Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 73 (2013) 136 – 144

Procedia
Social and Behavioral Sciences

The 2nd International Conference on Integrated Information

Graphical Representation and Exploratory Visualization for Decision Trees in the KDD Process

Mr Wilson A. Castillo Rojas[a], Mr Claudio J. Meneses Villegas[b]*

[a]*Arturo Prat University, Engineering - Area Computing and Informatics, Av. Arturo Prat 2120, Iquique - Postcode: 110-0000, Chile*
[b]*North Catholic University, Systems and Computer Engineering, Av. Angamos 0610 Postcode: 1280, Antofagasta, Chile*

Abstract

This article presents a proposal of representation and scheme of exploratory visualization for Decision Trees in the KDD (*Knowledge Discovery in Database*) process, specifically in the data mining stage. With this, the improvement of the understandability of the internal operation of the model is pursued. This exploratory visualization is based on the well-known technique named Treemap (maps of trees) that allows representing hierarchical structures like the Decision Trees, being used grids to represent the nodes of the Decision Trees. The proposed visualization represents the number of instances or weight associated to a node with a scale of colors in degradation. In this way it is managed to heighten the rules of the Decision Trees in a 2D and 3D graphical representation of this visualization. Finally, a first attempt of subjective evaluation, based on for criteria, of the proposed visualization, is made. In this sense, this work pursues to introduce new schemes of visualizations that allow specifically understand how the data mining models work internally.

Selection and peer-review under responsibility of The 2nd International Conference on Integrated Information

Keywords: Visualization of Decision Trees; Visual Data Mining Models; Schemes of Visualization; Visualization for Data Mining; Data Mining.

1. Introduction

The Knowledge Discovery in Database term (KDD) is defined as "*the nontrivial process of identifying valid patterns, novel, potentially useful and understandable data*" [1]. Therefore, KDD is the overall process of information analysis and knowledge extraction, which covers the stages from selecting the data to analyze, until eventually the end user gets a new knowledge. While data mining (DM) is a typical stage of this process and is

* Corresponding author. Tel.: +56-57-394403 ; fax: +56-57-394472 .
E-mail address: wilson.castillo@unap.cl

1877-0428
Selection and peer-review under responsibility of The 2nd International Conference on Integrated Information
doi:10.1016/j.sbspro.2013.02.033

responsible for extracting knowledge models that are to be submitted and verified by the user based on data previously prepared for this [2].

At this stage, various machine learning techniques are used, although the learning obtained through these techniques, regardless of type of problem, it is always hypothetical, i.e. patterns obtained do not stop being a hypothesis, and can be refuted by evidence future [3]. To achieve models that can respond appropriately to all types of evidence or sets of comprehensive data, and thus reduce the possibility that the patterns generated are subsequently rejected, we propose in this paper the need for graphical display to improve the comprehensibility of models. So, it is looking for an exploratory analysis and iterative models before generating a final model in the mining stage of the KDD process, i.e., do not apply only visualization on the stage of interpretation and final consolidation, which is now known and Visual DM.

Therefore, we argue that using ad hoc schemes to visualize models, a better knowledge of the inner workings of the model can be obtained, before generating and selecting a final one. Thus, it seeks to support the choice of most appropriate model to solve a problem, by adjusting its parameters or other variables that affect their behavior. For this, it is essential to count early with a visual graphic representation, with appropriate interaction mechanisms that allow for one side to tour the model and, on the other hand, internally observing their behavior. For example, visualizing the structure or components of the model for iteration, grouping the data or isolating outlier data. Many authors [4] [5] [6] raise the desirability of combining multiple DM techniques to make this internal analysis of the models at this stage of exploration. Thus, to evaluate the behavior obtained in order to estimate their degree of validity and compare them with others. This seeks to increase the expressiveness of patterns that are generated by applying a data-mining algorithm.

The technique widely used in DM to achieve these models is the Decision Trees (DT, this is due to its ease of use and understanding, because they support discrete and continuous attributes, have a good handle of not significant attributes and, through mechanisms of pruning, noise data can be eliminated and also missing values can be ignored. One of the main problems of DM techniques is its visual representation and understanding the inner workings of the model by the user, which in the case of DT is much more complex when it comes: large trees, the set of data to be analyzed has a high dimensionality, and due to its property of hierarchical structure [1].

