International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 11 (November 2014) www.ijirae.com

Application of Parallel Co-ordinate System to Visualize Multivariate Data Sets

Prabhakar Chakraborty*
Department of Computer Science
St. Xavier's College(Autonomous)
Kolkata,Inidia

Asoke Nath
Department of Computer Science
St. Xavier's College(Autonomous)
Kolkata, India

Abstract— The huge amount of multidimensional data everyday becomes bigger and more complex. For these reasons, data analysis of multivariate data becomes very difficult. In this paper, we present a visualization technique for multidimensional data sets named Parallel Coordinates as a technique for exploratory data analysis. The paper describes the technique and its applications. To accumulate large, multivariate data sets has far exceeded the ability to effectively process them in search of patterns, anomalies, and other features. Generally the conventional multivariate visualization techniques do not scale well with respect to the size of the data set. A multivariate data set consists of a collection of N-tuples, where each entry of an N-tuple is an ordinal value corresponding to an independent or dependent variable. Several methods have been proposed to display multivariate data. They can be categorized as (a) Axis reconfiguration techniques (such as parallel coordinates), (b) Dimensional embedding techniques (such as dimensional stacking), (c) Dimensional sub-setting (such as scatterplots), (d) Dimensional reduction techniques (such as principal component analysis. In the present paper the authors focus on how the interactive visualization of large multivariate data sets can be done using parallel coordinates display technique. How parallel coordinates can be used to display multivariate dataset is also shown here using xdat version 2.1.

Keywords— multidimensional, parallel co-ordinates, visualization, multivariate, scatterplots

I. INTRODUCTION

Researchers require more and more efficient ways to analyse and interpret large amount of information, as large multivariate datasets become increasingly common. The visualization methods are very efficient when displaying multidimensional and multivariate datasets. The multivariate dataset is an N-dimensional set E with elements described by ei = (x_{i1}, x_{i2},..., x_{in}). Each observation x_{ij} is usually independent of the other observations and, observations, in nature, may be discrete or continuous, or may take nominal values. Several techniques have been proposed for multivariate data representation. They include axes reconfiguration techniques (such as Parallel Coordinates and glyphs); dimensional embedding techniques (such as dimensional stacking and worlds within worlds); dimensional sub setting (such as scatter-plots) and dimensional reduction techniques (such as multidimensional scaling, self organizing maps and principal component analysis). In this paper we describe the multidimensional visualization technique called Parallel Coordinates. The idea for this technique comes from multi-dimensional geometry, which frustrates researchers by the absence of visualization. The question was "How to see geometry without the benefit of the picture?" The 2D or 3D Descartes coordinate systems couldn't solve the multidimensional problems. Prof. Alfred Inselberg proposed the use of a multidimensional coordinate system based on Parallel Coordinates. Researchers have been working on improving this technique for better data investigation and easier user-friendly interaction by adding data clustering [3], brushing [1; 8], etc. With these improvements, Parallel Coordinates becomes very efficient technique for visualization relationships between designated neighbouring dimensions. By using parallel axes for dimensions, the parallel coordinate techniques can represent N-dimensional data in a 2-dimensional space. Parallel coordinates can be considered as special node link diagrams, in which the nodes are on the parallel axes and the edges are those poly-lines linking the nodes on two neighbouring axes. The visual presentation and examination of large data sets from the physical and natural sciences often require the integration of terabyte or gigabyte distributed scientific databases. Genetic algorithms, radar range images, materials simulations, and atmospheric and oceanographic measurements generate large multidimensional multivariate data sets. The varied data come with different data geometries, sampling rates, and error characteristics. The display and interpretation of the data sets employ statistical analyses and other techniques in conjunction with visualization. Information visualization also includes visualizing information retrieved from large document collections (such as digital libraries), the World Wide Web, and text databases. This information is completely abstract, so the data must be mapped into a physical space representing the relationships contained in the information as accurately and efficiently as possible. An individual parallel coordinate axis represents a 1D projection of the data set. Thus, separation between or among sets of data on one axis represents a view of the data of isolated clusters.

II. APPLICATION OF PARALLEL CO-ORDINATES

The multivariate dataset is an *N*-dimensional set **E** with elements described by $e_i = (x_{i1}, x_{i2}... x_{in})$. Each observation x_{ij} is usually independent of the other observations and, observations, in nature, may be discrete or continuous, or may take nominal values. Several techniques have been proposed for multivariate data representation. In this paper we describe the multidimensional visualization technique called Parallel Coordinates.

