Chapter 7

Visualization Techniques: Which is the Most Appropriate in the Process of Knowledge Discovery in Data Base?

Maria Madalena Dias, Juliana Keiko Yamaguchi, Emerson Rabelo and Clélia Franco

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/50163

1. Introduction

Applying visual representation in the KDD process aims to facilitate the understanding over its results. Thus, visualization techniques can be integrated into the process of KDD in three different ways: to preview the data to be analyzed; to help in understanding the results of data mining, or to understand the partial results of the iterations inherent in the process of extracting knowledge [2].

However, the exploration and analysis of data using visualization techniques can bring new and enough knowledge exempting the use of other data mining techniques. Furthermore, the visualization is a powerful tool for conveying ideas, due to the vision plays an important role in human cognition [7].

In the visualization process, it is relevant to consider the choice of the best technique to be used in a certain application or situation. The inadequate use of visualization techniques can generate insufficient or even incorrect results, caused by graphic representation mistakes. In the attempt to solve that kind of problem, an evaluation of visualization techniques was carried out in the representation of data, which is shown in this study. Such evaluation provides subsidies for KDD users and system analysts when searching for the most appropriate visualization.

When visualization techniques are used, first off all, it should be observed the relevant characteristics of the data such as: data type, dimensionality (number of attributes) and scalability (number of records). The tasks that the user can perform during data exploration may also be another factor when deciding for a visualization technique. This paper aims to show how the characteristics of the data can influence the choice of visualization techniques, establishing guidelines to selecting them with the purpose of represent data in the best way.

In the next section, we present what mechanisms were used to find out the parameters which help to select more appropriated visualization techniques. The Keim's classification [26] was considered in associating these parameters with the visualization techniques.

In this paper two approaches about the usage of visualization techniques are presented: when the graphical representation is itself a tool for knowledge discovery [60, 61] and when they are applied on results of data mining, e.g., using K-means algorithm [45]. Here, both approaches are numerated as Approach 1 and Approach 2, respectively.

In the both approaches, the mainly idea is evaluating visualization techniques following some criteria intending to highlight its features according to the represented data. Approach 1 aims to defiine parameters to choose suitable visualization techniques according to the data characteristics. From these parameters, the Approach 2 comes to evaluate them in the data mining context for the clustering task results using the K-means algorithm. The next section presents how these criteria were defined and how the evaluation of visualization techniques were done.

2. Research methodology

The both early mentioned approaches are concerned in detecting the factors which guide the data analyst to choose the best visualization techniques to improve the understanding about data. The Keim's classification of visualization techniques was adopted in both approaches to establish a standard evaluation of visualization techniques. Keim [26] distinguishes five classes of visualization techniques: (1) standard 1D - 3D graphics, (2) iconographic techniques, (3) geometric techniques, (4) pixel-oriented techniques, (5) based on graphs or hierarchical techniques.

The Approach 1 differs from Approach 2 in the focus of application of visualization techniques. In the first one, they were used as a knowledge discovery tool; in the second one, they were applied in KDD process context at data mining stage.

In the Approach 1, the Grounded Theory (GT) methodology was used to identify parameters which are relevant in the choice of visualization techniques. When using GT as a research methodology, you have to do a systematic analysis of the data. In other words, it is necessary adopt procedures for codification of the data collected during the research. There are three stages of coding: open coding, axial and selective coding [1]. These steps can be performed cyclically and without a defined order until the primary collected data become organized in a classification well structured [19]. The organization of identified categories and the meaning of the association among them based an emergent theory that explain why this form of organization was reached and, in addition, brings a new hypothesis [34].

Based on this methodology, the literature related to visualization techniques was used as a data source, in which information about visualization were selected and analyzed. These gathered information were deeply compared until reach at the parameters to be considered to choose visualization techniques. This procedure was the GT's open coding step. These parameters are: data type, user-task type, scalability, dimensionality and position of the attributes in the graph.

Through analysis of relationship among the parameters and the visualization techniques, it was observed that each technique type have a certain configuration of parameters that reflect the characteristics of data and the objectives of the use of visualization. Based on this fact, the theory generation was described in form of guidelines. They serves to choose visualization techniques according to the parameters that influence this choice in function to the characteristics of the data. These guidelines were defined from the association between the identified parameters and the Keim's categories of visualization techniques, performing the GT's axial coding step. The GT's selective coding step is not considered in this approach, because there is not a core category that can be named as the main concept for the final formulation of the theory. Instead, there are a set of factors (parameters and types of techniques) that produce the indicative guidelines for choosing visualization techniques according to the characteristics presented by the dataset analyzed. The Figure 1 shows the association performed among the parameters and each categories of visualization techniques as a result of GT's axial coding.

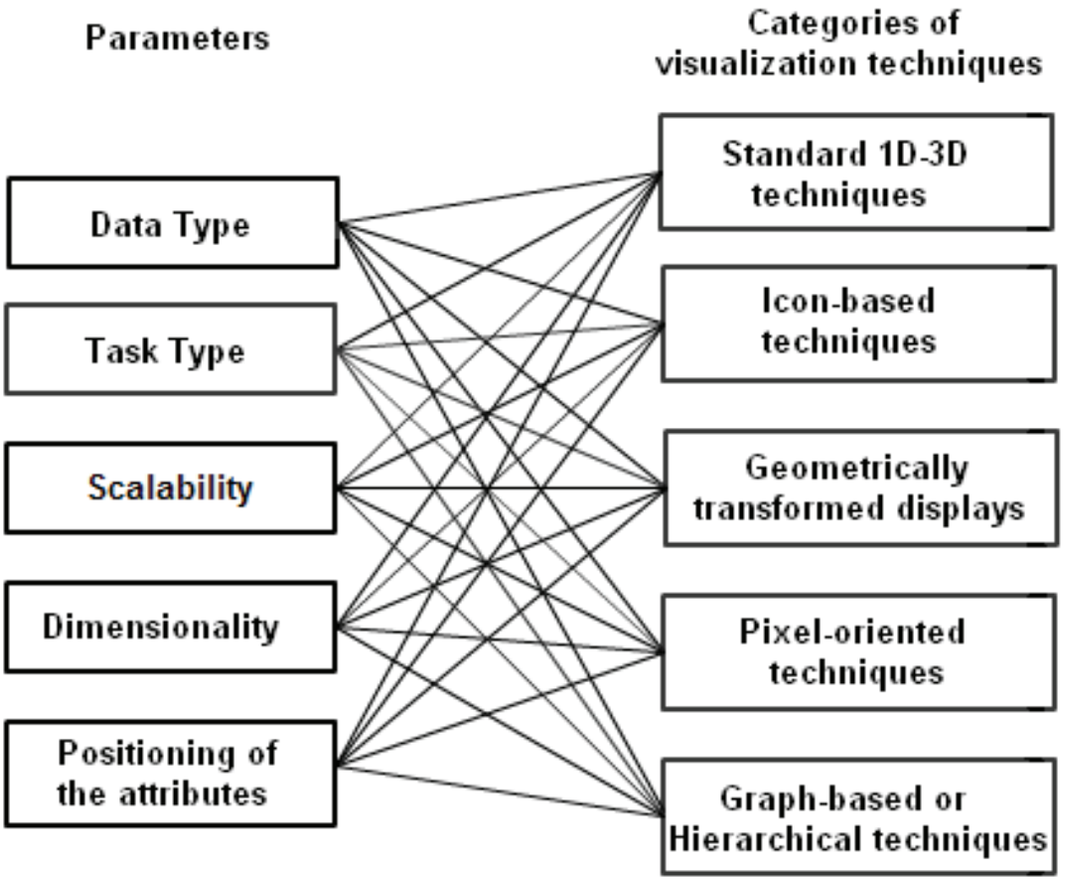

Figure 1. Association among the identified parameters and the categories of visualization techniques defined by [26].

The Approach 2 applies these guidelines to evaluate the best visualization techniques to improve the understanding about the results of clustering task in data mining. For this assay the evaluation technique known as data analysis techniques focusing on Analysis of Characteristics was chosen, as described in [44]. This technique consists of listing the

important characteristics of methods, processes or tools and then attributing scores to them - e.g. score 1 (it does not satisfy the need); up to 5 (it satisfies the need completely) - in this case using the visual representation of information visualization techniques. This evaluation is a way to reinforce the applicability of the guidelines obtained from the Approach 1.

In the Approach 2, geometric and iconographical techniques were analyzed because they were appropriate for the projection of results of the K-means algorithm. Two databases (USarrests and Mtcars) were analyzed with this techniques, both available in R language context [48]. The USarrests database contains statistical data on apprehensions carried out to each 100,000 residents in states of the United States of America in 1973, regarding assault, murder and violation. A Mtcars database contains data on fuel consumption and on ten aspects related to the project and performance of 32 vehicles produced from 1973 to 1974. Some matrixes were created in R language, in order to represent databases of high scalability and high dimensionality.

The process of getting the parameters done in the Approach 1 are described in the next section.

3. Getting the parameters

The parameters to consider when selecting visualization techniques emerged from GT's open coding process. The literature related to visualization techniques served as the data source in which we used the key points encoding method [1]. The result of this processes is illustrated in Table 1.

In the first column of this table, the expressions was taken from the main related works, whose references are in the next column, for each one was assigned the concepts, described in the third column. Thus it was possible to identify the parameters: data type, task type, scalability, dimensionality and position of the attributes in the graph, which compose the aspects to be considered in the decision to adopt visualization techniques to represent data.

