J Bus Econ (2017) 87:317–335
DOI 10.1007/s11573-016-0822-8

ORIGINAL PAPER

Visualizing association rules in hierarchical groups

Michael Hahsler[1] · **Radoslaw Karpienko**[2]

Published online: 7 May 2016

Abstract Association rule mining is one of the most popular data mining methods. 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. Although visualization has a long history of making large amounts of data better accessible using techniques like selecting and zooming, most association rule visualization techniques are still falling short when it comes to large numbers of rules. In this paper we introduce a new interactive visualization method, the *grouped matrix* representation, which allows to intuitively explore and interpret highly complex scenarios. We demonstrate how the method can be used to analyze large sets of association rules using the R software for statistical computing, and provide examples from the implementation in the R-package **arulesViz**.

Keywords Association rules · Visualization · Shopping baskets · Exploratory analysis

JEL Classification M3 · C6 · C8

1 Introduction

Businesses nowadays collect and store unprecedented amounts of customer data on a daily basis, and the so-called ‘data explosion’ has been identified as one of the major challenges for marketers in both, online and offline channels (Leeflang et al.

✉ Radoslaw Karpienko
rkarpien@wu.ac.at

[1] Department of Engineering Management, Information and Systems, Southern Methodist University, Dallas, USA

[2] Department of Marketing, Vienna University of Economics and Business, Vienna, Austria

2014). A recent discussion by Day (2011) pointed out that there is an emergent gap between increasing data complexity and the capacity of marketing departments to cope with the realities of the digital era.

In response, recent marketing publications provided quite diverse tools, able to cope with complex data structures—for instance—relationship data (Fader et al. 2010), user generated content (Lee and Bradlow 2011; Netzer et al. 2012), and scanner data (Rooderkerk et al. 2013). Although making meaning from ever-growing data warehouses requires state-of-the-art analytical techniques, many authors stressed the need for accessible and practical marketing tools which support managerial decision making (Fader et al. 2010; Netzer et al. 2012). Correspondingly, a number of publications have presented data visualization techniques as a means to extract meaningful results from highly complex settings (Lee and Bradlow 2011; Netzer et al. 2012).

We contribute to this stream of research, by providing a highly flexible integrated framework for post-processing and visualization of association rules—one of the most popular techniques in data mining (Agrawal et al. 1993). Association rule mining often results in a very large number of extracted rules, typically leaving the analyst with the cumbersome task to revise rules, and to identify the most interesting and important patterns by hand. In the present paper, we demonstrate how visualization techniques can be used to intuitively interpret even settings with vast amounts of extracted rules. In particular, we apply our proposed framework to a common marketing problem—modeling of shopping baskets from scanner data.

A 'shopping basket' represents the consumer's observable choices of products or categories during a shopping trip (Manchanda et al. 1999). The underlying notion of shopping basket analysis is that the observable choices of products during a shopping trip are interdependent. Hence, the overall composition of a shopping basket can be inferred by observing only few product choices, because the choice of one product affects all subsequent choices.

There are two basic cases where product choices in a shopping basket are considered *not* to be stochastically independent: (1) complements, when products co-occur more frequently than expected, and (2) substitutes, when products appear less frequently than expected (Hruschka 2012). From the marketer's perspective, both cases are particularly interesting, because they allow to utilize promotions in one category to drive profits through cross selling (Russell and Kamakura 1997). Hence, shopping basket analysis facilitates the implementation of well-founded and efficient promotion strategies (Natter et al. 2007; Breugelmans et al. 2010). In marketing literature, several approaches have been proposed to model shopping baskets (i.e. interrelations between categories), which can be generally summarized into explanatory or exploratory techniques (Mild and Reutterer 2003; Boztuğ and Silberhorn 2006).

The most popular explanatory methods are multivariate logit (MVL) and multivariate probit models (MVP). These models are widely adopted methods to analyze scanner data, and were among the first methods used to model shopping baskets (Manchanda et al. 1999; Mild and Reutterer 2003). The main argument for explanatory methods is that they can directly incorporate marketing variables such as (e.g.) customer demographics and advertising efforts. However, they are computationally expensive, and impractical for larger numbers of categories.

Therefore, explanatory methods are typically restricted to small numbers of categories (see Hruschka 2012 for a thorough discussion). Few extensions of the basic MVL and MVP models have been proposed in the literature in order to cope with these limitations. These extensions either involved tweaking the model itself (and particularly the employed estimation techniques) to handle more categories, or employing a more strategic stepwise procedure in which the actual explanatory modeling part was preceded by a data-compression step (Boztug and Reutterer 2008; Breugelmans et al. 2010).

Exploratory methods on the other hand, typically focus on extracting sets of interrelated categories—complements and substitutes—from large assortments. The most commonly used exploratory techniques are proximity based methods like hierarchical clustering, and multi-dimensional scaling. However, both methods can lead to results which are difficult to interpret, especially in applications with, e.g., sparse data or presence of dominant categories (Mild and Reutterer 2003; Boztuğ and Silberhorn 2006).

More recently, marketing publications have proposed the use of social network graphs to uncover interrelations between products and brands (Netzer et al. 2012; Lee and Bradlow 2011). These studies have shown that network analysis techniques were capable of dealing with large data, and particularly underlined the *value of visualizing* extracted patterns in a comprehensible way. In fact, one of the most prominent arguments for using network graphs is that the approach facilitates 'eyeballing' patterns within the data, and allows the analyst to interpret even highly complex data structures (Newman 2003). Correspondingly, the authors have mentioned the possibility of *zooming* into specific relations, which quite literally referred to taking a closer look into relations of connected categories. For instance, Netzer et al. (2012) have shown that zooming into connections between brands could provide information about which dimensions are commonly used, when consumers evaluate and compare brands. However, the zooming-in step can become quite cumbersome, because relations typically have to be evaluated one by one manually. Hence, although techniques offer quite rich information, the generated results remain at a rather aggregated level in practice.