The idea for this technique comes from multi-dimensional geometry. The 2D or 3D Descartes coordinate systems couldn't solve the multidimensional problems. Inserberg proposed the use of a multidimensional coordinate system based on Parallel Coordinates. N copies of the real line R (labeled x_1, x_2, ..., x_n) are placed equidistant and perpendicular to the X-axis(Figure 1). They correspond to the axes of parallel Coordinates system that represents the N-dimensional space [1]. All axes have the same positive orientation as the y-axis. The complete polygonal line C={C_1, C_2, … , C_n} is represented by the segments between the axes. In this way, a 1-1 correspondence is established between points in the plane and planar polygonal lines in Parallel coordinate system. This is the efficient way to place a large number of axes and to visualize the multivariate relations. In the early 90's, Parallel Coordinates was used as a two-dimensional technique for multidimensional data sets representation [2]. The technique has been enhanced during the next few years. Researchers have been working on improving this technique for better data investigation and easier user-friendly interaction by adding data clustering [], brushing [], etc. With these improvements, Parallel Coordinates becomes very efficient technique for visualization relationships between designated neighbouring dimensions.

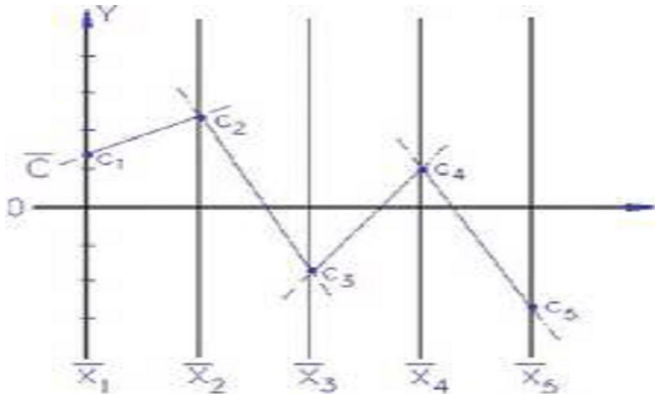

Figure 1. Representation of a lineC in parallel coordinate system C= {C_1, C_2, C_3, C_4, C_5}

Parallel coordinates can be used to show multidimensional points that reside at the same poly line which intersect at specific points between the vertical axes. It is very useful in preventing collisions such those in the air traffic and anytime where positive or negative correlations can be assumed. The first and more widespread application of Parallel Coordinates is Exploratory Data Analysis (EDA) for discovering of data subset relations. If the dataset have M items the subsets may be one of the $2^{M.}$ Our eyes can discriminate in a good data representation and navigate the discovery process. Good representation of datasets with M variables should preserve information and give good results for M rather any number of variables. Parallel coordinates transform multivariate relations into 2D patterns. These patterns are suitable for analysis and data exploration-searching for a clue in many dimensions. Even there are usually developed specialized queries to find patterns, they still could not handle all encountered situations. The requirement for successful exploratory data analysis needs to have an informative representation without loss of data, good choice of queries and skilful interaction with the display. The Parallel Coordinates technique can be considered as a generalization of two-dimensional Cartesian technique. The axes in Parallel Coordinates are drawn parallel to each other. We can draw as many axes as we want, so we can represent the points of dimensionality larger than three. Instead of using a "dot" to represent the location, a "line" is used to connect the coordinates of the point on the axes. In this way the points become lines. In Parallel Coordinates plots, the dual points are lines and the dual lines are points. A point in the Cartesian coordinates becomes a line in Parallel Coordinates. In general the point conic in Cartesian coordinates becomes a line conic in Parallel Coordinates. Also, the rotation in the Cartesian becomes Translation in Parallel Coordinates. Parallel Coordinates representations can provide statistical data interpretations. In the statistical setting, the following interpretations can be made: For highly negative correlated pairs, the dual line segments in Parallel Coordinates tend to cross near a single point between the two Parallel Coordinates axes. Parallel or almost parallel lines between axes indicate positive correlation between variables. In this way, most common objectives to this technique are to represent the dependency on the order to the axes to identify the relations between variables.