Data type is a determining factor for choosing visualization techniques. This parameter is also used as a criterion for classification of visualization techniques. Shneiderman [53], for example, classifies data according to the number of attributes (dimensions) and its nature which can be quantitative (numerical data) or qualitative (categorical data). Keim [26] adds to this classification categories that refer to documents or hypertext, algorithms or software, or hierarchical data described by graphs. In this work we considered only the nature of data . Dimensionality is treated separately in this study as a factor for the classfication of visualization techniques, as defined in the encoding process presented in the Section 4.4.

Task type is another criterion to classify visualization techniques. It refers to activities that user or analyst can perform according to goals in the use of a graphical representation as noted in the literature [26, 42, 53, 57]. For practical purposes, the most common tasks were considered in this work, such as:

- Overview data: view the whole data collection;
- Correlation among attributes: the degree of relationship among variables can reveal patterns of behavior and trends;
- Identification of patterns, standards and important characteristics;

Key points	References	Code
Visualization techniques can be classified, among other criteria, by data type	[26, 53]	Data type
Task type is one of the aspects considered in classification of visualization techniques, which provides means of interation between the analyst and the display	[26, 42, 53]	Task type
Visualization techniques are subject to some limitations, such as the amount of data that a particular technique can exhibit	[23, 40, 45]	Scalability
Visualization techniques can also be classified according to the number of attributes	[15, 26, 40, 53]	Dimensionality
In some category of visualization techniques, distribution form of attributes on the chart can influence the interpretation about the representation, such as correlation analysis, in which the relative distance among the plotted attributes is relevant for observation	[2, 21, 28, 40]	Positioning of attributes

Table 1. Key point coding method applied on collected data

- Clusters identification: attributes with similar behavior;
- Outliers detection: data set with atypical behavior in comparison for the rest of data.

Scalability and dimensionality are characteristics to be observed in the data before applying a visualization technique. To facilitate the analysis of these parameters, a convention was established to classify the scalability and dimensionality of data [4, 5, 40].

Thus, as to scalability, data can be classified as small (10^1 to 10^2 = 10 to 100), medium (10^3 to 10^5 = 1000 to 10000) or large volume (10^6 to 10^7 = 1000000 to 10000000), according to the magnitude orders. As to dimensionality (number of attributes), data with up to four attributes are defined as low-dimensional, with five to nine attributes, medium dimensionality and with more than ten attributes, high dimensionality. In this context, the Approach 2 illustrates ahead the limitations of geometric and iconographic techniques in data with varied scalability and dimensionality.

For some types of techniques, such as Stick Figures, Parallel Coordinates, and Mosaics Plot, the position of the attributes or the order of elements representing attributes in a graphic is an important aspect for interpretation of visualizing data. Thus, this parameter is more related to specific techniques or computational tools in which the display is influenced by the order of attributes arranged in the chart.

The next section comprises the guidelines for the choice of the best visualization techniques according to the data characteristics and the technique's purpose. These guidelines were based on parameters established in the Approach 1, which were analyzed below. The

Approach 2 contributes with examples of applicability of these guidelines to represent graphically the results of K-means algorithm over examples of databases.

4. Analysis of the parameters

In Approach 1, the GT's axial coding is the next step after the identification of the parameters. It was done based on analyzes about the relationship between the parameters and categories of visualization techniques, according to the classification suggested by Keim keim02, who distinguishes five classes of techniques: (1) standard 1D-3D graphics; (2) iconographic techniques; (3) geometric techniques; (4) pixel-oriented techniques; and (5) based on graphs or hierarchical techniques.

Standard graphics are commonly used in statistic to view an estimate of certainty about a hypothesis or the frequency distribution of an attribute or to view a data model. For example, Histograms and Scatter Plot.

In pixel-oriented techniques, each value of attribute is mapped to a pixel color and it is placed on the display screen, divided into windows, each corresponding to an attribute. In the end, they are arranged according to different purposes keim02.

Data with a naturally structure of relationships among its elements, as hierarchical or as simple network, may be represented by hierarchical or graph-based techniques, such as: the own graph [38], Cone Trees [9], Treemaps [54], Mosaic Plot [18], Dimensional Stacking [30, 56].

In iconographic techniques, data attributes are mapped into properties of an icon or glyph, which vary depending on the values of attributes. For instance, the icons in format of faces of the Chernoff Faces [8, 13], the icons as stars of the Star Glyphs [39] and the icons in stick shape of the Stick Figures [43].

In geometric techniques, multidimensional data are mapped into a two-dimensional plane providing an overview of all attributes. As examples can be cited Matrix of Scatter Plot [6], 3D Scatter Plot [29] and Parallel Coordinates [20, 21].

In the next sections, each parameter are analyzed in association with categories of visualization techniques in the context of Approach 1, presenting the respective guidelines. Some examples of applicability of the guidelines establishment are presented by the Approach 2.

4.1. Analysis on the data type parameter

Techniques of standard 1D-3D category, generally represents from one to three attributes. In most cases, they are used for analysis of quantitative data. All graphs considered in this class are able to display quantitative data. To represent qualitative data, alternative techniques are more limited. In this case, the histogram is an example whereby is possible to represent these both data type [37].

In literature are found examples of usage of pixel-oriented techniques on quantitative data. Query-independent techniques, for example, were applied to represent temporal data, and query-dependent techniques are commonly used to represent continuous quantitative data

[24]. The author of this technique [25], states that they are not recommended for displaying qualitative data.

Hierarchical or graph-based techniques are ideal for displaying data when they have a structure of relationships among themselves or with a structure of hierarchy or simple network. Data type like as documents, texts (available on web or stored on disk), are likely to be viewed by specific tools found. For example, see [17, 27, 35, 55]. Algorithms and software are also data for which there are visualization tools developed specifically for this data type. For example see [51, 58, 62].

Iconographic techniques are more appropriate for quantitative data because icon features vary with the values of represented attributes. In Chernoff faces, the shapes of each facial properties are changed; in Star Glyph, the components of the star are modified; in Stick Figure, the format of segments are different according to the value of attributes. Qualitative data representation presents more technical difficulties, which can be circumvented by using the appearance properties such as icon color [45].

Geometric techniques are more flexible, being able to represent quantitative and qualitative data. This applies to the Parallel Coordinates technique, which can display attributes of these two data types. Due to the Scatter Plot Matrix is formed by a set of scatter plots, it is more suitable for continuous quantitative data.

In the Approach 2 two databases were used in a data mining process and the R language was applied to illustrate the results through geometric and iconographic visualization techniques. When evaluating the techniques regarding information visualization by using R language, a problem was found in relation to the entry parameter for the qualitative type of datum. R language does not allow that type of data entry in some information visualization. To solve this problem, the code operation, presented by Goldschmidt and Passos [16], in which the qualitative values were substituted by numeric values, was carried out.

All the techniques of information visualization evaluated allow the representation of quantitative data (continuous and discrete). However, for nominal qualitative data, the icon-based techniques evaluated do not enable a good representation.

The insertion of visualization properties may turn the technique of information visualization more effective, when appraised in relation to the characteristic types of qualitative data. In [33] is proposed an ordination of priorities when using the visualization properties, that is, considering the most perceptible and the least perceptible priorities in relation to the types of quantitative and qualitative data (ordinal and nominal). It is possible to add qualitative data by using colors in the visualization components of geometric projection techniques.

The Table 2 summarizes the discussion about the analysis on data type parameter for each class of technique, taking into consideration the nature of data domain.

4.2. Analysis on the task type parameter

Generally, some techniques are better for certain tasks than others. A task type execution depends on whether it is implemented by the tool in use according to the goals in improve the exploitation data activity.

Category	Technique	Quantitative data		Qualitative data	
		continuous	discrete	nominal	ordinal
1D to 3D	Histogram	X	X	X	X
	Box Plot	X			
	Scatter Plot	X			
	Contour Plot	X	X		
Icon-based	Chernoff Faces	X	X		
	Star Glyphs	X	X		
	Stick Figure	X	X		
Geometrically transformed displays	Scatter Plot Matrix	X			
	Parallel Coordinates	X	X	X	X
Pixel oriented	Query-independent techniques	X	X		
	Query-dependent techniques	X	X		
Graph-based orHierarchical	Graph	X	X	X	X
	Cone Tree	X	X	X	X
	Treemap	X	X	X	X
	Dimensional Stacking	X	X	X	X
	Mosaic Plot	X	X	X	X

Table 2. Visualization techniques and the respective data type that they can best represent

Standard 1D-3D techniques serve, in general, to view an estimate of certainty about a hypothesis or the frequency distribution about an attribute, such as the usage of histogram. This class also provides graphs to make comparisons and data classifications (to this case, for example, can be used box plot), and also to determine the correlation between attributes. Different statistical graphs can be used in data analysis, in order to discover patterns and structures in data and identify outliers that can be observed, for example, through using box plot or scatter plot.

Iconographic techniques represent each data entry individually, allowing verification of rules and behavior patterns of the data. Icons with similar properties can be recognized and thus form groups and it be analyzed in particular. A representation with a discrepant format if compared to the other may characterize an outlier. For example, technique of icon-based visualization that works with multidimensional data is the Star Glyphs visualization. Johnson and Wichner [22] stated that this sort of visualization is useful to standardize certain information and they use Star Glyphs to determine the similarity within clusters. Lee et al. [31] also stated that icon-based visualizations are multidimensional points that make use of useful dimensional space to detect clusters and outliers.

