GUIDELINES FOR THE CHOICE OF VISUALIZATION TECHNIQUES APPLIED IN THE PROCESS OF KNOWLEDGE EXTRACTION

Juliana Keiko Yamaguchi, Maria Madalena Dias and Clélia Franco
Department of Informatic, State University of Maringá, Av. Colombo - 5.790, Maringá, Brazil

Keywords: Data visualization, Visualization techniques, Knowledge discovery, Data visualization parameters.

Abstract: Visualization techniques are tools that can improve analyst's insight into the results of knowledge discovery process or to directly explore and analyze data. They allows analysts to interact with the graphical representation to get new knowledge. The choice of visualization techniques must follow some criteria to guarantee a consistent data representation. This paper presents a study based on Grounded Theory that indicates parameters for select visualization techniques, which are: data type, task type, data volume, data dimension and position of the attributes in the display. These parameters are analyzed in the context of visualization technique categories: standard 1D - 3D graphics, iconographic techniques, geometric techniques, pixel-oriented techniques and graph-based or hierarchical techniques. The analysis over the association among these parameters and visualization techniques culminated in guidelines establishment to choose the most appropriate techniques according to the data characteristics and the objective of the knowledge discovery process.

1 INTRODUCTION

Represent the gained information in a visual way is a solution for facilitate the understanding of analized data. Thus, visualization techniques can be integrated into the process of KDD, so much to preview data to be analyzed or help in understanding the results of data mining, so much to understand the partial results of the iterations in the process of extracting knowledge (Ankerst, 2001).

However, exploration and analysis of data using visualization techniques directly applied to data can bring useful and new knowledge, enough to exempt other data mining techniques. Furthermore, visualization is a powerful tool for conveying ideas, due to vision plays an important role in human cognition (Nascimento and Ferreira, 2005).

When visualization techniques are chosen to data analysis, some criteria must be considered so that the graphical representation really helps in understanding data. First of all, it should be observed relevant characteristics of the data, such as data type, dimensionality (number of attributes) and volume.

Tasks that users can perform during data exploration may also be another factor in this decision. Basically, the following tasks are the most common mentioned in the literature: data overview, verification of correlation among attributes, identification of new rules or patterns, cluster analysis and outliers detection. Furthermore, depending on the visualization technique used, positioning of the attributes in the graph can be significant in interpreting the behavior of data.

This paper presents a study whose research methodology was based on Grounded Theory to establish guidelines for choosing most suitable visualization techniques according to the characteristics of analyzed data.

So, in next section Grounded Theory methodology is briefly presented. Following, the items: data type, dimensionality, volume, task type and positioning of attributes in the graphic are the named parameters identified through this methodology and they are described in sequence.

Next, an analysis on the association of each parameter to different types of visualization techniques is discussed. After this, general guidelines for choosing visualization techniques according to the identified parameters are outlined.

Keiko Yamaguchi J., Madalena Dias M. and Franco C..
GUIDELINES FOR THE CHOICE OF VISUALIZATION TECHNIQUES APPLIED IN THE PROCESS OF KNOWLEDGE EXTRACTION.
DOI: 10.5220/0003469901830189
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (ICEIS-2011), pages 183-189
ISBN: 978-989-8425-53-9

Finally, last section presents the conclusion of this work.

2 RESEARCH METHODOLOGY

Grounded Theory (GT) was proposed in 1967 by both sociologists, Glaser and Strauss, as an alternative to traditional scientific method that it normally consists of problem formulation, definition and verification of the hypothesis and conclusion. GT, in turn, does not start from any hypothesis.

In this work was employed the following steps of GT for the guidelines preparation for choosing visualization techniques, based on GT versions of the authors: Rodon and Pastor (2007) and Orlikowski (1993), which are illustrated in Figure 1 and described as follows.

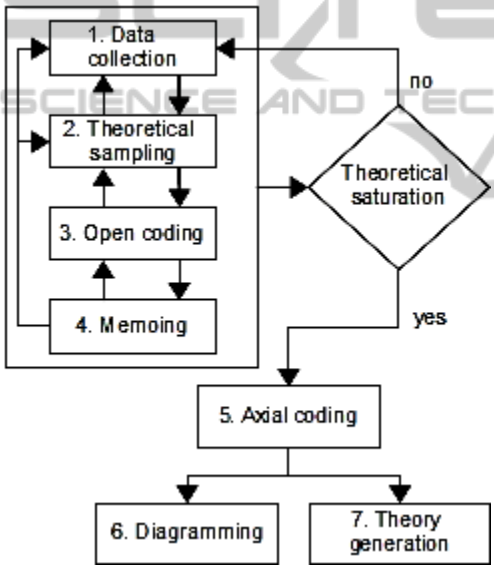

Figure 1: The Grounded Theory Process followed.

