Expert Systems With Applications 74 (2017) 70–81

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Construction and evaluation of structured association map for visual exploration of association rules

Jun Woo Kim

Department of Industrial and Management Systems Engineering, Dong-A University, Busan, Korea

ARTICLE INFO

Article history:
Received 27 December 2015
Revised 6 January 2017
Accepted 7 January 2017
Available online 9 January 2017

Keywords:
Visual exploration
Data mining
Association rule mining
Hierarchical clustering
Structured association map
Health examination

ABSTRACT

The association rule mining is one of the most popular data mining techniques, however, the users often experience difficulties in interpreting and exploiting the association rules extracted from large transaction data with high dimensionality. The primary reasons for such difficulties are two-folds. Firstly, too many association rules can be produced by the conventional association rule mining algorithms, and secondly, some association rules can be partly overlapped. This problem can be addressed if the user can select the relevant items to be used in association rule mining, however, there are often quite complex relations among the items in large transaction data. In this context, this paper aims to propose a novel visual exploration tool, structured association map (SAM), which enables the users to find the group of the relevant items in a visual way. The appearance of SAM is similar with the well-known cluster heat map, however, the items in SAM are sorted in more intelligent way so that the users can easily find the interesting area formed by a set of associated items, which are likely to constitute interesting many-to-many association rules. Moreover, this paper introduces an index called S2C, designed to evaluate the quality of SAM, and explains the SAM based association analysis procedure in a comprehensive manner. For illustration, this procedure is applied to a mass health examination result data set, and the experiment results demonstrate that SAM with high S2C value helps to reduce the complexities of association analysis significantly and it enables to focus on the specific region of the search space of association rule mining while avoiding the irrelevant association rules.

1. Introduction

As the living standards have improved, the preventive healthcare became a major public concern, and the mass health examination (MHE) plays an important role in monitoring and evaluating the individuals' health levels (Barnes et al., 2006). The primary purposes of the MHE are early detection and prevention of disease. In addition, the MHE result data provides a robust foundation for developing health policies or strategies, both at the national and individual levels (Kweon et al., 2014; Oyebode and Mindell, 2014).

However, analyzing the data sets collected from the MHE is difficult task, since they comprise many interrelated variables. Consequently, the MHE result data sets are often used to obtain simple descriptive statistics, though the additional useful information can be included within them. In this context, the initial goal of this research was to develop an effective exploration tool for the MHE result data sets.

The conventional MHE result data sets contain many binary variables indicating the occurrences of the specific symptoms, diseases or behaviors, and the author particularly focused on exploring the cause-and-effect relations among them. In other words, the MHE result data sets can be viewed in part as transaction data where various symptoms, diseases and behaviors represents the discrete items, and the author aimed to mitigate the difficulties faced by analyzers in exploring the association rules within high dimensional transaction data.

Let i_1, i_2, i_3, …, i_m denote the binary variables indicating m distinct items, and a transaction is denoted by a row vector [i_1 i_2 i_3 … i_m] where $i_k = 1$ (k = 1, 2, 3, …, m) if and only if the item k is included in the transaction. Then, the association rules take the form "X → Y " where X and Y are non-empty itemsets and they indicate that "If the values of i_xs (x ∈ X) are 1, then the values of i_ys ($y \in$ X) are also likely to be 1". Moreover, the X and Y are called the antecedent and the consequent of the association rule, respectively (Agrawal & Srikant, 1994; Agrawal, Imieliń ski, & Swami, 1993; Tan, Steinbach, & Kumar, 2005).

In general, the "usefulness" of an association rule can be measured by using the interestingness measures such as support, confidence and lift, etc., and the goal of association rule mining is to discover all useful association rules that have high support and confidence (Ju, Bao, Xu, & Fu, 2015; Tan et al., 2005). Due to the

E-mail address: kjunwoo@dau.ac.kr

http://dx.doi.org/10.1016/j.eswa.2017.01.007
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

repetitive data scan, this procedure can be time-consuming and the association rule mining algorithms such as well-known Apriori and its variants focused on extracting the useful association rules in efficient manner (Chen, Cai, Song, & Zhu, 2011; Kotsiantis & Kanellopoulos, 2006; Liu, Zhai, & Pedrycz, 2012). Another important issue in association rule mining is that such algorithms can find too many rules including redundant ones, and it is difficult for the human beings to interpret and exploit the plethora of the rules (Gu et al., 2003; Kim, 2015; Liu, Hsu, & Ma, 1999). Therefore, mining association rules from high dimensional data can be very challenging, and the visualization techniques has recently emerged as a promising tool for addressing this issue (Djenouri, Drias, & Bendjoudi, 2014; Lent, Swami, & Widon, 1997; Liu et al., 1999; Sekhavat & Hoeber, 2013). In general, the visualization techniques for the association analysis are used to represent the association rules in visual forms such as tables, charts and graphs, etc. Especially, the visualization techniques can provide significant aids to the analyzer if he or she is not an expert in data analysis.

This paper introduces a novel visualization method called structured association map (SAM) for summarizing the relations among the binary variables within high dimensional transaction data. The proposed method is based on the conventional data mining techniques such as association rule mining and cluster analysis, and it enables the users to easily find the interesting area formed by a set of associated binary variables, which are likely to constitute interesting many-to-many association rules. In order to avoid the irrelevant association rules, the users can choose an interesting area to be explored by association rule mining algorithms. That is, SAM enables to focus on the relevant items while filtering out the other ones. In this context, SAM provides significant benefits to the association rule based intelligent systems for recommendation or diagnosis, which have to find the consequent items associated with the given antecedent ones (Liao, Chu, & Hsiao, 2012). For example, SAM can reduce the execution time of the decision making procedure if the association rules have to be extracted in on-line manner, while it relieves the burden of maintaining a large amount of association rules in the systems such that extracts the association rules in off-line manner. Also, the visualization techniques such as SAM can contribute to reducing the development cycles of inductive expert systems in that they facilitate effective acquisition of domain knowledge (Kerdprasop & Kerprasop, 2014).

Fundamentally, SAM represents the relations among the binary variables (items) in the form of the data matrix combined with the item dendrograms. This structure is very similar with the traditional cluster heat map (Wilkinson & Friendly, 2009), however, SAM generates the item dendrograms in more suitable way for exploring the association rules. Although this idea was roughly introduced in the previous work (Kim, 2015), this paper provides refined concepts and enhanced procedures required to create and utilize SAM: (i) The items are classified into two groups, factor items and response items, in order to analyze the relations among them more systematically. (ii) Various ordering methods are incorporated into the item dendrogram construction procedure, while SAM was constructed in ad-hoc way in previous work. (iii) A quality measure which can be used to obtain an optimized SAM is developed. (iv) SAM based association rule mining procedure that adopts the concept of interesting area is proposed.

The remainder of this paper is organized as follows: Section 2 provides a brief literature review on the visualization techniques for the association rule mining. Section 3 is devoted to the construction procedure of the SAM, which is followed by the evaluation method for the visualization results. Section 4 provides the experiment results in the context of the case study with MHE data set and discusses the key findings. Finally, the concluding remarks and the future research directions are given in Section 5.

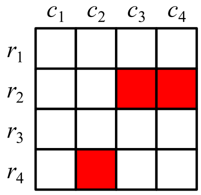

Fig. 1. Matrix based visualization for the items in transaction data.

2. Visualization techniques for association rule mining

The conventional association rule mining algorithms find all association rules with high support and confidence, so too many rules can be discovered. Moreover, the antecedents and consequents of association rules are defined on the power set of the set of all items, and they represent many-to-many relationships among the items (Yang, 2005). It is pointed out that the visualization techniques can help to deal with such large volumes and complexities of the plethora of the rules (De Oliveira & Levkowitz, 2003; Keim, 2002).

The most common and simple way for representing a large number of the association rules is to list them in a table. This table-based view is used in many conventional data mining softwares due to its simplicity, and the rules in such tables are typically sorted by an interestingness measure such as confidence or lift. If too many association rules are discovered, however, the analyzers still have trouble in interpreting the list and finding the interesting rules from such tables (Romero, Luna, Romero, & Ventura, 2011; Sekhavat & Hoeber, 2013).

There are more sophisticated methods for providing some visual aids to the analyzers. First of all, a wide range of charts and visual representations can be used to summarize the relations among the items in transaction data (Bornelöv, Marillet, & Komorowski, 2014; Techapichetvanich & Datta, 2005). For example, the scatter plot can be used to represent the association rules as the points in the coordinate plane, where the coordinates of the rules are determined by their interestingness measures (Bayardo & Agrawal, 1999; Hahsler & Chellubonia, 2011a). In parallel coordinates, the association rules can be represented as the polygonal lines that intersect the multiple vertical axes which represent the associated items. Reducing the tangled intersections of the polygonal lines is the primary issue in the parallel coordinate based approaches (Buono & Constabile, 2005; Itoh, Kumar, Klein, & Kim, 2016; Usman and Usman, 2016; Yang, 2005; Yang, 2008). Also, there are several directed graph based approaches that use the nodes representing the items or itemsets and the directed edges pointing from the antecedents to the consequents. The main issue of such techniques is the layouts of the graphs (Buono & Constabile, 2005; Romero et al., 2011; Sekhavat & Hoeber, 2013).