In the present paper, we present a framework of post-processing techniques for association rule mining (Agrawal et al. 1993). We discuss various tools for the graphical representations of association rules, which are able to capture the interrelations between categories in great detail. Furthermore, we introduce post-hoc clustering of association rules. We argue that our proposed framework is capable of capturing patterns beyond the mere coincidence of products categories. In particular, our method facilitates contrasting *entire sets* and *subsets* of complement and substitute categories. As clustering large numbers of association rules leads to a intricate hierarchical structure of results, we introduce a new interactive visualization technique—the *grouped matrix* representation (Hahsler et al. 2015)—which allows to intuitively explore such complex scenarios. Furthermore, we point to the implementation of the proposed methods in open source software packages **arules** and **arulesViz**.

The remainder of this paper is structured as follows. In Sects. 2 and 3, we provide a review of the theory on association rule mining, and discuss common visualization

techniques for association rules. In Sect. 3.2, we demonstrate how network graphs can be used to short-list and interpret the most critical patterns within complex data sets. In Sect. 4, we introduce post-hoc clustering, and the grouped matrix-based representation of association rules. Furthermore, we demonstrate how the method can be used to explore and understand vast amounts of extracted association rules. Finally, we provide an outlook on future applications of our proposed framework, and discuss managerial implications in Sect. 5.

2 Association rules and related techniques

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 et al. (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 $\mathcal{D} = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the *database*. Each transaction in $\mathcal{D}$ has a 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 $\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$. 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 σ and δ are the minimum support and minimum confidence, respectively. Note that both minimum support and minimum confidence are related to statistical concepts. Finding itemsets which surpass a minimum support threshold can be seen as a simplification of the unsupervised statistical learning problem called 'mode finding' or 'bump hunting' (Hastie et al. 2001), where each item is seen as a variable and the goal is to find prototype values so that the probability density evaluated at these values is sufficiently large. Minimum confidence can be interpreted as a threshold on the estimated conditional probability $P(Y|X)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS (seee.g., Hipp et al. 2000).

Another popular measure for association rules used throughout this paper is *lift* (Brin et al. 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 ($\gg 1$) 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 example, let us assume that we find the rule {milk, bread} $\Rightarrow$ {butter} with support of 0.2, confidence of 0.9 and lift of 2. Now we know that 20 % of all transactions contain all three items together, the estimated conditional probability of seeing butter in a transaction under the condition that the transaction also contains milk and bread is 0.9, and we saw the items together in transactions at double the rate we would expect under independence between the itemsets {milk, bread} and {butter}. For a more detailed treatment of association rules we refer the reader to the introductory paper for the R-package **arules** (Hahsler et al. 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, users increase the minimum support threshold σ to keep the number of association rules found at a manageable size. However, this has the disadvantage that it removes potentially interesting rules with lower support. Therefore, the need to deal with large sets of association rules is unavoidable when applying association rule mining in a real setting.

3 Visualization techniques for association rules

Many researchers introduced visualization techniques like scatter plots (Bayardo and Agrawal 1999), mosaic and double decker plots (Hofmann et al. 2000), and parallel coordinates plots (Yang 2003) to analyze association rules (a thorough overview is provided by Bruzzese and Davino 2008). However, most existing visualization techniques are not suitable for displaying large sets of rules.

This paper introduces a new method called 'grouped matrix-based visualization' which is based on a novel way of creating nested groups of rules (more specifically antecedent itemsets) via clustering. The nested groups form a hierarchy which can be interactively explored down to the individual rule.

In the remainder of this section we discuss two popular approaches to visualizing association rules. First, we introduce matrix based visualization techniques, which

are related to the new method presented in this paper. Subsequently, we discuss graph-based techniques (Klemettinen et al. 1994; Rainsford and Roddick 2000; Buono and Costabile 2005; Ertek and Demiriz 2006), which can be used to visualize the most important extracted association rules using vertices and directed edges.

3.1 Matrix-based visualization

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 an 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

$$\mathcal{R} = \{\langle X_1, Y_1, \theta_1\rangle, \ldots, \langle X_i, Y_i, \theta_i\rangle, \ldots, \langle X_n, Y_n, \theta_n\rangle\}$$

where X_i is the antecedent, Y_i is the consequent and θ_i is the selected interest measure for the i-th rule, $i = 1, \ldots, n$. In $\mathcal{R}$ we identify the set of A unique antecedents and C unique consequent. We create a $A \times C$ matrix $\mathbf{M} = (m_{ac})$, $a = 1, \ldots, A$ and $c = 1, \ldots, C$, with one column for each unique antecedent and one row for each unique consequent. We populate the matrix by setting $m_{ac} = \theta_i$ where $i = 1, \ldots, n$ is the rule index, and a and c correspond to the position of X_i and Y_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.

Ong et al. (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 et al. 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 illustrate matrix-based visualization using the package **arulesViz** (Hahsler et al. 2015) for the R software for statistical computing, an extension for the package **arules** (Hahsler et al. 2015). For illustration of the presented methods, we use the "Groceries" data set which is included in the **arules** package.

Groceries contains sales data from a local grocery store with 9835 transactions and 169 items (product groups). The data sets most popular item is 'whole milk' and the average transaction contains less than 5 items. Next we mine association rules using the Apriori algorithm implemented in **arules**. We use $\sigma = 0.001$ and $\delta = 0.5$, which results in a set of 5668 association rules. The rules contain 4097 unique antecedent, and 25 unique consequent itemsets (see Table 1). The top three rules with respect to the lift measure are presented in Table 2.

