Mining Patterns for Visual Interpretation in a Multiple-Views Environment

José F. Rodrigues Jr., Agma J.M. Traina, and Caetano Traina Jr.

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
C.P. 668 - CEP 13560-970 - São Carlos, SP, Brazil
{junio,agma,caetano}@icmc.usp.br

Abstract. This chapter introduces a novel systematization aiming at extending the application range of Information Visualization and Visual Data Mining. We present an innovative framework named Visualization Tree in order to integrate multiple data visualizations assisted by novel visual exploration techniques. These exploration techniques are named Frequency Plot, Relevance Plot and Representative Plot, and are integrated according the proposed Visualization Tree framework. The systematization of visualization techniques enabled by these concepts defines a Visual Data Mining environment where multiple presentation workspaces are kept together, linked according to analytical decisions taken by the user. Our emphasis is on developing an intuitive and versatile multiple-views system that helps the user to identify visual patterns while interpreting multiple data subsets. In this context, the analyst is able to draw and summarize several subsets that are inspected simultaneously each in a dedicated workspace.

1 Introduction

Information Visualization (Infovis) embodies many techniques [7] [20] designed to help analysts to visually explore large multidimensional datasets [17]. These techniques take advantage of the fact that it is easy for humans to understand the properties present in a dataset when these properties are represented graphically. In the process of finding a needle in a haystack, Infovis includes techniques to identify trends, outliers, correlations and clusters. They can be used to validate and formulate hypothesis and to highlight interesting properties hidden in the data. Such techniques enlarge the value of stored data, aiding the decision-making processes [16].

For Infovis techniques, the dimensions (attributes) of a multi-dimensional dataset can be seen as points in a k-dimensional space, where k is the number of dimensions. Therefore, the task of data visualization becomes mapping the k-dimensional dataset into the two dimensions of the computer display. However, due to spatial limitations of display devices and due to human sensorial limitations, existing Information Visualization techniques are usually limited to displaying roughly a thousand items [8]. In fact, most of the visualization techniques do not scale well with respect to the number of objects in a dataset [9], generating scenes with a reduced number of just noticeable differences. In this scenario, the challenge for the Infovis science is to create robust

S.J. Simoff et al. (Eds.): Visual Data Mining, LNCS 4404, pp. 196–214, 2008.
© Springer-Verlag Berlin Heidelberg 2008

visualizations that satisfy the limit of visual stimuli at the same time that the user is not overwhelmed by messy imagery.

Grinstein and Ward [11] affirm that, to bypass these drawbacks, the limitations in screen resolution and color perception can be treated with the help of multiple linked visualizations (multiple-views). Besides multiple-views, Infovis can also count on the support provided by interaction and by data mining techniques. The combination of such tools comprehends procedures that define the Visual Data Mining specialty, that is, the use of data mining methods coupled with interactive visual presentation. Interaction techniques provide mechanisms – as, for example, zooming, distortion and panning – to handle complexity in visual presentations. This complexity is caused primarily by the overlapping of graphical items and by visual cluttering. Data mining techniques, in turn, range from simple statistical summarization to complex pattern-discovery algorithms. Hence, in order to effectively aid an analyst in discovering information hidden in the data, it is desirable that a visualization environment provides: heterogeneous visualization techniques, a uniform set of interaction techniques and access to a systematic set of mining operations to assist the analytical process.

In this work we present a visual environment benefiting from rich interaction, visualization and summarization facilities, and from multiple linked views in order to amplify the possibilities of the Information Visualization practice. The goal is to allow the user to visually find patterns that lead her/him to useful interpretations in the context of her/his application domain. To do so, we utilize three dynamical approaches for statistical-based presentation in response to user interaction. These approaches are named *Frequency Plot*, *Relevance Plot* and *Representative Plot*. Frequency Plot intends to tackle the excessive population of data items and the consequent overlapping of graphical elements in visualization scenes. Relevance Plot describes a way to present a dataset according to what was stated as important by the user in accordance to an interactively defined set of properties. Representative Plot benefits from color and size emphasis in order to present statistical summarization data over a given visualization workspace. These methodologies are joined in a rich-interaction environment fused by an innovative multiple-views integration named Visualization Tree. In the Visualization Tree, several visualizations (workspaces) are linked in a structure that unfolds in tree-like manner in response to user interaction.

The remainder of the paper is organized as follows. Section 2 presents works on visualization techniques and other related concepts involved in this paper. Section 3 presents the main concepts proposed in this paper, including the Visualization Tree environment and the operations of visualization pipeline and composition that are employed in the VisTree environment. Section 4 further details the exploration techniques employed by the analyst in order to summarize the datasets being visualized. Section 5 presents experiments that show the proposed techniques being used to discover knowledge from a real world dataset. Finally, Section 6 concludes the paper.

2 Related Work

Highly populated databases usually have overlapping values, or a too spread distribution of data, which leads to visualizations with overlapping graphical items. This fact

degenerates many multivariate visualization techniques as the Parallel Coordinates [12] and the Scatter Plots [21]. These shortcomings have been dealt by the computer science community in many works. Artero *et al* [3] proposed two methods named Interactive Parallel Coordinates Frequency Plots and Interactive Parallel Coordinates Density Plots. These methods create bi-dimensional frequency histograms for each pair of attributes to be exhibited, at consecutive axes, over Parallel Coordinates. Although their work promotes a great advance on the use of Parallel Coordinates, it has not been extended for Infovis techniques in general. Fua *et al* [10] utilize hierarchical clustering to generate visualizations while expressing aggregation information. They proposed a complete navigation system to allow the user to achieve the intended level of details. The proposal is introduced as an implementation named Xmdv Tool [19], which comprises many multivariate visualization schemes. The drawback of this system is its complex navigation interface. Wong and Bergeron [22] used wavelets to present data in lower resolutions without losing its original behavior. This technique takes advantage of the wavelets intrinsic property of image details separation, which also leads to reduction of details if desired. Although there is a predicted data loss that might degrade the analytical capabilities, the use of this tool can enhance the interactive filtering activity. Keim [11] claims that "in exploring large datasets, interactive filtering is important to interactively partition the dataset into segments and focus on interesting subsets". Following that principle, many authors developed tools aiming at the interactive filtering goal, as the Magic Lenses [5] and the Dynamic Queries [1] techniques.

An interesting approach in selective exploration is presented in the VisDB tool [14], which includes an interface to specify a query whose results will be the basis for the visualization scene. A color scheme determines the hue of the visual items according to their proximity to the items returned by the query. The analysis depends on the user capability to join information obtained from multiple-views, since each data dimension is presented in a separate scene. On multiple-views systems, Ahlberg and Shneiderman [2] emphasize the importance of "progressive refinement of search parameters, continuous reformulation of goals and visual scanning to identify results". The idea is to perform one operation a number of times in more than one exhibition window. Extending such concept, Boukhelifa and Rodgers [6] discusses the importance of coordinating the data presentation in multiple-views visualization systems. In such systems, a natural demand is the Link & Brush functionality. Link & Brush [15] works propagating the user's interaction operations through various, and probably distinct, visualization scenes in order to integrate the advantages of different visual approaches.