Geometric techniques provide a good overview of the data, assigning no priorities to represent its attributes. Furthermore, verification of correlation among them may be more discerning

when using techniques of this class, such as the scatter plot matrix. This category of techniques also allows the identification of patterns, rules and behaviors. Therefore, outliers may also be detected, characterized by behaviors outside the common standard. The analyst may choose to analyze a group of data that can be detached using the tool. However, groups are not usually immediately identified by techniques of this class.

Pixel-oriented techniques can be used in the analysis of relationships among data attributes, so rules and patterns may be identified through observing the correlations among them. Furthermore, the pixels can be arranged to finding clusters.

Hierarchical techniques are useful for exploitation of data arranged in a hierarchical or simple relationship. Through techniques of this class is possible to obtain an overview of the data structure and analyze the relationship among the elements. Techniques of this category also allow grouping data, such as Treemaps [54]. Table 3 summarizes the most representative tasks for each class of visualization technique, as previously discussed.

Category	Technique	Tasks				
		overview	correlation	patterns	clustering	outlier
1D to 3D	Histogram			X		X
	Box Plot			X		X
	Scatter Plot		X	X		X
	Contour Plot		X			
Icon-based	Chernoff Faces			X	X	X
	Star Glyphs			X	X	X
	Stick Figure				X	
Geometrically transformed displays	Scatter Plot Matrix		X	X		X
	Parallel Coordinates	X	X	X		X
Pixel oriented	Query-independent techniques		X	X	X	
	Query-dependent techniques		X	X	X	
Graph-based orHierarchical	Graph	X	X			
	Cone Tree	X				
	Treemap	X	X	X	X	
	Dimensional Stacking			X	X	X
	Mosaic Plot	X		X		

Table 3. Most representative tasks for each category of visualization technique

The next section brings an example of analysis on the correlation task using geometrical visualization techniques (matrix of scatter plot and parallel coordinates) done in the Approach 2. It uses the R language to visually represent the results of clustering task of data mining.

4.2.1. Analysis of correlation task into a case study

Correlation is the association or interdependence among the database attributes, used to show if there is or not a relationship among attributes of interest. It is inherent to the association and clustering tasks within the data mining. When referring to correlation, the most emphasized

by literature is the visualization of Scatter Plot [10, 11], which supplies a positive or negative correlation measure according to the position of scattering.

The visualization of Scatter Plot supplies a 'cloud' of points in a Cartesian plane, by using axes 'x,y', being very useful to identify the linear correlation [10]. Correlation is identified in visualization according to the position of points. If the points, in the diagram, have a straight increasing line as 'image', it is linear positive, but, on the other hand, if the points form an 'image' as a descending straight line, it is considered linear negative. However, if the points are dispersed, not showing any clear defined 'image', that leads to the conclusion that there is no relationship among the attributes being studied [10, 11].

To show such characteristic, USarrests test database was used in the visualization of the Matrix of Scatter plot, with a function created in the R language, which calculates the correlation, shows the values calculated and builds lines that follow the scattering (the red lines in Figure 2). As it can be observed in that visualization, the correlation is interpreted as a positive correlation when it is found a large correlation coefficient between the attributes 'murder' and 'assault' - it, in other words, the number of deaths increases as the number of assaults increases. A negative correlation is found between the attributes 'assault' and 'urbanpop'.

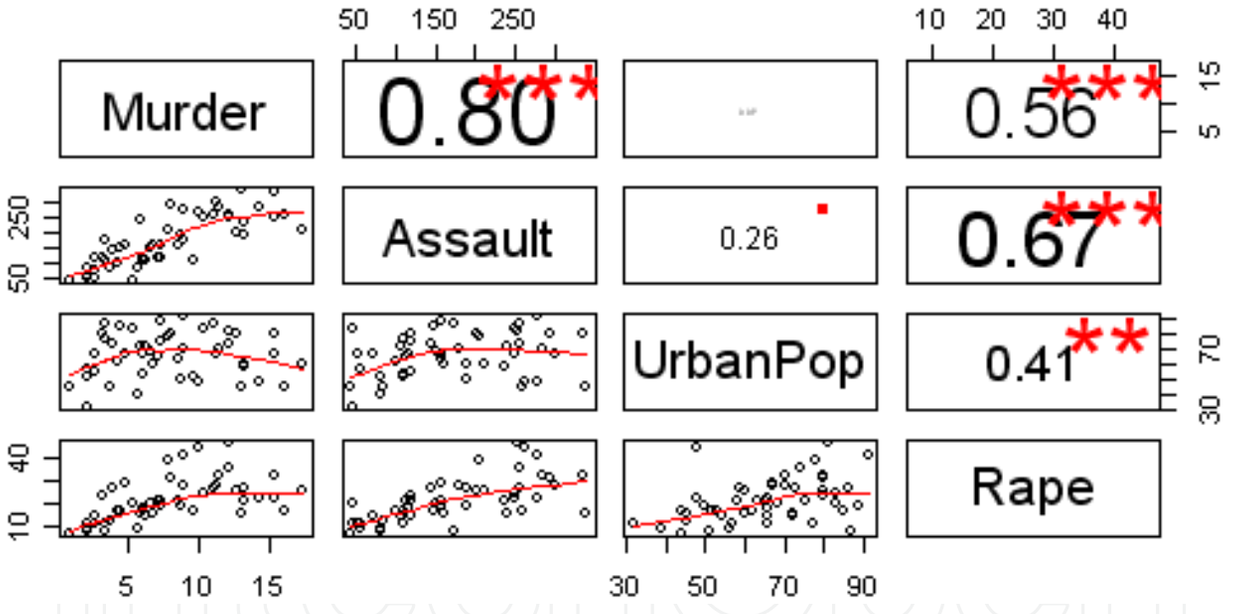

Figure 2. Correlation in the visualization Matrix of Scatter Plot - 'Usarrests' database (R language).

Figure 3 represents the Matrix of Scatter Plot with seven attributes of 'mtcars' database and with the three clusters generated by the K-means algorithm applied to that base. The colors (red, black and green) represent the clusters formed when executing the K-means algorithm. Figure 3 enables to interpret that there is a clear division within the clusters, which is determined by the attribute value named number of cylinders 'cyl'.

Another technique of geometric projection here evaluated was the visualization of Parallel Coordinates that project the relationship among the attributes of the database in bidimensional space. It allows to interpret characteristics as the difference in distribution and the correlation among attributes [20, 59]. Figure 4 represents the visualization of the parallel

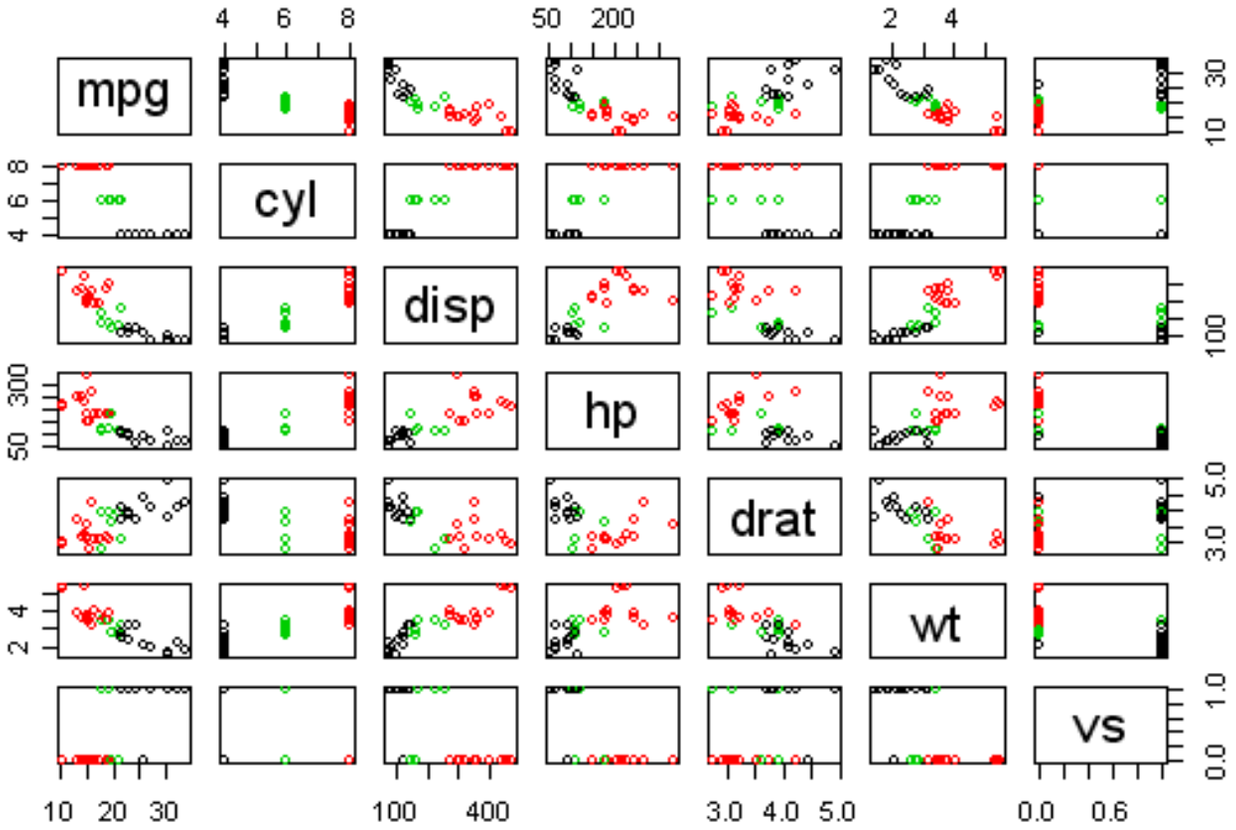

Figure 3. Matrix of Scatter Plot representing the 'Mtcars' database (R language).

coordinates, showing the distribution of records with the attributes of 'Mtcars' test database and the clusters formed by the K-means algorithm through colors (red, green and black).