- Data collection - in this stage, the literature was used as a data source, in which information about visualization were selected and analyzed in order to replace original data could be obtained from interviews, questionnaires and forms (Dick, 2005).
- Theoretical sampling - it is done alongside with data collection, gathering all analyzed data and comparing them with new informations.
- Open coding - from theoretical sampling the data are classified according to their similarities. Each class is identified for a code which, in this case, represents the parameters to be considered to choose visualization techniques. It was used the key point coding method (Allan, 2003).
- Memoing - consists of analyst's annotations about encoding process and the analyzed data, providing extra information for construct theory.
- Axial coding - corresponding to arrange the categories defined in open coding and the concepts connecting them for contextualize the theory. Thus, in this step the guidelines are defined from the association between the identified parameters and the categories of visualization techniques.
- Diagramming - comprises constructing a diagram to illustrate the concepts and categories in order to facilitate the understanding of the theory or phenomenon that concludes the study.
- Theory generation - this work has as resulted guidelines for choosing visualization techniques based on parameters that influence this choice according to the characteristics of the data, being the guidelines considered as the generated theory in this step.
- Theoretical saturation - when new data will not influence the organization and structure of categories and concepts previously defined, it is an indication that the theoretical saturation point was reached and, consequently, the theoretical sampling represents the scope of research.

The identified parameters through open coding process are described in next.

3 GETTING THE PARAMETERS

The parameters to consider in select visualization techniques emerged from open coding by analyzing the literature related to visualization techniques, in which we used the key points encoding method. The result is illustrated in Table 1.

In the first column of this table, the expressions was taken from the 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, volume, 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.

Table 1: Key point coding method applied on collected data.

Key points	References	Code
Visualization techniques can be classified, among other criteria, by data type	Shneiderman (1996) Freitas et al. (2001) Keim (2002)	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	Shneiderman (1996) Keim (2002) Pillat et al. (2005)	Task type
Visualization techniques are subject to some limitations, such as the amount of data that a particular technique can exhibit	Keim e Kriegel (1996) Oliveira e Levkowitz (2003) Rabelo et al. (2008)	Volume
Visualization techniques can also be classified according to the number of attributes	Shneiderman (1996) Grinstein et al. (2001) Keim (2002) Oliveira e Levkowitz (2003)	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	Ankerst (2001) Oliveira e Levkowitz (2003) Inselberg (2008) Klippel et al. (2009)	Positioning of attributes

4 GUIDELINES FOR SELECTION OF VISUALIZATION TECHNIQUES

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 (2002), 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. In iconographic techniques data attributes are mapped into properties of an icon or glyph, which vary depending on the values of attributes. In geometric techniques, multidimensional data are mapped into a two-dimensional plane providing an overview of all attributes. 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 (Keim, 2000). Data with a naturally structure of relationships among its elements, hierarchical or simple network, may be represented by hierarchical or graph-based techniques.

It was chose some visualization techniques of each category to illustrate the analysis of the parameters. Among standard charts, it was selected: the Histogram; the Box Plot; the Scatter Plot and the Contour Plot. Among Icon-based: the Chernoff Faces; the Star Glyphs and the Stick Figure. Among Geometrically transformed displays: the Scatter Plot Matrix and the Parallel Coordinates. Among Pixel-oriented displays: the Query-dependent and the Query-independent techniques. Among Graph-based or Hierarchical: the Graph; the Cone Tree; the Treemap; the Dimensional Stacking and the Mosaic Plot. Next, each parameter is analysed with the techniques mentioned above.

4.1 Analysis on the Data Type Parameter

Techniques of standard 1D-3D category generally represents from one to three attributes and are used for analysis of quantitative data in most cases. All charts considered in this class are able to display quantitative data. To represent qualitative data, alternative techniques are more limited. From the selected techniques of this class only the Histogram is able to plot qualitative data (Myatt, 2007). 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.

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 scatterplot matrix is formed by a set of scatterplots, it is more suitable for continuous quantitative data.

In literature are found examples of usage of pixel-oriented techniques on quantitative data. Query-independent techniques were applied to represent temporal data, and query-dependent techniques are commonly used to represent continuous quantitative data. Keim (2001) 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.