III. RESULTS AND DISCUSSION

There are only few notable software publicly available to convert databases into parallel coordinates graphics. Notable software are ELKI, GGobi, Macrofocus High-D, Mondrian, and ROOT. Libraries include Protovis.js, D3.js provide basic examples, while more complex examples are also available. D3.Parcoords.js (a D3-based library) and Macrofocus High-D API (a Java library) specifically dedicated to ||-coords graphic creation have also been published. We have used the software XDAT version 2.1 to generate the parallel coordinates from a dataset. It's easy to generate parallel coordinate in xdat from a given table. We just insert the table from a notepad and the parallel coordinate will be generated from that. Figure 1 shows such a table where we put some random headings and generate a parallel coordinate from that.

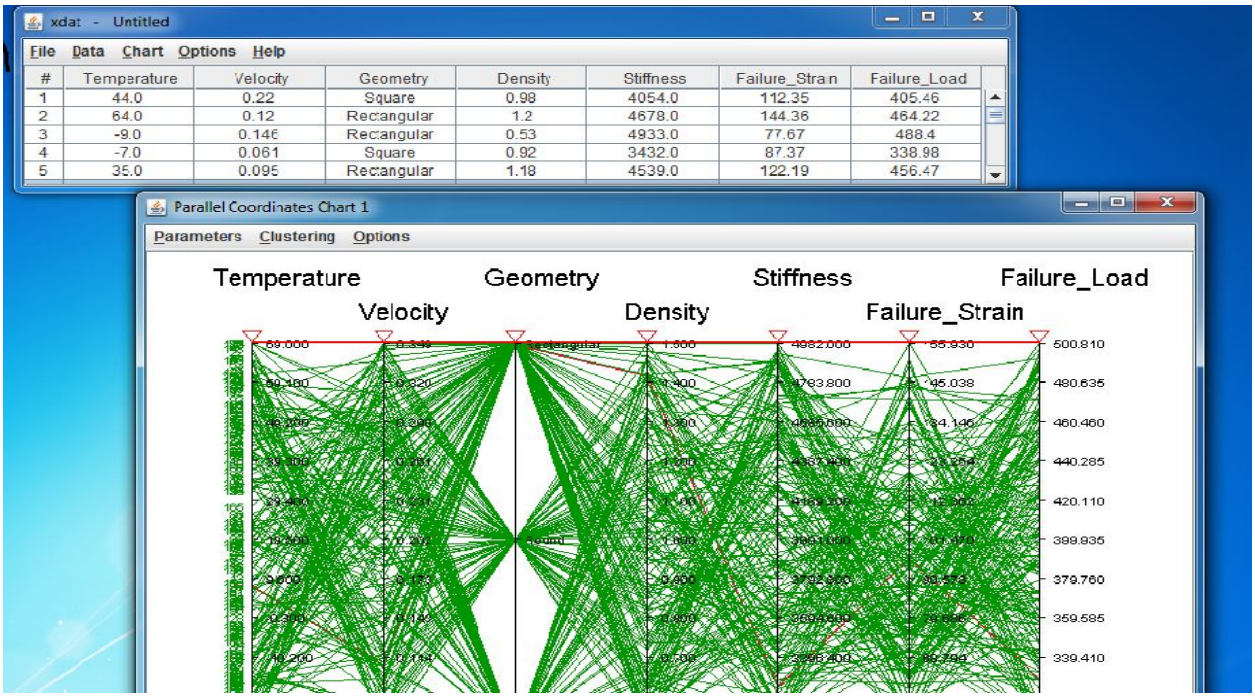

Figure 2. Generation of parallel coordinate from a data set.