Table 1 Example for extracted antecedent and consequent itemsets

	Itemsets in antecedent (lhs)		Itemsets in consequent (rhs)
1	{honey}	1	{whole milk}
2	{tidbits}	2	{rolls/buns}
3	{cocoa drinks}	3	{other vegetables}
4	{pudding powder}	4	{bottled beer}
5	{cooking chocolate}	5	{root vegetables}
.	.	.	.
.	.	.	.
.	.	.	.
4096	{tropical fruit, other vegetables, whole milk, yogurt, rolls/buns}	24	{pastry}
4097	{root vegetables, other vegetables, whole milk, yogurt, rolls/buns}	25	{beef}

Table 2 Top three rules with respect to lift

Itemsets in antecedent (lhs)	Itemsets in consequent (rhs)	Support	Confidence	Lift
{Instant food products, soda}	{hamburger meat}	0.001220132	0.6315789	18.99565
{soda, popcorn}	{salty snack}	0.001220132	0.6315789	16.69779
{flour, baking powder}	{sugar}	0.001016777	0.5555556	16.40807

These rules represent easy to explain purchasing patterns. However, it is clear that going through all the 5668 rules manually is not a viable option. Therefore, we create a matrix-based visualization using shaded squares and 3D bars. The resulting plots are shown in Figs. 1 and 2.

Since there is not much space for long labels in the plot, we only show numbers as labels for rows and columns (x and y-axis) and the complete itemsets are printed to the terminal for look-up.

The visual impression can be improved by reordering rows and columns in the matrix such that rules with similar values of the interest measure are presented closer together. This removes some of the fragmentation in the matrix display and therefore makes it easier to see structure. In the resulting plot in Fig. 3 we see the emergence of two large blocks of rules with two different consequents and then smaller blocks for the rest. Obviously matching the labels to the entries on the x and y-axis is cumbersome. In order to be able to print the complete labels on the axes we would have to reduce the number of rules significantly to typically less than 100 rules. Alternatively, rules in the plot can be interactively selected to reveal the rule's antecedent and consequent itemsets, but the plot is so crowded, that it is almost impossible to select a specific rule. Hahsler et al. (2015) discussed several