Another prevailing approach for the visualizing the cause-and-effect relations among the items is the use of matrix with rows and columns representing the antecedent and consequent items, respectively. Each element (tile) of this matrix, e_{ij} specifies the interestingness of the association rule "$\{r_i\} \rightarrow \{c_j\}$" by using numeric values or different colors, where r_i is the ith antecedent item and c_j is the jth consequent item, etc (Hahsler & Chellubonia, 2011b; Lei et al., 2016; Sekhavat & Hoeber, 2013; Wong, Whitney, & Thomas, 1999). For example, Fig. 1 illustrates a simple matrix based visualization for the set of antecedent items $I_{row} = \{r_i | i = 1, 2, 3, 4\}$ and the set of consequent items $I_{col} = \{c_j | j = 1, 2, 3, 4\}$. Let us assume that each matrix tile represents the lift of the corresponding association rule and the red-colored tiles indicate high lift values. Then, it is straightforward that the association rules "$\{r_2\} \rightarrow \{c_3\}$", "$\{r_2\} \rightarrow \{c_4\}$" and "$\{r_4\} \rightarrow \{c_2\}$" are interesting in Fig. 1.

Table 1
Characteristics of conventional visualization approaches for association rules.

Visualization approach		Suitability for many-to-many relationships	Ease of implementation	Ease of interpretation
Table based approach		Low	High	High
Graph based approach		Partly	Low	High
Parallel coordinate based approach		Partly	Partly	Low
Matrix based approach	Simple matrix	Low	High	High
	Cluster hear map	Partly	High	High

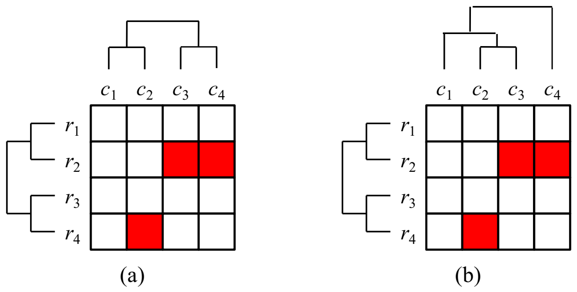

Fig. 2. The structure of cluster heat map.

Although the matrix based approaches provide quite intuitive visual aids, it is worth noting that only one-to-one relationships are explicitly represented by them, which is one of the most important limitations of the visualization methods for association rules. In more detail, the problem is that the analyzers tend to misunderstand that the Fig. 1 also suggests that the association rule "$\{r_2\} \to \{c_3, c_4\}$" is interesting, which might not be true. In other words, it is difficult to make any inferences about the interestingness of many-to-many association rule by using the traditional matrix based approaches.

In order to address this problem, the row and the column items should be ordered in appropriate manner so that the adjacent items have highly positive correlation. At the same time, the matrix based representations should convey some information about the "intra-correlations" within I_{row} or I_{col}, as well as the "inter-correlations" between I_{row} and I_{col} represented by e_{ij}. These objectives can be partially achieved by using the well-known cluster heat map. The cluster heat map can be viewed as the data matrix combined with the dendrograms showing intra-correlations (Wilkinson & Friendly, 2009), and Fig. 2 shows the basic structure of cluster heat map which enables the analyzers to get more insight into the many-to-many relationships. For example, the panel (a) of Fig. 2 suggests that the association rule "$\{r_2\} \to \{c_3, c_4\}$" is likely to be interesting, since the consequent items c_3 and c_4 are highly correlated in that they are merged at lower level in the dendrogram for consequent items. On the contrary, in the panel (b) of Fig. 2, the correlation between c_3 and c_4 is not strong and this suggests that the interestingness of the rule "$\{r_2\} \to \{c_3, c_4\}$" is probably not significantly high. The dendrograms can be generated by applying the conventional hierarchical clustering algorithms (Day & Edelsbrunner, 1984; Guenoche, Hansen, & Jaumard, 1991), and the cluster heat map based visualization techniques can provide two additional benefits, in comparison with the simple matrix based techniques: (i) The items are to some extent ordered according to the structures of the dendrograms. (ii) The dendrograms provide additional visual insight into the larger frequent itemsets and many-to-many relationships among them.

However, the dendrograms of conventional cluster heat map are inherently designed to represent the "intra-correlations" within I_{row} or I_{col} rather than many-to-many relationships between I_{row} and I_{col}, which must be considered in association analysis.

In Table 1, the characteristics of the important visualization approaches for association rules are compared from three perspectives: (i) If a many-to-many relationship can be appropriately implied by explicit connections or adjacencies of individual items, a visualization approach is suitable for representing many-to-many relationships. For example, table based approaches are typically not suitable for many-to-many relationships, since the itemsets must be directly connected in order to represent many-to-many relationships. (ii) A visualization approach is easy to implement if the entities required to visualize the association rules can be displayed in a straightforward manner. For example, matrix based approach can be easily implemented if the items are appropriately sequenced. On the contrary, the additional features such as positions and adjacencies of nodes must be carefully determined in graph based approach. (iii) A visualization approach should provide easy-to-interpret visual aids so that users can identify the interesting entities and understand their meaning in a convenient way.

In order to overcome the limitations of conventional approaches in Table 1, several researchers have recently proposed enhanced visualization techniques that enable the users to select some entities or items and examine them in more detail (Kosara et al., 2006; Itoh et al., 2016; Lei et al., 2016; Trevisan, Sanchez-Pi, Marti, & Garcia, 2015; Usman and Usman, 2016). However, such recent techniques are still based on conventional visualization approaches, not tailored to association rules, and users can have some difficulties in selecting entities or items to be investigated.

On the contrary, this paper proposes a novel visualization method called SAM, a variant of cluster heat map tailored for association rules. Since SAM is based on matrix based approach, it is easy to implement and interpret. Compared to classical matrix based techniques, however, SAM is more suitable for representing the many-to-relationships among the given items. Moreover, SAM provides a convenient way for selecting some interesting items to be investigated in more detail.

3. Structured association map

Fig. 3 depicts the overall SAM based association rule mining procedure, where the upper part describes the SAM construction phase while the lower part summarizes the SAM utilization phase. In construction phase, we can see that SAM is obtained by combining a matrix with two dendrograms constructed in different ways. Once created, SAM is used to visually identify the interesting areas and the interesting groups to be explored by association rule mining algorithms.

3.1. Item classification

Most association rule mining applications assume that a certain item within the given transaction data can appear either in the antecedents or in the consequents of the extracted association rules. On the contrary, the SAM classifies the items into two mutually exclusive categories, *factor items* and *response items*, and focuses on the association rules such that antecedents include only the former items while consequents consist of the latter ones. In

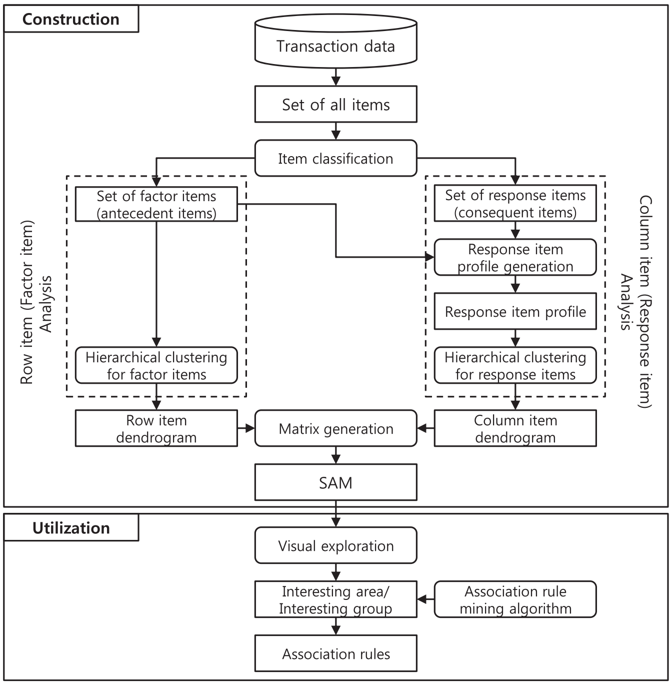
Construction
Transaction data
Set of all items
Item classification
Set of factor items (antecedent items)
Set of response items (consequent items)
Response item profile generation
Response item profile
Hierarchical clustering for factor items
Hierarchical clustering for response items
Row item (Factor item) Analysis
Column item (Response item) Analysis
Row item dendrogram
Matrix generation
Column item dendrogram
SAM
Utilization
Visual exploration
Interesting area/ Interesting group
Association rule mining algorithm
Association rules

Fig. 3. SAM based association rule mining procedure.

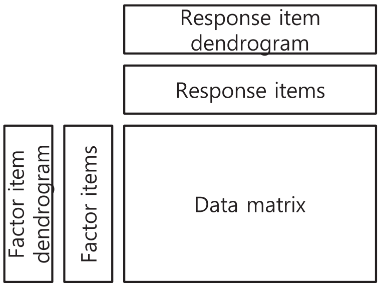

Fig. 4. Structure of SAM.

other words, the analyzers should select some items that affect other ones as the factor items, while the items affected by the factor items should be selected as the response items. Note that some items within the given transaction data may not be used with SAM. In addition, the terms ‘factor’ and ‘response’ came from the domain of quality management.