In Figure 4 it is possible to observe a concentration of colors in the horizontal axes that crosses with the vertical axis of attribute 'cyl' and, which irradiates towards 'disp' and 'hp' vertical axes. It can be concluded that the three clusters generated by the algorithm do not contain the same values for 'cyl' and 'disp' attributes and, the number of cylinders (cyl) is proportional to the values of 'disp' and 'hp' attributes.

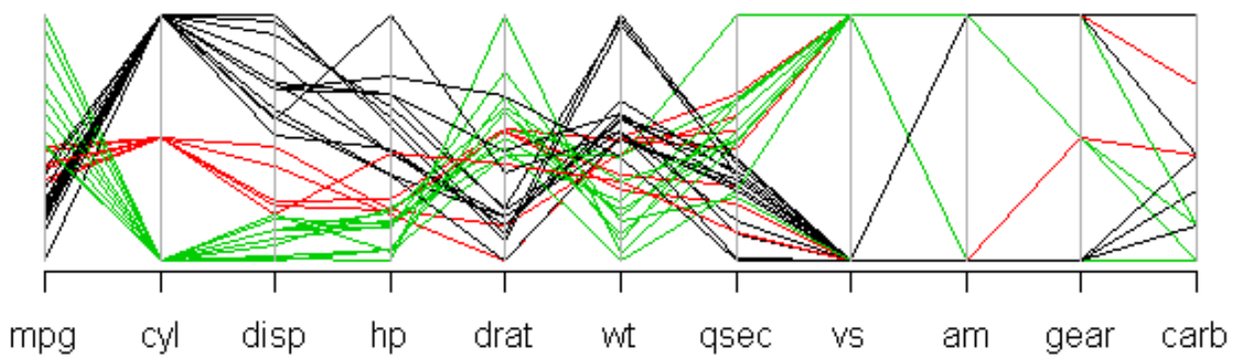

Figure 4. Parallel Coordinates representing the 'Mtcars' database - (R language).

Regarding the icon-based techniques, no type of evaluation was carried out, due to the difficulty to interpret the visualization of correlation characteristic.

4.3. Analysis on the scalability parameter

Implementations of visualization techniques must take in consideration the limits of dimensionality and scalability of data to hold in way that the tool be capable of providing a clear overview of data to the analyst. As described in [3], scalability refers to the computational complexity on the number of records in an array, as well as to the number of attributes. The amount of records that can be simultaneously presented is one of the limitations of the visualization techniques. With high number of records, the result shows a considerable degree of disorder [49].

Standard graphics has low dimensional, because they are intended to represent data with one to three attributes. In addition, they support the view of a small volume of data because, in general, they come from statistical studies, resulting of a sample or of percentages.

Iconographic techniques are able to handle a larger number of attributes in comparison to the standard graphics; however, the visualization generated is best for a small amount of data due to the space occupied by the icons in the screen. This is the same statement found in Approach 2, in which the iconographic techniques evaluated (Star glyphs and Chernoff Faces) were classified as low scalability (support to display an amount of data).

Geometric techniques, in turn, may work with an increased number of dimensions and volume when compared to standard 1D-3D graphics and iconographic techniques. But they are outweighed by the pixel-oriented techniques for their capability to represent the largest volume.

Hierarchical techniques or graph-based techniques are usually used to represent the relationship among data, regardless of dimensionality, which can be high or low, but have the same space constraints like that presented by iconographic techniques, being the visualization clearer if the amount data is not bulky.

However, visualization tools can offer features like zoom, select, filter, among others, to improve the interactivity with the visualization, mitigating the limitations of each technique. Table 4 summarizes, in general, these two parameters, scalability and dimensionality, for each class of visualization technique.

In [23], are presented the limitations of some visualization techniques in relation to the number of records in dataset and these authors state that the visualization of parallel coordinates can represent approximately 1000 records. They also affirm that the geometric techniques quickly reach the limits of what can be considered comprehensible. This happens because there are overlapping records mapped in the same position or close to each other, thus presenting 'blurs', or areas totally filled out. Shimabukuru [52] defends that the visualization of great volumes of data need the integration of the technique with appropriate interaction operations, which can enable the selection and filtering of items considered of interest.

The areas totally filled out, 'blurs' of parallel coordinates, generate incomprehensible visualizations. However, it is possible to notice that the use of colors can help the visualization

Category	Technique	Scalability	Dimensionality
1D to 3D	Histogram	Small	Low
	Box Plot		
	Scatter Plot		
	Contour Plot		
Icon-based	Chernoff Faces	Small	Low to Medium
	Star Glyphs	Small	
	Stick Figure	Medium	
Geometrically transformed displays	Scatter Plot Matrix	Medium	Medium to High
	Parallel Coordinates		
Pixel oriented	Query-independent techniques	Large	Medium to High
	Query-dependent techniques		
Graph-based or Hierarchical	Graph	Small to Medium	High
	Cone Tree	Small to Medium	High
	Treemap	Medium	High
	Dimensional Stacking	Medium	Medium
	Mosaic Plot	Medium	Medium

Table 4. Visualization techniques and their representations of scalability and dimensionality of data

of patterns. The Approach 2 brings an example to show this fact: matrixes with different amounts of records were created and used as entrance parameters in the execution of K-means algorithm. Then, the results obtained were plotted in the technique of parallel coordinates, where the lines represent the matrix attributes, and the colors represent the clusters, as shown in Figure 5a and Figure 5b. The colors appear as blurs, making possible the visualization of the patterns of each cluster.

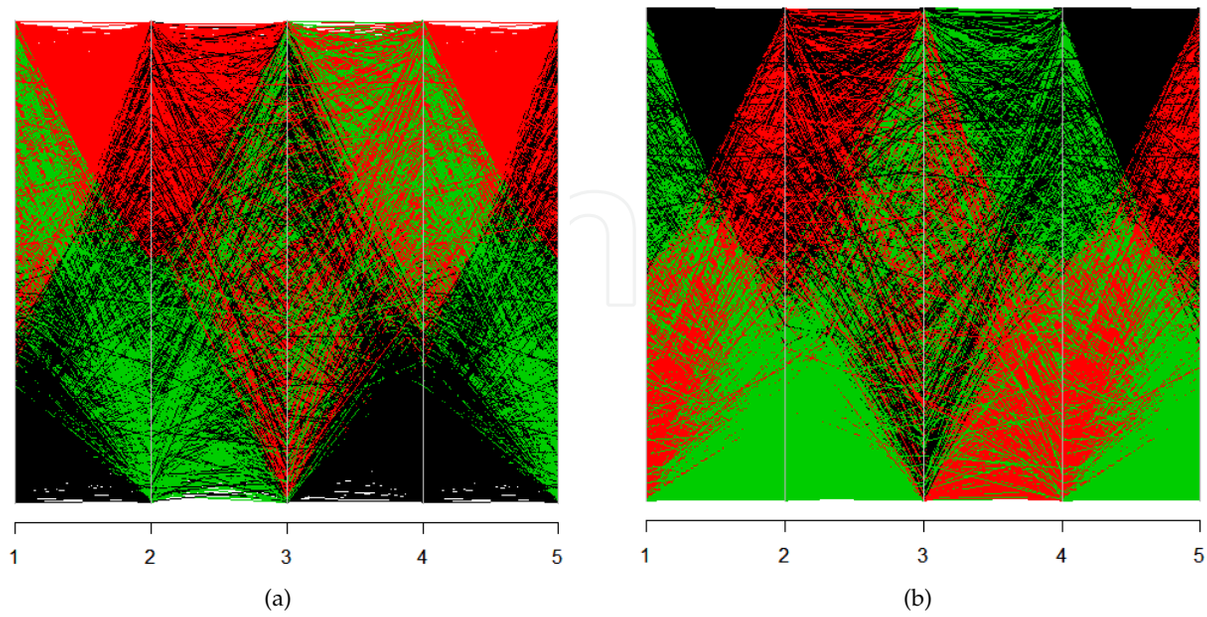

Figure 5. Parallel coordinates: a) 10,000 records, b) 100,000 records

As shown in Figure 5a and Figure 5b, the increase in the number of records from 10,000 to 100,000 generates blurs that show the patterns. In this example, the algorithm has generated three clusters, but, depending on the application domain and on the amount of records, it may be necessary to generate more clusters and, consequently, use a higher number of colors.

To reinforce the difficulty of visualizing a great amount of records by using geometric techniques, Figure 6a, Figure 6b and Figure 6c show the visualization of Scatter Plot in three-dimensional projection (3D Scatter Plot) of matrices with lines size: 100, 1,000 and 10,000, respectively, and with five columns.