4.2 Analysis on the Task Type Parameter

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

Task type refers to activities that user or analyst can perform according to goals in the use of a graphical representation as noted in the literature (Keim, 2002; Shneiderman, 1996; Pillat et al., 2005). 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 behaviour and trends;
- Identification of rules, standards and important characteristics;
- Clusters identification: attributes with similar behaviour;

Outliers detection: data set with atypical behaviour in comparison for the rest of data.

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 (in this case, 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 scatterplot.

Iconographic techniques represent each data entry individually, allowing verification of rules and behaviour patterns of the data. Icons with similar properties can be recognized and thus form groups and it be analysed in particular. A representation with a discrepant format if compared to the other may characterize an outlier.

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 scatterplot matrix. This category of techniques also allows the identification of patterns, rules and behaviours and may also detect outliers, characterized by behaviours outside the common standard. The analyst may choose to analyse a group of data that can be detached from the tool in use but, in principle, groups are not immediately identified by techniques of this class.

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

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 analyse the relationship among the elements. Techniques of this category also allow grouping data, such as Treemaps (Shneiderman, 2006).

4.3 Analysis on the Volume and Dimensionality Parameters

Implementations of visualization techniques must take in consideration the limits of dimensionality and volume of data to hold in way that the tool be capable of providing a clear overview of data to the analyst.

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 (Rabelo, 2008), 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.

4.4 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 (Shneiderman, 2006), Mosaic Plots (Hofmann, 2008), Dimensional Stacking (LeBlanc et al., 1990). 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.

In general, for 1D-3D standard graphics, positioning of attribute do 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.

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 (Pickett and Grinstein, 1988).

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 (Morris et al., 2000; Lee et al., 2003). Likewise it is for Star Glyph technique, for which once established the order of the best mapping of attributes (Peng et al., 2004; Klippel et al., 2009), it remains the same for all the icons representing a record data per star.

In the works of (Inselberg, 2008) and (Wegman, 1990), it is explained how the position of the attributes in the graph may influence the correlation detection in Parallel Coordinates. Scatteplot Matrix is, on the other hand, composed of a set of scatterplots, for this reason nor is influenced by the change of attributes positioning, since their main objective is to evaluate the correlation between attributes.

Keim (2000) 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 (Keim, 1997).

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 (Cockburn and McKenzie, 2000; Robertson et al., 1991).

5 CONCLUSIONS

The Grounded Theory provided a methodology for the identification of the parameters and guidelines for choose visualization techniques, set forth through the stages theoretical sampling, coding, diagramming and formulation of the theory.

During the development of this work, five parameters were identified: data type, task type, volume and dimensionality of the data and position of the attributes in the graph. Subsequently these parameters were analyzed in relation to the

categories of visualization techniques distinct among 1D-3D standard graphics, iconographic techniques, geometric techniques, pixel-oriented techniques, and hierarchical techniques or graph-based techniques.

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.

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 a priori used. Qualitative data, for example, will be hardly understood if they were represented by a technique developed to represent quantitative data and vice versa. Furthermore, it was verified in this study that there are more options for visualization techniques to represent quantitative data than qualitative data.

The task type to be performed corresponds to the goals of the analyst during the data exploration. In literature are found classifications of visualization techniques based on this parameter. 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.

Both volume 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. In the case of parallel coordinates, positioning of attributes influences the way data are displayed in polygonal lines; in star glyphs technique, the order of distribution of attributes to the icon properties can ease grouping task, considering that arranging them in different orders may generate diverse icon formats.

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

It should be noted that the guidelines were established based on the strongest features identified during the coding phase. This does not mean the invalidation of the use of a visualization technique for other purposes that differ from those established by the guidelines.

Therefore, visualization is a beneficial tool in understanding the knowledge, which may be achieved through data mining algorithms or by visual exploration performed directly on the data.

ACKNOWLEDGEMENTS

This work was supported by the *Fundação Araucária*.

REFERENCES

Allan, G. (2003). A critique of using grounded theory as a research method. *Electronic Journal of Business Research Methods*, 2(1):1–10.

Ankerst, M. (2001). Visual data mining with pixel-oriented visualization techniques. In *Proceedings of the ACM SIGKDD Workshop on Visual Data Mining*. Citeseer.

Cockburn, A. and McKenzie, B. (2000). An evaluation of cone trees. *People and Computers*, pages 425-436.