3 Multiple-Views within the Visualization Tree

This work introduces the concept of *Visualization Tree* (VisTree for short), which encompasses an environment composed of multiple-views organized as a hierarchical structure. In the VisTree system, each view (workspace) integrates three elements: the data to be visualized, a visualization technique, and an exploration technique. The data to be visualized focus on multidimensional datasets. The visualization technique can be any of the existing ones for multidimensional datasets, such as the well-known Parallel Coordinates, Scatter Plots, Star Coordinates [13] or Table Lens [18] (those which are

available in VisTree current implementation). The exploration technique can be one of three methods that define the second contribution of this work, as explained in section 4.

In order to work with VisTree system, the user initially chooses a visualization technique to fully render a dataset of interest. This dataset then is presented in the root workspace of the system. By interacting with this workspace, the user selects subsets of data using interactive filtering, and sends the results to create derived workspaces through a derivation operation. This operation can be performed iteratively and recursively so that the VisTree progressively unfolds into a tree-like structure where each sub tree can be manipulated either as a whole or as individual workspaces. Along this chapter we describe the VisTree functionalities and show their usefulness to identify visual patterns. In this section, specifically, we present the mechanisms that support the VisTree, which are denominated *pipeline* and *composition*.

We illustrate the VisTree technique using the Cars dataset (obtained from `http://stat.cmu.edu/datasets`), which has 406 records and 8 dimensions. The dimensions are: fuel consumption in miles per U.S. gallon (MPG), number of cylinders (CYLINDERS), engine displacement in cubic inches (DISPLACEMENT), output of the engine in horsepower (HORSEPOWER), vehicle weight in U.S. pounds (WEIGHT), time to accelerate from 0 to 60 mph (ACCELERATION), model year (YEAR), and origin of the car (ORIGIN). The last attribute organizes the dataset into three categories: American cars (1), Japanese cars (2) and European cars (3).

3.1 Visualization Pipeline

The visualization pipeline operation enables the analyst to derive new visualization workspaces based on subsets of the data visually selected. As seen in Figure 1(a), following a pipeline, the selected data is propagated to a new visualization workspace that presents the selected elements in a new visual environment. In this new workspace, the boundaries at each dimension are rescaled according to the subset being plotted. An example is shown in Figure 1(b) using the Parallel Coordinates visualization technique. The resulting effect presents the graphical elements expanded into a wider visual space, providing better perception of details.

In figure 1 we show that, in VisTree system, one visualization scene is used to create a descendant new scene in order to present the data items of interest in more details. To perform the pipeline operation the user needs to visually select interesting items via interactive filtering. Then, to create a new workspace, it is necessary to choose the visualization technique of the next workspace. This new visualization will be fed by the data items interactively selected.

3.2 Visualization Composition

The visualization composition stands for the combination of two or more visual workspaces. That is, it stands for the gathering of the data elements being graphically presented in a given set of workspaces. Figure 2 demonstrates this operation through the composition of two distinct visualization workspaces. Figure 2(a) shows a workspace based on the Star Coordinates technique presenting the American cars with 4 cylinders. Figure 2(b) presents a Scatter Plots visualization of European cars with 3 cylinders. In

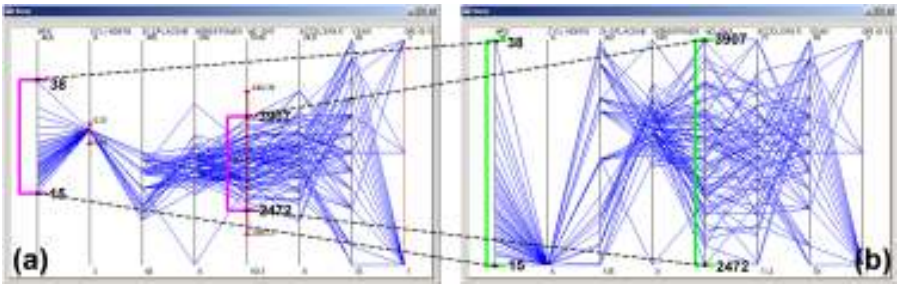

Fig. 1. Example of the pipeline operation. The visual query of a workspace can be better observed in a dedicated workspace: (a) the 6 cylinders cars weighting from 2472 to 3907 pounds are visually selected, (b) the information from the previous workspace feeds the visualization in the new workspace.

the sequence, Figure 2(c) presents the composition of both figures in a workspace based on the Parallel Coordinates technique. The composition is obtained executing a logical *OR* and it is triggered under user request. The final workspace shows that European cars (ellipse at the bottom), even with fewer cylinders, are as powerful (horsepower) as American cars with 4 cylinders (ellipse at the top).

3.3 The System

In this section we describe how the concepts of pipeline, composition and multiple visualizations are integrated into our system. The main constituent of the VisTree systematization is the visual pipeline concept, which promotes progressive refinements according to user interaction. After selecting a region of interest, a new workspace can be created (pipelined) and the exploration continues. The VisTree system is responsible for tracking the user-driven analysis and to present it in a tree-like scheme.

Figure 3 shows how pipelining and composition can be joined into a unique exploration environment. As it can be seen, pipelining proceeds horizontally and composition proceeds vertically. The pipeline determines the creation of levels that constitute the browsing tree, whereas compositions can occur only among two levels of the tree. Pipelined scenes progressively advance toward the right part of the screen while composed scenes are placed at the lower part. As the scene is populated, panning and zooming are necessary to focus or overview the environment.

Figure 4 shows the interactive aspect of the proposed multiple-views methodology. Given a data source, each cycle of the scheme determines the reading of different data portions. The reading is followed by the generation of a new visualization workspace or by an interactive transformation of an existing workspace. The workspaces, in turn, can be interactively explored by zooming/panning and by visual queries. After visual querying, pipeline refinement or composition can be used to determine a new node in the browsing tree. This node will be loaded with the selected data items.