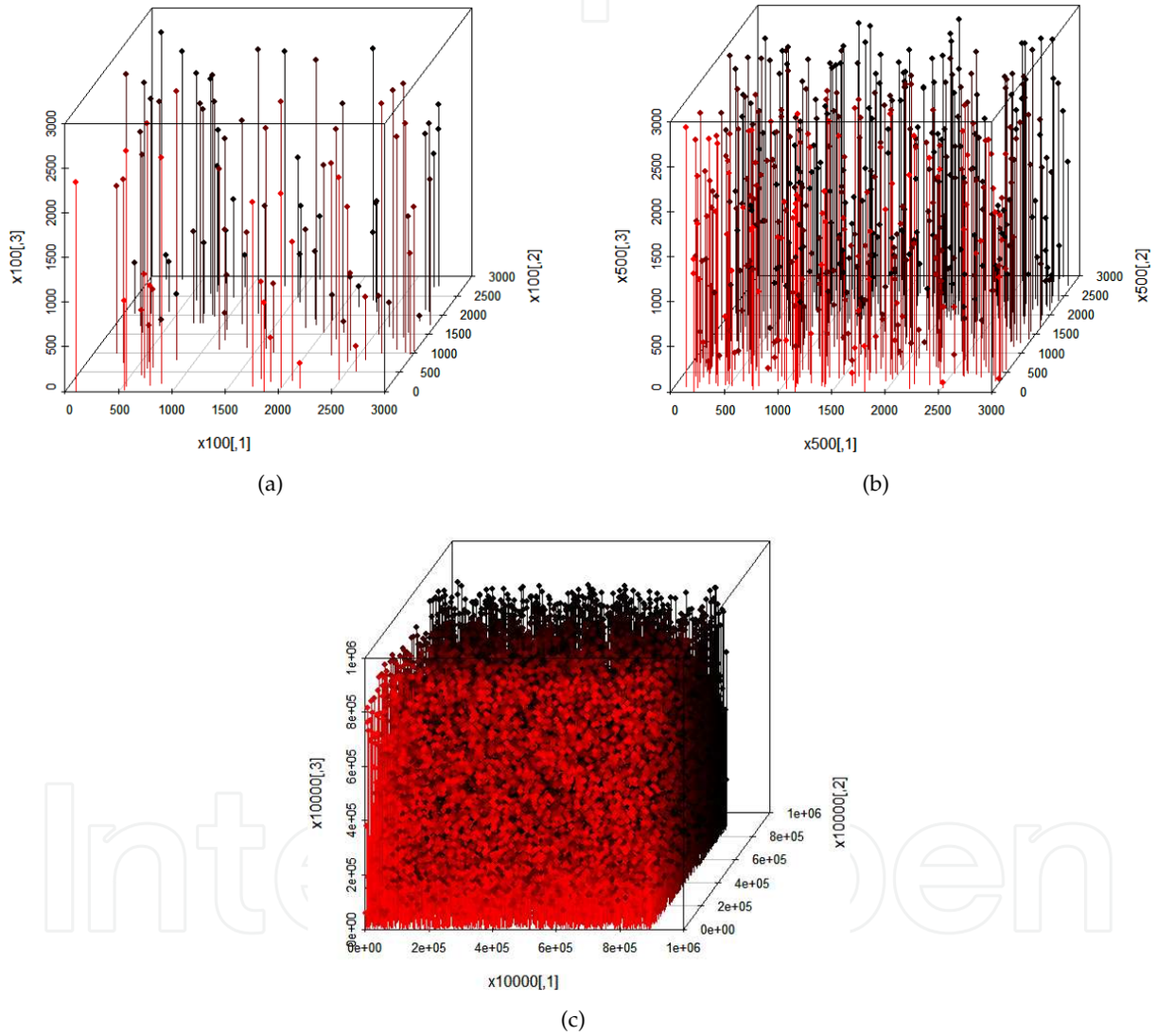

Figure 6. 3D Scatter Plot: a) 100 records, b) 1,000 records, c) 10,000 records

In these figures it is possible to observe that, as the number of records increases, the visualization becomes incomprehensible or difficult to be visualized. The techniques based on icons enable the representation of a small number of records due to the size of the graphical elements [47]. Chernoff Faces is the visualization that has the greatest limitation in relation to scalability, because it just allows the representation of a small amount of records.

According to Shimabukuru [52], the visualization of Stick Figure makes possible to represent great volumes of data. Such visualization technique uses the two dimensions of the screen to map two attributes of data, whereas the other attributes are mapped according to angles and/or segment lengths.

4.4. Analysis on the dimensionality parameter

Dimensionality characteristic is related to the visualization technique capacity of representing attributes. [Keim 02] mentions that, usually, in visualizing information, a great amount of records is used. And each one has many attributes, for instance: a physical experience can be described with five attributes or hundreds of attributes.

In that interpretation, it should be taken into account the capacity of human perception, or the conceptual limit of the dimensionality that, according to Rodrigues [47], may be between low and high dimensionality. However, there is not a consensus on what may be considered low and high dimensionality which standards limits were defined as discussed in the Section 3. Using these conceptual limits as base, several matrixes were created in the Approach 2 with different columns (representing attributes). After the creation, each matrix was plotted in the techniques of information visualization used in this investigation.

The literature revised is unanimous regarding parallel coordinates for representation of multidimensional data [14, 20, 23, 38, 47, 52, 59]. This technique maps each attribute to a line by connecting points in the axes. Figure 7 shows three visualizations using the technique of parallel coordinates, with different amounts of attributes (10, 34 and 100, respectively) but with the same amount (100) of records. The limit of attributes that the parallel coordinates can support is restricted to the resolution of the computer screen. As it can be observed, the increase of attributes can cause blurs, which hinder the visualization, even hindering the recognition of patterns.

The Matrix of Scatter Plot is another technique for visualizing geometric projection, which is able to represent high dimensionality. For the visualization of 3D Scatter Plot, [12] suggest the possibility of using icons for representing data attributes, thus allowing an increase in the number of dimensions to be explored in the visualization. It can be considered that such visualization has good representation in the dimensional characteristic.

The technique of icon-based visualization is one of the most used. In such technique, figures are used as geometric encoders, by taking advantage of attributes, which are perceptible visually, such as color, forms and texture [32].

Chernoff Faces, classified as an icon-based visualization technique, can also be used to visualize multidimensional data. Although such technique is highly useful to exhibit multidimensional data, the records are presented separately, because they do not transmit any information on the real value arrays. However, Chernoff Faces enables to illustrate tendencies and focus on data to be highlighted [50]. Specific literature on the subject does not limit the amount of characteristics that may be used in Chernoff Faces' visualization, but in [22] is suggested using up to 18 attributes. In R language, the function 'faces' enables a maximum representation of 15 attributes.

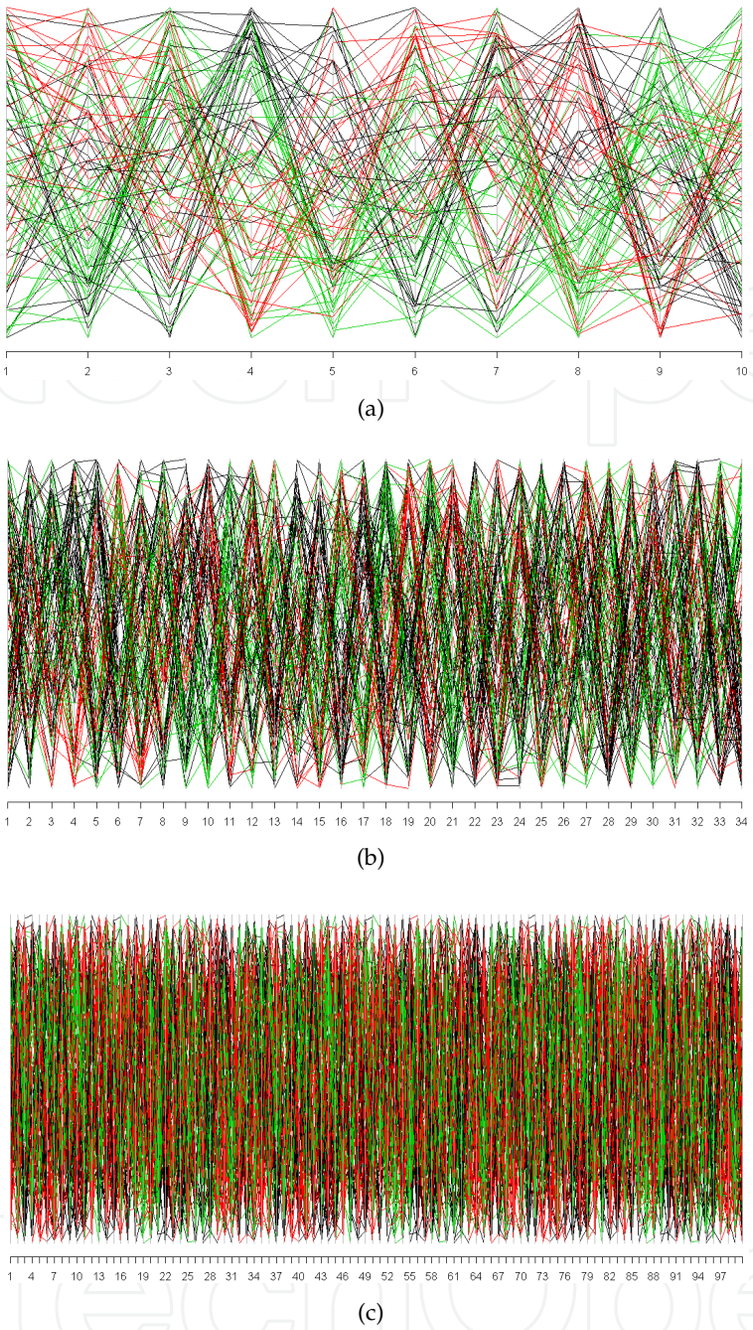

Figure 7. Parallel Coordinates a) 10 attributes, b) 34 attributes, c) 100 attributes (R language)

Star Glyphs visualization allows a larger number of attributes if compared to Chernoff Faces. That can be seen in Figure 8, which shows that the possibility of representation is around 80 attributes (Figure 8c). However, as shown in Figure 8d, only blurs are visualized when there is a great amount of attributes.