Let $I_F = \{F_1, F_2, \ldots, F_{m_f}\}$ and $I_R = \{R_1, R_2, \ldots, R_{m_r}\}$ denote the set of factor items and the set of response items, respectively. SAM focuses on the association rules of the form “X → Y” such that $X \subset I_F$ and $Y \subset I_R$, and Fig. 4 illustrates the basic structure of SAM, where a data matrix is combined with two dendrograms. In data matrix, each row represents a factor item, and each column represents a response item. The rows are combined with factor item dendrogram showing the intra-correlations among the factor items. Similarly, the response item dendrogram combined with the columns of data matrix shows the intra-correlations among the response items. Moreover, the factor/response item dendrograms are generated by *Row item (Factor item) analysis* and *Column item (Response item) analysis* in Fig. 3, which will be explained in more detail in the next sub-sections.

3.2. Row item (factor item) analysis

The objective of the row item analysis is to generate the factor item dendrogram by applying hierarchical clustering algorithms to the factor items. To this end, similarity or distance (dissimilarity) between two items must be defined. The *affinity* in (1) is the similarity measure for two items *a* and *b*, while the *Jaccard distance* in (2) can be used to measure the distance between them where sup(*X*) denotes the support of itemset *X* (Aggarwal, Procopius, & Yu, 2002; Gupta, Strehl, & Ghosh, 1999; Hahsler, Buchta, Grün, Hornik, & Borgelt, 2015). In this paper, the distance measure in (2) is used to generate the factor item dendrogram.

$$A(a, b) = \frac{\sup(\{a, b\})}{\sup(\{a\}) + \sup(\{b\}) - \sup(\{a, b\})} \qquad (1)$$

$$J_d(a, b) = 1 - A(a, b) \qquad (2)$$

Then, D_F, the square distance matrix for the factor items can be obtained as follows, where m_f is the number of the factor items and $df_{ij}=J_d(F_i, F_j)$. Note that $df_{ij} = 0$ if $i=j$ and $df_{ij} = df_{ji}$.

$$D_F = \begin{bmatrix} df_{11} & df_{12} & \ldots & df_{1m_f} \\ df_{21} & df_{22} & \ldots & df_{2m_f} \\ \ldots & \ldots & \ldots & \ldots \\ df_{m_f1} & df_{m_f2} & \ldots & df_{m_fm_f} \end{bmatrix} \qquad (3)$$

There are two prevailing approaches for hierarchical clustering analysis, agglomerative and divisive algorithms. In this paper, the agglomerative hierarchical clustering algorithm is applied to the distance matrix D_F in order to generate the dendrograms.

There are still two remaining issues regarding dendrogram generation. Firstly, the agglomerative clustering algorithm needs the linkage criteria for determining the distance between two clusters (itemsets). The commonly used linkage criteria are single-linkage (SL), complete-linkage (CL), average-linkage (AL) and Ward’s criterion (WC) (Tan et al., 2005). Secondly, a single dendrogram can be displayed in different forms, since a dendrogram can be modified by switching the positions of two sub-trees at any merge point without any change in the hierarchy within the dendrogram. In this paper, three methods for ordering the sub-trees of the dendrograms are considered: (i) OM1: the ordering method used in *hclust* function of the well-known data analysis software, R (Zhao, 2012). At each merge point, this function orders the sub-trees so that the left one has merged at lower level. (ii) OM2: GW-method, the ordering method proposed by Gruvaeus and Wainer (1972). At each merge point for two clusters, this method orders the sub-trees so that the distance between the closest items is minimized. In addition, the GW-method can be applied by using the *reorder.hclust* function in the *gclus* package of R (Hurley & Hurley, 2012). (iii) OM3: the last ordering method is based on the support measure. This is a simple top-down ordering method, which starts at the highest merge point. At each merge point, this method finds which sub-tree has an item with highest support, and put it on the left side (upper side) of the dendrogram. That is, this method makes the frequent itemsets to appear on the left side (upper side).

Note that we have 4 × 3 = 12 options for configuration of a single dendrogram, since 4 linkage criteria and 3 ordering methods are considered in this paper.

3.3. Column item (response item) analysis

Next, we have to generate the response item dendrogram by applying the agglomerative hierarchical clustering algorithm to the distance matrix for response items, D_R, which is obtained in a more sophisticated manner.

Let’s assume that both “$\{F_i\}\to\{R_j\}$” and “$\{F_i\}\to\{R_k\}$” are interesting association rules and we have to determine if a rule “$\{F_i\}\to\{R_j, R_k\}$” is likely to be interesting. Of course, $J_d(R_j, R_k)$ may be helpful in part, however, the distance measures such as Jaccard distance are typically based on the simple frequency of the co-occurrence of two items. This paper argues that more intelligent approach is required for such decisions, and proposes a response item profile based approach. What is important is that all factor items and their influences on the response items R_j and R_k should be considered in measuring the similarity or distance between those two response items. In other words, the distance between R_j and R_k is small if and only if they are similarly influenced by the factor items. In this context, the profile of a response item R_j is defined as follows:

$$PF(R_j) = [L_{1j} \quad L_{2j} \quad \dots \quad L_{m_f j}], \tag{4}$$

where L_{ij} is the influence of F_i on R_j. The interestingness measures of the rule "$\{F_i\}\rightarrow\{R_j\}$" can be used as L_{ij}, and this paper computes L_{ij} by using the lift measure as follows:

$$LIFT(\{F_i\} \rightarrow \{R_j\}) = \frac{\text{conf}(\{F_i\} \rightarrow \{R_j\})}{\text{sup}(\{R_j\})}, \tag{5}$$

where conf($\{F_i\} \rightarrow \{R_j\}$) is the confidence of the association rule "$\{F_i\}\rightarrow\{R_j\}$". Taking the distribution of the influences of the factor items into account, the distance between R_j and R_k, dr_{jk} is computed as follows:

$$dr_{jk} = 1 - \frac{PF(R_j) \cdot PF(R_k)}{\left|PF(R_j)\right| \times \left|PF(R_k)\right|}, \tag{6}$$

where $PF(R_j) \cdot PF(R_k)$ is the inner product of two profile vectors, $PF(R_j)$ and $PF(R_k)$, and $|PF(R_j)|$ is the length of $PF(R_j)$. Note that the second term in the right side of Eq. (6) is the cosine similarity between the two profile vectors. Finally, we have the distance matrix for the response items, D_R as follows:

$$D_R = \begin{bmatrix} dr_{11} & dr_{12} & \dots & dr_{1m_f} \\ dr_{21} & dr_{22} & \dots & dr_{2m_f} \\ \dots & \dots & \dots & \dots \\ dr_{m_f 1} & dr_{m_f 2} & \dots & dr_{m_f m_f} \end{bmatrix} \tag{7}$$

The two additional issues mentioned in previous sub-section, linkage criteria and ordering methods, are also relevant to the response item dendrogram, and we have 12 options for the response item dendrogram configuration.

3.4. Matrix generation

The objective of this step is to combine the dendrograms with the data matrix in order to obtain the completed SAM. At first, the factor/response items in rows/columns of the data matrix should be sorted according to the position in the associated dendrograms. Let $F_{(i)}$ denote the factor item at the ith row and $R_{(j)}$ denote the response item at the jth column. Then, the element e_{ij} of the data matrix corresponds to the association rule "$\{F_{(i)}\}\rightarrow\{R_{(j)}\}$", and e_{ij} should represent the interestingness measure of this rule. Similarly with the conventional matrix based visualization techniques, the values of interestingness measures, specific colors, and graphical icons can be assigned to e_{ij}. In this paper, the e_{ij}s are colored according to the lift value of corresponding association rules, and this will be illustrated in Section 4.

3.5. Evaluation of SAM

As mentioned above, 12 options for configuration of a single dendrogram are considered in this paper, and this means that we theoretically have 12 × 12 = 144 options for a single SAM combined with two dendrograms. Hence, we have to evaluate how well a single configuration of SAM summarizes the relations among the items. In this context, this paper proposes an interestingness based evaluation measure, the sum of the 2 × 2 rule contribution (S2C), which is based on the notion that an interesting 2 × 2 association rule should consist of the "adjacent" items.

Let $O_{row}(F_i)$ and $O_{col}(R_j)$ denote the index of the row indicating F_i and the index of column indicating R_j, respectively. Then, S2C measure suggests that if a 2 × 2 association rule "$\{F_a, F_b\}\rightarrow\{R_c, R_d\}$" ($a \neq b$ and $c \neq d$) is interesting, SAM should be constructed in a way that minimizes $|O_{row}(F_a) - O_{row}(F_b)|$ and $|O_{col}(R_c) - O_{col}(R_d)|$. For example, the association rule is easily identified from SAM if $|O_{row}(F_a) - O_{row}(F_b)| = |O_{col}(R_c) - O_{col}(R_d)| = 1$.

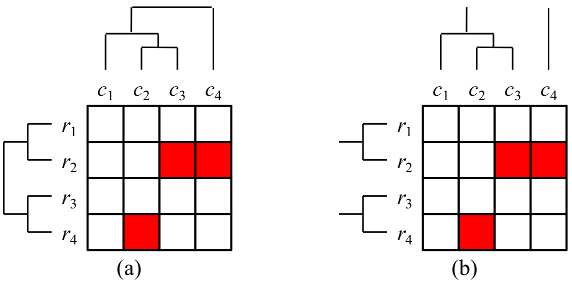

Fig. 5. SAM and reduced SAM.