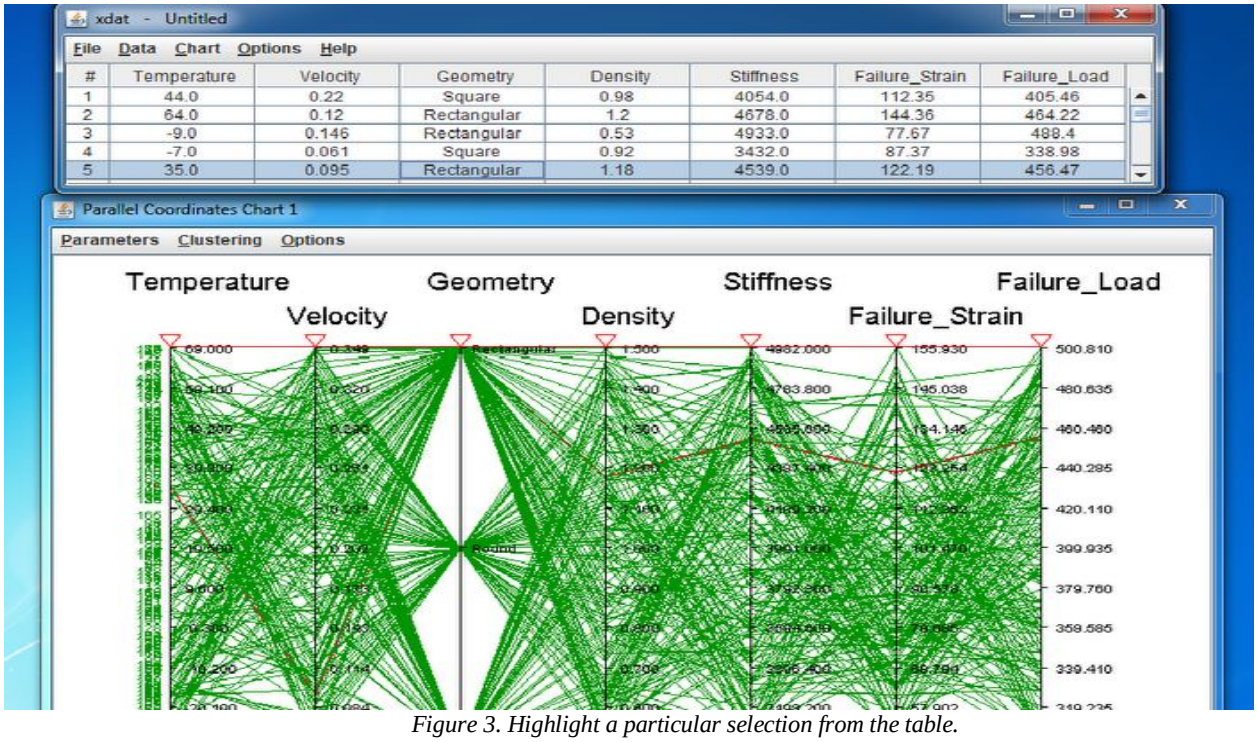

Figure 3. Highlight a particular selection from the table.

In the above figure we have selected a tuple from the table and the red line in the graph is highlighting our selection. In the above figure the result of semesters in which total no of students appeared, students who got first class, and students who got 2nd class are represented. A particular year 2003 is selected which is highlighted using a red line.

To represent a large data set and showing a particular tuple from the table is done using this software. But the problem with large data set is the graph become so congested that any data retrieval is quite difficult from there. To overcome this problem BRUSHING is applied to the graph representation.