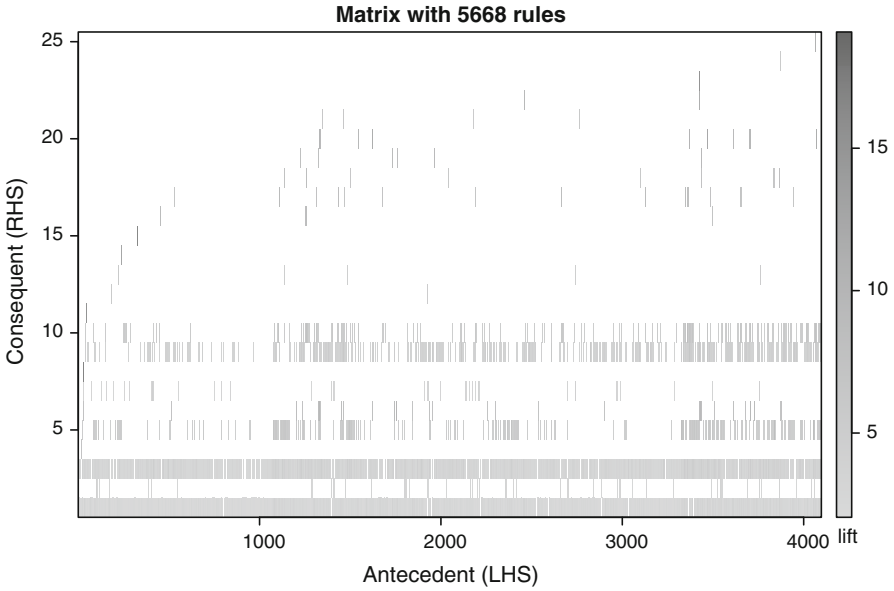

Fig. 1 Matrix-based visualization with colored squares

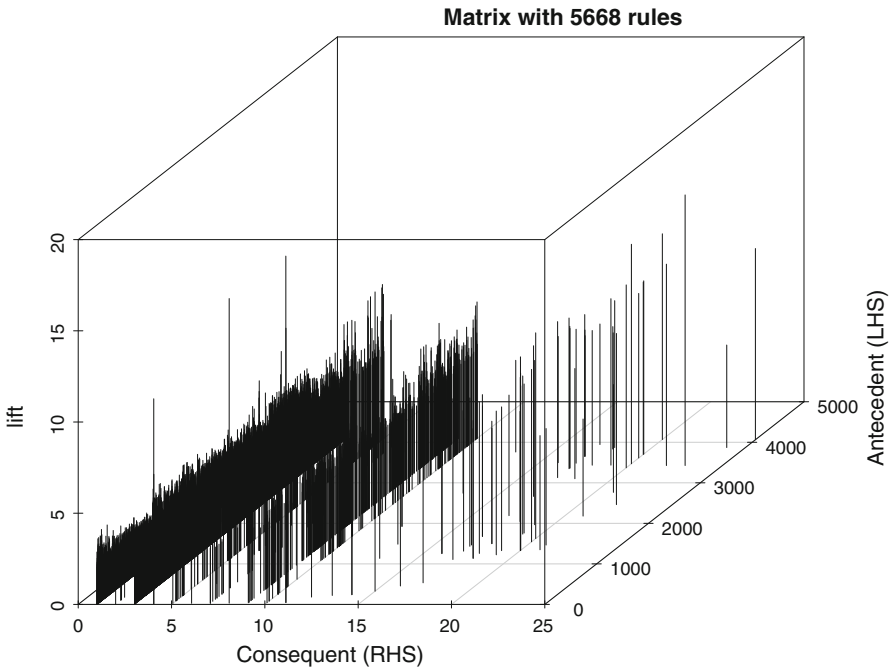

Fig. 2 Matrix-based visualization with 3D bars

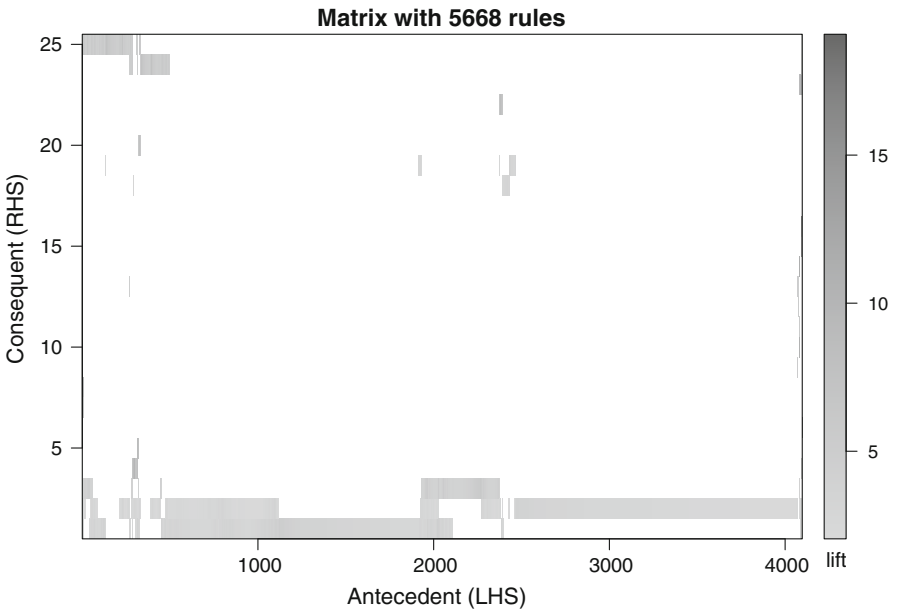

Fig. 3 Reordered matrix-based visualization with colored squares

reordering strategies to improve the plots usefulness for large number of rules, but only with very limited success. Hence, this illustration clearly shows that, even with reordering, the usefulness of simple matrix-based visualization is very limited when facing large rule sets.

3.2 Graph-based visualizations

Graph-based visualization is particularly well suited when the analyst is interested in an aggregated perspective on the most important rules. Graph-based techniques (Klemettinen et al. 1994; Rainsford and Roddick 2000; Buono and Costabile 2005; Ertek and Demiriz 2006) visualize association rules using vertices and edges, where vertices typically represent items or itemsets and edges indicate relationships in terms of rules. Interest measures are typically added to the plot as labels on the edges or by color or width of the arrows displaying the edges. Hence, the method is closely related to recent marketing publications, in which network analysis techniques have been used to extract patterns from user generated content (Netzer et al. 2012; Lee and Bradlow 2011).

In the context of association rule mining, graph-based visualization techniques offer a very clear representation of rules for relatively small sets of most important rules, which can be easily selected based on their corresponding lift scores (see Sect. 2). Figure 4 presents the graph visualization for the most important extracted association rules. In the network graph, itemsets are represented as vertices, whereas

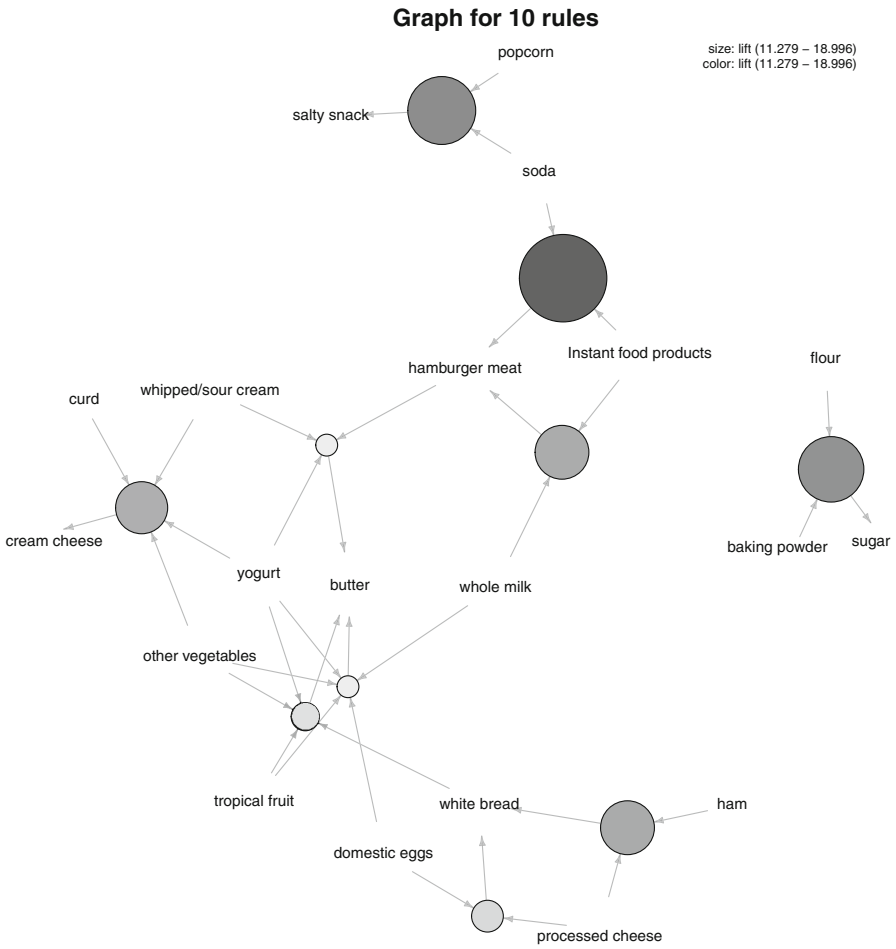

Fig. 4 Graph-based visualization with itemsets as vertices

rules are represented as directed edges between itemsets. For illustration purposes, we select the 10 rules with the highest lift scores.

We generated the presented network graph using the **arulesViz** package, which contains several graph-based visualization methods through interfaces to other network analysis packages. Specifically, graphs can be rendered using either the *igraph* library via the package **igraph** (Csardi and Nepusz 2006), or the the *GraphViz* software in the package **Rgraphviz** (Gentry et al. 2010). Note that the graph visualization presented in Fig. 4 provides a 'shortlist' of the most important association rules in a highly intuitive form.