In order to focus on the lower parts of dendrograms, the S2C measure is computed by using the reduced SAM obtained by deleting the upper α% of merge points ($0 \leq \alpha \leq 100$). What is important is that if "$\{F_a, F_b\}\rightarrow\{R_c, R_d\}$" is interesting, the factor items F_a and F_b should be connected in the partial factor item dendrogram of reduced SAM, and the response items R_c and R_d also should be connected in the partial response item dendrogram. For convenience, we set $\alpha = 50$ in this paper, and Fig. 5 illustrates the original SAM in panel (a) and the reduced SAM in panel (b). Note that the performance of poor SAM can be overestimated if α is too low, while too high α can lead analyzers to underestimate the performance of good SAM. In addition, if the number of merge points n_{merge} is odd, $(n_{merge} - 1)/2$ merge points are deleted in creating reduced SAM.

Next, the S2C measure is computed as follows:

$$S2C = \sum_{i=1}^{m_f - 1} \sum_{j=1}^{m_r - 1} CN(\{F_{(i)}, F_{(i+1)}\} \rightarrow \{R_{(j)}, R_{(j+1)}\}), \tag{8}$$

where $CN(\{F_{(i)}, F_{(i+1)}\} \rightarrow \{R_{(j)}, R_{(j+1)}\})$ is the contribution of the rule "$\{F_{(i)}, F_{(i+1)}\}\rightarrow\{R_{(j)}, R_{(j+1)}\}$". If both antecedent and consequent items are connected in dendrograms of the reduced SAM, this rule is *closed* and its contribution is computed as follows:

$$CN(\text{closed rule}) = LIFT(\text{closed rule}) \tag{9}$$

Note that while the lift is used to compute the contributions of the association rules in this paper, other interestingness measures also can be used. The CN(closedrule) in (9) indicates that the closed rules should have high interestingness values in that the analyzers tend to expect they are interesting rules.

If neither antecedent nor consequent items are connected in dendrograms of the reduced SAM, an association rule is *opened* and its contribution is obtained as follows:

$$CN(\text{opened rule}) = \begin{cases} LIFT(\text{opened rule})^{-1}, & \text{if } LIFT(\text{opened rule}) > 0 \\ M, & \text{otherwise} \end{cases} \tag{10}$$

The CN(openedrule) in (10) implies that the opened rules should have low interestingness values. Hence, the contribution of the opened rule is the reciprocal of its lift, if it has positive lift value. If the lift of an opened rule is 0, the contribution of the rule is set to an arbitrary value M, and $M = 1$ in this paper.

If either antecedent or consequent items are connected in the reduced SAM, this rule is *half-closed* and its contribution is 0. That is, the half-closed rules are not considered by the S2C measure. Consequently, the high values of S2C measure indicate that the SAM summarizes the relations among items in appropriate manner.

Table 2
Variables in MHE result data set.

Category	Sub-category	# of variables
Basic check-up		9
Dental check-up	Dental diseases	14
Dental inquiry	Subjective symptoms / Life styles	10
General inquiry	Subjective symptoms	29
	Life styles	24

3.6. SAM utilization

Once created, SAM can be used to visually explore the relations among the binary variables within the given transaction data. Especially, SAM helps the users find the interesting area to be investigated more thoroughly by applying association rule mining algorithm.

Let $S(a, b, p, q)$ denotes a $p \times q$ submatrix of SAM consisting of the rows indicating $F_{(a)}$, $F_{(a+1)}$, …, $F_{(a+p-1)}$ and the columns indicating $R_{(b)}$, $R_{(b+1)}$, …, $R_{(b+q-1)}$ where $p, q \geq 2$. $S(a, b, p, q)$ is called an interesting area if following two conditions are satisfied: (i) Almost all e_{ij}s in $S(a, b, p, q)$ indicates positive correlation between $F_{(i)}$ and $R_{(j)}$, or almost all e_{ij}s in $S(a, b, p, q)$ indicates negative correlation between $F_{(i)}$ and $R_{(j)}$ $(a \leq i \leq a+p-1, b \leq j \leq b+q-1)$. (ii) $F_{(a)}$, $F_{(a+1)}$, …, $F_{(a+p-1)}$ are merged at lower level in the factor item dendrogram, and $R_{(b)}$, $R_{(b+1)}$, …, $R_{(b+q-1)}$ are merged at lower level in the response item dendrogram.

If $S(a, b, p, q)$ is an interesting area, the set of items related to this submatrix, $G(a, b, p, q)$, is called interesting group where

$$G(a,b,p,q) = \{F_{(a)}, \quad F_{(a+1)}, \dots, F_{(a+p-1)}\} \cup \{R_{(b)}, R_{(b+1)}, \dots, R_{(b+q-1)}\} \tag{11}$$

In addition, an interesting area $S(a, b, p, q)$ and an interesting group $G(a, b, p, q)$ imply that the association rules of the form "$X \rightarrow Y$" such that $X \subset \{F_{(a)}, F_{(a+1)}, \dots, F_{(a+p-1)}\}$ and $Y \subset \{R_{(b)}, R_{(b+1)}, \dots, R_{(b+q-1)}\}$ are likely to be interesting. Hence, we should apply the conventional association rule mining algorithms to $G(a, b, p, q)$ so that the useful association rules are efficiently extracted from $S(a, b, p, q)$ while avoiding the irrelevant rules.

4. Application to MHE result data set

For illustration, the SAM based association rule mining procedure was applied to MHE result data set, and this section provides the experiment results and some discussions.

4.1. MHE result data set

The raw data is collected from a MHE for 278 adolescents in Korea. As shown in Table 2, the variables within the data are classified into 4 categories, *Basic check-up, Dental check-up, Dental inquiry* and *General inquiry*. Among them, the variables in the first two categories indicate the examinee's physical status or medical history, checked by the dedicated medical staffs. On the contrary, the variables in the other categories indicate the subjective symptom and personal life style, based on the examinee's statement about perceived health status.

Note that this MHE was designed with focus on the dental health, and the categories *Dental check-up* and *Dental inquiry* are the core part of the data summarized in Table 2. In this context, the author aimed to visualize the relations among the variables related to the personal dental health by using SAM. Moreover, this paper assumes that the *Life styles* variables in categories *Dental inquiry* and *General inquiry* affect both *Subjective symptoms* and *Dental diseases* variables, while the *Subjective symptom* variables have influences on the *Dental diseases* variables, as shown in Fig. 6. Therefore, the *Life styles* variables are selected as the factor items, while *Subjective symptoms* and *Dental diseases* variables are used as the response items, in order to analyze the association rules of the form "$\{I_{life}\} \rightarrow \{I_{symptom} \cup I_{disease}\}$" where I_{life}, $I_{symptom}$ and $I_{disease}$ are the sets of some *Life styles, Subjective symptoms* and *Dental diseases* variables.

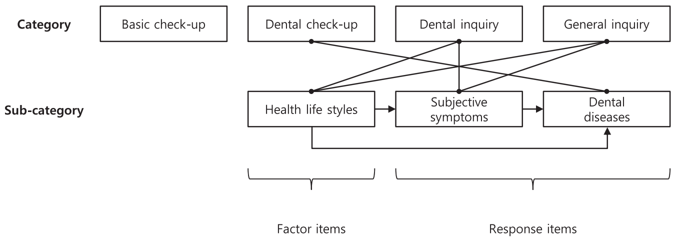

Fig. 6. Item classification for the MHE result data set.

Table 3
The factor items (life styles).

Item	Description
F1	Dental clinic visit during last year
F2	Preference for sugary foods or carbonated beverage
F3	Fluoride toothpaste usage
F4	Regular meal
F5	Preference for milk or milk products
F6	Preference for vegetables or fruits
F7	Refraining from sugary or salty foods
F8	Preference for fast foods
F9	Hand washing after going out
F10	Brushing teeth more than twice a day

Table 4
The response items (subjective symptoms and diseases).

Item	Description
R1	Dental caries
R2	Dental caries risk
R3	Oral hygiene
R4	Gingival bleeding
R5	Tartar
R6	Tooth fracture
R7	Dental pain triggered by cold or hot foods and beverages
R8	Dental pain
R9	Bleeding gums
R10	Glossalgia
R11	Halitosis

In more detail, 10 factor items and 11 response items were selected for SAM based association analysis, and they are listed in Tables 3 and 4, respectively. Note that most of them were inherently binary variables in the raw data set, while some of them have been binarized before SAM based association analysis. Moreover, several items, such as Stomatitis and Paradental cyst in *Dental check-up* category have been discarded since there support values are 0.

4.2. SAM construction and evaluation

Once the items are classified, the next step is to generate the factor item dendrogram. To this end, the agglomerative hierarchical clustering algorithm is applied to the distance matrix for factor items, D_F, obtained by using Eq. (2). Figs. 7 and 8 illustrate the row item dendrograms obtained from D_F shown in Table 5, and we can see that the structures and leaf orderings of the dendrograms vary according to the linkage criterion and the ordering method.