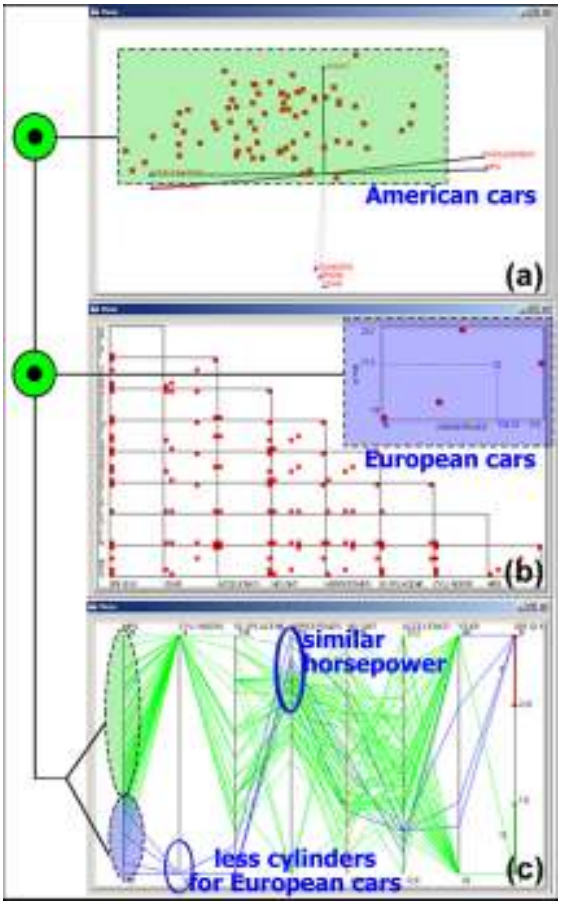

Fig. 2. Example of the visualization composition operation

3.4 Features of the VisTree Methodology

The main goal of our multiple-views technique is to allow a broad range of exploration facilities to guide the analyst into a better usage of information visualization techniques. These facilities include:

- Explorative memory – the VisTree visualization structure allows the user to keep track of the decision steps that guided to a specific visual configuration;
- Heterogeneous visualizations – enhanced data analysis due to the integration of different visualization techniques;
- Unlimited detailing – the pipeline refinement of a workspace enables a more detailed observation of the graphical elements. The repetition of this process allows the exploration to proceed until a single data element is visualized in a dedicated workspace;

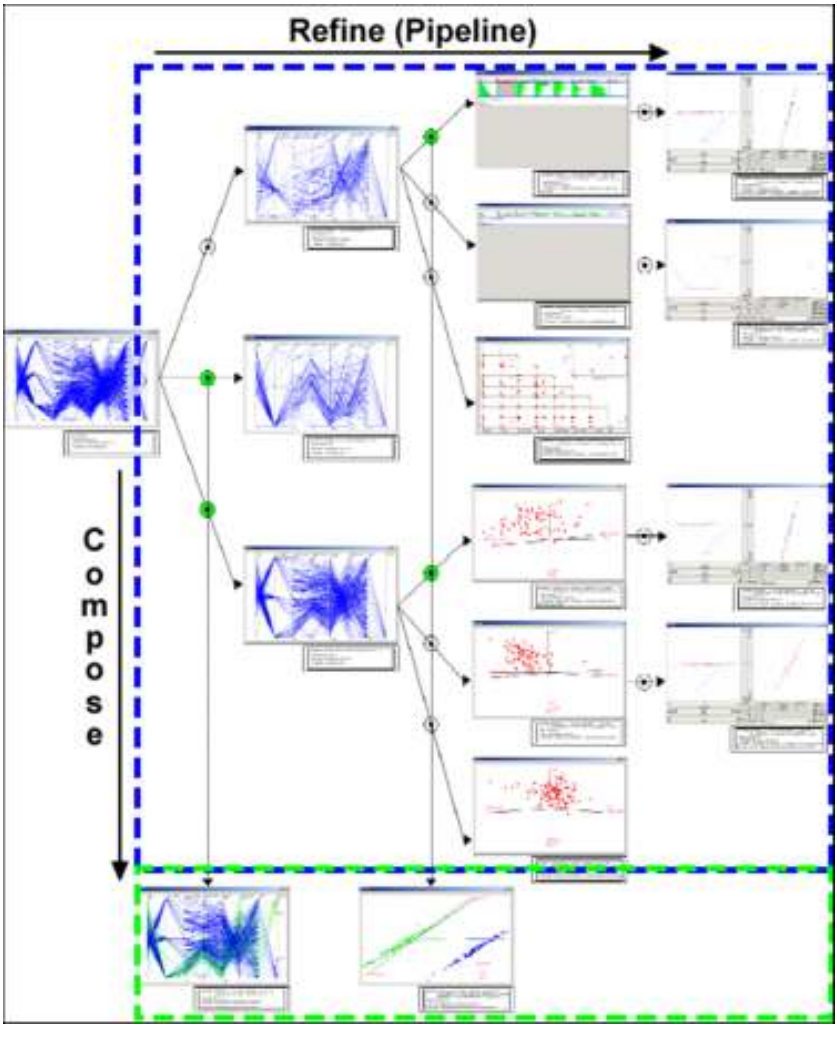

Fig. 3. An example of the Visualization Tree system showing three levels of horizontal visual pipeline refinement (central dashed square), and two vertical visualization compositions (dashed square at the bottom). The arrows of the tree delineate the user decisions. The compositions embody the workspaces indicated by the color-filled circular selectors.

- Undo functionality – when the analyst redefines a visualization, it is easy to lose track of the former configuration, reducing the efficacy of the exploration in case of misleading decisions. The tree-like structure stores the former steps taken by the user, who can interactively undo any sequence of previous steps;
- Common user actions – integrating several visualization techniques into a single interactive environment provides for a better usability of each technique. In VisTree, user actions supported by one visualization are common to all the others;

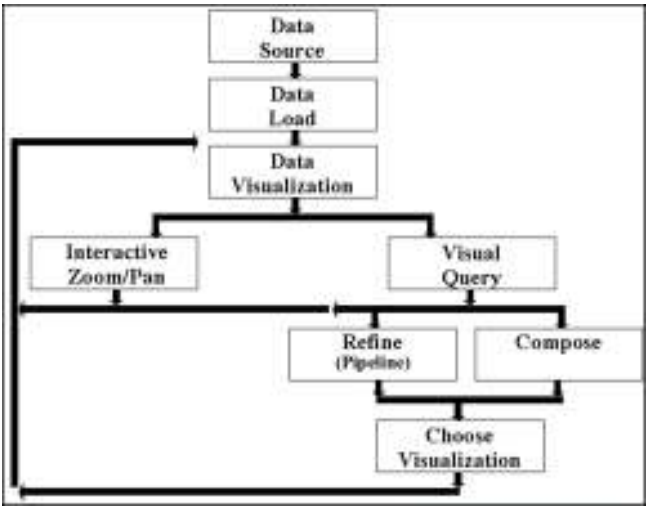

Fig. 4. The structure of the proposed interaction systematization

- Enhanced visualization – the overall benefit of the proposed system is a generalized enhancement of the visualization process. The interactivity allows the use of multiple heterogeneous visualizations bringing new possibilities for the analyst. The scalability offered by the system exceeds that of previous single-visualization applications.

4 Exploration Techniques

This section presents a complementary part of our contribution. We present three exploration techniques, Frequency Plot, Relevance Plot and Representative Plot, which aim at visually summarizing the contents of the workspaces that compose the VisTree. In the VisTree environment, such functionality enables the user to have several visualizations represented by just a few graphical items that can be analyzed alone or comparatively. As will be illustrated along the examples, the visual summarization of multiple workspaces results on macro visualization scenes where analytical perception is condensed.