Besides the icon-based visualization techniques described, there is also the Stick Figure visualization that, in spite of representing high scalability, possesses certain limitation regarding dimensionality, which is in the order of approximately a dozen attributes [26].

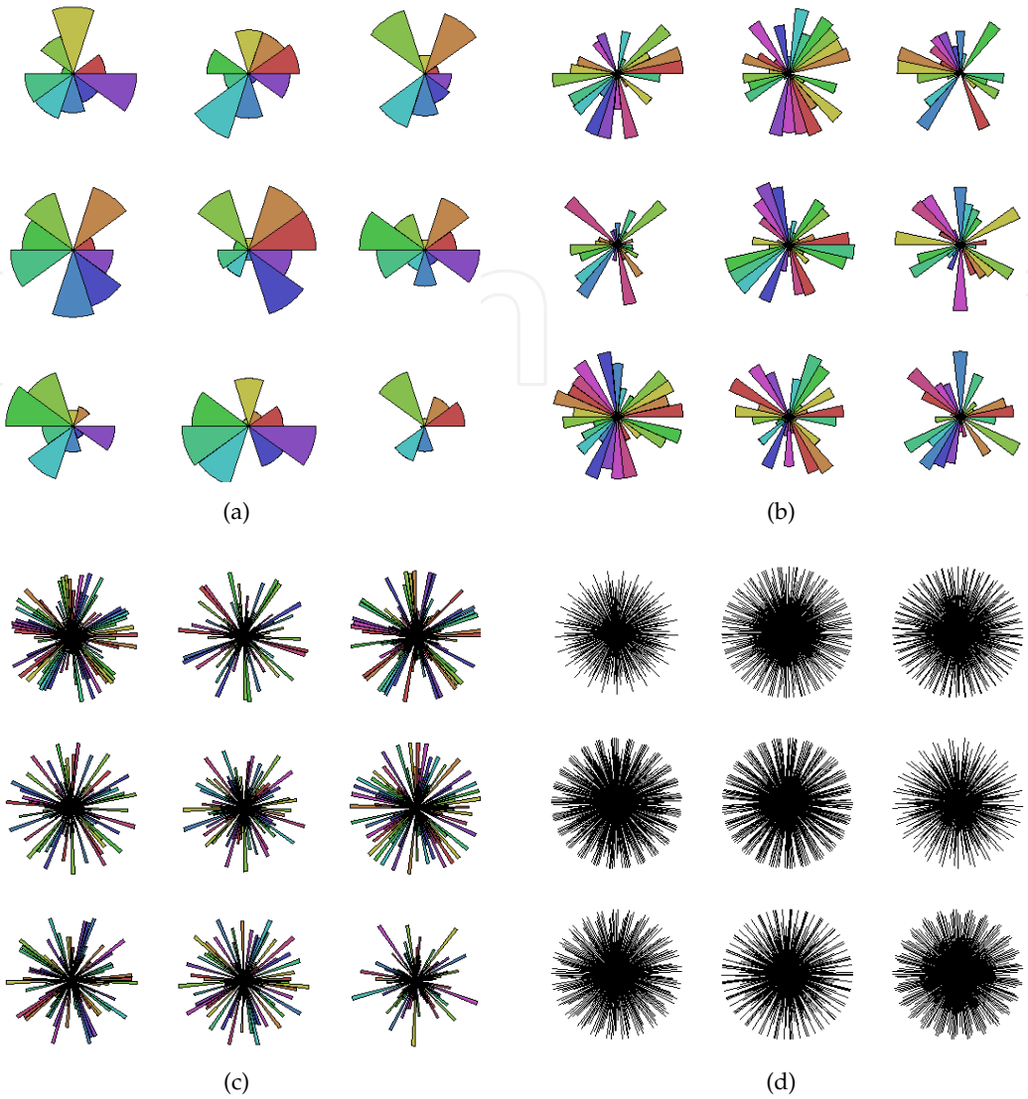

Figure 8. Star Glyphs a) 10 attributes, b) 30 attributes, c) 80 attributes, d) 500 attributes (R language)

4.5. Analysis on the positioning of attributes parameter

Although it is not a parameter directly linked to the characteristics of data, it is an important factor in visual data exploration for some techniques as, among others, Treemaps [54], Mosaic Plots [18], Dimensional Stacking [30]. This parameter depends on the technique or tool used to generate the visualization, which should allow the change of the positions of the attributes in the graph, producing different views that can reveal new patterns.

Thus, the categories of visualization techniques followed by its respective analyzed techniques are described in Table 5. In the third column, it is indicated if the positioning of attribute can influence or not in the interpretation of graphical data representation for this technique.

In general, for 1D-3D standard graphics, positioning of attribute does not change the interpretation of results due to the low dimensionality of the data that might be represented. Moreover, the goal of using techniques of this class is to analyze the behavior of a given attribute, or the correlation among two or three attributes.

Category	Technique	Positioning
1D to 3D	Histogram	It does not influence
	Box Plot	
	Scatter Plot	
	Contour Plot	
Icon-based	Chernoff Faces	It does not influence
	Star Glyphs	It does not influence
	Stick Figure	It influences
Geometrically transformed displays	Scatter Plot Matrix	It does not influence
	Parallel Coordinates	It influences
Pixel oriented	Query-independent techniques	It influences
	Query-dependent techniques	
Graph-based or Hierarchical	Graph	It influences
	Cone Tree	It does not influence
	Treemap	It influences
	Dimensional Stacking	It influences
	Mosaic Plot	It influences

Table 5. Interference of the positioning of attributes and over each visualization techniques

Among the iconographic techniques presented in Table 5, Stick Figures is an example in which the position of the attributes can influence the visual data exploration according to the icon type used, derived from the variation of the mapping of data attributes into icon properties [43, p. 516].

Chernoff faces, in turn, have a fixed structure for its icon, since it corresponds to the human face characteristics and thus, the change of the positioning of attributes is not a relevant aspect for this technique.

But there are studies about which icon properties may be more representative for the interpretation of results, such as the eyes size and the shape of the face are aspects that draw attention [31, 36]. Likewise it is for Star Glyph technique, for which once established the order of the best mapping of attributes [28, 41], it remains the same for all the icons representing a record data per star.

In the works of Inselberg [21] and Wegman [59], it is explained how the position of the attributes in the graph may influences the correlation detection in Parallel Coordinates. Scatteplot Matrix is, on the other hand, composed of a set of Scatter Plots, for this reason nor is influenced by the change of attributes positioning, since their main objective is to evaluate the correlation between attributes.

Keim [24] presents techniques for the placement of pixels on the display, which can influence the interpretation of the visualization to identify patterns and relationships among the represented attributes.

The query-independent technique, for example, may have the pixels arranged by recursive pattern technique. When using the query-dependent technique, the pixels can be arranged in the window using spiral technique [24].

Hierarchical techniques or graph-based techniques are in general influenced by the attributes positioning, due to its elements naturally hold a relationship structure, therefore, the assignment of variables in the graph should be made carefully, especially when there is a hierarchy between the elements. The exception is for the Cone Trees technique, which represents a defined tree structure (as files and directories structures in a hard disk), providing only interactive features such as animation to navigate among the tree nodes [9, 46].

The Approach 2 brings a deeper analysis on the positioning of attributes focusing on the relationship among attributes in geometrical visualization techniques, which is presented below.

4.5.1. Relationship among attributes

Likewise the Matrix of Scatter Plot, another technique of geometric projection that shows the relationship among attributes is the visualization of Parallel Coordinates, which is shown in Figure 9. When generating a plane representation, it transforms multi-varied relationships into bidimensional patterns, being possible to visualize many attributes [59].

The relationship among the attributes it is found on the vertical axes, that means that, closer the axes, better the visualization of the relationship. For example, the attribute named 'assault' related with the attribute 'urban pop' is shown through the position of the horizontal lines that exhibit the meaning in the relationship, as shown in Figure 9a. In the relationship among the attributes 'assault' and 'rape', which are separated by the attribute 'urban pop', it is necessary to create the relationship mentally, or to remove the attribute, according to Figure 9b.

Regarding the icon-based techniques, it was not possible to determine the existence of relationships among attributes, thus turning the evaluation for such a characteristic impossible.

5. Discussion about the results

In the Approach 1, the five parameters (data type, task type, scalability and dimensionality of the data, and position of the attributes in the graph) were identified by means of GT and subsequently they were analyzed in relation to the categories of visualization techniques classified by Keim [26]. Through analysis of relationship among the parameters and the visualization techniques, it was observed that each technique type have a certain configuration of parameters that reflect the characteristics of data and the objectives of the use of visualization as seen below:

- Data type must be the first parameter to be considered. It is the type of data that determines what kind of visualization technique can be prior used. Qualitative data, for example, will be hardly understood if they were represented by a technique developed to represent quantitative data and vice versa.
- The task type to be performed corresponds to the goals of the analyst during the data exploration. For tasks related to statistical analysis, for example, the graphics 1D-3D may be sufficient; for tasks of correlation verification may be used visualization techniques of geometric category, and so on.