Then, we can generate the profiles of the response items by using (4)–(6), and these profiles are used to compute the distance matrix for the response items, D_R, shown in Table 6. Again, the

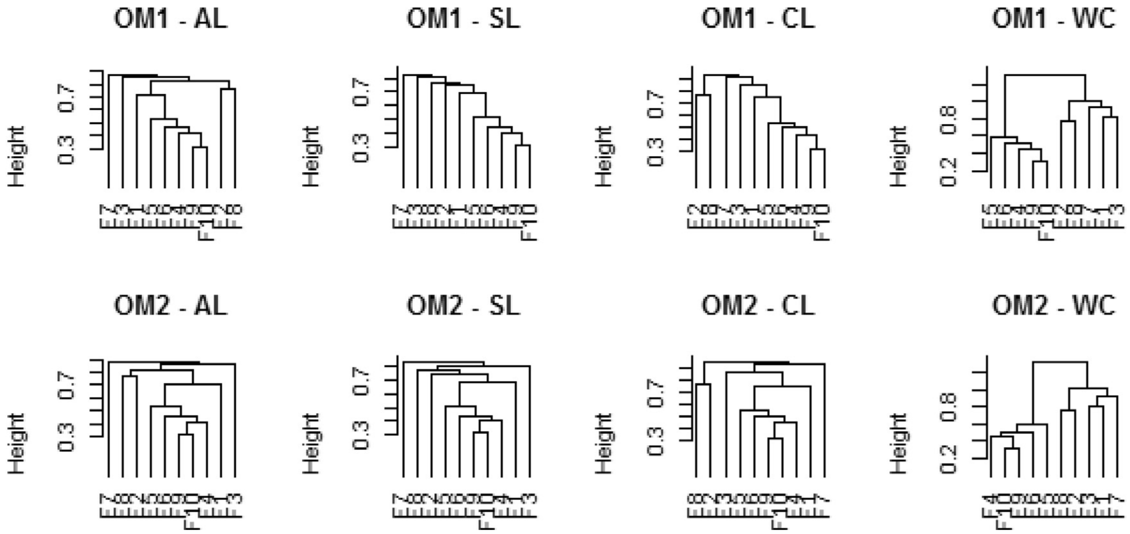

Fig. 7. Factor item (row item) dendrograms generated by using OM1 and OM2.

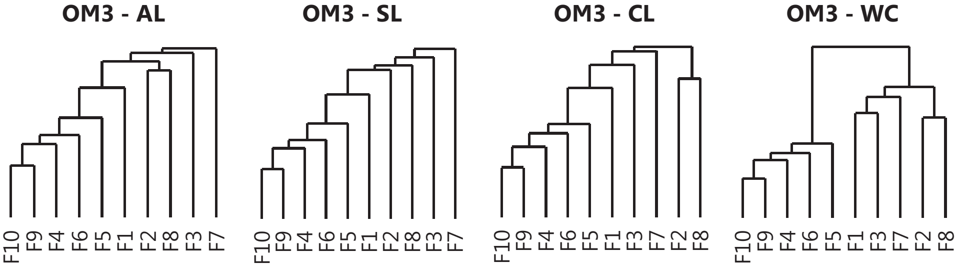

Fig. 8. Factor item (row item) dendrograms generated by using OM3.

Table 5
Distance matrix for the factor items.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
F1	0.0000	0.8201	0.8142	0.7171	0.7487	0.7245	0.8824	0.9048	0.6906	0.6838
F2	0.8201	0.0000	0.9074	0.8182	0.7614	0.8122	0.9459	0.7677	0.7841	0.7470
F3	0.8142	0.9074	0.0000	0.8639	0.8025	0.8409	0.9259	0.9302	0.8396	0.8381
F4	0.7171	0.8182	0.8639	0.0000	0.5392	0.5000	0.8324	0.8724	0.4375	0.4000
F5	0.7487	0.7614	0.8025	0.5392	0.0000	0.5098	0.8667	0.8148	0.5462	0.5285
F6	0.7245	0.8122	0.8409	0.5000	0.5098	0.0000	0.8400	0.8444	0.4397	0.4559
F7	0.8824	0.9459	0.9259	0.8324	0.8667	0.8400	0.0000	0.9167	0.8278	0.8374
F8	0.9048	0.7677	0.9302	0.8724	0.8148	0.8444	0.9167	0.0000	0.8372	0.8266
F9	0.6906	0.7841	0.8396	0.4375	0.5462	0.4397	0.8278	0.8372	0.0000	0.3144
F10	0.6838	0.7470	0.8381	0.4000	0.5285	0.4559	0.8374	0.8266	0.3144	0.0000

response item dendrograms are obtained by applying the hierarchical clustering algorithm, and Figs. 9 and 10 represent the consequent response item dendrograms with varying structures and leaf orderings.

Since we have 12 factor item dendrograms and the same number of response item dendrograms, 144 different SAMs can be constructed. The performances of them can be evaluated by using the S2C measure described in (8)–(10), and the evaluation results are summarized in Table 7 and Fig. 11. In Table 7, each row indicates the ordering method and the linkage criterion of a factor item dendrogram, while each column specifies the ordering method and the linkage criterion of a response item dendrogram.

An element in the *i*th row and *j*th column of Table 7 represents S2C value of a SAM constructed by combining *i*th factor item dendrogram and *j*th response item dendrogram, and we can see that the performance of SAM vary significantly according to the structures of the combined dendrograms. Among the 144 different SAMs listed in Table 7, we have to choose the SAM with OM2-WC\OM3-AL (factor item dendrogram structure \ response item dendrogram structure), since it maximizes the value of S2C measure. Consequently, we can obtain the optimized SAM shown in Fig. 12, where each tile e_{ij} is colored according to the lift value of the association rule "$\{F_{(i)}\} \rightarrow \{R_{(j)}\}$" as follows: (i) A blue-colored tile indicates lift value higher than 1 (positive correlation), while a red-colored one implies that lift value is lower than 1 (negative correlation). (ii) A deeper colored tile represents the stronger correlation, and a white-colored tile indicates lift value about 1 (week correlation between $F_{(i)}$ and $R_{(j)}$).

4.3. Visual exploration and utilization

Inherently, only information about one-to-one association rules can be explicitly displayed in the conventional data matrix, however, SAM is designed to provide some additional insights on many-to-many association rules. Especially, SAM enables the analyzers to visually identify the interesting areas, helpful for extracting the useful association rules while avoiding the irrelevant ones.

Table 6
Distance matrix for the response items.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11
R1	0.0000	0.8980	0.6163	0.8810	0.8511	0.7946	0.8031	0.8878	0.8475	0.9570	0.8393
R2	0.8980	0.0000	0.8750	0.9143	0.9167	0.9178	0.9091	0.9608	0.8889	0.8611	0.9118
R3	0.6163	0.8750	0.0000	0.8605	0.7885	0.8608	0.8737	0.9333	0.8625	0.9600	0.8961
R4	0.8810	0.9143	0.8605	0.0000	0.8485	0.9683	0.9750	0.9737	0.9000	0.9200	0.9474
R5	0.8511	0.9167	0.7885	0.8485	0.0000	0.9605	0.9091	0.8958	0.9041	0.9750	0.9429
R6	0.7946	0.9178	0.8608	0.9683	0.9605	0.0000	0.7941	0.8732	0.8737	0.9206	0.9022
R7	0.8031	0.9091	0.8737	0.9750	0.9091	0.7941	0.0000	0.7875	0.8077	0.9241	0.7835
R8	0.8878	0.9608	0.9333	0.9737	0.8958	0.8732	0.7875	0.0000	0.8088	0.8333	0.8308
R9	0.8475	0.8889	0.8625	0.9000	0.9041	0.8737	0.8077	0.8088	0.0000	0.9219	0.8276
R10	0.9570	0.8611	0.9600	0.9200	0.9750	0.9206	0.9241	0.8333	0.9219	0.0000	0.9138
R11	0.8393	0.9118	0.8961	0.9474	0.9429	0.9022	0.7835	0.8308	0.8276	0.9138	0.0000

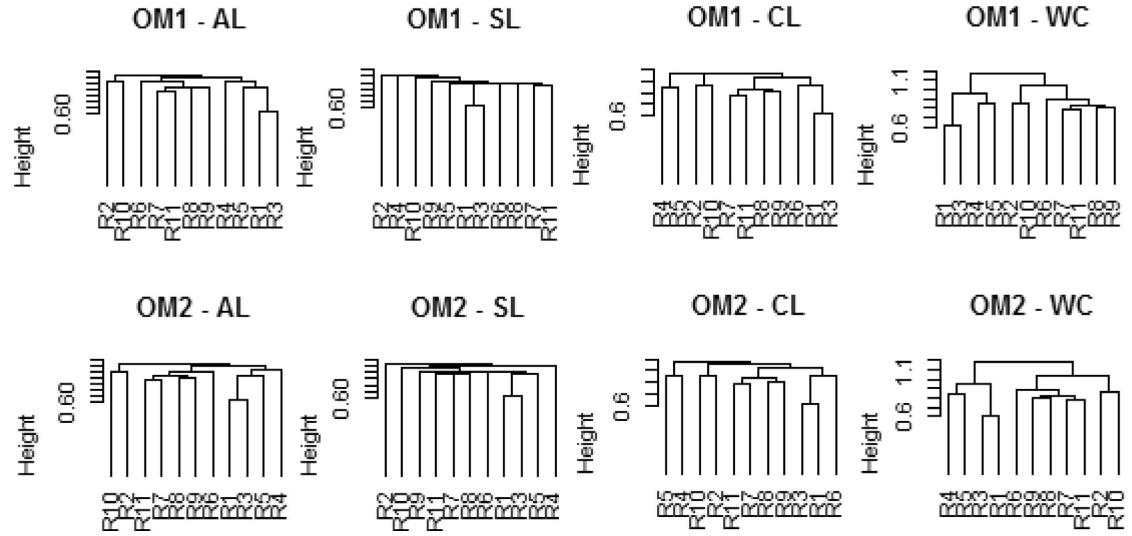

Fig. 9. Response item (column item) dendrograms generated by using OM1 and OM2.