This article describes a proposal for a graphical representation and visualization scheme for DT exploration in the mining phase of KDD process, based on a technique called treemap [7-8], which allows represent hierarchical structures such as DTs.

The paper is organized as follows: Section 2 describes a set of existing visual representations of DT. Section 3 explains some criteria found on the evaluation of treemaps visualizations. Section 4 describes the proposed DT visualization based on the treemap technique and supplemented with the use of colors on a scale of degradation, to enhance the visualization in 2D and 3D representation. Finally, section 5 presents the conclusions of the work done and future work planned on it.

Nomenclature

DT	Decision Tree
DM	Data Mining
KDD	Knowledge Discovery in Database

2. Visual Representations of Decision Trees

In DM, DTs deliver patterns induced from a data set, however these are not useful to quantify this categorization and thus reporting how these data behave. The learning process of DTs to generate patterns of

behavior is a simple method that has been successfully used in inductive learning tasks, it's by approximation of functions robust to the presence of spurious data and capable of learning disjunctive expressions or rules.

There is a whole family of DT learning algorithms sometimes referenced as TDIDT (Top Down Induction of Decision Trees), among these the best known are: ID3, Cart, Assistant Chaid, C4.5 and C5.0. They are characterized by searching a completely expressive hypothesis space that avoids the difficulties of restricted hypotheses; and their inductive bias is to prefer small trees to large ones [5]. This inductive learning starts from particular cases (samples) and produces general cases obtained through the establishment of rules or models, which generalize or abstract the evidence [9]. The generation of a DT is through an induction process, i.e. given a dependent variable or class, the objective is to set the class for new cases [10].

2.1. Visual Representation of Decision Trees Based on Treemaps

The treemap technique is one of the visualization schemes used for DTs, and it is a method that divides an area of deployment in a nested sequence of rectangles; where each of these generated rectangles correspond to an attribute of the data set. The rectangle that represents an item is cut in each recursive step by parallel lines in smaller rectangles representing its children. At each level of recursion, the orientation of the lines changes from horizontal to vertical [11]. When the data are 2D, one of which is correlated with the area of rectangles, as returned by the algorithm, while the other is made to correspond to a color for the rectangle. The treemaps can be designed to be applying hierarchically, so that any rectangle within it may contain, in turn, another treemap, recursively. There are two properties that are true in most of the algorithms used: First, the rectangles must be kept as square as possible, i.e. the ratio width/height should be close to 1. Very long and narrow rectangles (figure 1) could be obtained, if not taken into account this principle, making more difficult the perception of users. Second, the order of the data should be maintained, so to make it easier to find items.

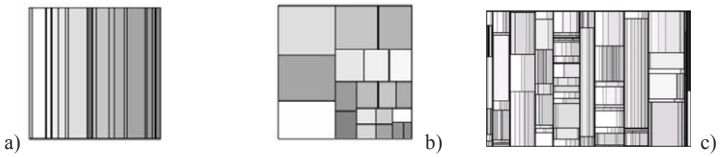

Fig. 1. Graphical display of a treemap: a) long, b) narrow rectangles easy to see, and c) representing hierarchical levels (children) of a DT [7]

In [8] shown a visualization of hierarchical data structure called hierarchical clustering (figure 2), similar to a treemap, and is based on a recursive division of a rectangular area that displays the individual items as a cluster.

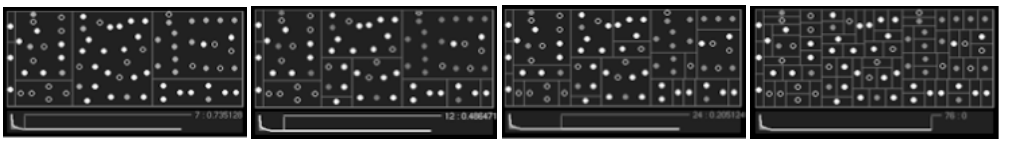

Fig. 2. Visualization of hierarchical clustering with 7, 12, 24 and 76 grids [8]

3. Criteria for Evaluating Treemap Visualizations

It was noted that the treemap technique allows information to visually represent hierarchical structures in 2D, and is capable of representing large collections of data [9].

3.1. TreeMap characteristics

There is a broad family of algorithms [11] to partition the display area of a treemap. In all these cases, the input is a set of *n* numbers and *1* rectangle. For each number *n*, creating a new partition of the original rectangle. The algorithm ensures that all the available area is filled, and the rectangles have an area proportional to the values of the data set. Treemap have been widely used to represent DT, so as to establish a flat display their hierarchical structure, where each node in the tree, starting from its root, has a grid-representation.