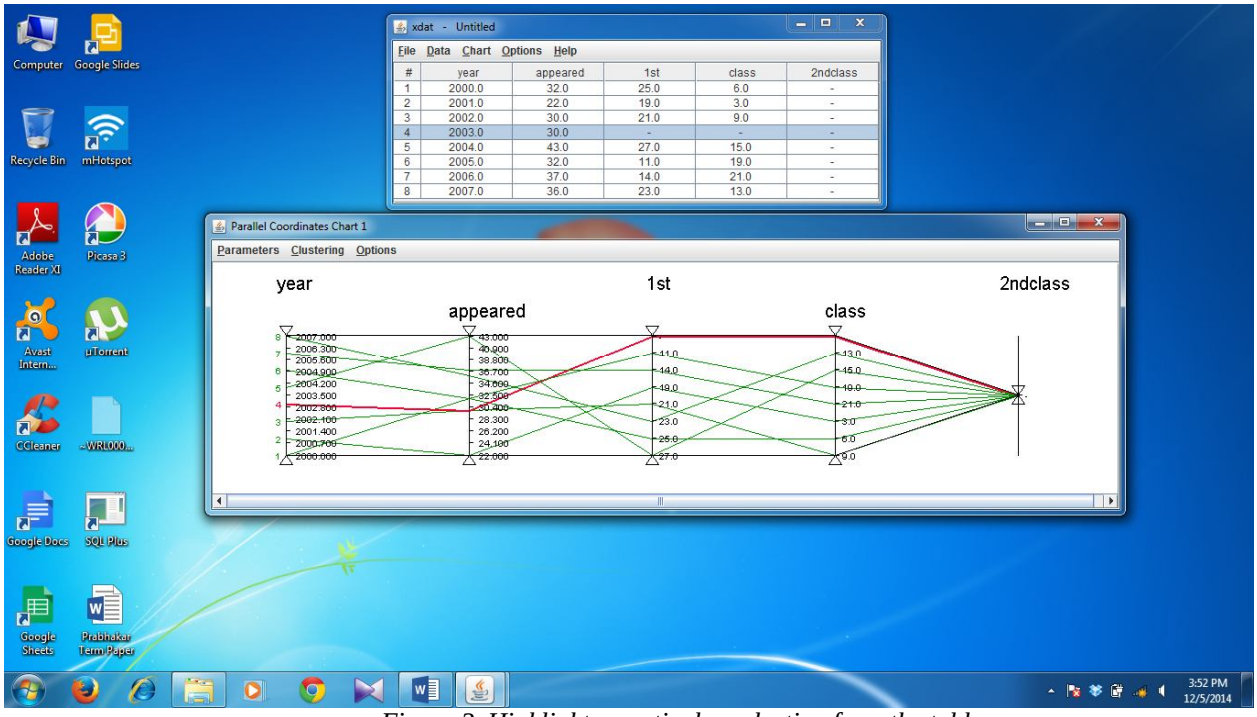

Figure 3. Highlight a particular selection from the table.

IV. ADVANTAGES AND DISADVANTAGES

The main advantage of parallel coordinates technique is that the number of data dimensions is restricted only by the horizontal resolution of the screen, but as the axes get closer it may become more difficult to perceive structures or data relations. Another advantage is that the correlations between variables in the dataset can be spotted easily.

1. A reduction of the amount of useful information when the level of disorder is present in the visualization is one disadvantage of the technique. As for many visualization techniques, readability as well as efficiency of Parallel Coordinates plots suffer when large datasets are displayed. Novotny [Nov04] names three major problems which high density displays like Parallel Coordinates confronted with:

- A loss of speed and interaction that is noticeable when displays become more populated
- Occlusion may happen, means that two or more visual elements overlap, making it impossible to recognize single items.
- Having high density displays, the viewer is confronted with aggregation in a way that objects are drawn over each other. The actual number of objects then cannot be perceived any more.

2. In parallel coordinates, each axis can have at most two neighboring axes (one on the left, and one on the right). For a d-dimensional data set, at most d-1 relationships can be shown at a time. In time series visualization, there exists a natural predecessor and successor; therefore in this special case, there exists a preferred arrangement. However when the axes do not have a unique order, finding a good axis arrangement requires the use of heuristics and experimentation. In order to explore more complex relationships, axes must be reordered.

V. CONCLUSION AND FUTURE SCOPE

The effective exploratory data analysis of great volume of multidimensional data demands some specific data visualization techniques for analysing a complex and huge databases. Because of the possibility to pose many axes in 2-D surface and some interactive technique, Parallel Coordinate is a very useful technique for exploratory data analysis. In the present paper the authors have mainly described the working principle of parallel coordinate. The generation of parallel coordinate is done using the software XDAT version 2.0. But the generation of parallel co-ordinates and the retrieval of data elements from the chart can be done in an efficient way using MATLAB. The authors are presently working in presenting the parallel coordinates using MATLAB.

ACKNOWLEDGMENT

The authors are very much thankful to Prof. Shalabh Agarwal, Head, Department of Computer Science, St. Xavier's College(Autonomous), Kolkata for giving inspiration to research work in the area of parallel co-ordinates.