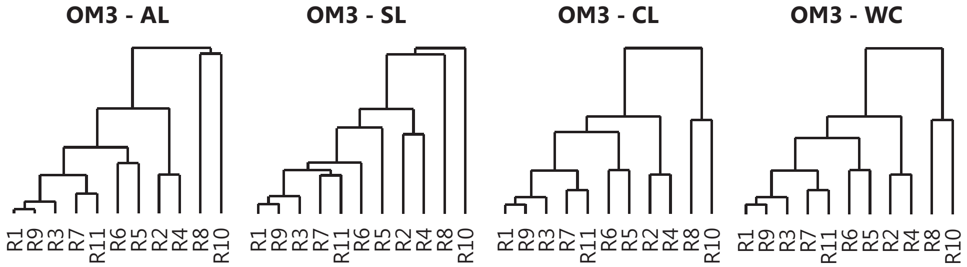

Fig. 10. Response item (column item) dendrograms generated by using OM3.

Table 7
S2C values of different SAMs.

Row	Col.											
	OM1-AL	OM1-SL	OM1-CL	OM1-WC	OM2-AL	OM2-SL	OM2-CL	OM2-WC	OM3-AL	OM3-SL	OM3-CL	OM3-WC
OM1-AL	38.46	38.40	37.28	38.10	40.37	37.07	38.46	38.46	40.71	37.41	36.87	36.87
OM1-SL	39.66	38.96	**34.46**	36.52	40.40	37.30	36.85	36.85	41.30	38.19	37.14	37.14
OM1-CL	39.38	38.87	36.98	37.95	41.15	37.85	38.77	38.77	41.40	38.11	37.71	37.71
OM1-WC	46.47	49.32	45.28	45.37	47.64	47.33	45.08	45.08	51.37	51.07	45.50	45.50
OM2-AL	40.31	40.18	37.81	39.03	40.12	39.24	39.30	39.30	41.03	40.15	38.78	38.78
OM2-SL	38.23	38.94	39.49	40.62	39.90	36.56	39.38	39.38	40.80	37.46	37.99	37.99
OM2-CL	40.11	40.15	39.08	40.33	40.84	38.74	39.91	39.91	41.75	39.65	38.47	38.47
OM2-WC	45.60	48.73	44.08	45.60	47.48	46.38	44.73	44.73	**51.42**	50.32	46.11	46.11
OM3-AL	39.66	38.96	**34.46**	36.52	40.40	37.30	36.85	36.85	41.30	38.19	37.14	37.14
OM3-SL	39.66	38.96	**34.46**	36.52	40.40	37.30	36.85	36.85	41.30	38.19	37.14	37.14
OM3-CL	38.51	37.93	36.98	37.95	41.15	36.91	38.77	38.77	41.40	37.17	36.97	36.97
OM3-WC	44.50	47.26	44.87	45.27	47.43	45.25	44.25	44.25	51.41	49.24	45.22	45.22
Mean = 40.59, Standard deviation = 3.93, Max. = 51.42, Min. = 34.46												

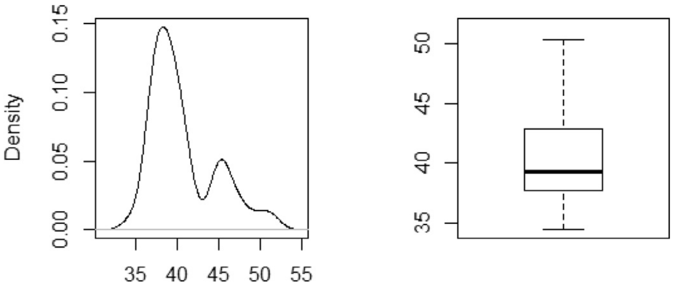

Fig. 11. The distribution of S2C values.

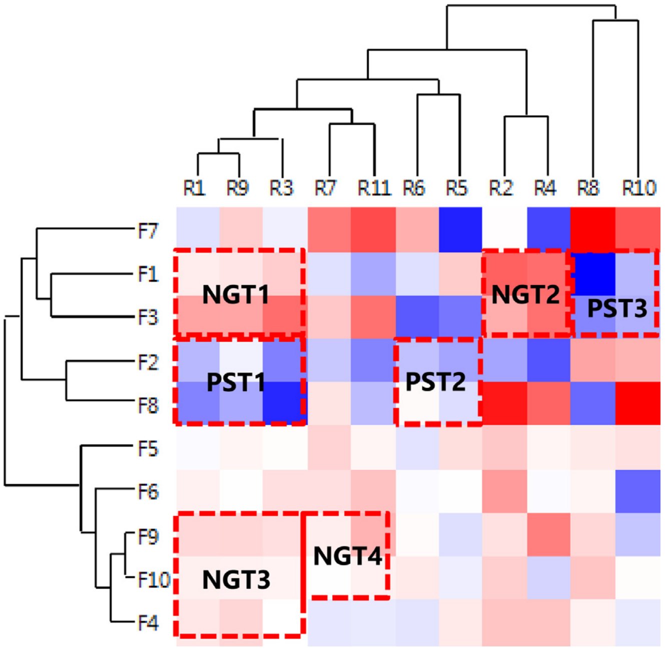

Fig. 12. SAM with the highest S2C (SAM with OM2-WC\OM3-AL).

Table 8
Positive association rules suggested by the SAM with the highest S2C.

Interesting area	Association rules	Lift
PST1	$\{F_2, F_8\} \rightarrow \{R_1, R_9\}$	2.6878
	$\{F_2, F_8\} \rightarrow \{R_1, R_3\}$	2.1980
	$\{F_2, F_8\} \rightarrow \{R_3, R_9\}$	3.2929
PST2	$\{F_2, F_8\} \rightarrow \{R_5, R_6\}$	**0.0000**
PST3	$\{F_1, F_3\} \rightarrow \{R_8, R_{10}\}$	2.2037

For example, we can identify several ‘positive’ interesting areas consisting of blue-colored tiles (PST1, PST2 and PST3) and ‘negative’ interesting areas consisting of red-colored tiles (NGT1, NGT2, NGT3 and NGT4) in Fig. 12 that shows an optimized SAM with the highest S2C value, 51.42.

In order to investigate the usefulness of the interesting area, the 2×2 association rules extracted from the 7 interesting areas shown in Fig. 12 are listed in Tables 8 and 9. In Table 8, we can see that almost 2×2 rules extracted from the positive interesting areas have quite high lift values indicating the strong positive correlations. One exception is the rule “$\{F_2, F_8\} \rightarrow \{R_5, R_6\}$”, of which lift is 0. Indeed, the lift values of the association rules “$\{F_2, F_8\} \rightarrow \{R_5\}$” and “$\{F_2, F_8\} \rightarrow \{R_6\}$” are 1.3947 and 1.3688, and this means that both R_5 and R_6 are affected by the factor variables F_2 and F_8. However, the variable R_5 (tartar) has quite low support ($=0.09$), and $\sup(\{F_2, F_8, R_5, R_6\}) = 0$. Therefore, an interesting area have to be carefully interpreted if it contains infrequent items such as R_5.

Table 9
Negative association rules suggested by the SAM with the highest S2C.

Interesting area	Association rules	Lift
NGT1	$\{F_1, F_3\} \rightarrow \{R_1, R_9\}$	0.0000
	$\{F_1, F_3\} \rightarrow \{R_1, R_3\}$	0.4010
	$\{F_1, F_3\} \rightarrow \{R_3, R_9\}$	0.0000
NGT2	$\{F_1, F_3\} \rightarrow \{R_2, R_4\}$	0.0000
NGT3	$\{F_9, F_{10}\} \rightarrow \{R_1, R_9\}$	0.7403
	$\{F_9, F_{10}\} \rightarrow \{R_1, R_3\}$	0.8378
	$\{F_9, F_{10}\} \rightarrow \{R_3, R_9\}$	**1.1161**
	$\{F_4, F_9\} \rightarrow \{R_1, R_9\}$	0.4575
	$\{F_4, F_9\} \rightarrow \{R_1, R_3\}$	0.7489
	$\{F_4, F_9\} \rightarrow \{R_3, R_9\}$	0.3741
	$\{F_4, F_{10}\} \rightarrow \{R_1, R_9\}$	0.5951
	$\{F_4, F_{10}\} \rightarrow \{R_1, R_3\}$	0.9181
	$\{F_4, F_{10}\} \rightarrow \{R_3, R_9\}$	0.4856
NGT4	$\{F_9, F_{10}\} \rightarrow \{R_7, R_{11}\}$	0.7318

Similarly, almost association rules in Table 9 have lift values lower than 1, indicating the negative correlations. Therefore, we can conclude that the optimized SAM shown in Fig. 12 provides appropriate visual supports for the analyzers, and the interesting areas deserve to be intensively explored.