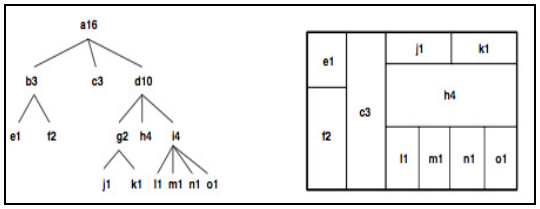

Fig. 3. A tree and its representation in a Treemap

As can be seen in figure 3, the hierarchical nature of the DT to establish decision rules generated by a data analysis model can easily be represented and interpreted by the quadrants defined by a treemap. Using these combined vertical and horizontal partitions at each level in a treemap is achieved data represent hierarchical manner similar to the branches of a DT.

3.2. Evaluation Criteria

This treemap technique has been evaluated in many studies, one of these is made by Barlow and Neville [12], comparing this technique with three displays hierarchical data in the context of a DT, and has resulted be less in terms of time and accuracy of reaction given tasks as well as user preference. Similarly, Stasko and et al. [13] compared the treemap technique with another technique for visualizing hierarchical structures in the context of searching for files and directories, the operation of the treemap was lower in these tasks, especially in the initial use, but improves over time. Also the user's preference was lower in the case of this technique. Card et al. [14] establishes that the treemap technique is especially effective to display variable data where quantitative values and large is important. Moreover, Johnson in [8] treemaps are evaluated empirically and it is showed that the technique is used efficiently after 15 minutes of training in their use.

Currently there are few studies providing assessment tools or methods of visualization techniques, and thus able to establish empirically the capabilities of each technique to answer business questions and tasks of DM. Marghescu in 2008 [4], performs an evaluation of nine techniques of visualization, among which is the treemap technique, in the context of a problem in financial analysis comparing different companies. In this study, they are evaluated and compared the visualization techniques on four criteria:

1) Visual efficiency: the ability of techniques to answer questions and DM tasks, formulated for the problem.
2) Information of high/low level: the ability of techniques to display data items or data models.
3) Type of data: the type of data used as input to the visualization technique (original or normalized data).
4) The expressiveness: the extent to which the variables are represented graphically.

Table 1 summarizes the evaluation of the nine techniques considered in [4], which shows the capacity of each regarding business questions arising in DM tasks, based on the four criteria listed above.

Table 1. Summary of the Marghescu Evaluation

Technique / tasks of Data Mining	Detection outliers values	Analysis of Dependencies (Relationships)	Clusters	Clusters Description	Class	Class Description	Comparison between items
Multiple line graph	X	X			X	X	X
Permutations Matrix	X	X					X
Survey Plot	X	X			X	X	X
Matrix diagram Dispersion	X	X					X
Parallel coordinates	X	X					X
Treemap	X				X	X	X
PCA (Principal Components Analysis)	X	X	X	X			X
Sammon's Mapping					X	X	
SOM (All views)	X	X	X	X	X	X	X

This assessment concluded that the nine techniques evaluated provide an overview of the data set analyzed, and these views can be displayed simultaneously obtained all classes and most of the variables in the data table. Also provide information on the data structure and its characteristics. For example, you can see if there are clusters in the data, regardless of whether the variables are correlated, if there are outliers or outside edge, and if there are similarities and differences between the classes of interest.

- Regarding the criterion of visual efficiency, treemap proved to be able to: describe characteristics of the class, to facilitate comparison between different objects, and allow revealing values outside edge.
- For the second criterion of evaluation of this study, the treemap technique, it allows to easily deploy data items, as these items represent only data the user has to use their perceptual abilities to distinguish patterns of interest. When a data model is rendered, this is automatically generated and deployed by the technique.
- For the criterion of the type of data processing, treemap only represents the original and not standardized data.
- The fourth criterion referred to the level of expressiveness of the technique, where these techniques are evaluated and compared for the degree to which all the variables are involved in the construction of the visual representation of data, it was concluded that using the technique treemap another user can change the variables used to construct the graph.

