

## ***Design and Evaluation of Visualization Support to Facilitate Association Rules Modeling***

**Yan Liu**

School of Industrial Engineering, Purdue University

**Gavriel Salvendy**

School of Industrial Engineering, Purdue University  
and Department of Industrial Engineering, Tsinghua University, Beijing, P.R. China

Association rules mining is a popular data mining modeling tool. It discovers interesting associations or correlation relationships among a large set of data items, showing attribute values that occur frequently together in a given dataset. Despite their great potential benefit, current association rules modeling tools are far from optimal. This article studies how visualization techniques can be applied to facilitate the association rules modeling process, particularly what visualization elements should be incorporated and how they can be displayed. Original designs for visualization of rules, integration of data and rule visualizations, and visualization of rule derivation process for supporting interactive visual association rules modeling are proposed in this research. Experimental results indicated that, compared to an automatic association rules modeling process, the proposed interactive visual association rules modeling can significantly improve the effectiveness of modeling, enhance understanding of the applied algorithm, and bring users greater satisfaction with the task. The proposed integration of data and rule visualizations can significantly facilitate understanding rules compared to their nonintegrated counterpart.

### **1. INTRODUCTION**

Data mining (DM), also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown, and potentially useful information from data in databases (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Association rules mining is a popular DM modeling tool. It discovers interesting associations or correlation relationships among a large set of data items, showing attribute values that occur frequently together in a given dataset. Although the induced rules do not necessarily indicate causation, they are helpful starting points for further exploration, making association rules a popular tool for

understanding data (Appice, Ceci, Lanza, Lisi, & Malerba, 2003; Buchter & Wirth, 1999; Motwani & Rajput, 2002). Despite their great potential benefit, current association rules mining tools are far from optimal. One of the most important and difficult issues in association rules modeling is to control the derivation of rules from typically large, complex, and nonhomogeneous datasets (Matsumoto & Hashimoto, 1998; Srikumar & Bhasker, 2004; Wang, He, & Han, 2003). Currently, most association rules modeling tools simply support an automatic modeling process (Liu & Salvendy, 2005), during which users have little interaction with the computing machine, other than specifying confinements for some statistical measures of derived rules at the beginning of the process. Only derived rules are presented to users, either visually or nonvisually. If users are not satisfied with the induced rules, they manipulate parameter values and restart the modeling process. Such iteration continues until the results are satisfactory. Selection of these parameter values can be arbitrary if users do not have much information about data characteristics, which may lead to an unproductive modeling process, such as missing some important rules and generating many uninteresting rules. In addition, once association rules have been derived, they need to be presented in a form that allows users to assess their qualities and applicability. However, few association rules mining tools have addressed this concern adequately.

Visualization can help to address the issues just mentioned and thus facilitate the association rules modeling process. First, appropriate visualization techniques can provide users insights to better examine derived patterns. Second, visualization can be integrated with nonvisual techniques to greatly enhance the effectiveness and efficiency of the modeling process by allowing users to take advantage of their skills and direct the modeling process when a situation warrants (Ankerst, Keim, & Kriegel, 1996; Keim, 2002; Kreuzeler & Schumann, 2002). Furthermore, being more involved in the modeling process may improve users' understanding of applied algorithms and give them a stronger sense of independence and responsibility, which in turn, according to Hackman and Oldham's (1975) job characteristic model, may bring them greater satisfaction.

Within this context, this article studies how visualization techniques can be applied to facilitate the association rules modeling process, particularly what visualization elements should be incorporated and how they can be displayed. We organize the remainder of this article as follows. Section 2 gives a brief overview of the fundamental concept of association rules algorithms. A conceptual model of visualization support to association rules modeling is proposed in section 3. The proposed visualization of end results and visualization of rule derivation process for supporting interactive visual association rules modeling (IVAR) are presented in sections 4 and 5, respectively. In section 6 we describe two experiments that have been conducted to test four hypotheses based on the proposed IVAR and integration of data and rule visualizations. In section 7 we offer conclusions of this study.

## **2. OVERVIEW OF ASSOCIATION RULES ALGORITHMS**

Association rules algorithms differ in content from classification rule induction techniques. Whereas classification rule induction techniques concentrate on find-

ing rules that predict a single, preselected class variable, association rules algorithms are motivated by identifying rules that capture increased frequency of an attribute value, or a collection of attribute values, without limitations on the values that may appear in the consequent of a rule (Webb, 2000).

Let  $I = \{I_1, I_2, \dots, I_m\}$  be a set of  $m$  distinctive items and  $D$  be a database consisting of subsets of  $I$ . Items are binary attributes, because they are either present or absent in a record. Each record (tuple),  $T_i \in D$  ( $i = 1, 2, \dots, n$ ), contains a set of items such that  $T_i \subseteq I$ . An association rule is expressed in the form of  $X \Rightarrow Y$ , where  $X, Y \subset I$  are called *antecedent* and *consequent* of the association rule, respectively, and  $X \cap Y = \emptyset$ . An  $n$ -itemset is a combination of  $n$  single items.

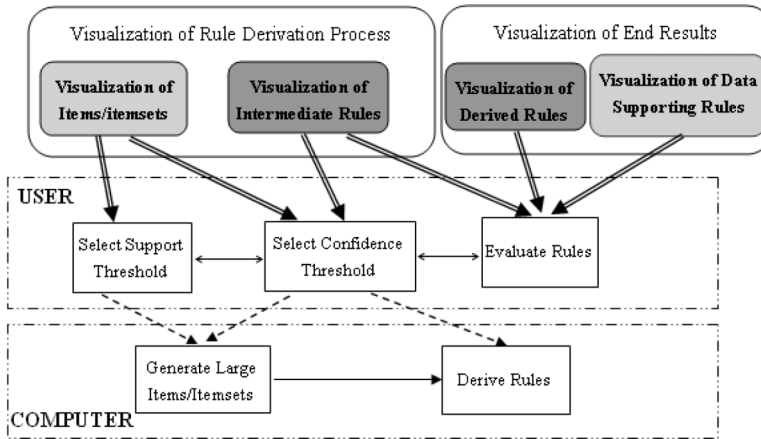
These rules are computed from data, and, unlike the if-then rules of logic, association rules are probabilistic in nature. One of the major problems with association rules mining is the overwhelming number of rules that can be generated, many of which may not be of interest (Agrawal, Imielinski, & Swami, 1993; Klemettinen, Mannila, Ronkainen, Toivonen, & Verkamo, 1994). To address this issue, two important measures are often used to constrain derived rules. One measure is called the support of a rule. It is defined as the probability that a randomly selected data record will contain all items in both the antecedent and consequent of the rule. Support is a measure of the frequency of use of a rule. Similarly, support of an item or itemset  $X$  is the fraction of records in  $D$  that contain  $X$ . The other measure is the confidence of a rule, which is defined as the conditional probability that a randomly selected data record contains all items in the consequent of the rule, given all items in the antecedent of the rule are present; it measures the strength of a rule. With a given database, an association rules algorithm aims at discovering rules that meet the user-specified minimum support (support threshold) and minimum confidence (confidence thresholds) levels.