In order to extract association rules concerning specific items, the analyzers can select an interesting area to be explored. The set of the items involved in the selected interesting area is called interesting group, and the conventional association rule mining algorithms can be used to extract the association rules from the reduced transaction data which consists of the items within the interesting group. In other words, SAM can be used as a feature selection method for association rule mining.

For example, Table 10 shows the interesting groups obtained from the positive interesting areas in Fig. 12. The first interesting group indicates that the sugary foods, carbonated beverage and fast foods are likely to cause dental caries, dental caries risk and bleeding gums. Similarly, the second interesting group implies that the preferences for sugary foods, carbonated beverage and fast foods have positive correlation with the occurrences of tartar and tooth fracture. Finally, the third interesting group suggests that the adolescents tend to take care of their dental health after perceived symptoms such as dental pain and glossalgia. In this manner, SAM also can be used to summarize the association rules indicating many-to-many relationships among the variables within given transaction data.

4.4. Comparisons of the visualization results

Next, Fig. 13 shows the SAM with OM1-SL\OM1-CL, one of the three SAMs with the lowest S2C value, 34.46, in Table 7. From Figs. 12 and 13, we can see that some interesting areas in the optimized SAM are also identified in a poor SAM with low S2C value. Note that PST1, PST2, NGT3 and NGT4 are identified in both Figs. 12 and 13. This means that a poor SAM can be helpful to some extent in association analysis. However, it is more difficult to visually identify PST1 and PST2 from Fig. 13, where the factor items F_2 and F_8 are merged at higher level in factor item dendrogram. On the contrary, the optimized SAM in Fig. 12 provides better visual aids for identifying PST1 and PST2 in that F_2 and F_8 are merged at lower level. Moreover, PST3, NGT1 and NGT2 are not included within Fig. 13. Therefore, we can conclude that a poor SAM with low S2C value is less helpful and we need to generate an optimized SAM which maximizes S2C measure.

In more detail, the performances of the ordering methods for the dendrograms are compared in terms of S2C measure in Table 11, and we can see that the factor item dendrograms based

Table 10
Interesting groups derived from Fig. 12.

Area	Factor items	Response items
PST1	F_2: Preference for sugary foods or carbonated beverage F_8: Preference for fast foods	R_1: Dental caries R_3: Dental caries risk R_9: Bleeding gums
PST2	F_2: Preference for sugary foods or carbonated beverageF_8: Preference for fast foods	R_5: Tartar R_6: Tooth fracture
PST3	F_1: Dental clinic visit during last year F_3: Fluoride toothpaste usage	R_8: Dental pain R_{10}: Glossalgia

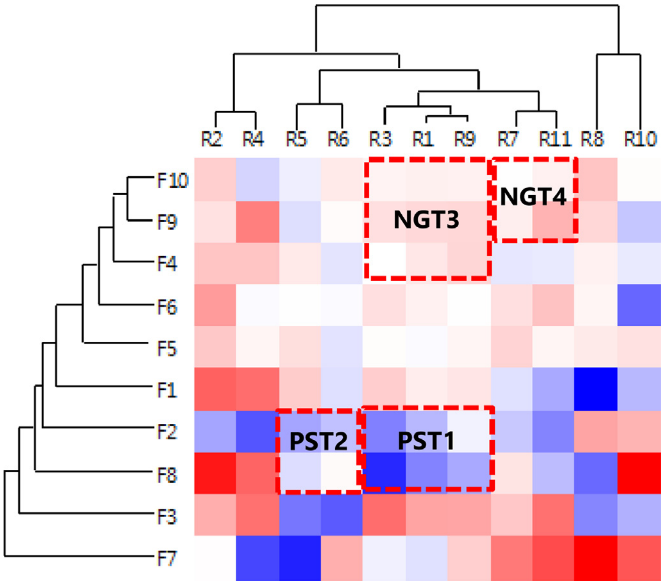

Fig. 13. SAM with the lowest S2C (SAM with OM1-SL\OM1-CL).

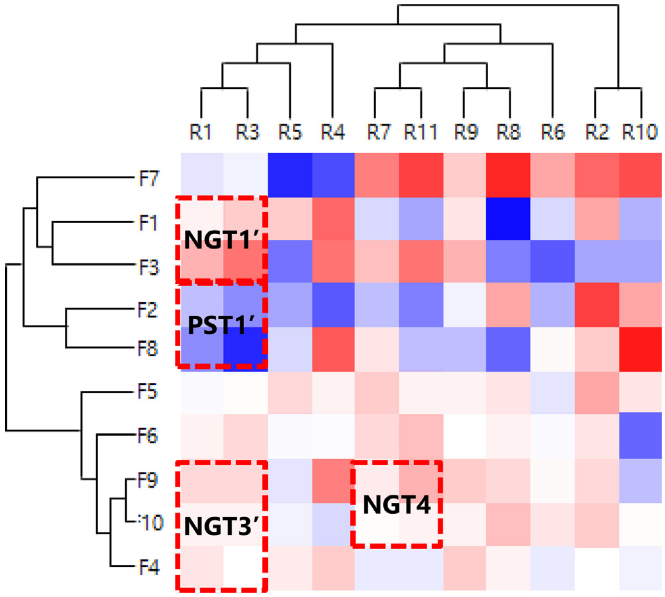

Fig. 14. Cluster heat map with response item dendrogram based on simple Jaccard distance.

Table 11
Comparisons of the ordering methods.

Ordering method	Factor item dendrogram			Response item dendrogram		
	Average	Max.	Min.	Average	Max.	Min.
OM1	40.48	51.37	34.46	40.25	49.32	34.46
OM2	**41.24**	**51.42**	**36.56**	40.48	47.64	36.56
OM3	40.07	51.41	34.46	**41.05**	**51.42**	**36.87**

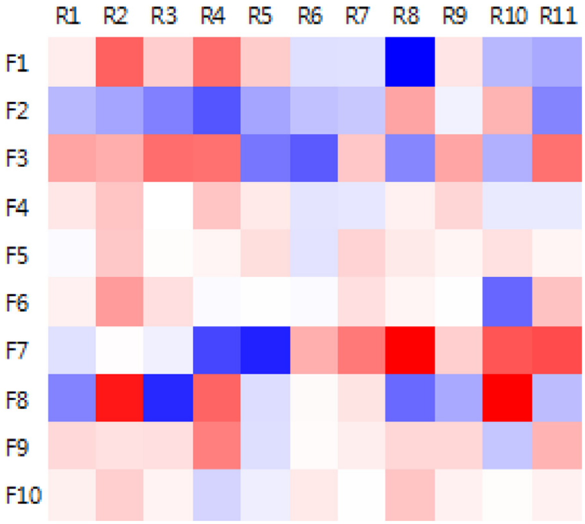

Fig. 15. Simple data matrix without dendrogram and item ordering.

on OM2 and response item dendrograms based on OM3 tend to produce better SAMs with higher S2C values.

On the other hand, Fig. 14 depicts a cluster heat map where the response item dendrogram is constructed by using simple Jaccard distance in (2), instead of profile based distance in (6). In addition, the ordering methods and the linkage criteria of the cluster heat map in Fig. 14 are the same as those of the optimized SAM in Fig. 12, OM2-WC (factor item dendrogram) and OM3-AL (response item dendrogram).

In Fig. 14, four interesting areas are specified, where PST1', NGT1' and NGT3' are subareas of PST1, NGT1 and NGT3 in Fig. 12, respectively. Moreover, only NGT4 is identified in both Figs. 12 and 14, and we can see that the conventional cluster heat map is less helpful in discovering the interesting areas and identifying the interesting association rules. For example, it is not easy to identify the interesting association rules "{F_2, F_8} → {R_1, R_9}", "{F_2, F_8} → {R_3, R_9}" and "{F_1, F_3} → {R_8, R_{10}}" in Table 8 from Fig. 14, even though they have high life values.

Lastly, Fig. 15 represents a simple data matrix with no dendrogram, where the items are arranged in simple numerical order. We can see that none of the interesting areas in Fig. 12 can be found in Fig. 15. That is, such simple data matrix contains little information on the many-to-many association rules, and the SAM proposed in this paper may be a promising approach for visual exploration of association rules in large transaction data sets.

4.5. Discussions

In summary, the experiment results provide several important insights into SAM: (i) SAM is a novel visualization technique enables to extract association rules with focusing on the specific items. Especially, the interesting area is very helpful to extract useful many-to-many association rules while avoiding the irrelevant ones. (ii) The quality of SAM can be evaluated by S2C measure, and the analyzers can find a larger number of interesting areas more easily from a SAM with higher S2C value. In order to obtain the optimized SAM which maximizes S2C measure, the structures and the ordering methods of factor and response item dendrograms must be carefully determined. (iii) SAM can be used for both pre-processing and post-processing purposes in association analysis. In other words, SAM can be used not only to select a subset of items before applying association rule mining algorithms, but also to summarize the extracted many-to-many association rules.

5. Conclusions and further remarks

The relationships among the binary variables, items, within the transaction data set are typically analyzed by using the association rule mining techniques. However, the analyzers are often confused by a plethora of the association rules extracted from high dimensional transaction data. Moreover, this problem is also relevant to the expert and intelligent systems that utilize the association rules, since complex relationships among the items cause much inefficiencies in the procedures for extracting, maintaining and deploying association rules. Visualization techniques can help to deal with such complexities in association analysis.