Together with the VisTree systematization, these techniques intend to deal with the problems caused by the limited amount of elements that can be simultaneously presented in traditional visual exploration environments. Their principle is to benefit from interactive filtering added with visual color effects that highlight or that modify the way that the scene is rendered, aiding user perception with summarization data.

For our work, we assume datasets D with n elements and k-dimensions. Hence, $D = \{V_0, V_1, ..., V_{n-1}\}$ and each element $V_w = \{v_{(0,w)}, v_{(1,w)}, ..., v_{(k-1,w)}\}, \forall V_w \in D$. Alternatively, D can be viewed as an ordered set of k collections, $D = \{D_0, D_1, ..., D_{k-1}\}$, where each D_x corresponds to a column in dataset D so that $|D_x| = n, \forall D_x \in D$. Thus, for any value $v_{(x,w)}$, it is true that $v_{(x,w)} \in D_x$ and that $v_{(x,w)} \in V_w$.

4.1 Frequency Plot

The *Frequency Plot* exploration technique provides an interaction mechanism that combines filtering and dynamical analysis in order to visually summarize the contents of a given dataset. We define "Frequency" as how commonly a given value can be found inside a collection of values (duplicates are allowed), as follows.

Definition 1. [*Frequency*] — *Given a collection of n values* $D_x = \{v_{(x,0)}, v_{(x,1)}, ..., v_{(x,n-1)}\}$, *function* $q(v_{(x,w)}, D_x) \rightarrow N$ *is a function that counts how many times* $v_{(x,w)} \in D_x$ *occurs in collection* D_x. *We call this computation the frequency of item* $v_{(x,w)}$.

We also state function $m(D_x) \rightarrow D_x$, $m(D_x)$ as the statistical mode value of collection D_x – that is, the value or item occurring most frequently in D_x. This function is necessary for the computation of what we call Frequency Coefficient.

Definition 2. [*Frequency Coefficient*] — *The Frequency Coefficient assumes a value between* 0.0 *and* 1.0 *that indicates how frequently a given value* $v_{(x,w)}$ *is found inside a non-empty collection* D_x. *Using functions q and m, the frequency coefficient of a value* $v_{(x,w)} \in D_x$ *is given by:*

$$f(v_{(x,w)}, D_x) = \frac{q(v_{(x,w)}, D_x)}{q(m(D_x), D_x)}, \; for \; |D_x| > 0 \tag{1}$$

Definition 3. [*Vector of Frequencies*] — *Given a k-dimensional element* $V_w = \{v_{(0,w)}, v_{(1,w)}, ..., v_{(k-1,w)}\}, V_w \in D$, *the correspondent vector of frequencies is given by:*

$$F(V_w, D) = \{f(v_{(0,w)}, D_0), f(v_{(1,w)}, D_1), ..., f(v_{(k-1,w)}, D_{k-1})\} \tag{2}$$

Equation 2 enables us to calculate the frequency for any *k*-dimensional element of a given dataset. The idea of the Frequency Plot technique is to exhibit this calculus with the aid of visual effects of color and size.

Following, we demonstrate the ability of Frequency Plot in summarizing the information contained in the Breast Cancer dataset [4]. This dataset comprises 5,027 attribute values spread into 457 records, each with 11 dimensions. It includes a numeric identifier (attribute 0) and a classifier (attribute 11), which indicates the tumor type (0 for benign and 1 for malign). The remaining fields come from tests over patients' tissue samples carried on clinical laboratories. The tests' results, from the 2nd to the 10th attribute, are intended to determine the nature of breast cancer occurrences.

Using the Parallel Coordinates technique, figure 5(a) presents an overview of the complete Breast Cancer dataset. In Figure 5(b) we have pipelined only the malign cancer records and in Figure 5(c) only the benign cancer records. By comparing Frequency Plots in Figures 5(b) and 5(c) it is possible to see how each attribute contributes to determine the cancer nature. While attributes 3, 4, 7 and 9 clearly follows a distribution dictated by the cancer nature, attributes 2, 5, 6, 8 and 10 are either sparse along their domain or potentially indicate false positives, notably attribute 10.

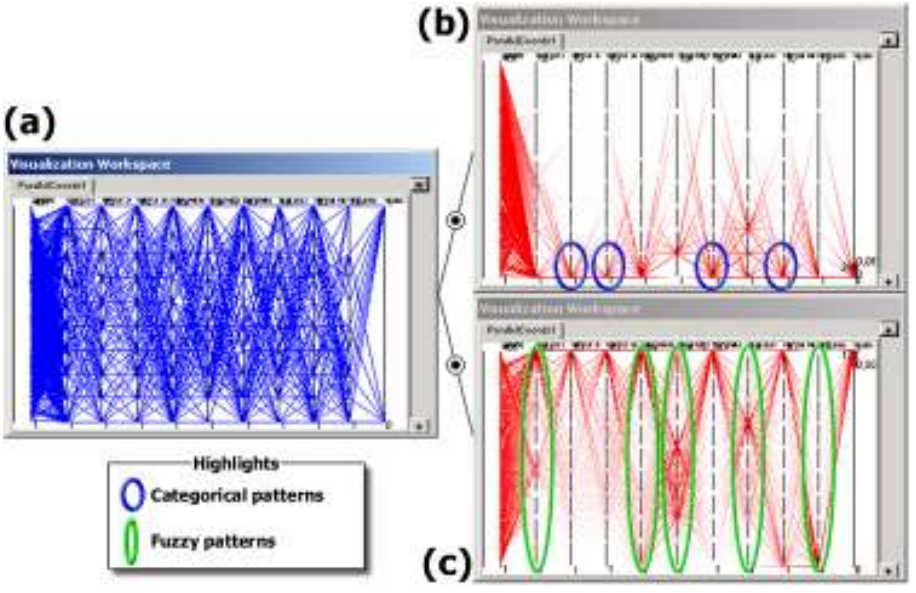

Fig. 5. Using a VisTree to analyze the Breast Cancer dataset. (a) Complete dataset. (b) Frequency Plot of benign cancer records. (c) Frequency Plot of malign cancer records. (Circular marks added for highlighting are not part of the visualization.)

4.2 Relevance Plot

The Relevance Plot exploration technique is based on the concept of presenting the information according to the user's perception of relevance and interestingness in the parts of the data. Following guidance provided by the user, the goal is to reduce the amount of presented information without losing the contextual meaning of the visualization. If the data is close to what the user defined as more interesting in the visualization scene, the correspondent visualization ought to stress this fact. Otherwise, data that is not relevant must have its visibility de-emphasized.