The problem of discovering association rules can be broken down into two phases, as stated in Agrawal et al. (1993):

1. Generate all  $n$ -itemsets ( $n \geq 2$ ) whose support values meet the user-specified support threshold. These itemsets are called *large itemsets*. The itemsets with support values below the specified support threshold are called *small itemsets*.
2. Then, from the large itemsets, generate all rules that satisfy the specified confidence threshold.

### **3. CONCEPTUAL MODEL OF VISUALIZATION SUPPORT TO ASSOCIATION RULES MODELING**

Figure 1 shows the proposed conceptual model of visualization support to association rules modeling in which the double solid arrow lines, dash arrow lines, and solid arrow lines represent visualization support, inputs from the user to the computer, and the process flow, respectively. An association rules modeling process consists of two main steps: rule derivation and rule evaluation. As described in section 2, users need to specify support and confidence thresholds to derive rules. Modeling can be thought of as a type of problem solving, which was defined in Newell and Simon (1972) as the cooperative processes of understanding and



**FIGURE 1** Proposed conceptual model of visualization support to association rules modeling.

searching. The understanding process generates a person's internal representation of the problem (mental model), and the search process generates a solution. As revealed in Figure 1, visualization can facilitate association rules modeling via visualizations of the rule derivation process and end results. In Figure 1, the light- and dark- shaded squares with round corners represent data visualization and rule visualization components, respectively.

Visualizing the rule derivation process serves not only to help users understand how rules are generated using the applied algorithm but also to provide them the ability to discern the derivation process, adjusting the support and confidence thresholds using their background knowledge and pattern recognition capacities. Studies in Willemain (1995) showed that an important characteristic of experienced modelers, in contrast to novice ones, was their continual self-monitoring of progress; they often suspended an activity at one stage and returned to an "earlier" stage. Two components are of importance in visualization of the rule derivation process: visualization of items/itemsets and visualization of intermediate rules. Data visualization assists users in discovering and understanding relationships within data, which leads to creation or enhancement of a mental model. This captures the implications of the data relative to the modeling objective (Crapo, Waisel, Wallace, & Willemain, 2000). In this sense, visualization of items/itemsets can facilitate selecting appropriate support and confidence thresholds. Because of the typically large number of rules derived from association rules modeling, it is very inefficient for humans to examine the results without appropriate visualization (Bruzzese & Davino, 2003; Hofmann & Wilhelm, 2001).

The purpose of visualizing end results is to enable users to identify quickly the important features of derived rules and understand them. Understanding DM output means more than just comprehending the meaning of the results; it also involves context. The modeling process consists of the "context of discovery" and the "context of justification" (Morris, 1967). The context of discovery is the circumstance under which a hypothesis is formed and the context of justification is the cir-

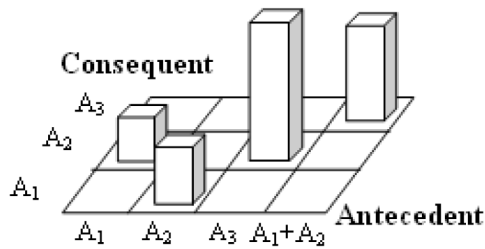
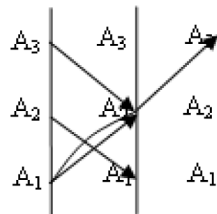
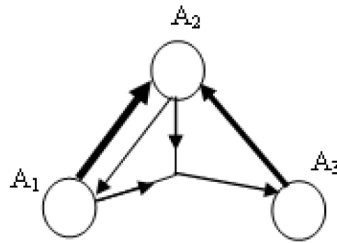
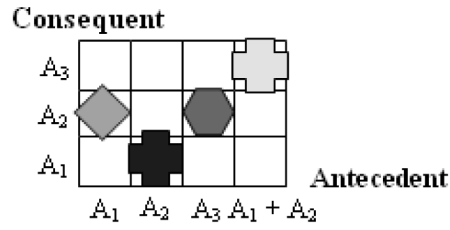
cumstance under which the hypothesis is tested. Data visualization provides the context of modeling, that is, the situation in which the model is built. Integration of visualizations of data and derived models enables users to relate the models to the data from which they are built (Thearling et al., 2002). In association rules modeling, for example, it might be helpful to know the characteristics of the records that support rules so that some actions can be taken to target the corresponding groups. One typical application is market segmentation. A data record supports a rule if it contains both the antecedent and consequent items of the rule. Due to these considerations, other than visualization of rules, visualization of data supporting rules is also important for understanding rules.

## **4. VISUALIZATION OF END RESULTS**

### **4.1. Visualization of Rules**

Visualization of rules purposes to provide immediate insights into the primary characters of derived rules, which can facilitate evaluation of these rules and direct further surveying if needed. Visualization systems for information inquiry should follow the paradigm “Overall view first, zoom and filter, then details on demand” (Shneiderman, 1994). Therefore, rules should be represented in forms that allow users to easily identify their features. The principle information of association rules includes mappings of the involved items and the interestingness measures of these rules. Mappings of items can be one-to-one, one-to-many, many-to-one, or many-to-many. Several techniques have been developed to visualize association rules, including matrix views, node-link views, parallel coordinates views, and information landscape views, as illustrated in Figure 2 in which four rules  $A_1 \Rightarrow A_2$ ,  $A_2 \Rightarrow A_1$ ,  $A_3 \Rightarrow A_2$ , and  $A_1 + A_2 \Rightarrow A_3$  are shown. Table 1 lists four desirable features that a visualization of association rules should possess. The first two features in the table look at the important content information of rules. Searching for rules associated with specific items of interest is a typical application of association rules (Agrawal et al., 1993). As stated earlier, it is typical that numerous rules are generated from association rules modeling; therefore, a visualization technique that can handle a large number of rules and items is critical in real applications.

In the matrix view, the screen is divided into cells. Rows and columns can represent antecedent and consequent items, respectively, and the support and confidence values of a rule are indicated by the color and shape of the icon placed on its cell (SAS Enterprise Miner; SAS Institute, Inc., n.d.), as illustrated in Figure 2a. Shapes and colors, however, are not effective in coding quantitative information (Cleveland & McGill, 1987). In addition, one major problem with the item-to-item representation is that when there are rules containing more than one antecedent or consequent item, replications of these items in rows and columns result in the loss of their identities, which makes it difficult to search rules associated with these items. In Ong, Ong, Ng, and Lim (2002), the rows and columns of the matrix represented the support and confidence levels, respectively, and the description of a rule could be viewed by placing the mouse over its cell. The elimination of item identities in this approach ruled out examining rules associated with specific items of in-



**FIGURE 2** Visualization of association rules (Four rules are shown:  $A_1 \Rightarrow A_2$ ,  $A_2 \Rightarrow A_1$ ,  $A_3 \Rightarrow A_2$ ,  $A_1 + A_2 \Rightarrow A_3$ ). (a) Item-to-item matrix view. (b) Node-link view. (c) Parallel coordinates view. (d) Item-to-item information landscape view.