This paper proposes a novel visualization method called SAM, carefully designed to represent the complex relationships among the items within large transaction data. SAM is similar with classical cluster heat map in that a matrix is combined with two dendrograms. However, the dendrograms of SAM are constructed in more sophisticated way in order to avoid misunderstanding about many-to-many association rules. This paper refines the abstract concept of SAM introduced in previous work and develops detailed concepts and enhanced procedures for constructing the optimized SAM. From the visual data exploration perspectives, the concepts and the procedures in this paper provide two important implications. Firstly, it is difficult to visualize the association rules by using classical methods such as simple charts and matrices. Especially, a visualization technique for association rules must deal with the many-to-many relationships appropriately. Secondly, we can obtain the optimized visual aids if the quality of visualization results can be numerically evaluated. This paper proposes S2C measure which can be used to evaluate the quality of SAM, and we should construct an optimized SAM with the highest S2C value.

Compared to the previous visualization techniques for association rules, SAM has following important benefits. Firstly, SAM can provide useful insights into the many-to-many association rules which cannot be appropriately represented in previous visualization methods. Secondly, SAM is based on classical matrix based visualization, and it is easy to implement and interpret. In other words, the graphical representation of SAM is more straightforward than the previous methods based on graphs or parallel coordinates. Thirdly, users can conveniently find the interesting area and the interesting group from SAM in order to select the items to be investigated in more detail, while several recent visualization techniques allow the users to select appropriate items but do not provide convenient ways for identifying them. Fourthly, we can evaluate the quality of SAM by applying S2C measure and construct an optimized SAM suitable for visual exploration.

On the contrary, SAM in this paper also has several drawbacks which should be addressed by future research. Firstly, the optimized SAM was found by an exhaustive search, and this procedure can be time-consuming. Secondly, the users have to identify the interesting areas from SAM by themselves, but this procedure may be automated. Thirdly, only 2×2 association rules are considered by S2C measure, but larger association rules might be also helpful to evaluate the quality of SAM. Therefore, the author plan to apply SAM to various data sets and try to develop an enhanced version of SAM which addresses such drawbacks.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A1A1044834).

References

Aggarwal, C. C., Procopius, C., & Yu, P. S. (2002). Finding localized associations in market basket data. *IEEE Transactions on Knowledge and Data Engineering, 14*(1), 51–62.

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

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the international conference on very large databases* (pp. 125–131).

Barnes, G. J., Wilson, R. F., Phillips, K., Maynor, K., Hwang, C., Marinopoulos, S., et al. (2006). *Value of the periodic health evaluation*. US Department of Health and Human Services, Agency for Healthcare Research and Quality.

Bayardo, R. J., & Agrawal, R. (1999). Mining the most interesting rules. In *Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 145–154).

Bornelöv, S., Marillet, S., & Komorowski, J. (2014). Ciruvis: A web-based tool for rule networks and interaction detection using rule-based classifiers. *BMC Bioinformatics, 15*(1), 139–150.

Buono, P., & Constabile, M. F. (2005). Visualizing association rules in a framework for visual data mining. In M. Hemmje, C. Niederee, & T. Risse (Eds.), *From integrated publication and information systems to information and knowledge environments* (pp. 221–231). BerlinHeidelberg: Springer.

Chen, Z., Cai, S., Song, Q., & Zhu, C. (2011). An improved apriori algorithm based on pruning optimization and transaction reduction. In *Proceedings of the 2nd international conference on artificial intelligence, management science and eletronic commerce* (pp. 1908–1911).

Day, W. H. E., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering method. *Journal of Classification, 1*(1), 7–24.

De Oliveira M. C. , F., & Levkowitz, H. (2003). From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics, 9*(3), 378–394.

Djenouri, Y., Drias, H., & Bendjoudi, A. (2014). Pruning irrelevant association rules using knowledge mining. *International Journal of Business Intelligence and Data Mining, 9*(2), 112–144.

Gruvaeus, G., & Wainer, H. (1972). Two additions to hierarchical cluster analysis. *British Journal of Mathematical and Statistical Psychology, 25*(2), 200–206.

Gu, L., Li, J., He, H., Williams, G., Hawkins, S., & Kelman, C. (2003). Association rule discovery with unbalanced class distributions. *Lecture Notes in Computer Science, 2903*, 221–232.

Guenoche, A., Hansen, P., & Jaumard, B. (1991). Efficient algorithms for divisive hierarchical clustering. *Journal of Classification, 8*(1), 5–30.

Gupta, G. K., Strehl, A., & Ghosh, J. (1999). Distance based clustering of association rules. In *Proceedings of the 9th artificial neural networks in engineering conference* (pp. 759–764).

Hahsler, M., Buchta, C., Grün, B., Hornik, K., & Borgelt, C. (2015). *Package 'arules'* Available at http://cran.r-project.org/web/packages/arules/arules.pdf (accessed on 24 July 2015).

Hahsler, M., & Chellubonia, S. (2011a). Visualizing association rules: Introduction to the R-extension package arulesViz. *R Project Module*.

Hahsler, M., & Chellubonia, S. (2011b). Visualizing association rules in hierarchical groups. In *Proceedings of the 42nd symposium on the interface: Statistical, machine learning and visualization algorithms*.

Hurley, C., & Hurley, M. C. (2012). *Package 'gclus'* Available at http://www.icesi.edu.co/CRAN/web/packages/gclus/gclus.pdf (accessed on 24 July 2015).

Itoh, T., Kumar, A., Klein, K., & Kim, J. (2016). High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots. *ArXiv Preprint, arXiv*:1609.05268. Available at http://arxiv.org/abs/1609.05268 (accessed on 25 November 2016).

Ju, C., Bao, F., Xu, C., & Fu, X. (2015). A novel method of interestingness measures for association rules mining based on profit. *Discrete Dynamics in Nature and Society*.

Kerdprasop, N., & Kerdprasop, K. (2014). Visual data mining and the creation of inductive knowledge base. In *Proceedings of the 5th international conference on circuits, systems, control, signals* (pp. 3–5).

Keim, D. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics, 8*(1), 1–8.

Kim, J. W. (2015). Association rule visualization by structured association map. *Journal of Knowledge Information Technology and Systems, 10*(3), 305–317.

Kosara, R., Bendix, F., & Hauser, H. (2006). Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics, 12*(4), 558–568.

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 71–82.

Kweon, S., Kim, Y., Jang, M. J., Kim, Y., Kim, K., Choi, S., et al. (2014). Data resource profile: The Korea national health and nutrition examination survey (KNHANES). *International Journal of Epidemiology, 43*(1), 69–77.

Lent, B., Swami, A., & Widon, J. (1997). Clustering association rules. In *Proceedings of the 13th international conference on data engineering* (pp. 220–231).

Lei, H., Xie, C., Shang, P., Zhang, F., Chen, W., & Peng, Q. (2016). Visual analysis of user-driven association rule mining. In *Proceedings of the 9th international symposium on visual information communication and interaction* (pp. 96–103).

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications, 39*(12), 11303–11311.

Liu, B., Hsu, W., & Ma, Y. (1999). Pruning and summarizing the discovered associations. In *Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 125–134).

Liu, X., Zhai, K., & Pedrycz, W. (2012). An improved association rules mining method. *Expert Systems with Applications, 39*(1), 1362–1374.

Oyebode, O., & Mindell, J. S. (2014). A review of the use of health examination data from the health survey for England in government policy development and implementation. *Population, 8*(9).

Romero, C., Luna, J. M., Romero, J. R., & Ventura, S. (2011). RM-Tool: A framework for discovering and evaluating association rules. *Advances in Engineering Software, 42*(8), 566–576.

Sekhavat, Y. A., & Hoeber, O. (2013). Visualizing association rules using linked matrix, graph and detail views. *International Journal of Intelligence Science, 3*(1), 34–49.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Addison-Wesley.

Techapichetvanich, K., & Datta, A. (2005). A new technique for visualizing mined association rules. In X. Li, S. Wang, & Z. Y. Dong (Eds.), *Advanced data mining and applications* (pp. 88–95). Berlin, Heidelberg: Springer.

Trevisan, D. G., Sanchez-Pi, N., Marti, L., & Garcia, A. C. B. (2015). Big data visualization for occupational health and security problem in oil and gas industry. In *Proceedings of the international conference on human interface and the management of information* (pp. 46–54).

Usman, M., & Usman, M. (2016). Multi-level mining and visualization of informative association rules. *Journal of Information Science and Engineering, 32*(4), 1061–1078.

Wilkinson, L., & Friendly, M. (2009). The history of the cluster heat map. *The American Statistician, 63*(2), 179–184.

Wong, P. C., Whitney, P., & Thomas, J. (1999). Visualizing association rules for text mining. In *Proceedings of the 1999 IEEE symposium on information visualization* (pp. 120–123).

Yang, L. (2005). Pruning and visualizing generalized association rules in parallel coordinates. *IEEE Transactions on Knowledge and Data Engineering, 17*(1), 60–70.

Yang, L. (2008). Visual exploration of frequent itemsets and association rules. *Lecture Notes in Computer Science, 4404*, 60–75.

Zhao, Y. (2012). *R and data mining: examples and case studies*. Academic Press.