Relevance Plot, which is exemplified in Figure 6, requires that the analyst chooses values, or Relevance Points, from the dimensions being visualized. Hence, given a set of data items D with n elements and k dimensions assumed to be previously normalized (each dimension ranges from 0.0 to 1.0), the following definitions hold:

Definition 4. [*Relevance Point*] — *The Relevance Point (RP) of the x-th dimension, or* RP_x*, is the value belonging to the x-th dimension domain chosen by the user that must be considered to determine the data relevance in that dimension. Only one RP may be chosen per dimension and, for default, if the RP is not set, it takes value* -1.

Once the Relevance Points are set, all the database items must be analyzed relatively to them. For each of the dimensions where an *RP* was set, relevance analysis is performed by calculating the Euclidean distance between the values (attributes) and the relevance value of their respective dimension.

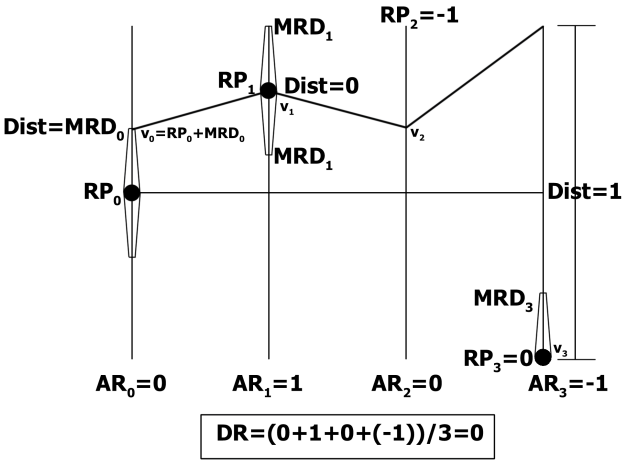

Fig. 6. Schematic for the Relevance Plot calculus

Definition 5. [*Distance to RP*] — *for the w-th k dimensional data record* $V_w = (v_{(0,w)}, v_{(1,w)}, \ldots, v_{(k-1,w)})$, *the distance of its x-th attribute to the x-th RP, or* $Dist(v_{(x,w)}, RP_x)$, *is given by:*

$$Dist(v_{(x,w)}, RP_x) = |v_{(x,w)} - RP_x| \tag{3}$$

Also, for each of the dimensions of the k-dimensional database, a maximum acceptance distance is defined. These thresholds are called *Max Relevance Distances*, or *MRDs*, and are used in the relevance rendering.

Definition 6. [*Max Relevance Distance*] — *The Max Relevance Distance of the x-th dimension, or* MRD_x, *is the maximum distance* $Dist(v_{(x,w)}, RP_x)$ *that a data attribute can assume before having its relevance decreased during relevance rendering. The MRDs take values within the range* $[0.0, 1.0]$.

Based on the *MRDs* and on the calculated distances $Dist(v_{(x,w)}, RP_x)$, a value named *Attribute Relevance* (*AR*) is computed for each attribute. A total of k ARs are computed for each of the n k-dimensional data records of the database.

Definition 7. [*Attribute Relevance*] — *The Attribute Relevance determines the contribution of the x-th attribute of the w-th data item,* $v_{(x,w)}$, *in the relevance analysis. It is given by:*

$$AR(v_{(x,w)}) = \begin{cases} 1 - \frac{Dist(v_{(x,w)}, RP_x)}{MRD_x} \; if \; Dist(v_{(x,w)}, RP_x) \leq MRD_x \\ \frac{Dist(v_{(x,w)}, RP_x)}{(1 - MRD_x)} \; if \; Dist(v_{(x,w)}, RP_x) > MRD_x \\ 0 \quad if \; RP_x = -1 \end{cases} \tag{4}$$

Equation 4 states that:

- For distances $Dist(v_{(x,w)}, RP_x)$ smaller or equal MRD_x, the equation assigns values ranging from 1.0 (where the distances are null) to 0.0 (for distances equal MRD_x);
- For distances $Dist(v_{(x,w)}, RP_x)$ larger than MRD_x, the equation linearly assigns values ranging from 0.0 to -1.0. This last value is assigned to the attributes whose calculated distance is the maximum from RP_x;
- In dimensions without a chosen *RP*, the *AR* assumes a value equal to 0, which does not affect the relevance computation process. Finally, after processing all the database, each k-dimensional data item will have a value computed. This value is called *Data Relevance* (*DR*).

Definition 8. [*Data Relevance*] — *Based on the Attribute Relevancies and on the Max Relevance Distances, the Data Relevance describes how relevant a given data item is. For a given data item* $V_w \in D$, *the DR is the average of its correspondent Attributes Relevancies. That is, for* V_w, *the Data Relevance is given by:*

$$DR(V_w) = \frac{\sum_{x=0}^{k-1} AR(v_{(x,w)})}{R}, v_{(x,w)} \in V_w \text{ and } 0 < R \leq k \tag{5}$$

where *R* is the number of Relevance Points. Hence, reflecting the user defined Relevance Points, the Data Relevance value directly denotes the importance of a data element. In order to visually explicit this fact, we use the *DRs* to determine the color and the size of the graphic elements presentation. Hence, lower values are represented by weaker saturations and smaller sizes, while the higher ones are represented by more stressed saturations and bigger sizes.

Figure 7 presents the Relevance Plot exploration technique over the breast cancer dataset. In order to generate this figure, we started selecting the Parallel Coordinates visualization of the complete dataset. Then we derived three workspaces, over which we defined nine Relevance Points for each of the correspondent visualizations (dimensions 2 to 10). In figure 7(a) the Relevance Points are set to the smallest values of each dimension, in figure 7(b) they are set to the maximum values of each dimension, and in figure 7(c) we set middle values. From figure 7(a) one might conclude that the lowest values at each dimensions' domain indicates class 0 (benign cancer). It also warns that this is not a final conclusion since the visualization reveals some records, in lower concentration, which are classified as 1 (malign cancer). It can be said that false malign cancer diagnosis is possible with this test set. In Figure 7(b) the highest values indicate the records of class 1. It can be seen that false benign cancer analysis can occur, but they are very unusual, since just a shadow of pixels heads to class 0 at the 11th (right most) dimension. Finally, in Figure 7(c) the Relevance Points were set to middle values in order to make an intermediate analysis. Through the visualization, one can conclude that this kind of test set is quite categorical, since just one record is positioned in the middle of the space determined by the dimensions' domains. In such cases, it maybe safer to classify the analysis as malign cancer and to proceed with more clinical exams. Comparing a visualization obtained with the traditional Parallel Coordinates technique (the root workspace) with the workspaces obtained using the Relevance Plot exploration technique, it is possible to reach extra insights allowing further knowledge discovery.