**Table 1: Desirable Features of Visualization of Rules and Evaluation of Six Visualization Techniques**

<i>Desirable Features</i>	<i>Visualization Techniques</i>					
	<i>Matrix View</i>			<i>Information Landscape</i>		
	<i>Item -to- Item</i>	<i>Measure -to- Measure</i>	<i>Node-Link View</i>	<i>Parallel Coordinates</i>	<i>Item -to- Item</i>	<i>Rule -to- Item</i>
Can display all possible mappings	+	+	+	+	+	+
Effective visualization of interestingness measures	-	+	-	-	-	-
Allow examination of rules containing specific items of interest (identities of items)	-	-	-	-	-	+
Scalable to large number of rules and items	+	-	-	-	Depends on the visual attributes used to code interestingness measures	

*Note.* + = positive; - = negative.

terest. Such a confidence-to-support representation also could not handle the situation when many rules have similar or the same support and confidence values.

As suggested from its name, the node-link view of association rules represents rules as nodes and links; the nodes correspond to items and the arrow lines symbolize relationships among items (IBM Intelligent Miner Rule Visualizer, n.d.). The width and/or colors of links indicate interestingness measures of rules, as illustrated in Figure 2b. One obvious shortcoming of this method is that when the number of rules increases, the representation can quickly become an intertwined picture making visualization difficult.

Although the parallel coordinates graph has been used mainly for visualizing multivariate data (Inselberg, 1985; Wegner, 1990), it was adopted in Yang (2005) to display association rules, as illustrated in Figure 2c. In Yang, all items were listed on each vertical coordinate in a two-dimensional (2D) space. The number of coordinates was the same as the maximum number of items involved in derived rules. A rule manifested itself as a polyline connecting all its items on coordinates, one item on one coordinate starting from the leftmost coordinate; the antecedent and consequent items of a rule were separated by an arrow line. However, information of interestingness measures was not depicted in Yang. Although the parallel coordinates representation showed each item of a rule on one coordinate, the same item might spread to different coordinates, which led to the problem of loss of identities of items. Another severe limitation of the parallel coordinates method was the overlapping of lines and connection points when the numbers of rules and items were large.

The information landscape view is essentially a 2D matrix view as the floor space with another variable extruded into the spatial dimension (Card, 2003). Depending on how items of rules are placed in the 2D floor space, two types of infor-



conveying quantitative information. Through interaction with the graph, users can view the exact support and confidence values of rules, highlight their involved items, and highlight rules associated with specific items of interest.

### 4.2. Visualization of Data Supporting Rules

Integration of data and rule visualizations seems to have been overlooked by most association rules mining tools despite the fact that some tools provide a visualization of an entire dataset (Liu & Salvendy, 2005). Drill-through, which allows users to select a subset of a model and visualize its corresponding underlying data, is an effective integration technique (Sprenger, Gross, Bielser, & Strasser, 1998; Thearling et al., 2002). In the approach proposed in this article, users can access visualization of the data that support a rule of interest by clicking its associated vertical bar in the rule visualization screen, illustrated in Figure 4.

The dataset used in the example demonstrated in Figure 4 is adapted from the flag dataset in the University of California Irvine (UCI) machine learning repository (Blake & Merz, 1998). Nine attributes are included in this dataset: Landmass, Zone, Religion, Red, White, Blue, Black, Icon, and Cross. The last six attributes are binary and rules have been derived among them. The remaining three attributes—Landmass, Zone, and Religion—have seven, four, and seven categories, respectively.

In Figure 4, the screen on the left is visualization of the data supporting the leftmost rule (Blue  $\Rightarrow$  White) in the rule visualization window on the right. In the data visualization window, each horizontal bar is associated with an attribute, and its entire length corresponds to the total number of data records in the specific domain. The total number of data records is shown at the top of the window. If an attribute has more than two categories, then the entire bar is divided into sections, each of which represents one category. The length of a section is proportional to the number of data records in its associated category. For binary attributes, the length of the shaded section is proportional to the number of records that contain it.

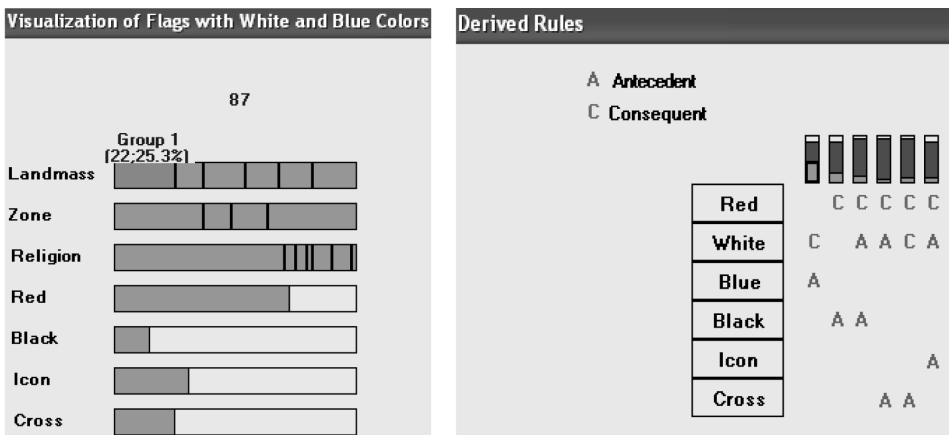


FIGURE 4 Integration of data and rule visualizations.

Length is used to visualize the number of records here because it is very effective in demonstrating quantitative information compared to other commonly used visual attributes such as areas, angles, colors, and shapes (Cleveland & McGill, 1987).

### 5. VISUALIZATION OF RULE DERIVATION PROCESS

As stated in section 3, visualization of the rule derivation process aims not only to help users understand how rules are derived but also to support IVAR. IVAR refers to a process in which users are able to adjust support and confidence thresholds and remove undesired itemsets and rules promptly by taking advantage of their knowledge about the task and data through visualizing the underlying data and intermediate outcomes during the process. Figure 5 shows the flow chart of the IVAR process proposed in this research, in which the dash arrow lines and solid arrow lines represent the information flow between the user and computer and the process flow, respectively. The light- and dark- shaded squares with round corners represent data and rule visualizations, respectively.