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. 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 et al. 2009). From **arulesViz**, graphs for sets of association rules can be exported in formats which are compatible with other tools that allow interactive exploration of the extracted rules.

However, graph-based visualizations tend to easily become cluttered when graphs need to include larger sets of rules of interest. In order to cope with such settings, the next chapter introduces the grouped matrix-based visualization—a new technique which can be used in settings where the analyst is interested in in-depth exploration of a large number of extracted rules.

4 Grouped matrix-based visualization

As discussed in Sect. 3.1, traditional 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 and consequents. Therefore, in this chapter we introduce a new visualization techniques that enhances matrix-based visualization by grouping rules through k-means clustering, in order to handle large sets of rules. Groups of rules are presented by aggregating rows and columns of the matrix. The groups are nested and organized hierarchically allowing the analyst to explore them interactively by zooming into groups.

4.1 Clustering association rules

A direct approach to clustering itemsets (and rules) is to define a distance metric between two itemsets X_i and X_j. The distance between two sets can be measured, for example, by the Jaccard distance defined as

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

This distance is based on the number of items that X_i and X_j have in common divided by the number of unique items in both sets and was called for clustering association rules *conditional market-basket probability* by Gupta et al. (1999). For a set of m rules we can calculate the $m(m-1)/2$ distances between the sets of all items in each rule and use them as the input for clustering. However, using clustering on the itemsets creates several problems. First of all, data sets typically mined for association rules are high-dimensional, i.e., contain many different items. This high dimensionality carries over to the mined rules and leads to a situation referred to as the 'curse 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 for clustering association rules and itemsets to address the dimensionality and sparseness problem were proposed in the literature. Toivonen et al. (1995) and Berrado and Runger (2007) propose clustering association rules by

looking at the number of transactions which are covered by the rules. A transaction is covered by a rule if it contains all the items in the rule's antecedent. 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, two subsets of a frequent itemset will cover many common transactions. This bias will lead to just rediscovering the already known frequent itemset structure from the set of association rules.

We pursue a completely different approach. We start with the matrix $\mathbf{M}$ defined in Sect. 3.1, which contains the values of a selected interest measure of the rules in set $\mathcal{R}$. The columns and rows are the unique antecedents and consequents in $\mathcal{R}$, respectively. Now grouping rules becomes the problem of grouping columns or rows in $\mathbf{M}$.

Since for most applications the consequents in mined rules are restricted to a single item there is no problem with combinatorial explosion and we can restrict our treatment to only grouping antecedents (i.e., columns in $\mathbf{M}$). However, note that the same grouping method can be used also for consequents.

We use the interest measure lift, but other interest measures can be used as well. The idea behind lift is that antecedents that are statistically dependent on the same consequents (i.e., have a high lift value) are similar and thus should 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 a similar dependence relationship with the same consequents (e.g., bread). Note that clustering based on shared items or common covered transaction cannot uncover this type of relationship.

For grouping we propose to split the set of antecedents into a set of k groups $\mathcal{S} = \{S_1, S_2, \ldots, S_k\}$ while minimizing the within-cluster sum of squares

$$\text{argmin}_{\mathcal{S}} \sum_{i=1}^{k} \sum_{\mathbf{m}_j \in S_i} ||\mathbf{m}_j - \boldsymbol{\mu}_i||^2,$$

where $\mathbf{m}_j$, $j = 1, \ldots, A$, is a column vector representing all rules with the same antecedent and $\boldsymbol{\mu}_i$ is the center (mean) of the vectors in S_i. Minimizing the stated loss function is known as the k-means problem which is *NP-hard* (Aloise et al. 2009). However, several good and fast heuristics exist which do not require a precomputed distance matrix. We use the k-means algorithm by Hartigan and Wong (11979) and restart it 10 times with random initialized centers to find a suitable solution.

A challenge with using the k-means algorithm is that $\mathbf{M}$ contains many missing values for rules which are not included in $\mathcal{R}$ since they do not pass the minimum support or minimum confidence threshold. Since most values will be missing, marginalization (i.e., remove antecedents/consequents with missing values) is not an option and we use imputation. Imputation strategies typically assume that the values are missing randomly which is not the case here. Values miss in our case systematically when rules do not meet the support and confidence thresholds and

thus are deemed not interesting. This means that we would like to group antecedents when they have many missing values with the same set of consequents in common. To achieve this we replace all missing lift values with 1, a value indicating that antecedent and consequent of the rule are statistically independent. This ensures that matching missing values will contribute positively for grouping while it will help to separate them from existing rules with most likely larger lift values.

4.2 Visualizing grouped rules

To visualize the grouped matrix we use a balloon plot with antecedent groups as columns and consequents as rows (see Fig. 5). The color of each balloon represents the aggregated interest measure in the group and the size of the balloon shows the aggregated support. Aggregation in groups can be achieved by several aggregation functions (e.g., maximum, minimum, average, median). In the examples in this paper we use the median to represent the group since it is robust against outliers. The number of rules and the most important (frequent) items in the group are displayed as the labels for the columns followed by the number of other items in that