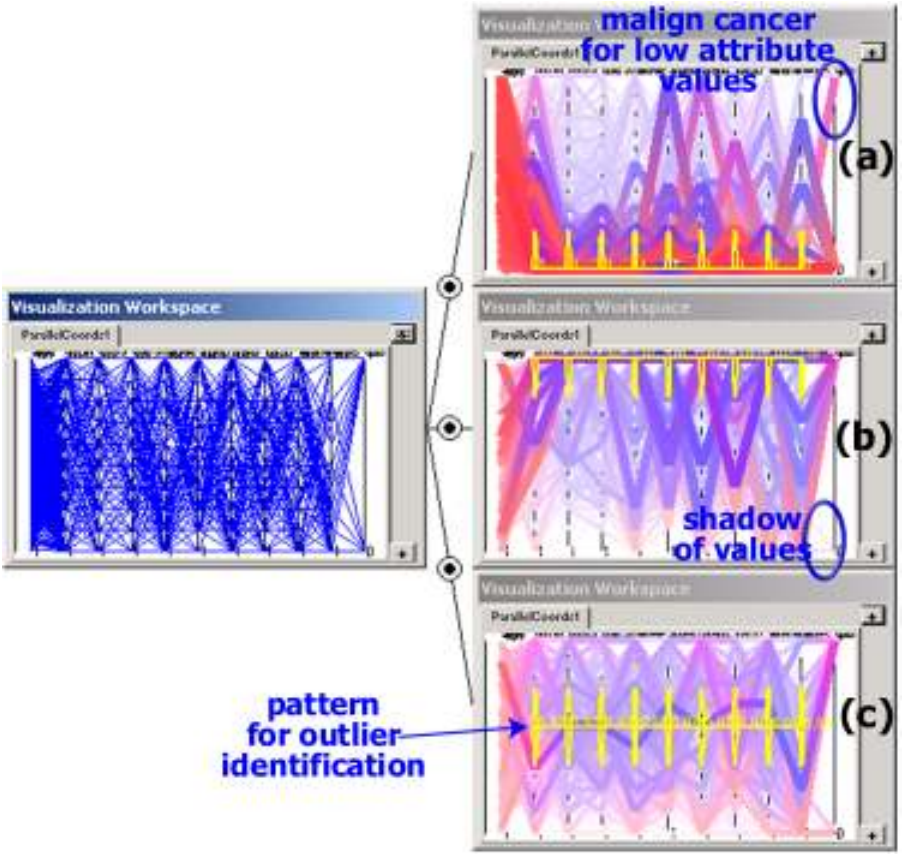

Fig. 7. The Relevance Plot over a Parallel Coordinates scene. (a) Relevance points set to the smallest values of their dimensions. (b) Relevance points set to the maximum values. (c) Relevance points set to middle values.

4.3 Representative Plot

For visual patterns identification, we also employ classical statistical summarization over the visualized data in order to obtain selected subsets represented by a reduced (or by a single) number of graphical entities. The VisTree system implements average, standard deviation, median and mode values, which are used as "representatives" of the summarized data. This practice leads to what we call the Representative Plot exploration technique. In the workspace nodes of VisTree, the raw visualization scene is rendered at the same time that representative information is used to draw extra graphic elements emphasized through color and size stimuli over the image.

In Figure 8 we present the Breast Cancer dataset using the Parallel Coordinates and the Star Coordinates visualization techniques. In such techniques, an array of values, which corresponds to the multidimensional data items under analysis, can be drawn as a polygon by joining each of the values projected onto the correspondent dimension

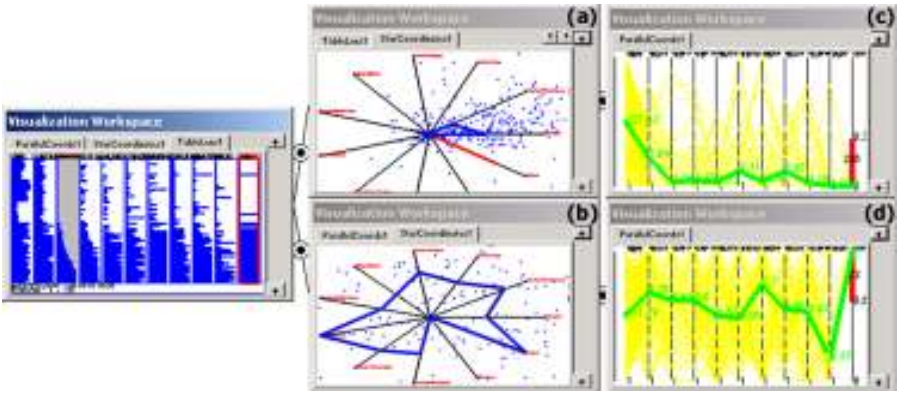

Fig. 8. Representative Plots of the Breast Cancer dataset (a) Median of the benign cancer exams presented over the Star Coordinates. (b) Median of the malignant cancer exams. (c) Average line of the benign cancer exams over the Parallel Coordinates. (d) Average of the malignant exams.

axes. Figure 8(a) presents the median values for the benign cancer exams, while figure 8(b) presents the median values for the malignant cancer exams. Figures 8(a) and 8(b) provide a clear notion of the records for each kind of cancer. The former records compose narrow polygons over the scene while the later records compose wider polygons. Intuitively, this visual perception translates into a familiar notion of the patterns for records that indicate benign and malign cancer. Correspondingly, the cancer records are presented according to narrower or wider patters readily perceptible by the user.

Still in Figure 8, we can see the same dataset using the Parallel Coordinates technique. The visualization is added with a polyline indicating the average values of the benign cancer exams in Figure 8(c) and the average values of the malignant cancer exams in Figure 8(d). From these two last figures, it is possible to conclude, based on the respective average lines, that exams indicating benign cancer can be more easily categorized than the exams that indicate the contrary.

5 Experiments

This Section aims at experimentally demonstrating the ideas introduced in this work together with the main features of the VisTree framework. To do so, let us illustrate the system resources using the real world NFL dataset (http://stat.cmu.edu/datasets). The NFL data defines a 6-dimensional dataset with 6,048 attribute values spread into 672 records. The records carry data from three seasons (1989, 1990, 1991) of the NFL league including oddsmakers predictions. The dataset allows observing the scoring of the teams and the accuracy of the oddsmakers. The attributes are, respectively, home/away (1 for "favorite plays at home" or 0 for "favorite plays away"), favorite team score, underdog team score, winner (1 for favorite or 0 for underdog), points spread (tells how favorite the favorite team is considered), favorite name, underdog name, year and week of the season.