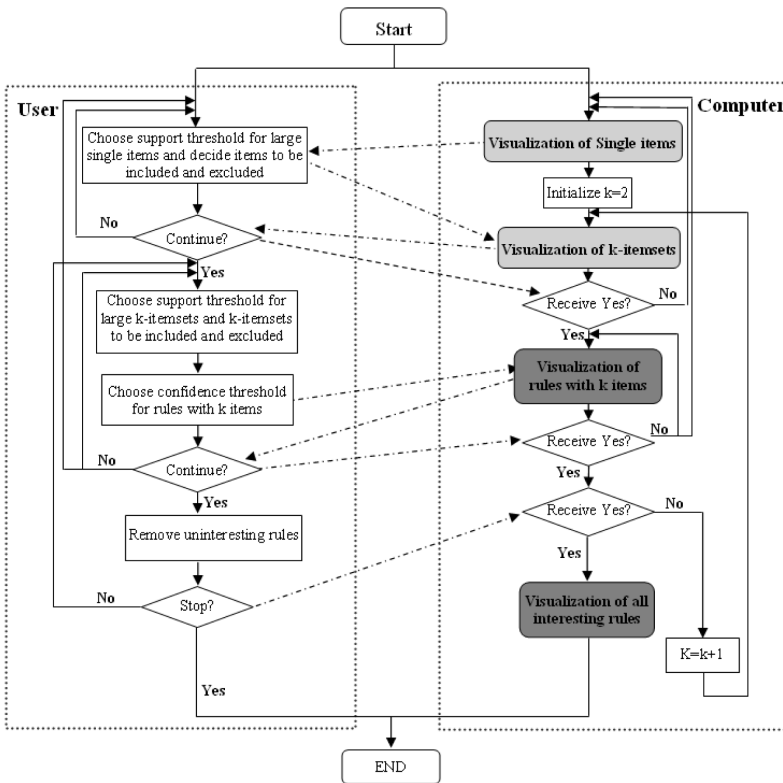
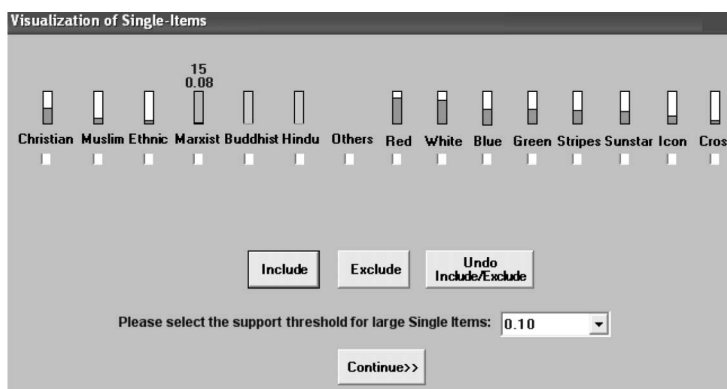


FIGURE 5 Flow chart of the proposed interactive visual association rules modeling process.

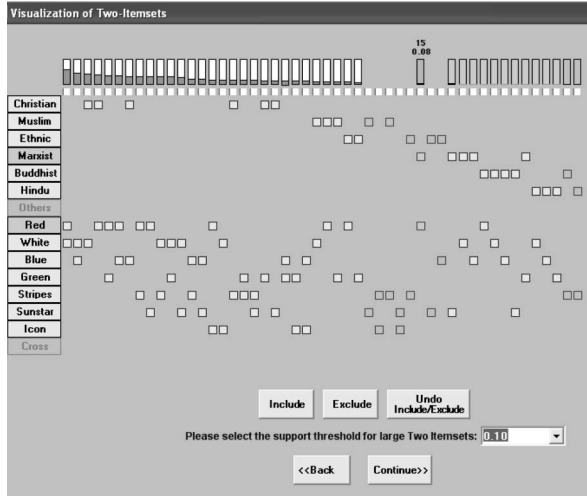
A dataset adapted from the same flag dataset in the UCI machine learning repository mentioned in section 4.2 is used to demonstrate the proposed IVAR process. Fifteen items are considered in this example. Seven items are religions: Christian, Muslim, Ethnic, Marxist, Buddhist, Hindu, and Others; these items are mutually exclusive. Four items are colors: Red, White, Blue, and Green. The remaining 4 items are pictures: Stripes, Sunstar, Icon, and Cross.

The process starts with visualization of single items (or individual items), as shown in Figure 6. In this interface, similar to the visualization of support values applied in the proposed rule visualization described in section 4.1, a vertical bar on the top of each item visualizes its support value. The entire height of the bar represents 1.00, and the height of the shaded section is proportional to the actual support value. Based on the visualization, users choose a support threshold for generating large single items. In real applications, datasets can be nonhomogenous in that some items are by nature less prevalent than others, yet they are interesting. However, if the support threshold is set low enough to include these rare attributes, enormous rules might be derived among the frequent items. To address this dilemma, the proposed process allows users to include the rare but interesting items in the set of large single items regardless of the specified support threshold. Similarly, users can remove items that have high support but are not interesting.

After the large single items are determined, the next interface presented to users is visualization of two-itemsets, which are combinations of two large single items, as shown in Figure 7. Visualization of two itemsets is shown in the form of an itemset-to-item matrix, in which each row corresponds to an item and each column a two-itemset. Such visual representation is consistent with the proposed 2D rule-to-item visualization of rules described in section 4.1, because consistent displays of graphics enhance their dissemination, accuracy, and ease of comprehension (Hix & Hartson, 1993). Similar to visualization of single items, a vertical bar is placed on the top of each two-itemset to visualize its support. Through interacting with the visual interface, users can view the exact support values, highlight involved items, and highlight itemsets associated with specific items of interest.



**FIGURE 6** Interface of visualization of single items (the mouse is over the vertical bar associated with item Marxist).



**FIGURE 7** Interface of visualization of two-itemsets (the mouse is over the vertical bar associated with two-itemset [Marxist+Red]).

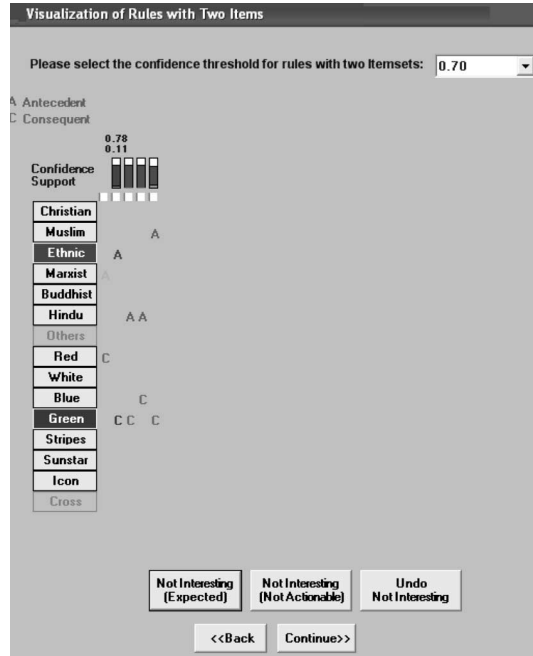
From these two-itemsets, large two-itemsets can be generated by selecting a support threshold. Users can also include and exclude any two-itemset as desired, and they can go back to the previous interface to make modifications if needed.

After the large two-itemsets are determined, rules with two items will be generated. In the conventional two-step association rules algorithm described in section 2, rules are derived after all large  $n$ -itemsets are generated. In this proposed IVAR process, however, rules with  $k$  ( $k \geq 2$ ) items are induced right after large  $k$ -itemsets are generated until the maximum number of items in rules is reached. This approach allows uninteresting itemsets and rules to be detected and removed promptly, which precludes the subsequent uninteresting itemsets and rules from being generated and thus can improve the effectiveness of the process. This strategy also enables users to choose different thresholds for rules with different numbers of items. After users choose a confidence threshold, rules that meet the threshold will be visualized in the form of the rule-to-item matrix view as depicted in section 4.1, shown in Figure 8. Not all rules that meet the support and confidence thresholds are interesting. Rules can be uninteresting because they are expected or not actionable (Silberschatz & Tuzhilin, 1996). Due to this consideration, users should be allowed to remove these uninteresting rules. By doing that, rules that are consequences of the eliminated rules will not be derived in later steps.

If rules with three or more items are desirable, then the same steps are followed until users decide to terminate the process. At the end of the process, visualization of all interesting rules is displayed.