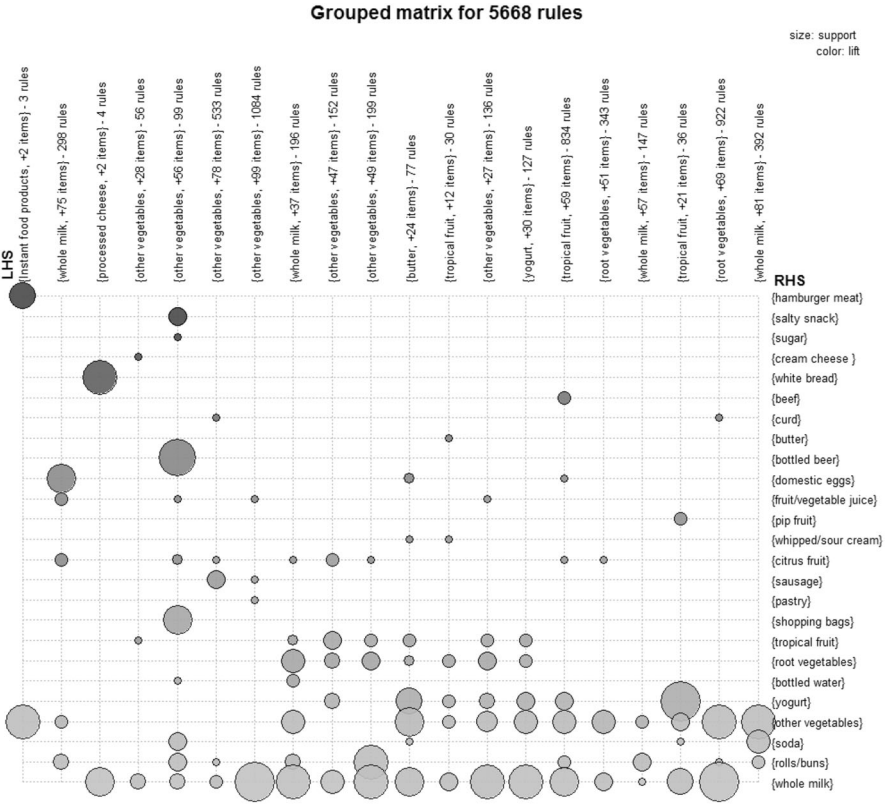

Fig. 5 Grouped matrix-based visualization

antecedent group. 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, directing the user to the most interesting group in the top left corner.

The resulting visualization for the 5569 rules used earlier with $k = 20$ groups is shown in Fig. 5. The group which contains the most interesting rules according to lift (which is the default measure) are shown in the leftmost column. The group contains 3 rules with two possible consequents, 'hamburger meat' and 'other vegetables'. The rules for 'hamburger meat' are stronger and are displayed in the top-left corner of the plot. The most frequent item in the LHS (antecedent) of the rules in the group can be found at the top end of the column and is 'Instant food products'. The antecedents of the rules also contain two additional items. The rules in the leftmost group are presented in Table 3. The first two rules with rather large lift values are represented in Fig. 5 by the upper balloon. While the third weaker rules is the second balloon in the figure.

To allow the user to explore the whole set of rules we can create a hierarchical structure of subgroups. This is achieved by creating a submatrix $\mathbf{M}_i$ for each group $S_i, i = 1, \ldots, k$, which only contains the columns corresponding to the elements in S_i. Now we can use the same grouping process again on a submatrix selected by the user. This allows the user to recursively 'drill down' into the rule set. An advantage of this process is that we only need to run the k-means algorithm on demand when the user wants to explore a group further.

The grouped matrix visualization can be used interactively to zoom into groups and inspect rules at a highly fine-granulated level. For example, Fig. 6 presents the interactive version zoomed into the 5th group of rules. This group contains 99 rules and the most common item in the antecedents is *other vegetables*. However, this is only the case because 'other vegetables' is a very frequent item in the data set. The header of the group in the plot also reveals that there are 56 other items in the antecedent of some of the rules in the group. After zooming in, we see that there are many subgroups with different antecedents (see Fig. 6). The subgroup with the highest lift (top-left corner) shows the strong relationship between 'soda' and 'salty snack'. The relationship with the highest support in the group (represented by the largest balloon) is between 'liquor' and 'bottled beer'. Most of these relationships are not surprising, but this is only the case because, for illustration purposes, we choose a data set where the relationship between most items are evident. For data sets with less well known relationships, the grouped matrix-based visualization with

Table 3 Example for inspecting antecedent groups

Itemsets in antecedent (lhs)	Itemsets in consequent (rhs)	Support	Confidence	Lift
{Instant food products, soda}	{hamburger meat}	0.001220132	0.631579	18.99565
{whole milk, Instant food products}	{hamburger meat}	0.001525165	0.500000	15.03823
{whole milk, Instant food products}	{other vegetables}	0.001525165	0.500000	2.58408