Dick, B. (2005). Grounded theory: a thumbnail sketch.

Freitas, C., Chubachi, O. M., Luzzardi, P. R. G., and Cava, R. A. (2001). Introdução à visualização de informações. *Revista de Informática Teórica e Aplicada*, 8(2):143–158.

Grinstein, G., Trutschl, M., and Cvek, U. (2001). High dimensional visualizations. In *Proceedings of the 7th Data Mining Conference-KDD*. Citeseer.

Hofmann, H. (2008). Mosaic plots and their variants. In Chen, C., Härdle, W., and Unwin, A., editors, *Handbook of Data Visualization*, pages 617–642. Springer.

Inselberg, A. (2008). Parallel Coordinates: Visualization, Exploration and Classification of High-Dimensional Data. In Chen, C., Härdle, W., and Unwin, A., editors, *Handbook of Data Visualization*, pages 643–680. Springer.

Keim, D. A. (1997). Visual techniques for exploring databases.

Keim, D. A. (2000). Designing pixel-oriented visual-ization techniques: Theory and applications. *Visualization and Computer Graphics, IEEE Transactions on*, 6(1):1–20.

Keim, D. A. (2001). Visual exploration of large data sets. *Communications of the ACM*, 44(8):38–44.

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

Keim, D. A. and Kriegel, H. (1996). Visualization techniques for mining large databases: A comparison. *IEEE Trans. on Knowl. and Data Eng.*, 8(6):923–938.

Klippel, A., Hardisty, F., and Weaver, C. (2009). Starplots: How shape characteristics influence classification tasks. *Cartography and Geographic Information Science*, 36(2):149–163.

LeBlanc, J.,Ward,M. O., andWittels, N. (1990). Exploring n-dimensional databases. In *Proceedings of the 1st conference on Visualization'90*, page 237. IEEE Computer Society Press.

Lee, M. D., Reilly, R. E., and Butavicius, M. E. (2003). An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data. In *Proceedings of the Asia-Pacific symposium on Information visualisation,* Volume 24, pages 1–10. Australian Computer Society, Inc.

Morris, C. J., Ebert, D. S., and Rheingans, P. (2000). Experimental analysis of the effectiveness of features in chernoff faces. In *Proc Spie Int Soc Opt Eng*, volume 3905, pages 12–17. Citeseer.

Myatt, G. J. (2007). *Making sense of data: a practical guide to exploratory data analysis and data mining*. Wiley-Blackwell.

Nascimento, H. A. D. and Ferreira, C. B. R. (2005). Visualização de informações – uma abordagem prática. In *XXV Congresso da Sociedade Brasileira de Computação, XXIV JAI*, São Leopoldo, RS, Brazil. UNISINOS.

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

Orlikowski, W. J. (1993). Case tools as organizational change: investigating incremental and radical changes in systems development. *MIS quarterly*, 17(3):309–340.

Peng, W., Ward, M. O., and Rundensteiner, E. A. (2004). Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 89–96. IEEE.

Pickett, R. M. and Grinstein, G. G. (1988). Iconographic displays for visualizing multidimensional data. In *Proc. IEEE Conf. on Systems, Man and Cybernetics, IEEE Press*, Piscataway, NJ, volume 514, page 519.

Pillat, R. M., Valiati, E. R. A., and Freitas, C. M. D. S.(2005). Experimental study on evaluation of multidimensional information visualization techniques. In *Proceedings of the 2005 Latin American conference on Human-computer interaction*, pages 20–30. ACM.

Rabelo, E., Dias, M., Franco, C., and Pacheco, R. C. S.(2008). Information visualization: Which is the most appropriate technique to represent data mining results?In *CIMCA '08: Proceedings of the 2008 International Conference on Computational Intelligence for Modelling Control & Automation*, pages 1228–1233, Vienna, Austria. IEEE Computer Society.

Robertson, G. G., Mackinlay, J. D., and Card, S. K. (1991). Cone trees: animated 3d visualizations of hierarchical information. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, pages 189–194. ACM.

Rodon, J. and Pastor, J. (2007). Applying grounded theory to study the implementation of an inter-organizational information system. *Electronic Journal of Business Research Methods*, 5(2):71–82.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–343, Boulder, Colorado. IEEE Computer Society.

Shneiderman, B. (2006). Discovering business intelligence using treemap visualizations. Technical report, B-Eye: Business Intelligence Network.

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85(411):664–675.