## 6. EXPERIMENTATION

Two experiments were administered to test four hypotheses regarding the proposed IVAR process and integration of data and rule visualizations.



**FIGURE 8** Visualization of rules with two items (the mouse is placed over rule [Ethnic  $\Rightarrow$  Green] and rule [Marxist  $\Rightarrow$  Red] is expected and thus removed).

### 6.1. Experiment 1

This experiment tested three hypotheses: (a) Compared to automatic modeling, IVAR improves effectiveness of modeling; (b) compared to automatic modeling, IVAR improves understanding of the applied algorithm; and (c) compared to automatic modeling, IVAR improves users' satisfaction with the modeling task.

#### Participants

Twenty participants (10 male and 10 female) between the ages of 22 and 34 ( $M = 27.1$ ,  $SD = 3.63$ ), were recruited for this experiment; all were graduate engineering students. Each participant had no previous knowledge of the applied association rules algorithm, was not color deficient, and had the capacity to conduct tasks on personal computers. To make sure they were not color deficient for the three colors used in the visual interfaces of this experiment (red, green, gold), the participants looked at objects with these colors and then stated the colors they had seen. These participants were randomly divided into two groups, 10 participants for each type of DM modeling process.

#### Apparatus

The visual interfaces were created using Microsoft Visual Basic.Net 2003. All participants performed tasks on a Dell UltraSharp 1905FP 19-in. flat monitor in the experiment.

### **Task**

The dataset of this task was the modified flag dataset exercised for demonstrating the IVAR process described in section 3. The goal of the task was to generate rules with two and three items and maximize the effectiveness of association rules modeling (EAR), which is defined in the Experiment 1 Dependent Variables section.

### **Procedure**

Five steps were followed in this experiment: tutorial, training, practice, formal task, and poststudy. At the beginning of the experiment, each participant was given a short tutorial of the applied association rules algorithm. This tutorial was intended not to teach participants the details of the algorithm but to introduce its fundamental concepts so that they would be able to perform the modeling task. After the tutorial, participants were asked to answer four questions regarding the knowledge learned from the tutorial to make sure they were ready to proceed. Next, participants were provided with a demonstration of the visualization features that would be used in the upcoming modeling task. The purpose of practice was to allow participants to become familiar with the demonstrated visualization features. Although no specific time limits were given, participants typically spent 3 to 4 min on the practice. The dataset utilized for the practice was just an example and was not related to the formal task. Each participant spent 20 min deriving rules during the formal task. After the formal task, participants filled out two questionnaires—the algorithm understanding questionnaire and the modeling satisfaction questionnaire, which are described in the XXX section—to measure their understanding of the applied algorithm and satisfaction with the task.

### **Experimental Design**

**Independent variables.** The independent variable in this experiment was the type of association rules modeling process IVAR, or automatic modeling.

The IVAR process was presented in section 5. The automatic modeling process carried out in this experiment resembled the design of most current association rules modeling tools, which accept users' inputs of parameters at the beginning of the process, execute the algorithm, and show derived rules to users. In this experiment, two interfaces were provided in the automatic modeling process. The first interface displayed the support value of each single item and allowed participants to include and exclude any items as desired. After users specified the support and confidence thresholds in the first interface, rules were derived and visualized in the second interface. Because the objective of this experiment was to test the potential benefits of having users more involved in the modeling process through visualizing the rule derivation process, to rule out the effects of the visual representation of derived rules on users' performance, the same rule-to-item matrix view of association rules employed in the IVAR process was applied in the automatic modeling process except for the functions for the removal of uninteresting rules.

**Dependent variables.** There were three dependent variables in this experiment: effectiveness of modeling, understanding of the applied algorithm, and users' satisfaction with the task.

**Effectiveness of modeling.** Typical measures of effectiveness of tasks are time-to-completion and task outcomes (Novick, 1997). However, because association rules modeling is an iterative process until users find the satisfactory results, task outcomes were used in this experiment as the measure of effectiveness of modeling. This also eliminates possible effects of task execution time on the other two dependent variables. Each participant was required to spend the same amount of time (20 min) to perform the task, and his or her best result was used for data analysis.

The important outcome measures of association rules modeling include the total number of derived rules and the number of derived interesting rules (Sahar, 1999). Silberschatz and Tuzhilin (1996) categorized the measures for interestingness of association rules into objective and subjective. In this experiment, support and confidence were objective interestingness measures. Expected rules were not interesting, and this is a subject measure. A measure of the EAR was defined in this experiment as Equation 1.

$$EAR = \sum_{i=1}^{n_I} r_i - \alpha \cdot n_{NI} \quad (1)$$

where  $n_I$  is the number of derived interesting rules,  $r_i$  is the reward score of the  $i$ th derived interesting rule,  $n_{NI}$  is the number of derived uninteresting rules, and  $\alpha$  is the penalty for deriving an uninteresting rule. Because participants might have distinct standards for objective and subjective interestingness measures, common criteria were necessary to evaluate their task performance fairly. Therefore, a list of expected uninteresting rules was given to participants before the modeling task started. Furthermore, to reduce the complexity of the task, it was assumed that the reward scores of interesting rules only depended on their confidence values as long as their support values were no less than 0.10 if they did not contain any rare yet interesting items, or if there were no less than 0.02 otherwise. Specifically, 1, 2, 3, 4, and 5 points were given to rules with support values that satisfied the criteria and that had confidence values between 0.70 and 0.75, between 0.75 and 0.80, between 0.80 and 0.85, between 0.85 and 0.90, and greater than 0.90, respectively.  $\alpha$  was assigned to 2, which meant each uninteresting rule caused a deduction of 2 points in EAR. Such assignments of reward scores and penalties were merely for the sake of easing evaluation of EAR for participants so that they could devote most of their attention to deriving rules. In reality, however, these numbers should depend on applications and users' goals.

**Understanding of the algorithm.** A variety of approaches have been applied in measuring program comprehension. Dunsmore and Roper (2000) classified the methods of measuring program comprehension into five groups: recall (or memo-

rization) test, maintenance operation, dynamic/mental simulation tasks, static questions, and subjective ratings. These approaches can be applicable for measuring comprehension of a DM algorithm as well, because understanding a computer program is to comprehend its functions, logic, and process flow, the same as understanding how an algorithm derives models. An experiment conducted in Dunsmore and Roper indicated that the dynamic/mental simulation is the most reliable and easily controllable approach. Dynamic/mental simulation tasks typically involve some kind of paper-based execution of the program via walkthrough or what-if questions, whereas static questions usually include identification of variables and functions. Therefore, an algorithm understanding questionnaire was developed in this experiment; it consisted of seven questions—five dynamic questions and two static questions. The static questions asked the definitions of support and confidence, and the dynamic questions asked the process flow of association rules modeling and how changes of thresholds might affect the results. Among these questions, six were multiple choice with one correct answer, and one required ordering steps involved in association rules modeling. Provided that each question had only one correct answer, the maximum score of this questionnaire was 7 points.