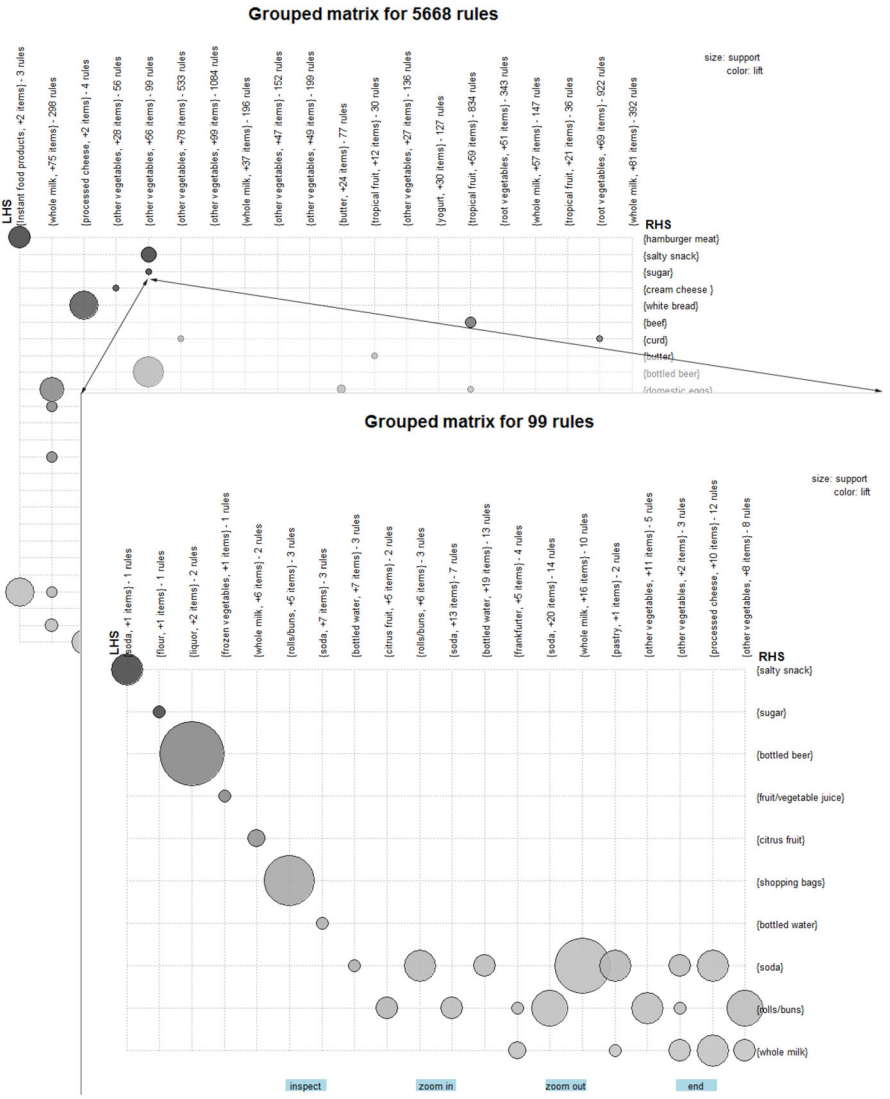

Fig. 6 Interactive grouped matrix-based visualization (zoomed into the 5th group in Fig. 5)

zooming into subgroups has the potential to enable marketers and analysts to discover previously unknown relationships faster and with less effort.

5 General discussion and implications

In the present paper, we contribute to the extant literature on exploratory data analysis by providing a highly flexible integrated framework for post-processing and visualization of association rules. We discussed tools for the graphical

representation of association rules, which summarize the *most critical* patterns in a highly parsimonious way. Furthermore, we introduce post-hoc clustering of association rules, which facilitates in-depth understanding of the underlying decision process which leads to specific compositions of shopping baskets. We argue that our proposed framework is capable of capturing patterns beyond the mere coincidence of products categories. In particular, our method facilitates contrasting *entire sets* and *subsets* and of complement and substitute categories. As clustering large numbers of association rules leads to a rather intricate hierarchical structure of results, we introduce a new interactive visualization method—the *grouped matrix* representation—which allows to explore such complex scenarios.

The method addresses the problem that sets of mined association rules are typically very large by grouping antecedents and allows the analyst to interactively explore a hierarchy of nested groups. Grouped matrix-based visualization is unique in the way that most other visualization methods (see Bruzzese and Davino 2008) are not able to efficiently deal with very large sets of association rules and that to our knowledge no other method can handle complementary categories.

From the marketing practitioner's perspective, our framework offers a number of additional benefits, which are crucial for modern marketing tools. First, association rule mining can handle extensive and highly complex data. We extend this property to capture and uncover highly detailed information about the nature of relations between rules (i.e. information about the underlying decision processes in the case of shopping basket analysis). Second, the interactive grouped matrix-based visualization is easy to use. Coloring and the position of elements in the plot almost automatically guide the analyst to the most interesting groups and rules. Finally, the presented approach is fully implemented in open source software packages for the R software for statistical computing. Hence, all presented methods, packages, and the corresponding documentation are free and easily accessible.

Interesting areas for future research include methodological contributions and applications to other areas of marketing research. From the methodological standpoint, it would be interesting to explore different ways to group antecedents and to look at grouping antecedents and consequents simultaneously (i.e., by co-clustering or two-mode clustering). Potential use-cases for follow-up studies include the extraction of patterns from user-generated content (textmining) and data on social interactions.

Acknowledgments Open access funding provided by Vienna University of Economics and Business (WU).

Appendix: Example R-code

We illustrate the discussed visualization techniques using the R-package **arulesViz**, an extension for package **arules** (Hahsler et al. 2015). For the examples in this paper we load the 'Groceries' data set which is included in **arules**.

```
> library("arulesViz")
> data("Groceries")
> Groceries
transactions in sparse format with
 9835 transactions (rows) and
 169 items (columns)
```

Next we mine association rules using the Apriori algorithm implemented in **arules**. We use $\sigma = 0.001$ and $\delta = 0.5$.

```
> rules <- apriori(Groceries, parameter=list(support=0.001, confidence=0.5), control=list(verbose=FALSE))
> rules
set of 5668 rules
```

The result is a set of 5.668 association rules. We inspect the top three rules regarding their lift score.

```
> inspect(head(sort(rules, by ="lift"),3))
    lhs                              rhs              support     confidence lift
53  {Instant food products,soda} => {hamburger meat} 0.001220132 0.6315789  18.99565
37  {soda,popcorn}               => {salty snack}    0.001220132 0.6315789  16.69779
444 {flour,baking powder}        => {sugar}          0.001016777 0.5555556  16.40807
```

Next, we create a matrix-based visualization using shaded squares, and 3D bars.

```
> plot(rules, method="matrix", measure="lift")
> plot(rules, method="matrix3D", measure="lift")
```

Note that in the resulting plots, labels for rows and columns (x and y-axis) are replaced by numbers and the complete itemsets are printed to the terminal for look-up. This may result in several thousand labels printed to the console.

