data sets typically mined for association rules are high-dimensional, i.e., contain many different items. This high dimensionality is carried over to the mined rules and leads to a situation referred to as the “course of dimensionality” where, due to the exponentially increasing volume, distance functions lose their usefulness. The situation is getting worse since minimum support used in association rule mining leads in addition to relatively short rules resulting in extremely sparse data.

Several approaches to cluster association rules and itemsets to address the dimensionality and sparseness problem were proposed in the literature. Toivonen, Klemettinen, Ronkainen, Hatonen, and Mannila (1995), Gupta, Strehl, and Ghosh (1999) and Berrado and Runger (2007) propose clustering association rules by looking at the number of transactions which are covered by the rules. Using common covered transactions avoids the problems of clustering sparse, high-dimensional binary vectors. However, it introduces a strong bias towards clustering rules which are generated from the same frequent itemset. By definition of frequent itemsets, two subsets of a frequent itemset will cover many common transactions. This bias will lead to mostly just rediscovering the already known frequent itemset structure from the set of association rules.

Here we pursue a completely different approach. We start with the matrix $M$ defined in Section 4 which contains the values of a selected interest measure of the rules in set $R$. The columns/rows are the unique antecedents/consequents in $R$, respectively. Now grouping antecedents becomes the problem of grouping columns in $M$. To group the column vectors fast and efficient into $k$ groups we use $k$-means clustering. The default interest measure used is lift. The idea is that antecedents that are statistically dependent on the same consequents are similar and thus can be grouped together. Compared to other clustering approaches for itemsets, this method enables us to even group antecedents containing substitutes (e.g., butter and margarine) which are rarely purchased together since they will have similar dependence to the same consequents.

The same grouping method can be used for consequents. However, since the mined rules are restricted to a single item in the consequent there is no problem with combinatorial explosion and such a grouping is typically not necessary.

To visualize the grouped matrix we use a balloon plot with antecedent groups as columns and consequents as rows (see Figure 7). The color of the balloons represent the aggregated
interest measure in the group with a certain consequent and the size of the balloon shows the
aggregated support. The default aggregation function is the median value in the group. The
number of antecedents and the most important (frequent) items in the group are displayed
as the labels for the columns. Furthermore, the columns and rows in the plot are reordered
such that the aggregated interest measure is decreasing from top down and from left to right,
placing the most interesting group in the top left corner.

The matrix visualization with grouped antecedents for the set of 5668 rules mined earlier can
be easily created by

```r
> plot(rules, method = "grouped")
```

The resulting visualization is shown in Figure 7. The group of most interesting rules according
to lift (the default measure) are shown in the top-left corner of the plot. There are 3 rules
which contain “Instant food products” and up to 2 other items in the antecedent and the
consequent is “hamburger meat.”

To increase the number of groups we can change \( k \) which defaults to 20.

```r
> plot(rules, method = "grouped", control = list(k = 50))
```

The resulting, more detailed plot is shown in Figure 8.

An interactive version of the grouped matrix visualization is also available.

```r
> sel <- plot(rules, method = "grouped", interactive = TRUE)
```

Here it is possible to zoom into groups and to inspect the rules contained in a selected group.

### 6. Graph-based visualizations

Graph-based techniques (Klemettinen, Mannila, Ronkainen, Toivonen, and Verkamo 1994;
Rainsford and Roddick 2000; Buono and Costabile 2005; Ertek and Demiriz 2006) visualize
association rules using vertices and edges where vertices annotated with item labels represent
items, and itemsets or rules are represented as a second set of vertices. Items are connected
with itemsets/rules using arrows. For rules arrows pointing from items to rule vertices indicate
LHS items and an arrow from a rule to an item indicates the RHS. Interest measures are
typically added to the plot by using color or size of the vertices representing the itemsets/rules.