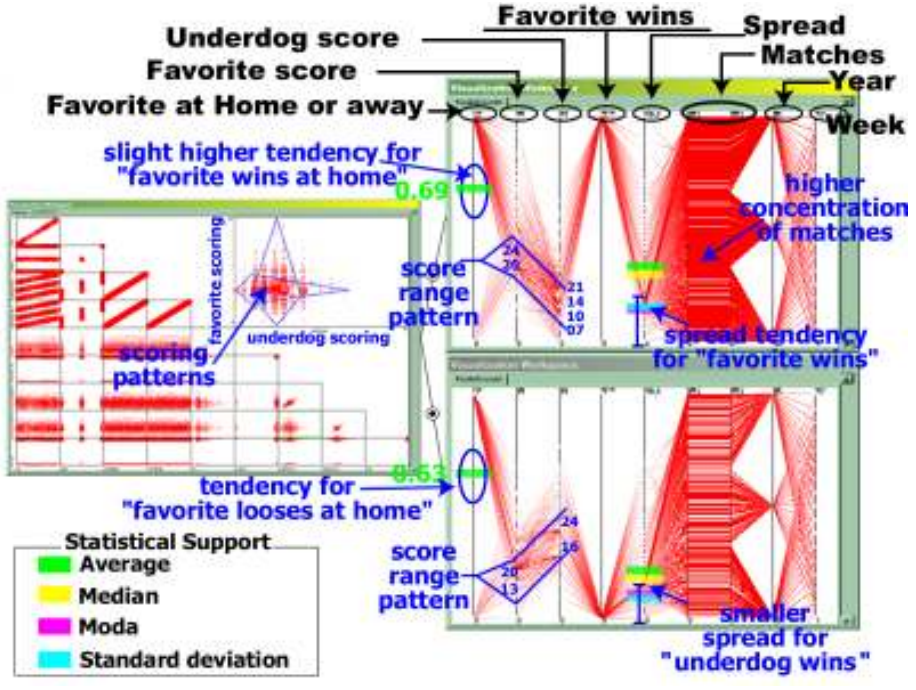

Fig. 9. Center-left shows the NFL dataset overview. Upper-right shows records satisfying "Favorite wins" and lower-right shows records satisfying "Underdog wins".

In Figure 9, at the left-hand side, we can see a Scatter Plot enhanced by the Frequency Plot exploration technique. In the scene, we emphasize with focus and zoom the crossing of dimensions "favorite team score" (vertical) and "underdog team score" (horizontal). From the visual entities, it is possible to visually interpolate geometrical forms that personify the scoring performance of what is set as favoritism and as underdog condition. At this initial analysis of the dataset, the overview of the patterns provides the notion of scoring efficiency, wider for favorites.

Still in Figure 9, the initial scene is refined into two other workspaces. These workspaces hold two semantical ideas, "Favorite wins", for records corresponding to football matches won by the favorite team, and "Underdog wins" for the matches won by the underdog team. There, one can see that the amount of matches is bigger for favorite winning than for underdog winning, stating that the oddsmakers have some credibility. The oddsmakers spread for matches satisfying "underdog wins" is notably smaller than that for "favorite wins", suggesting that, for such games, the oddsmakers had the perception that the favorites were not, actually, so favorite. Considering both cases, as observed from the first attribute, the factor "playing at home" is nearly uniformly distributed, diminishing the importance of such factor for the matches results. It is possible to perceive a slight concentration for "playing at home" in the case of "favorite wins". That is, playing at home makes a difference but this difference is not as

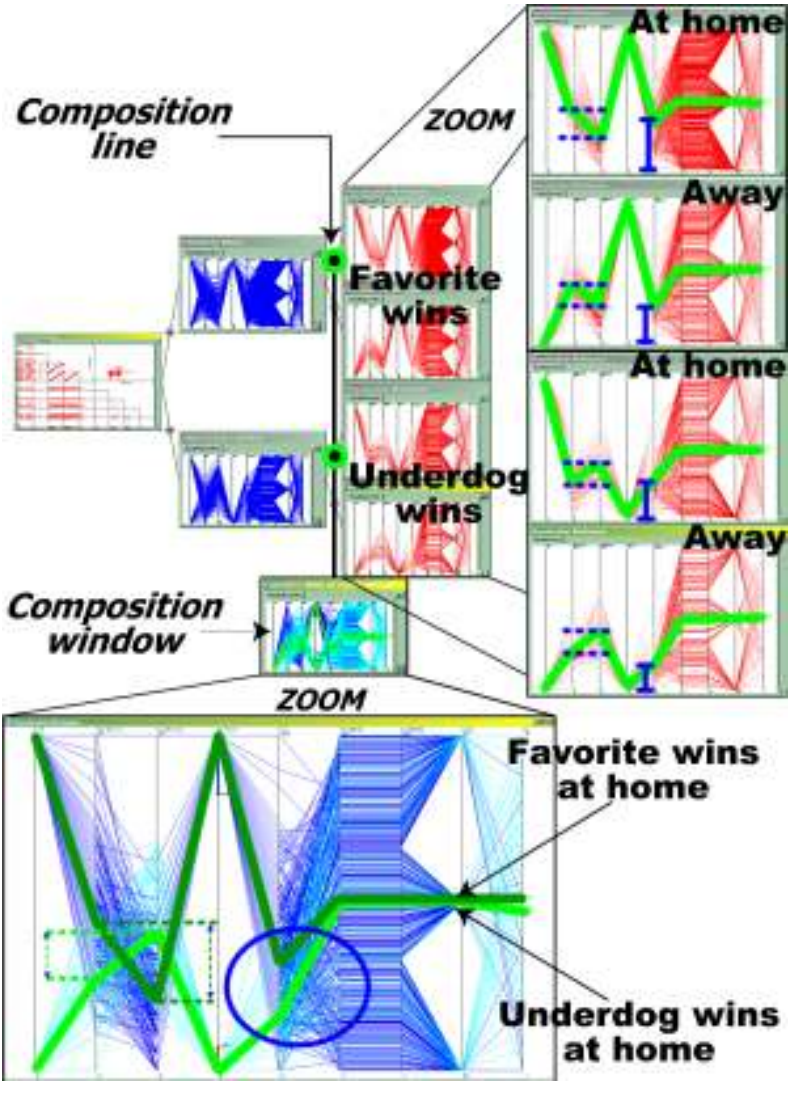

Fig. 10. Further refinement of the NFL dataset. Upper-right shows records satisfying "Favorite wins at home" and "Favorite wins away". Lower-right shows records satisfying "Underdog wins at home" and "Underdog wins away". Lower-left shows composition for records satisfying "Favorite wins at home" and "Underdog wins at home".

big as one would expect. The Frequency Plot over the two right-hand side scenes provides an idea for the scoring patterns. By interaction, we verified that for favorite wins, the most common scores are close to 20x14, 24x21, 20x10 and 20x07. Meanwhile, for underdog wins, the most common scores are close to 20x24 and 13x16. Situations when the underdog teams strive to diminish their non-favoritism.

In Figure 10, the workspaces are further refined into "Favorite wins at home" and "Favorite wins away", and into "Underdog wins at home" and "Underdog wins away". By checking the pairs of scenes, for "Favorite wins at home" and "Favorite wins away", the score difference (dashed line segments) is bigger in favor of the wining team when playing at home. However, it is not a pronounced difference, but rather a slight tendency in line with the conclusion that playing at home is not really such an advantage. The same scenes for "Underdog wins at home" and "Underdog wins away" reveal that when the underdog team wins, it does not matter where the team is playing, the score difference is very much the same. In contrast, by analyzing the four scenes, it seems that the oddsmakers care very much about where the team is playing, because the spread of points (highlighted with delimiters) differs significantly for the pairs of situations.