**Satisfaction with the task.** The Job Diagnostic Survey (JDS; Hackman & Oldham, 1975, 1980) is a widely used instrument to study the effects of job characteristics on people. It has been successfully applied in many areas (Eaton & Thomas, 1997; Medcof, 1996; Shikdar & Das, 2003). Most of its subscales have internal consistency of higher than or close to 0.7. It contains subscales to measure the characteristics of jobs (core job dimensions), cognitive feeling about job experiences (critical psychological states), and reactions to a job (personal and job outcomes). One of the reactions is job satisfaction, which was defined in Hackman and Oldham (1975) as personal, affective reactions or feelings a person obtains from performing a job.

In this experiment, a 7-point Likert scale modeling satisfaction questionnaire (MSQ) with nine questions was developed by modification of the Hackman and Oldham's JDS: two questions regarding general satisfaction, one question regarding trust in results, one question regarding satisfaction with worthwhile accomplishment, one question regarding satisfaction with independent thought and action, one question regarding satisfaction with the amount of challenge, and two questions regarding satisfaction with acquired knowledge about the algorithm.

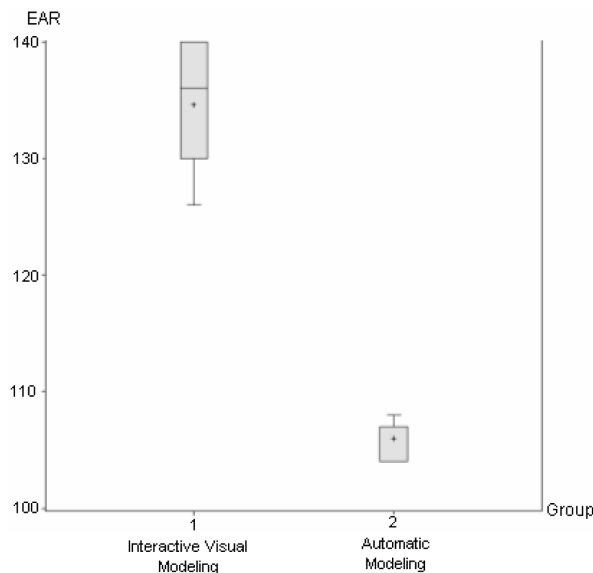
**Statistical design.** A one-way between-subject design was adopted in this experiment to rule out learning effects. Each participant was randomly allocated to only one type of association rules modeling process—IVAR or automatic modeling.

### **Experiment Results**

In the following paragraphs of this section, Groups 1 and 2 correspond to IVAR and automatic modeling processes, respectively. In each statistical test conducted

in this experiment, the normality and constant variance assumptions were examined using the Anderson–Darling and Bartlett’s tests, respectively; both assumptions seemed valid.

**Hypothesis 1: Compared to automatic modeling, IVAR improves effectiveness of modeling.** Figure 9 shows the box plots of EAR for the two types of modeling process. As indicated in Figure 9, the mean of EAR was about 134.6 ( $SD = 4.90$ ) and 106.8 ( $SD = 0.79$ ) for Groups 1 and 2, respectively. Three participants in Group 1 removed three large two-itemsets, which contained the items in the given list of expected uninteresting rules without realizing that they had not only eliminated the uninteresting rules but also interesting rules that contained these items. As a consequence, the EAR of these participants was much lower than that of other participants in the same group, because several interesting rules with high reward scores were removed due to participants’ inappropriate actions. This incident suggested that the increased flexibility offered in interactive modeling may also bring about more opportunities for humans’ inappropriate operations. One potential solution to this dilemma is to clearly explain the provided functions and make users aware of the consequences of their actions. To remedy the violation of constant variance assumption, data transformation was applied by taking the reciprocal of the original data. Analysis of variance (ANOVA) results indicated that, compared to automatic modeling, IVAR significantly,  $F(1, 18) = 397.57, p < .0001$ , improved the effectiveness of modeling, with the EAR increasing by about 26% in this experiment.



**FIGURE 9** Boxplots of EAR for the interactive visual associations rules modeling and automatic modeling.

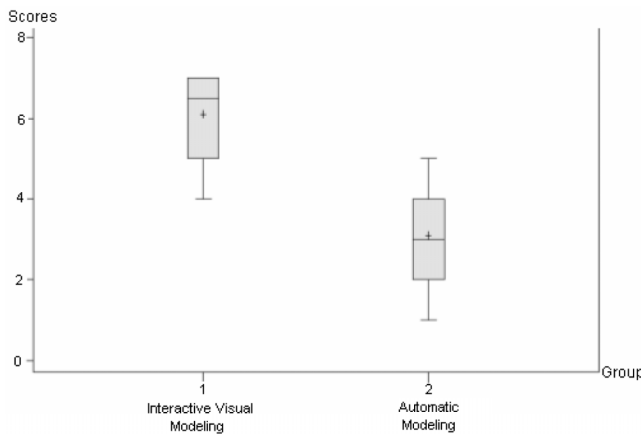
**Hypothesis 2: Compared to automatic modeling, IVAR improves understanding of the applied algorithm.** ANOVA results indicated that scores of participants in the first group were significantly higher than those in the second group,  $F(1, 18) = 29.14, p < .0001$ . The average increase was 3.1 points ( $SD = 0.55$ ) out of 7 points. The box plots of the total scores for these two groups are shown in Figure 10. Each question was compared individually between the two groups using a  $t$  test, and it was found that the correction rates of five questions were significantly different between the two groups, especially those asking about the process flow, which required relatively deeper understanding of the algorithm.

**Hypothesis 3: Compared to automatic modeling, IVAR improves users' satisfaction with the modeling task.** The Cronbach's alpha of the MSQ applied in this experiment was about 0.86. ANOVA results indicated that there were significant differences between the two groups in overall satisfaction,  $F(1, 18) = 5.63, p = .029$ ; satisfaction with independent thought and action,  $F(1, 18) = 8.71, p = .009$ ; satisfaction with amount of challenge,  $F(1, 18) = 4.46, p = .049$ ; and satisfaction with the knowledge gained about the algorithm,  $F(1, 18) = 27.00, p < .0001$ .

Figure 11 shows the correlations among the EAR, the total scores of the algorithm understanding questionnaire, and the six types of satisfaction questions, which implies significant correlations among these measures. Table 2 shows the summary of the results of this experiment.

## 6.2. Experiment 2

The second experiment tested the hypothesis that, compared to nonintegrated data and rule visualizations, integrated data and rule visualizations improves understanding of derived rules. In this article, understanding rules means identification of the characteristics of the data that support them.



**FIGURE 10** Boxplots of the total scores of the algorithm understanding questionnaire for the interactive visual associations rules modeling and automatic modeling.

**Participants**

Ten participants (5 female and 5 male) who had taken part in the first experiment participated in this experiment.

**Apparatus**

As in the first experiment, the visual interfaces in the second experiment were created using Microsoft Visual Basic.Net 2003. All participants performed tasks on a Dell UltraSharp 1905FP 19-in. flat monitor in the experiment.