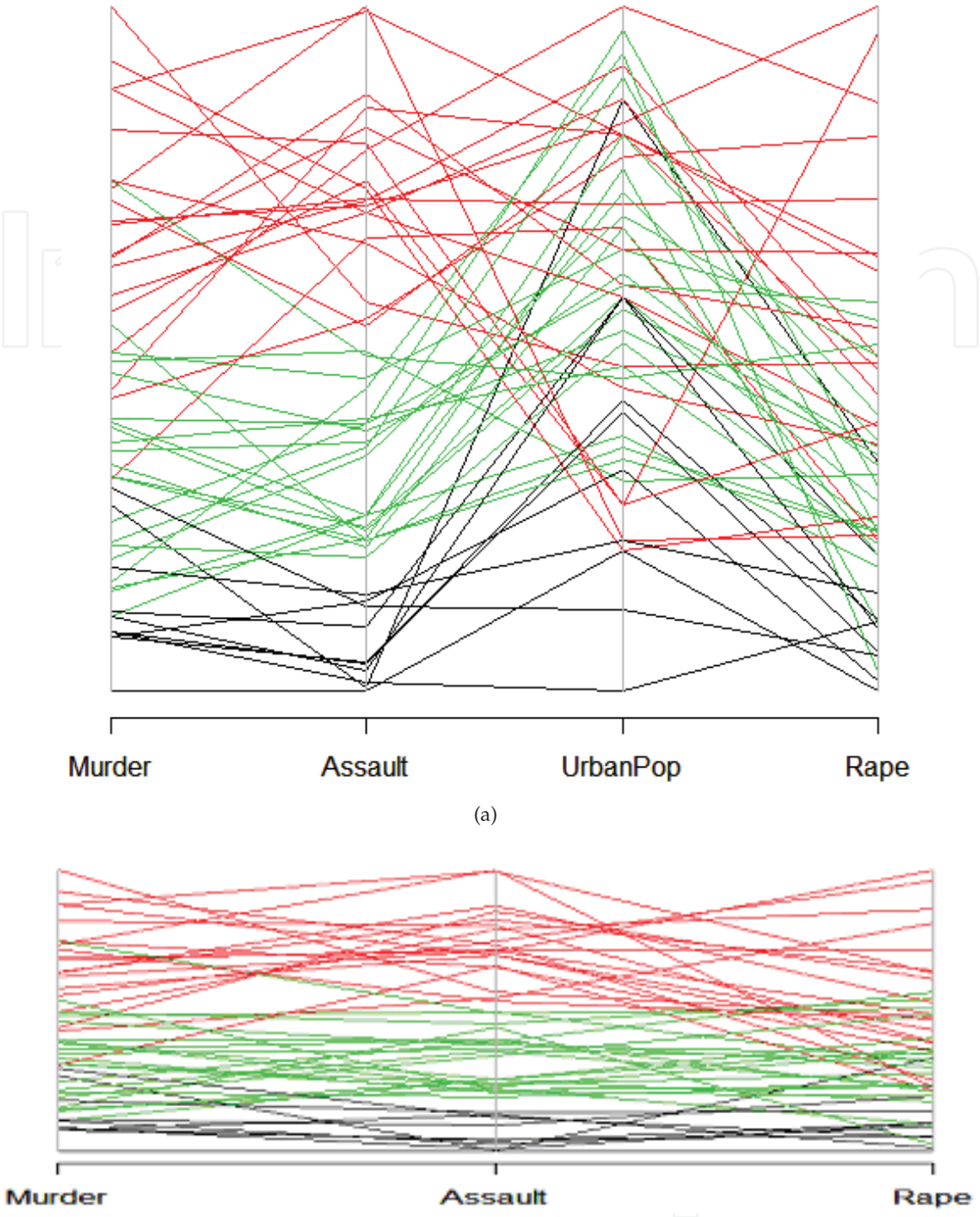

Figure 9. Parallel Coordinates - 'USarrests' database (R language)

- Both scalability and dimensionality of data are limiting factors for visualization techniques. Although most of them supports multidimensional data, usually these techniques differ in the ability to display a certain amount of dimensionality and volume of data. This is the case of categories of techniques iconographic, geometric and pixel-oriented. However,

other ways of interaction can be used during the visual exploration to minimize these limitations, for example, the functions of zooming, selection and filter.

- The positioning of the attributes is a factor more dependent on visualization technique to be used and, hence, on the tool that implements it. For some techniques (such as parallel coordinates and star glyphs), positioning of attributes is important for discovery new patterns or behaviors.

About the Approach 2, the Table 6 shows the results of the Analysis of Characteristics technique [44] by which iconographic and geometric techniques were appraised. The punctuations attributed are subjective, varying of score 1 (it does not satisfy the need) up to 5 (it satisfies the need completely), and it was based on practical experiments carried out in the language R and on the literature studied. In the Table 6 can be seen that the geometric projection techniques were valued by higher punctuations regarding the iconographical techniques for almost all the analyzed characteristics.

Characteristics	Geometric Projection			Iconographical Techniques		
	Matrix of Scatterplot	3D Scatter Plot	Parallel Coordinates	Star Glyphs	Stick Figure	Chernoff Faces
Scalability	5	2	5	1	5	1
Dimensionality	4	3	4	5	3	3
Nominal qualitative data	3	3	3	1	1	1
Ordinal qualitative data	4	3	4	5	5	5
Discreet quantitative data	5	5	5	5	5	5
Continuous quantitative data	5	5	5	5	5	5
Correlation	5	3	3	X	X	X
Relationship among attributes	5	3	3	1	1	1

Table 6. Analysis on characteristics of information visualization techniques

Another important point to consider is the analyst's familiarity with the analyzed data. This is what will awaken new interests or stimulate the user's curiosity during data exploration, forming new hypotheses that can be verified by means of visualizations, or simply comparing the results generated by graphical representations.

6. Conclusion

The aim of this work is minimizing difficulties in the selection of visualization techniques to represent data mining results or even to clarify the data structure throughout the knowledge discovery process. Following this purpose, two approaches were presented: the first one is visualization technique as itself a tool for knowledge discovery, and the second one is visualization techniques applied on results of data mining process.

In the first approach, through Grounded Theory methodology, as soon data have been obtained, they had undergone the coding process that allowed emergence of the concepts that led to the parameters: data type, task type, volume, dimensionality and positioning of the attributes in the graph. Then, each parameter was analyzed in conjunction with visualization techniques, and those most frequently found in the literature were selected, and separated by categories defined in the Keim's taxonomy. Through analysis of this relationship, it was observed that each technique type (1D to 3D standard graphics, icon-based displays, geometrically transformed displays, pixel-oriented displays and graph-based or hierarchical displays) have a certain configuration of parameters that reflect the data characteristics and the objectives of its use.

In the second approach, it was made an evaluation of geometric projection and iconographical information visualization techniques using Analysis of Characteristics technique. However, this technique of evaluation is subjective, because the evaluation reflects the analyst's tendency. For this reason, this evaluation served as an example of applicability of the guidelines established by Approach 1.

It should be noted that the guidelines were established based on two main items: the strongest characteristics of data, identified during the GT's coding phase in the Approach 1, and the features of visualization techniques. This does not mean the invalidation of the use of a visualization technique for other purposes that differ from those established by the guidelines. The meaning of the data analyzed to the analyst is very relevant. As the analyst's familiarity about data increases, greater is his/her incentive to make data exploration to get new hypotheses to be verified by visualizations, or just comparing the results generated by various graphical representations. Thus, the established guidelines intend to be helpful when the analyst is planning to use visualization techniques in the process of extracting knowledge from data.

Acknowledgements

This work was supported by the Fundação Araucária.

Author details

Dias Maria Madalena, Yamaguchi Juliana Keiko, Rabelo Emerson and Franco Clélia
State University of Maringá, Informatic Department, Paraná, Brazil

7. References

[1] Allan G (2003) A critique of using grounded theory as a research method. Electronic Journal of Business Research Methods. j. 2: 1-10.