Graph-based visualization offers a very clear representation of rules but they tend to easily
become cluttered and thus are only viable for very small sets of rules. For the following plots
we select the 10 rules with the highest lift.

```r
> subrules2 <- head(rules, n = 10, by = "lift")
```

arulesViz contains several graph-based visualizations rendered using either the igraph library
via package igraph (Csardi and Nepusz 2006) or the interface to the GraphViz software in
package Rgraphviz (Gentry, Long, Gentleman, Falcon, Hahne, Sarkar, and Hansen 2010). By
default igraph is used. The following plot represents items and rules as vertices connecting
them with directed edges (shown in Figure 9). This representation focuses on how the rules
are composed of individual items and shows which rules share items.
Figure 7: Grouped matrix-based visualization.
Items in LHS Groups

RHS (+15 not shown)

Figure 8: Grouped matrix with $k = 50$. 
> plot(subrules2, method = "graph")

An interactive visualization is available in arulesViz, however, the built-in graph based visualizations are only useful for small set of rules. To explore large sets of rules with graphs, advanced interactive features like zooming, filtering, grouping and coloring nodes are needed. Such features are available in interactive visualization and exploration platforms for networks and graphs like Gephi (Bastian, Heymann, and Jacomy 2009). From arulesViz graphs for sets of association rules can be exported in the GraphML format or as a Graphviz dot-file to be explored in tools like Gephi. For example the 1000 rules with the highest lift are exported by:

> saveAsGraph(head(rules, n = 1000, by = "lift"), file = "rules.graphml")

Figure 10 shows a screenshot of exploring these rules interactively. Rules can be explored by zooming, filtering and coloring vertices and edges.

7. Parallel coordinates plot

Parallel coordinates plots are designed to visualize multidimensional data where each dimension is displayed separately on the x-axis and the y-axis is shared. Each data point is represented by a line connecting the values for each dimension. Parallel coordinates plots were used previously to visualize discovered classification rules (Han, An, and Cercone 2000) and association rules (Yang 2003). Yang (2003) displays the items on the y-axis as nominal values and the x-axis represents the positions in a rule, i.e., first item, second item, etc. Instead of a simple line an arrow is used where the head points to the consequent item. Arrows only span enough positions on the x-axis to represent all the items in the rule, i.e., rules with less items are shorter arrows.

> plot(subrules2, method = "paracoord")

Figure 11 shows a parallel coordinates plot for 10 rules. The width of the arrows represents support and the intensity of the color represent confidence. It is obvious that for larger rule sets visual analysis becomes difficult since with an increasing number of rules also the number of crossovers between the lines increases Yang (2003).

The number of crossovers can be significantly reduced by reordering the items on the y-axis. Reordering the items to minimize the number of crossovers is a combinatorial problem with $n!$ possible permutations. However, for visualization purposes a suboptimal but fast solution is typically acceptable. We applies a variation of the well known 2-opt heuristic Bentley (1990) for travelers salesman problem to the reordering problem. The objective function is to minimize the number of crossovers. The simple heuristic uses the following steps:

1. Choose randomly two items and exchange them if it improves the objective function.
2. Repeat step 1 till no improvement is found for a predefined number of tries.

Reordering is achieved with reorder=TRUE.
Figure 9: Graph-based visualization with items and rules as vertices.
Figure 10: Visualization of 1000 rules with Gephi (Fruchterman Reingold layout, vertex and label size is proportional to the in-degree, i.e., the number of rules the consequent participates in).
Figure 11: Parallel coordinate plot.

Figure 12: Parallel coordinate plot (reordered).
> plot(subrules2, method = "paracoord", control = list(reorder = TRUE))

Figure 12 shows the parallel coordinates plot with reordered items to reduce crossovers.

8. Double Decker plots

A double decker plot is a variant of a mosaic plot. A mosaic plot displays a contingency table using tiles on a rectangle created by recursive vertical and horizontal splits. The size of each tile is proportional to the value in the contingency table. Double decker plots use only a single horizontal split.

Hofmann, Siebes, and Wilhelm (2000) introduced double decker plots to visualize a single association rule. Here the displayed contingency table is computed for a rule by counting the occurrence frequency for each subset of items in the antecedent and consequent from the original data set. The items in the antecedent are used for the vertical splits and the consequent item is used for horizontal highlighting.

We randomly choose a single rule and visualize it with a double decker plot.