	<b>EAR</b>	<b>Understanding</b>	<b>Overall</b>	<b>Trust</b>	<b>Accomplishment</b>	<b>Independence</b>	<b>Challenge</b>	<b>Knowledge</b>
<b>EAR</b>	1.000	0.865	0.467	0.392	0.430	0.570	0.476	0.764
<b>Understanding</b>	0.865	1.000	0.723	0.613	0.506	0.674	0.579	0.825
<b>Overall</b>	0.467	0.723	1.000	0.656	0.702	0.747	0.576	0.701
<b>Trust</b>	0.392	0.613	0.656	1.000	0.495	0.601	0.434	0.667
<b>Accomplishment</b>	0.430	0.506	0.702	0.495	1.000	0.711	0.781	0.594
<b>Independence</b>	0.570	0.674	0.747	0.601	0.711	1.000	0.744	0.684
<b>Challenge</b>	0.476	0.579	0.576	0.434	0.781	0.744	1.000	0.580
<b>Knowledge</b>	0.764	0.825	0.701	0.667	0.594	0.684	0.580	1.000

**FIGURE 11** Correlation among effectiveness of association rules modeling (EAR), the total scores of the algorithm understanding questionnaire, and the six types of satisfaction questions.

**Table 2: Summary of Experiment One**

<i>Modeling Process</i>	<i>Measures</i>							
	<i>EAR</i>		<i>Understanding</i>		<i>Overall Satisfaction</i>		<i>Trust</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Interactive Visual Modeling	134.6	4.90	6.1	1.10	5.6	0.70	5.5	0.85
Automatic Modeling	106.8	0.79	3.1	1.37	4.7	1.18	4.9	0.78
<i>F</i> value	395.57		29.14		5.63		2.84	
<i>p</i> value	< .0001		< .0001		.029		.109	
% increase	26.03		96.77		19.15		12.24	
<i>Modeling Process</i>	<i>Measures</i>							
	<i>Worthwhile Accomplishment</i>		<i>Independence</i>		<i>Challenge</i>		<i>Knowledge</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Interactive Visual Modeling	5.3	0.82	5.7	0.82	5.8	1.14	5.5	0.62
Automatic Modeling	4.4	1.18	4.6	0.84	4.7	0.95	4.1	0.67
<i>F</i> value	3.1		8.71		4.46		27	
<i>p</i> value	.095		.009		.049		< .0001	
% increase	20.45		23.91		23.40		34.15	

### **Task**

Participants were successively presented with five multiple-choice questions that asked about characteristics of the data that support some derived rules. They were asked to answer these questions correctly and quickly; they had 5 min at most for each question.

### **Procedure**

As in the first experiment, five steps were followed in this experiment: tutorial, training, practice, formal task, and poststudy.

### **Experiment Design**

**Independent variable.** The main independent variable was the type of visualization interface. Two types of visualization interfaces were compared. In the proposed form of integrated data and rule visualizations, visualization of the data that support a rule could be directly accessed as described in section 4.2. On the other hand, the design of nonintegrated data and rule visualizations adopted in this experiment simulated most current association rules modeling tools, in which only visualization of the entire dataset is provided other than the spreadsheet view of the original dataset. For the latter, a spreadsheet view of the dataset was provided using Microsoft Access because of its popularity.

The introduction of a second independent variable—type of task—was to check whether complexity of tasks would affect the users' performance. Five tasks, which represented the typical information of data characteristics, were given in each interface. These tasks included discovering the most dominant groups in some attributes of the data that supported some specified rules, identifying missing groups in some attributes of the data corresponding to some specified rules, and finding out the percentage that groups in a specific attribute accounted for the data supporting some specified rules.

**Dependent variable.** The dependent variable—understanding of rules—was measured by administering a questionnaire that consisted of five multiple-choice questions that corresponded to the tasks described in the last paragraph. Each question had one correct answer; therefore, the maximum score of the questionnaire was 5 points. Both their task completion time and correction rate were recorded.

**Statistical design.** A two-factorial within-subject design was applied in this experiment to reduce the random variation of the dependent variable. Complete counterbalancing and partial counterbalancing were used across the type of visualization interface and type of task, respectively, to reduce the possible carryover effects.

## Experiment Results

ANOVA results indicated that, compared to the nonintegrated visualization, integration of data and rule visualizations significantly reduced task completion time,  $F(1, 76) = 252.96, p < .0001$ , and increased correction rates,  $F(1, 76) = 3.79, p = .05$ . ANOVA results also suggested significant effect of the task type on task completion time,  $F(4, 76) = 31.42, p < .0001$ , and correction rate,  $F(4, 76) = 3.79, p = .007$ , and significant interaction effects between the type of interface and type of task on task completion time,  $F(4, 76) = 21.17, p < .0001$ , and correction rate,  $F(4, 76) = 5.39, p = .0007$ . Therefore, in the further analysis, paired  $t$  tests were applied to examine the differences of the correction rates and completion times between the two interfaces for each task individually, and Tukey's multiple comparison was conducted to compare the impact of the type of task on users' performances. These tests suggested that the integrated interface significantly reduced task completion time in all tasks (shown in Table 3) and increased correction rates for the most complex task (Task 5 in Table 3),  $t(9) = 1.96, p = .04$ , in the experiment. In addition, the advantage of the integrated interface over the nonintegrated one was more significant when the number of data records increased (Tasks 1 vs. 3 in Table 3) and the task became more complex (Task 5 in Table 3). All participants in both groups correctly answered the four questions corresponding to the first four tasks in Table 3. However, for the most complex task, all participants in the integrated interface completed it correctly, but 3 (of 10) participants in the other group gave wrong answers.

Participants' answers to the two satisfaction questions after the formal task also indicated that they were more satisfied with the integrated interface than the nonintegrated one in terms of understanding of rules; Cronbach's alpha of these questions was about 0.92.

## 7. CONCLUSION

In this article we presented how visualization techniques can be applied to facilitate association rules modeling, particularly what visualization elements should

**Table 3: Summary of Differences of Task Completion Time Between Two Interfaces in Experiment Two**

Difference of Task Completion Time Between Two Interfaces	Tasks				
	1	2	3	4	5
<i>M</i>	45.1	64	81.7	32.5	170.7
<i>SD</i>	6.59	6.99	10.27	8.74	14.37
<i>T</i> value	21.642	28.954	25.157	11.754	37.564
<i>p</i> value	< .0001	< .0001	< .0001	< .0001	< .0001
Tukey's test		B	B	C	A
	C	C			

*Note.* Tasks with the same letter shown in the Tukey's test were not significantly different in completion time difference between the integrated and nonintegrated interfaces.

be included and how to display them. We proposed original designs of visualization of rules, integration of data and rule visualizations, and IVAR process. Two experiments were conducted to test four hypotheses regarding the potential advantages of the proposed IVAR process and integration of data and rule visualizations.

It is not our argument, however, that IVAR process should replace its automatic counterpart. The automatic modeling takes much less time (Ankerst, Ester, & Kriegel, 2000; Ware, Frank, Holmes, Hall, & Witten, 2001) and thus is favorable if users want to get a quick overview of the possible rules that can be derived. They can then turn to the interactive visual modeling when simply manipulating parameters at the beginning of the process will not produce satisfactory results. In addition, as suggested by the experiment, the effectiveness of interactive visual modeling can vary greatly among users because of its relatively higher flexibilities. Hence, a clear understanding of the provided functions is critical to avoid unwise decisions.

Finally, we have a couple of comments on the applicability of the experiment results. Only graduate engineering students were recruited for the experiments because it was assumed that they would have the ability to understand the tutorial of the applied association rules algorithm in a short time at the beginning of the experiment. In addition, although participants were given some time to practice with the interfaces before the formal tasks started, their behaviors might not reflect those of experienced users due to the short duration of the experiments.