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

```
> subrules <- head(sort(rules, by="lift"), 10)

> subrules

set of 10 rules
```

The following plot represents itemsets as vertices and rules as directed edges between itemsets.

```
> plot(subrules, method="graph", control=list(type="items"))
```

Sets of association rules can be exported in the GraphML format or as a Graphviz dot-file to be further explored in external tools like Gephi.

```
> saveAsGraph(sort(subrules, by="lift"), file="rules.graphml")
```

Finally, the matrix visualization with grouped antecedents (*grouped matrix*) for the rules mined earlier can be created and explored interactively.

```
> plot(rules, method="grouped", control=list(k=20), interactive=TRUE)

Interactive mode.
```

References

Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on management of data, pp 207–216

Aloise D, Deshpande A, Hansen P, Popat P (2009) NP-hardness of Euclidean sum-of-squares clustering. Mach Learn 75:245–248. doi:10.1007/s10994-009-5103-0

Bastian M, Heymann S, Jacomy M et al (2009) Gephi: an open source software for exploring and manipulating networks. ICWSM 8:361–362

Bayardo RJ Jr, Agrawal R (1999) Mining the most interesting rules. In: KDD '99: proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 145–154

Berrado A, Runger GC (2007) Using metarules to organize and group discovered association rules. Data Mining Knowl Discov 14:409–431

Boztug Y, Reutterer T (2008) A combined approach for segment-specific market basket analysis. Eur J Oper Res 187:294–312

Boztuğ Y, Silberhorn N (2006) Modellierungsansätze in der Warenkorbanalyse im Überblick. Journal für Betriebswirtschaft 56:105–128

Breugelmans E, Boztug Y, Reutterer T (2010) A multistep approach to derive targeted category promotions. Marketing Science Institute Working Paper Series, p 10

Brin S, Motwani R, Ullman JD, Tsur S (1997) Dynamic itemset counting and implication rules for market basket data. In: SIGMOD 1997, proceedings ACM SIGMOD international conference on management of data, Tucson, Arizona, USA, pp 255–264

Bruzzese D, Davino C (2008) Visual mining of association rules. Visual data mining: theory. Techniques and Tools for Visual Analytics. Springer, New York, pp 103–122

Buono P, Costabile MF (2005) Visualizing association rules in a framework for visual data mining. In: From integrated publication and information systems to virtual information and knowledge environments, pp 221–231

Csardi G, Nepusz T (2006) The igraph software package for complex network research. Int J Complex Syst 1695:1–9
Day GS (2011) Closing the marketing capabilities gap". J Mark 75:183–195
Ertek G, Demiriz A (2006) A framework for visualizing association mining results. In: ISCIS, pp 593–602
Fader PS, Hardie BGS, Shang J (2010) Customer-base analysis in a discrete-time noncontractual setting. Mark Sci 29:1086–1108
Gentry J, Long L, Gentleman R, Falcon S, Hahne F, Sarkar D, Hansen K (2010) Rgraphviz: provides plotting capabilities for R graph objects, r package version 1.24.0
Gupta G, Strehl A, Ghosh J (1999) Distance based clustering of association rules. In: Intelligent engineering systems through artificial neural networks (Proceedings of ANNIE 1999). ASME Press, pp 759–764
Hahsler M, Chelluboina S (2015) arulesViz: visualizing association rules and frequent itemsets, R package version 1.0-4
Hahsler M, Buchta C, Grün B, Hornik K (2015) arules: mining association rules and frequent itemsets, R package version 1.3-1
Hahsler M, Grün B, Hornik K (2005) arules—a computational environment for mining association rules and frequent item sets. J Stat Softw 14:1–25
Hartigan JA, Wong MA (1979) A K-means clustering algorithm. Appl Stat 28:100–108
Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning (data mining, inference and prediction). Springer, New York
Hipp J, Güntzer U, Nakhaeizadeh G (2000) Algorithms for association rule mining—a general survey and comparison. SIGKDD Explor 2:1–58
Hofmann H, Siebes A, Wilhelm AFX (2000) Visualizing association rules with interactive mosaic plots. In: KDD, pp 227–235
Hruschka H (2012) Analyzing market baskets by restricted Boltzmann machines. OR Spectr 36:209–228
Klemettinen M, Mannila H, Ronkainen P, Toivonen H, Verkamo AI (1994) Finding interesting rules from large sets of discovered association rules. In: CIKM, pp 401–407
Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. J Mark Res 48:881–894
Leeflang PS, Verhoef PC, Dahlstrom P, Freundt T (2014) Challenges and solutions for marketing in a digital era. Eur Manag J 32:1–12
Manchanda P, Ansari A, Gupta S (1999) The 'shopping basket': a model for multicategory purchase incidence decisions. Mark Sci 18:95–114
Mild A, Reutterer T (2003) An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. J Retail Consumer Serv 10:123–133
Natter M, Reutterer T, Mild A, Taudes A (2007) Practice prize report—an assortmentwide decision-support system for dynamic pricing and promotion planning in DIY retailing. Mark Sci 26:576–583
Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: market-structure surveillance through text mining. Mark Sci 31:521–543
Newman M (2003) The structure and function of complex networks. SIAM Rev 45:167–256
Ong KH, Leong Ong K, Ng WK, Lim EP (2002) CrystalClear: active visualization of association rules. In: ICDM'02 international workshop on active mining AM2002
Rainsford CP, Roddick JF (2000) Visualisation of temporal interval association rules. In: IDEAL '00: Proceedings of the second international conference on intelligent data engineering and automated learning, data mining, financial engineering, and intelligent agents. Springer, New York, pp 91–96
Rooderkerk RP, Van Heerde HJ, Bijmolt TH (2013) Optimizing retail assortments. Mark Sci 32:699–715
Russell GJ, Kamakura WA (1997) Modeling multiple category brand preference with household basket data. J Retail 73:439–461
Toivonen H, Klemettinen M, Ronkainen P, Hatonen K, Mannila H (1995) Pruning and grouping discovered association rules. In: Proceedings of KDD'95
Wong PC, Whitney P, Thomas J (1999) Visualizing association rules for text mining. In: INFOVIS '99: proceedings of the 1999 IEEE symposium on information visualization. IEEE Computer Society, Washington, DC, p 120
Yang L (2003) Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates. In: Computational science and its applications—ICCSA 2003. Lecture Notes in Computer Science, pp 21–30