> oneRule <- sample(rules, 1)
> inspect(oneRule)

<table>
<thead>
<tr>
<th>lhs</th>
<th>rhs</th>
<th>support</th>
<th>confidence</th>
<th>coverage</th>
<th>lift</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>{citrus fruit, grapes, fruit/vegetable juice} =&gt; {tropical fruit}</td>
<td>0.0011</td>
<td>0.85</td>
<td>0.0013</td>
<td>8.1</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

> plot(oneRule, method = "doubledecker", data = Groceries)

Figure 13 shows the resulting plot. The area of blocks gives the support and the height of the “yes” blocks is proportional to the confidence for the rules consisting of the antecedent items marked as “yes.” Items that show a significant jump in confidence when changed from “no” to “yes” are interesting. This is captured by the interest measure difference of confidence defined by Hofmann and Wilhelm (2001).

9. Comparison of techniques

In this section, we compare the visualization techniques available in arulesViz based on the size of the rule set which can be analyzed, the number of interest measures which are shown simultaneously, if the technique offers interaction and reordering and how intuitive each visualization technique is. Note that most of these categories are only evaluated qualitatively here, and the results presented in Table 1 are only meant to guide the user towards the most suitable techniques for a given application.

Scatterplot (including two-key plots) and grouped matrix plot are capable to analyze large rule sets. These techniques are interactive to allow the analyst to zoom and select interesting rules. Matrix-based can accommodate rule sets of medium size. Reordering can be used to improve the presentation. To analyze small rule sets the matrix-based method with 3D bars,
Double decker plot for 1 rule

Figure 13: Double decker plot for a the rule citrus fruit, grapes, fruit/vegetable juice => tropical fruit.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Method</th>
<th>Rule set</th>
<th>Measures</th>
<th>Interactive</th>
<th>Reordering</th>
<th>Ease of use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scatterplot</td>
<td>&quot;scatterplot&quot;</td>
<td>large</td>
<td>3</td>
<td>✓</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>Two-Key plot</td>
<td>&quot;scatterplot&quot;</td>
<td>large</td>
<td>2 + order</td>
<td>✓</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>Matrix-based</td>
<td>&quot;matrix&quot;</td>
<td>medium</td>
<td>1</td>
<td>✓</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Matrix-b. (2 measures)</td>
<td>&quot;matrix&quot;</td>
<td>medium</td>
<td>2</td>
<td>✓</td>
<td>– –</td>
<td></td>
</tr>
<tr>
<td>Matrix-b. (3D bar)</td>
<td>&quot;matrix (engine 3d)&quot;</td>
<td>small</td>
<td>1</td>
<td>✓</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Grouped matrix</td>
<td>&quot;grouped&quot;</td>
<td>large</td>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>0</td>
</tr>
<tr>
<td>Graph-based</td>
<td>&quot;graph&quot;</td>
<td>small</td>
<td>2</td>
<td>✓</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>Graph-b. (external)</td>
<td>&quot;graph&quot;</td>
<td>large</td>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>+</td>
</tr>
<tr>
<td>Parallel coordinates</td>
<td>&quot;paracoord&quot;</td>
<td>small</td>
<td>1</td>
<td>✓</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Double decker</td>
<td>&quot;doubledecker&quot;</td>
<td>single rule</td>
<td>(2)</td>
<td></td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparison of visualization methods for association rules available in arulesViz.

graph-based methods and parallel coordinates plots are suitable. Graphs for large rule sets can be analyzed using external tools like Gephi. Finally, double decker plots only visualize a single rule.

The techniques discussed in this paper can also be categorized based on the number of interest measures simultaneously visualized. Most methods can represent two measures and scatter plots are even able to visualize three measures for each rule in one plot.

Scatter plot and graph based techniques are the most intuitive while matrix-based visualization with two interest measures, parallel coordinates and double decker require time to learn how to interpret them correctly.

10. Further Reading

Details on the grouped matrix plot are reported in Hahsler and Karpienko (2016). Interactive
visualization using htmlWidgets with examples is describe in Hahsler (2017).

11. Conclusion

Association rule mining algorithms typically generate a large number of association rules which poses a major problem for understanding and analyzing rules. In this paper we presented several visualization techniques implemented in arulesViz which can be used to explore and present sets of association rules. In addition we present a new interactive visualization method called grouped matrix-based visualization which can used to effectively explore large rule sets.

References


Affiliation:
Michael Hahsler
Engineering Management, Information, and Systems
Lyle School of Engineering
Southern Methodist University
P.O. Box 750123
Dallas, TX 75275-0123
E-mail: mhahsler@lyle.smu.edu
URL: http://lyle.smu.edu/~mhahsler