REFERENCES

[1] Zudilova-Seinstra E., Adriaansen T., Van Liere R., Trends in interactive Visualization, Springer, 2009-09-12
[2] Brandst¨atter A., Visualization of Online Sales Databases, Wien, 02-2007
[3] [AWS92] C. Ahlberg, C. Williamson, and B. Shneiderman, "Dynamic queries for information exploration: an Implementation and evaluation", in: *Proceedings ACM CHI'92*, ACM Press, New York, 1992, pp. 619-626
[4] [B83] J. Bertin, *Semiology of Graphics. Diagrams, Networks, Maps*, The University of Wisconsin Press, Madison, 1983
[5] A. Waddell and R. Oldford. Visual clustering of high-dimensional data by navigating low-dimensional space. 58th Congress of the International Statistical Institute, STS 57. Dublin, Ireland, 2011.
[6] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In Proceedings of: IEEE Symposium on Information Visualization, pages 157–164, 2005.
[7] Solka J.L., Marchette J.D., Adams M.L., Applications of Statistical Visualization to Computer security, NSWC, DTD, 2002
[8] Rundensteiner E.,Ward M., Xie Z., Cui Q, Wad C, Yang D, Huang S., XmdvTool: Quality-Aware Interactive Data Exploration, ACM 978-1-59593-686-8/07/0006, (http://davis.wpi.edu/~xmdv/)
[9] Parallel coordinate graphics using MATLAB - http://isomap.stanford.edu/IsomapR1.tar
[10] GGoby (http://www.ggobi.org)
[11] http://www.cs.uta.fi/~hs/pce/
[12] Savoska S., Loskovska S., Dimitrovski I., Information Visualization from the Public Utilities Databases of Local Municipality for Municipalities Managers, Proceedings, ITI, Cavtat, 2008
[13] Hauser H. , Ledermann F., Doleisch H., Angular Brushing of Extended Parallel Coordinates, http://www.VRVis.at/vis/
[14] M. Novotn´y, "Visually effective information visualization of largedata," in Proceedings of Central European Seminar on Computer Graphics (CESCG), 2004
[15] http://vis.lbl.gov/Events/SC05/Drosophilia/index.html (Drosophila Gene Expression Data Exploration and Visualization)
[16] J. LeBlanc, M.Ward, and N.Wittels. Exploring n-dimensional databases. *Proc. of Visualization '90, p. 230-7*, 1990.
[17] H. Lee and H. Ong. Visualization support for data mining. *IEEE Expert Vol. 11(5), p. 69-75*, 1996.
[18] Y. Leung and M. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *ACM Transactions on Computer-Human Interaction Vol. 1(2), June 1994, p. 126-160*, 1994.
[19] A. Martin and M. Ward. High dimensional brushing for interactive exploration of multivariate data. *Proc. of Visualization '95, p. 271-8*, 1995.
[20] A. Mead. Review of the development of multidimensional scaling methods. *The Statistician, Vol. 33, p. 27-35*, 1992.
[21] G. Nielson, B. Shriver, and L. Rosenblum. *Visualization in Scientific Computing*. IEEE Computer Society Press, 1990.
[22] R. Rao and S. Card. Exploring large tables with the table lens. *Proc. of ACM CHI'95 Conference on Human Factors in Computing Systems, Vol. 2, p. 403-4*, 1995.
[23] W. Ribarsky, E. Ayers, J. Eble, and S. Mukherjea. Glyphmaker: Creating customized visualization of complex data. *IEEE Computer, Vol. 27(7), p. 57-64*, 1994.
[24] B. Shneiderman. Tree visualization with tree-maps: A 2d space-filling approach. *ACM Transactions on Graphics, Vol. 11(1), p. 92-99*, Jan. 1992.
[25] M. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. *Proc. of Visualization '94, p. 326-33*, 1994.
[26] M. Ward and K. Theroux. Perceptual benchmarking for multivariate data visualization. *Proc. Dagstuhl Seminar on Scientific Visualization*, 1997.
[27] E. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association, Vol. 411(85), p. 664*, 1990.
28] E.Wegman and Q. Luo. High dimensional clustering using parallel coordinates and the grand tour. *Computing Science and Statistics, Vol. 28, p. 361-8.*, 1997.
[29] S. Weinberg. An introduction to multidimensional scaling. *Measurement and evaluation in counseling and development, Vol. 24, p. 12-36*, 1991.
[30] G. Wills. An interactive view for hierarchical clustering. *Proc. of Information Visualization '98, p. 26-31*, 1998.
[31] INSELBERG A., DIMSDALE B.: Parallel coordinates: a tool for visualizing multi –dimensional geometry. [Nov04] M. Novotny. Visually Effective Information Visualization of Large Data. In 8th Central European Seminar on Computer Graphics (CESCG 2004), Proceedings, 2004.