At the composition workspace in Figure 10, we contrast "Favorite wins at home" against "Underdog wins at home". In the resulting workspace (lower-left in the figure), we highlight the score differences for "favorite wins" and "underdog wins". We also highlight with a circle the average for the oddsmakers' spread of points. By analyzing these two patterns, it is possible to see that the oddsmakers statements are very accurate as, when they reduce the points spread, the favorite teams are beaten by the underdog teams. It is also possible to see that the score difference for the matches won by underdog teams are not so different from when the favorite teams win. This fact shows that the underdog teams carry a considerable potential, however, their irregular performance prevents them to triumph.

6 Conclusions

We have presented the Visualization Tree (VisTree) framework that, in a multiple-views environment, provides an innovative presentation methodology integrating several techniques of information visualization and statistical-based exploration. The VisTree framework is presented along with the introduction of standard data exploration techniques Frequency Plot, Relevance Plot and Representative Plot, all integrated to the VisTree multiple-views systematization. Such combination led us to the creation of a versatile visual data mining environment. In our system, the user can benefit from operations of pipeline and composition in order to keep track and manage his visual refinement tasks. Such tasks, distributed over multiple visualization workspaces, are then presented in a tree-like structure that reflects the sequence of analytical decisions taken by the user.

Besides presenting the VisTree, we have also worked on several datasets in order to demonstrate how the proposed exploration techniques, based on frequency, relevance and representativeness, contribute for the exploration of datasets. The idea of such methods is to dynamically convert datasets into simpler visual patterns. It is possible, then, to interpret a set of data items in a glimpse, guiding the analyst to further hypothesis-making. At the same time, the VisTree multiple-views integration enables presenting several workspaces at once, each of which with a different setting regarding the data visualized, the visualization technique used to render the scene and the exploration technique used to further refine and evolve the visual analytical process. In such environment, the user can observe many visualizations in parallel or in a single

scene. The effect is like the creation of macro visualizations over which one can browse different aspects of the analytical process.

Aknowledgements. This research has been supported by Brazilian research agencies Fapesp (Fundação de Amparo à Pesquisa do Estado de São Paulo), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

References

1. Ahlberg, C., Shneiderman, B.: Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In: Human Factors in Computing Systems. Conference Proceedings, pp. 313–317 (1994)
2. Ahlberg, C., Shneiderman, B.: Visual information seeking: tight coupling of dynamic query filters with starfield displays. In: CHI 1994: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 313–317. ACM Press, New York (1994)
3. Artero, A.O., de Oliveira, M.C.F., Levkowitz, H.: Uncovering clusters in crowded parallel coordinates visualizations. In: IEEE Symposium on Information Visualization, pp. 81–88 (2005)
4. Bennett, K.P., Mangasarian, O.L.: Robust linear programming discrimination of two linearly inseparable sets. In: OptimizationMethods and Software, pp. 23–34. Gordon & Breach Science Publishers (1994)
5. Bier, E.A., Stone, M.C., Pier, K., Buxton, W., DeRose, T.D.: Toolglass and magic lenses: The see-through interface. In: Computer Graphics. Annual Conference Series, vol. 27, pp. 73–80 (1993)
6. Boukhelifa, N., Rodgers, P.J.: A model and software system for coordinated and multiple views in exploratory visualization. Information Visualization 2(4), 258–269 (2003)
7. Faloutsos, C., Lin, K.: Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In: ACM Int'l Conference on Data Management (SIGMOD), Zurich, Switzerland, pp. 163–174. Morgan Kaufmann, San Francisco (1995)
8. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: Knowledge discovery and data mining: Towards a unifying framework. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, USA, pp. 82–88. AAAI Press, Menlo Park (1996)
9. Fekete, J.-D., Plaisant, C.: Interactive information visualization of a million items. In: INFOVIS, p. 117 (2002)
10. Fua, Y.-H., Ward, M.O., Rundensteiner, A.: Hierarchical parallel coordinates for visualizing large multivariate data sets. In: IEEE Symposium on Information Visualization, pp. 43–50 (1999)
11. Grinstein, G.G., Ward, M.O.: Introduction to data visualization. In: Fayyad, U., Grinstein, G.G., Wierse, A. (eds.) Information Visualization in Data Mining and Knowledge Discovery, pp. 21–45. Morgan Kaufmann Publishers, San Francisco (2002)
12. Inselberg, A., Dimsdale, B.: Parallel coordinates: A tool for visualizing multidimensional geometry. In: IEEE Visualization, vol. 1, pp. 361–370. IEEE Computer Press, Los Alamitos (1990)
13. Kandogan, E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: KDD 2001: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 107–116. ACM Press, New York (2001)

14. Keim, D.A., Kriegel, H.-P.: Visdb: Database exploration using multidimensional visualization. IEEE Computer Graphics and Applications 14(5), 40–49 (1994)
15. Keim, D.A., Kriegel, H.-P.: Visualization techniques for mining large databases: A comparison. IEEE Transactions in Knowledge and Data Engineering 8(6), 923–938 (1996)
16. Keim, D.A., Mansmann, F., Schneidewind, J., Ziegler, H.: Challenges in visual data analysis. In: 10th Intl Conference on Information Visualisation (IV 2006), London, England, pp. 9–16. IEEE Computer Society, Los Alamitos (2006)
17. Keim, D.A., Schneidewind, J.: Scalable visual data exploration of large data sets via multiresolution. Journal of Universal Computer Science 11(11), 1766–1779 (2005)
18. Rao, R., Card, S.K.: The table lens: Merging graphical and symbolic representation in an interactive focus+context visualization for tabular information. In: Proc. Human Factors in Computing Systems, pp. 318–322 (1994)
19. Rundensteiner, E.A., Ward, M.O., Yang, J., Doshi, P.R.: Xmdvtool: visual interactive data exploration and trend discovery of high-dimensional data sets. In: SIGMOD Conference, p. 631 (2002)
20. Seo, J., Shneiderman, B.: Knowledge discovery in high-dimensional data: Case studies and a user survey for the rank-by-feature framework. IEEE Transactions on Visualization and Computer Graphics 12(3), 311–322 (2006)
21. Ward, M.O.: Xmdvtool: integrating multiple methods for visualizing multivariate data. In: VIS 1994: Proceedings of the conference on Visualization 1994, pp. 326–333. IEEE Computer Society Press, Los Alamitos (1994)
22. Wong, P.C., Bergeron, R.D.: Multiresolution multidimensional wavelet brushing. In: VIS 1996: Proceedings of the 7th conference on Visualization 1996, p. 141. IEEE Computer Society Press, Los Alamitos (1996)