The framework of the evaluation of this study was based on the model described by Soukup and Davidson [5] mapping of a business problem to business questions, the tasks of DM and visualization techniques. The evaluation shows that there is not a technique of visualization that is capable, by itself, to answer all the DM tasks, and therefore, suggests that multiple or combined visualizations are better suited to be implemented in a tool benchmarking [15]. This is also emphasized by other evaluations [16][6], performed on sets of artificial data or test for various types of tasks. A limitation of this study is that users only assessed static visualizations of data, while adding interaction in the visualizations; these could have influenced the views of users about the capabilities of the techniques. Finally, considering these evaluation criteria identified, this technique and its variant treemap [11][17], do not increase their capacity to represent the DT, if not incorporate mechanisms for user interaction [2].

4. Proposal of Visualization for DT

The main objective of this work at an early stage is to introduce a new form of graphical representation and visualization exploratory of a DT, for use in the mining phase of KDD process, through a form of partition simple gridding technique based on the treemap technique.

4.1 Design of the Proposal

For this proposal of graphic representation for a DT, it was based on a square representing the root node, and then partitions it according to the levels of DTs, using their weight as the number of instances of each node or child, to establish the size or area of each grid. Then, defines a color scale on a selected tone or shade with different degrades, allowing reflect this weight or size.

The proposal is a basic condition that the DT to consider either a small or simply have been previously reduced through some mechanism of pruning. Also, it is then quantitatively identify each node of the tree at different levels, through the use of their respective weights. For generation of the graphic form quantitative of the DT, must meet a number of steps. The first is to be applied to the data set a classification system based on DT to deliver a set of rules. Arm the DT from these rules, and if there are redundant rules or a very large number of levels (e.g. over 10), apply an algorithm for pruning trees. Although recommendation is set as a DT small, or with a number of visually manageable levels, this is not strictly the case, as can be followed by partitioning the grids with more levels deep. However, in a DT with many levels it becomes difficult to display on a screen of standard size and configuration, and the display is designed to have an overview of the rules at higher levels. To display deeper levels of a tree, can be combined with another technique and tour these levels, and add interaction mechanisms for zooming (zoom in) in and out (zoom out) so as to observe details of deeper levels.

After having achieved a tree "clean", i.e. without redundancy and a number of appropriate levels can begin construction of the graph. To graph requires that each node have a weight, which is obtained from the original data from which each occurrence of the rule provides a unit of weight to the nodes. The DT generated and measured; draws a square where each level represented by one side of the same, starting from the left side clockwise of clock to the underside (figure 4 a).

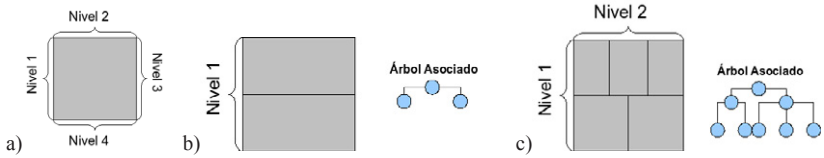

Fig. 4. a) Initial square corresponding to the DT root node, b) representation of the first level, c) and the second level of a DT

Depending on the number of nodes that have at each level, is made a uniform division of the square, the first node and the lower division always following the clockwise order of the clock (figure 4 b). From this first division, the following levels are generated and distributed to the subdivision of the spaces allocated to each parent node, depending on the number of children you have each (figure 4 c). After the division of all levels in the square, take the extreme values "evident" from left to right. These extreme values are the leaves of the DT with the minor and major weight. These extreme values allow creating a scale of colors, and in this way to colorize each node on the square (figure 5 a).

As indicated above it is proposed to establish a scale of colors or shades of a shade gradient selected to represent the number of instances (weight or size) of each tree node. The tone and color saturation is calculated proportionally to the weight of the node, a node that is heavier, will have direct bearing on the color scale, with a stronger tone of high saturation and dark, and in the same way, the node will have a lower weight smoother tone with low saturation and clear, on the same hue selected to scale.

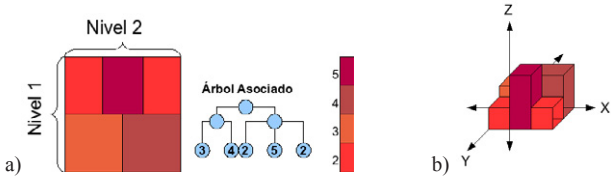

Fig. 5. a) Full graphic 2D representation of the DT, b) and 3D representation of a DT

Should not lose the perspective of the relationship of sizes that exist in each colored box, at first glance may seem that such is more significant either half, but we must not forget that this is only a 2D representation of a is 3D graph, as can be seen in this figure 5 b).