[2] Ankerst M (2001) Visual Data Mining with Pixel-oriented Visualization Techniques. ACM SIGKDD Workshop on Visual Data Mining, San Francisco, CA.
[3] Barioni C M, Botelho E, Faloustsos C, Razente H, Traina A J M, Traina Jr, C (2011) Data Visualization in RDBMS. Proc. International Conference on Information Systems and Databases, Tokyo, Japan. pp. 367-063.
[4] Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When is "Nearest Neighbor" Meaningful?. Proc. 7th International Conference on Database Theory, Jerusalem, Israel. pp. 217-235
[5] Böhm C, Kriegel H P (2000) Dynamically Optimizing High-Dimensional Index Structures. Proc. 7th International Conference on Extending Database Technology, Konstanz, Germany. pp. 36-50.
[6] Carr D B, Littlefield R J, Nicholson W L, Littlefield J S (1987) Scatterplot matrix techniques for large n. Journal of the American Statistical Association. j. 82:424-436.
[7] Chen C, Härdle W, Unwin A (2008) Handbook of Data Visualization. Springer, 2008. 936 p.
[8] Chernoff H (1973) The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association. j. 68:361-368.
[9] Cockburn A, McKenzie B (2000) An evaluation of cone trees. People and Computers. j. pp. 425-436.
[10] Crespo A A (2011) Estatística Fácil. Saraiva, São Paulo.
[11] Downing, D., Clark, J (2002) Estatística Aplicada. Saraiva, São Paulo.
[12] Ebert D S, Rohrer M R, Shaw D C, Panda P, Kukla M J, Roberts A D (2001) Procedural Shape Generation for Multi-dimensional Data Visualization. Computers & Graphics. j. 24:375-384.
[13] Flury B, Riedwyl H (1981) Graphical representation of multivariate data by means of asymmetrical faces. Journal of the American Statistical Association. j.76:757-765.
[14] Gershon N, Eick S G (1997) Information Visualization. IEEE Computer Graphics and Applications. j. 17:29-31.
[15] Grinstein G, Trutschl M, Cvek U (2001) High dimensional visualizations. Proceedings of the 7th Data Mining Conference-KDD. Citeseer.
[16] Goldschmidt R, Passos E (2005) Data Mining: um Guia Prático. Campus, Rio de Janeiro, Brazil.
[17] Havre S, Hetzler E, Whitney P, Nowell L (2002) Themeriver: Visualizing thematic changes in large document collections. Visualization and Computer Graphics, IEEE Transactions on. j. 8:9-20.
[18] Hofmann H (2003) Constructing and reading mosaicplots. Computational Statistics & Data Analysis. j. 43: 565-580.
[19] Hunter K, Hari S, Egbu C, Kelly J (2005) Grounded Theory: Its Diversification and Application Through two Examples From Research Studies on Knowledge and Value Management.The Electronic Journal of Business Research Methodology. j. 3:57-68. Available:
www.ejbrm.com. Accessed 2010 Feb 14.
[20] Inselberg A, Dimsdale B (1990) Parallel Coordinates: a Tool for Visualizing Multidimensional Geometry. Proc. Conference on Visualization, San Francisco, Los Alamitos. p. 23-26.

[21] Inselberg A (2008) Parallel Coordinates: Visualization, Exploration and ClassiïňĄcation of High-Dimensional Data. In: Chen C, Härdle, W, Unwin A, editors. Handbook of Data Visualization. Springer. pp. 643-680.
[22] Johnson A R, Wichner W D (2011) Applied Multivariate Statistical Analysis. Prentice-Hall, New Jersey.
[23] Keim D, Kriegel H P (1996). Visualization Techniques for Mining Large Databases: a Comparison. IEEE Transactions on Knowledge and Data Engineering. j. 8:923-938.
[24] Keim D (2000). Designing pixel-oriented visualization techniques: Theory and applications. Visualization and Computer Graphics, IEEE Transactions on. j. 6:1-20.
[25] Keim D (2001) Visual exploration of large data sets. Communications of the ACM. j.44:38-44.
[26] Keim D (2002) Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics. j. 7: 100-107.
[27] Kienreich W, Sabol V, Granitzer M, Klieber W, Lux M, Sarka W (2005) A visual query interface for a very large newspaper article repository. Database and Expert Systems Applications. IEEE Proceedings. Sixteenth International Workshop on. pp. 415-419.
[28] Klippel A, Hardisty F, Weaver C (2009) Starplots: How shape characteristics influence classification tasks. Cartography and Geographic Information Science. j. 36:149-163.
[29] Kosara R, Sahling G, Hauser H (2004) Linking Scientific And Information Visualization With Interactive 3d Scatterplots. International Conference In Central Europe On Computer Graphics, Visualization And Computer Vision Short Communication. j. 12: 133-140.
[30] LeBlanc J, Ward M O, Wittels N (1990) Exploring n-dimensional databases. Proceedings of the 1st conference on Visualization'90. IEEE Computer Society Press. p.237.
[31] Lee M D, Reilly R E, Butavicius M E (2003) An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data. Proceedings of the Asia-Pacific symposium on Information visualisation. Australian Computer Society, Inc. j. 24: 1-10.
[32] Levkowitz H (1991) Color Icons: Merging Color and Texture Perception for Integrated Visualization of Multiple Parameters. Proc. IEEE International Conference on Visualization, San Diego, USA . pp. 164-170.
[33] Mackinlay J (1986) Automating the Design of Graphical Presentations of Relational Information. ACM Transactions on Graphics. j. 5: 110-141.
[34] Matavire R, Brown I (2008) Investigating the use of "Grounded Theory" in information systems research. SAICSIT '08: Proceedings of the 2008 annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries. j. 139-147.
[35] Mao Y, Dillon J, Lebanon G (2007) Sequential document visualization. IEEE transactions on visualization and computer graphics. pp.1208-1215.
[36] Morris C J, Ebert D S, Rheingans P (2000) Experimental analysis of the effectiveness of features in chernoff faces. Proc Spie Int Soc Opt Eng. Citeseer. 3905:12-17.
[37] Myatt G J (2007) Making sense of data: a practical guide to exploratory data analysis and data mining. Wiley-Blackwell.
[38] Nascimento H A D, Ferreira C B R (2005) Visualização de informações - uma abordagem prática. XXV Congresso da Sociedade Brasileira de Computação, XXIV JAI, São Leopoldo, RS, Brazil.
[39] NIST/SEMATECH (2003) Nistsematech e-handbook of statistical methods. Available: http://www.itl.nist.gov/div898/handbook/. Accessed 2011 Nov 14.

[40] Oliveira M C F, Levkowitz H (2003) From visual data exploration to visual data mining: A survey. IEEE Transactions on Visualization and Computer Graphics. j. 9:378-394.
[41] Peng W, Ward M O, Rundensteiner E A (2004) Clutter reduction in multi-dimensional data visualization using dimension reordering. IEEE Symposium on Information Visualization. pp. 89-96.
[42] Pillat R M, Valiati E R A, Freitas C M D S (2005) Experimental study on evaluation of multidimensional information visualization techniques. Proceedings of the 2005 Latin American conference on Human-computer interaction, ACM. j. pp. 20-30.
[43] Pickett R M, Grinstein G G (1988) Iconographic displays for visualizing multidimensional data. Proc. IEEE Conf. on Systems, Man and Cybernetics. IEEE Press, Piscataway, NJ. 514:519
[44] Pfleeger L S (2004) Engenharia de Software: Teoria e Prática. Pearson Prentice-Hall: São Paulo.
[45] Rabelo E, Dias M M, Franco C, Pacheco R C S (2008) Information Visualization: Which is the most Appropriate Technique to Represent Data Mining Results?. Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation, Viena - Austria. pp. 1218-1223.
[46] Robertson G G, Mackinlay J D, Card S K (1991) Cone trees: animated 3d visualizations of hierarchical information. Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology. ACM. pp. 189-194.
[47] Rodrigues J F (2003) Desenvolvimento de um Framework para Análise Visual de Informação Suportando Data Mining. Phd Thesis. Universidade de São Paulo, São Paulo, Brazil.
[48] (2011) The R Project. Available: http://www.r-project.org/. Accessed 2011 Nov 14.
[49] Rundensteiner E A, Ward M O, Yang J, Doshi P R (2002) XmdvTool: Visual Interactive Data Exploration and Trend Discovery of High-dimensional Data Sets. Proc. ACM SIGMOD International Conference on Management of Data, Madison pp.631.
[50] Russo S C, Gros P, Abel P (1999) Visualização Tridimensional de Grandes Volumes de Informação. Proc. Congresso Luso-Moçambicano de Engenharia, Maputo, Mozambique. 2:73-87
[51] Sensalire M, Ogao P, Telea A (2008) Classifying desirable features of software visualization tools for corrective maintenance. Proceedings of the 4th ACM symposium on Software visualization. pp. 87-90.
[52] Shimabukuru H M (2004) Visualizações Temporais em uma Plataforma de Software Extensível e Adaptável. Phd Thesis. Universidade de São Paulo, São Paulo, Brazil.
[53] Shneiderman B (1996) The eyes have it: A task by data type taxonomy for information visualizations. VL'96: Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, Colorado. j. pp. 336-343.
[54] Shneiderman B (2006) Discovering business intelligence using treemap visualizations. Technical report, B-Eye: Business Intelligence Network.
[55] Starre L V D, Vries T (2005) Visualizing documents: analysis and evaluation.
[56] Taylor A L, Hickey T J, Prinz A A, Marder E(2006) Structure and visualization of high-dimensional conductance spaces. Journal of neurophysiology. j. 96:891-905.
[57] Valiati E R A (2008) Avaliação de usabilidade de técnicas de visualização de informações multidimensionais. PhD thesis, Universidade Federal do Rio Grande do Sul.

[58] Voinea L, Telea A (2007) Visual data mining and analysis of software repositories. Computers & Graphics. j. 31:410-428.
[59] Wegman E J, Luo Q (2011) High-Dimensional Clustering Using Parallel Coordinates and the Grand Tour. Computing Science and Statistics. j. 28:352-360.
[60] Yamaguchi J K, Dias M M, Franco C (2011) Guidelines For The Choice of Visualization Techniques Applied in the Process of Knowledge Extraction. 13th International Conference on Enterprise Information Systems - ICEIS 2011, Beijing, China. pp. 183-189.
[61] Yamaguchi J K, Dias M M (2011) A Study about Influenceable Parameters in the Choice of Visualization Techniques Based on Grounded Theory. IADIS International Conferences Computer Graphics, Visualization, Computer Vision and Image Processing 2011, Rome, Italy. pp.177-184.
[62] Zeckzer D, Kalcklösch R, Schröder L, Hagen H, Klein T (2008) Analyzing the reliability of communication between software entities using a 3d visualization of clustered graphs. Proceedings of the 4th ACM symposium on Software visualization. pp. 37-46.