## REFERENCES

- Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in very large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 207–216.
- Ankerst, M., Ester, M., & Kriegel, H.-P. (2000). Towards an effective cooperation of the computer and the user for classification. *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD '2000)*, 178–188.
- Ankerst, M., Keim, D. A., & Kriegel, H.-P. (1996). Circle segments: A technique for visually exploring large multidimensional data sets. *Proceedings of IEEE Visualization 96*.
- Appice, A., Ceci, M., Lana, A., Lisi, F. A., & Malerba, D. (2003). Discovery of spatial association rules in geo-referenced census data: A relational mining approach. *Intelligent Data Analysis*, 7, 541–566.
- Blake, C. L., & Merz, C. J. (1998). *UCI Repository of machine learning databases*. Irvine, CA: University of California, Department of Information and Computer Science. Retrieved April 15, 2005 from <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- Bruzzese, D., & Davino, C. (2003). Visual post analysis of association rules. *Journal of Visual Languages and Computing*, 14, 621–635.
- Buchter, O., & Wirth, R. (1999). Exploration of ordinal data using association rules. *Knowledge and Information Systems*, 1, 393–414.
- Card, S. K. (2003). Information visualization. In J. A. Jacko & A. J. Sears (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications* (pp. 544–582). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Chakravarthy, S., & Zhang, H. (2003). Visualization of association rules over relational DBMSs. *Proceedings of 2003 ACM Symposium on Applied Computing*, 922–926.

- Cleveland, W. S., & McGill, R. (1987). Graphical perception: The visual decoding of quantitative information on statistical graphs (with discussion). *Journal of the Royal Statistical Society Series A*, 150, 192–229.
- Crapo, A. W., Waisel, L. B., Wallace, W. A., & Willemain, T. R. (2000). Visualization and the process of modeling: A cognitive–theoretic view. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 218–226.
- Davis, S. (1995). A guessing measure of program comprehension. *International Journal of Human–Computer Studies*, 42, 245–263.
- Dunsmore, A., & Roper, M. (2000). *A comparative evaluation of program comprehension measures* (Tech. Rep. No. EfoCS 35–2000). Department of Computer Science, University of Strathclyde, Glasgow, United Kingdom.
- Eaton, N., & Thomas, P. (1997). Job diagnostic surveys of pediatric nursing: An evaluative tool. *Journal of Nursing Management*, 5, 167–174.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 1–36). Menlo Park, CA: AAAI Press.
- Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 60, 159–170.
- Hackman, J. R., & Oldham, G. R. (1980). *Work redesign*. Boston: Addison Wesley.
- Hix, D., & Hartson, H. R. (1993). *Developing user interfaces: Ensuring usability through product and process*. New York: Wiley.
- Hofmann, H., & Wilhelm, A. (2001). Visual comparison of association rules. *Computational Statistics*, 16, 399–415.
- IBM. (n.d.). *IBM Intelligent Miner for Data*. Retrieved August 20, 2005 from <http://www306.ibm.com/software/data/iminer/>
- Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer*, 1, 69–97.
- Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8, 1–8.
- Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A. (1994). Finding interesting rules from large sets of discovered association rules. *Proceedings of the Third International Conference on Information and Knowledge Management*, 401–407.
- Kreusel, M., & Schumann, H. (2002). A flexible approach for visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8, 39–51.
- Liu, Y., & Salvendy, G. (2005). Visualization support to facilitate association rules mining process: A survey. *Ergonomia—An International Journal of Ergonomics and Human Factors*, 27, 11–23.
- Matsumoto, K., & Hashimoto, K. (1998). Association rule filter for data mining in call tracking data. *IEICE Transactions on Communications*, E81-B, 2481–2486.
- Medcof, J. W. (1996). The job characteristics of computing and non-computing work activities. *Journal of Occupational & Organizational Psychology*, 69, 199–212.
- Morris, W. T. (1967). On the art of modeling. *Management Science*, 13, B707–717.
- Motwani, M., & Rajput, A. (2002). Mining multiple level generalized association rules in large databases. *Ultra Scientist of Physical Sciences*, 14, 401–406.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.
- Novick, D. (1997). What is effectiveness? *CHI'97 Workshop on HCI Research and Practice Agenda Based on Human Needs and Social Responsibility*.
- Ong, H.-H., Ong, K.-L., Ng, W.-K., & Lim, E.-P. (2002). CrystalClear: Active visualization of association rules. *International Workshop on Active Mining*.
- Purple Insight. (n.d.). *Purple Insight Mineset 3.1*. Retrieved August 20, 2005 from <http://www.purpleinsight.com/products/mineset/index.html>

- Sahar, S. (1999). Interestingness via what is not interesting. *Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 332–336.
- SAS Institute, Inc. (n.d.). *SAS Enterprise Miner*. Retrieved August 20, 2005 from <http://www.sas.com/>
- Shikdar, A. A., & Das, B. A. (2003). Strategy for improving worker satisfaction and job attitudes in a repetitive industrial task: Application of production standards and performance feedback. *Ergonomics*, 46, 466–481.
- Shneiderman, B. (1994). Dynamic queries for visual information seeking. *IEEE Software*, 11, 70–77.
- Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8, 970–974.
- Sprenger, T. C., Gross, M. H., Bielser, D., & Strasser, T. (1998). Ivory—An object-oriented framework for physics-based information visualization in Java. *Proceedings of IEEE Information Visualization '98*, 79–86.
- Srikumar, K., & Bhasker, B. (2004). Efficiently mining maximal frequent sets in dense databases for discovering association rules. *Intelligent Data Analysis*, 8, 171–182.
- Thearling, K., Becker, B., DeCoste, D., Mawby, B., Pilote, M., & Sommerfield, D. (2002). Visualizing data mining models. In U. M. Fayyad, G. Grinstein, & A. Wierse (Eds.), *Information visualization in data mining and knowledge discovery* (pp. 205–222). San Francisco: Morgan Kaufmann.
- Tory, M., Moller, T., Atkins, M. S., & Kirkpatrick, A. E. (2004). Combining 2D and 3D views for orientation and relative position tasks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 73–80.
- Wang, K., He, Y., & Han, J. (2003). Pushing support constraints into association rules mining. *IEEE Transactions on Knowledge and Data Engineering*, 15, 642–658.
- Ware, M., Frank, E., Holmes, G., Hall, M., & Witten, T. H. (2001). Interactive machine learning: Letting users build classifiers. *International Journal of Human Computer Studies*, 55, 281–292.
- Webb, G. I. (2000). Efficient search for association rules. *Proceedings of the Sixth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining*, 99–107.
- Wegner, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85, 664–675.
- Willemain, T. R. (1995). Model formulation: What experts think about and when. *Operations Research*, 43, 916–932.
- Wong, P. C., Whitney, P., & Thomas, J. (1999). Visualizing association rules for text mining. *Proceedings of the 1999 IEEE Symposium on Information Visualization*, 120–127.
- Yang, L. (2005). Pruning and visualizing generalized association rules in parallel coordinates. *IEEE Transactions on Knowledge and Data Engineering*, 17, 60–70.