4.2 Use of Color in the Proposal

In this proposal, it was decided to use color as a mechanism and discriminatory element of the rules generated in a DT, from the processed data. The justification for the use of color is based on two visions or studies supporting this decision: The first study considers the psychic parameters of color perception are brightness, hue, and saturation [18], and it is added it is recognized that certain colors have a gradual impact on the mood of people, and about 70% of the cells of our body that allow residents feel sensations in the eyes, and as a result there an intimate connection between seeing and thinking [19]. The most important use of color is to distinguish one element from another, ie the data label. The second study [20] considers several variables to justify the use of color in visualization, as an important tool in this task, as well as using a color scale with degradation.

4.3. Application to an Example

Once this representation and graphic display scheme to represent and quantify the rules extracted from a DT, we proceeded to apply it to a preliminary example, and for this we used the data set known as weather. To generate the DT of the use case, was entered this dataset to the Orange program, rated J48 tree, occupying the same data as training set, and obtained the DT shown in figure 6 a). As can be seen in this figure, it achieves a DT on two levels with their respective weights in the leaves, thus obtaining the weights visually ends to the nodes (at least 2 and maximum 4), from which generates the scale of colors in appropriate degradation. It performs the division of the square defined according to the proposal, and starts from the left side by dividing the total square into three equal parts, where in the first node is the lower division. The second level to the division is performed from the top of the square, then divide the subsections which correspond, not the total of the square, in this case it divides the first and third segment into two equal parts.

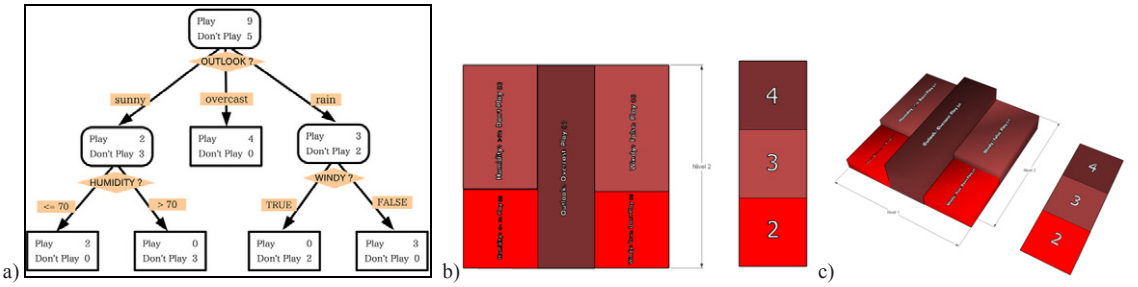

Fig. 6. a) DT built from the weather dataset, b) its 2D representation and, c) and 3D using color scale.

Finally, establishing a scale of colors in gradient, which represents the relative weight of each node as shown in figure 6 b). This graph can be represented visually quantifying the DT presenting a view completely in simple and understandable for the end user.

Comparing this graphical representation flat 2D of the DT with the hierarchical representation of figure 6 a), can be easily seen in this picture that there is a rule or pattern and is highlighted by the color scale in degradation associated with increased weight that stands out from the other two end nodes of the same level, and that in turn these two nodes are similar with respect to the number of instances or weight. In contrast, only observing this hierarchical representation the user must resort to the values of the nodes to determine which rule has a greater number of instances of the other. If this visualization is reinforced by the 3D representation of the DT of figure 6 c), is much clearer by the volume associated with weight and height of the intermediate node, that this rule is more prevalent, or at least as many instances in the model.

4.4. Preliminary Assessment of the Proposal

Where as it is appropriate to use a DT with a manageable number of levels or subjected to a pruning

mechanism, 3D visualization allows easily see the extent and size of the nodes, noting that rule with the highest number of instances in the tree. Using a color scale in gradations also helps facilitate the recognition of the rule through the visual perception of the weight of each of its nodes, both in 2D and 3D representation of the proposed visualization, unlike treemap, which only provides a hierarchical representation through no uniform grids.

5. Conclusions and Future Work

The simplicity of the DT representation can clearly visualize small trees in both two and three dimensions, and understand with ease the distribution of nodes at each level. The use of color can enhance the weights of the nodes in DT, specifically in the flat 2D graphical representation, which allows the user to visually identify the number of instances of each rule in a simple and easy, with respect to the traditional hierarchical representation. Also, in the 3D graphical representation, this proposal allows the rules by enhancing its support (weight or size). The proposed DT appears displaying a successful completion of the evaluation criteria used and established in [15], and comparison with the treemap allows more intuitively to observe the weight of each rule or pattern DT on the number of instances, the latter supported by the color scale in degradation and the combination of 2D and 3D representations.

Work is currently in the design and construction of a prototype application that allows implementing this representation and graphic display automatically; it is contemplated to finish defining the design and implementation aspects of the type of displays that will consider the final tool.

References

[1] Hernández, J.; Ramírez, M.; Ferri, C. (2004). Introduction to Data Mining. Ed. Pearson Education S.A. Spain. ISBN: 84-205-4091-9.
[2] Sierra, B. (2006). Machine learning: Basic and advanced concepts. Editorial Pearson Education S. A. Spain. ISBN: 84-8322-318-X.
[3] Witten, I.; Eibe, F. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Ed. Morgan Kaufmann, USA, ISBN: 0-12-088407-0.
[4] Soukup, T.; Davidson, I. (2002). Visual Data Mining: Techniques and Tools for Data Visualization and Mining. Ed. Wiley Publishing Inc., USA, ISBN: 0-471-14999-3.
[5] Hand, D.; Mannila, H.; Smyth, P. (2001). Principles of Data Mining. A Bradford Book The MIT, Massachusetts London-England, USA, ISBN: 0-262-08290-X.
[6] Fayyad, U. M., Piatestky, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. The MIT Press, pp. 1-34.
[7] Bederson, B. B., Shneiderman, B. and Wattenberg, M. (2002). Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies. ACM Transactions on Graphics (TOG), Vol. 21, No. 4, pp. 833-854.
[8] Wills, G. J. (1998). An Interactive View for Hierarchical Clustering. Proceedings of Infovis1998, IEEE Symposium on Information Visualization, North Carolina, USA, pp. 26-34.
[9] Shneiderman, B. (2002). Treemaps for space-constrained visualization of hierarchies. HCI Lab, University of Maryland.
[10] Siirtola, H. (2007). Interactive Visualization of Multidimensional Data. Dissertations in Interactive Technology, Number 7, TAMPERE, Acta Electronica Universitatis Tamperensis 618 ISBN 978-951-44-6939-8, ISSN 1456-954X.
[11] Barlow, T. and Neville, P. (2001). A comparison of 2D visualizations of hierarchies. In Proc. of the IEEE Symposium on Information Visualization (Infovis'01). IEEE Computer Society, Washington, DC, 131.
[12] Stasko, J., Catrambone, R., Guzdial, M., and McDonald, K. (2000). An evaluation of space-filling information visualizations for depicting hierarchical structures. In Int. J. Human-Computer Studies, 53: 663-694.
[13] Card, S.K., Mackinlay, J.D, and Shneiderman, B. (1999). Readings in Information Visualization - Using Vision to Think. Morgan Kaufmann Publishers, p. 151.
[14] Johnson, B. (1993). Treemaps: Visualizing Hierarchical and Categorical Data. Unpublished Ph.D. Thesis, University of Maryland.
[15] Marghescu, D. (2008). Evaluating Multidimensional Visualization Techniques in Data Mining Tasks. ISBN 978-952-12-2152-1, ISSN 1239-1883, Painosalama Oy – Turku, Finland.
[16] Hoffman, P. E. (1999). Table Visualizations: A Formal Model and Its Applications. PhD Thesis, University of Massachusetts Lowell.
[17] Keim, D. A. (2002). Information Visualization and Visual Data Mining. IEEE transactions on Visualization and Computer Graphics, Vol. 7., No.1, pp. 100-107.
[18] Escalera, A. (2001). Visión por computador: Fundamentos y métodos. Editorial Pearson Education S. A., España. ISBN: 84-205-3098-0.
[19] Few, S. (2005). Uses and Misuses of Color. DM Review, v15 n11 p62(2). ISSN: 1067-3717.

[20] Silva, S.; Sousa, B.; Madeira, J. (2011). Using color in visualization: A survey. An International Journal of Systems & Applications in Computer Graphics. Computers & Graphics, Volume 35, Issue 2, pages 320-333. ISSN: 0097-8493.