INTERACTIVE INFORMATION VISUALIZATION OF ENTITY-RELATIONSHIP-DATA

Thomas Rieger, Francesca Taponecco

Technische Universität Darmstadt, Interactive Graphics Systems Group, 3D Graphics Computing
Rundeturmstraße 6, 64283 Darmstadt, Germany
Email {rieger, ftapone}@gris.informatik.tu-darmstadt.de
http://www.gris.informatik.tu-darmstadt.de

ABSTRACT

This work presents methods for interactive analysis of unknown data and relations. It supports the selection of relevant parts of the data and the visualization of relations using different layout strategies for 2d visualizations and for 3d visualizations. Moreover, the user can select additional mappings to display extra information, with the help of supplementary visual variables in an easy way. Hereby the user can interactively create new visualizations and new sights and sift through the data in order to reveal discrepancies, to uncover trends or to provide vital operational insights.

Keywords: Data Mining, KDDB, DBMS, Entity-Relationship-Data, Spiral, Layout, Mapper, 3d Space

1. INTRODUCTION

The visualization of relational networks becomes ever more important in the context of *DATA MINING* and the search of information in data bases (*Knowledge Discovery in Databases, KDDB*).
A special area of application of this technique is the analysis of complex networks, which results in the area of investigating organized crime. The fight against organized criminality often requires extensive relational networks to be analyzed. Many objects, as persons, events, places, cars, telephones or things such as weapons, drugs, and thief property output information are usually linked together and characterized by having various relations. In this context two-dimensional graphs have been used for quite a while.
A problem of such 2d visualizations is the limited amount of information, which can be represented at the same time on the display area. Large data records, as they typically result in this area of application, can be hardly presented completely in this way. Therefore, a general attempt to produce diagram reductions is to offer functions, in the 2d-visualization, like fading out, aggregations (normally coupled with abstractions) and selection. For a first outline of the original data, an automatic dissolution reduction is necessary [HEDM98].
An elegant way to enlarge the representation area is achieved by including a progressive rate of information along the third dimension in the information and graph visualization.
Thereby, an appropriate diagram properly addresses the spatial perception abilities of a viewer. Generally, such a representational form must be also combined with appropriate navigation and interaction possibilities. Here, 3d navigation represents the largest problem. In most cases, at least in standard Desktop systems, 3d navigation has to be performed using 2d input devices. In this report the concepts and the relative solutions are described; as a result the importance of interactive two-dimensional and three-dimensional visualization it is pointed out. With this intention, the fundamental aspects of information visualization are summarized and introduced. The state of the art within the area of information visualization is described. On the basis of selected exemplary scenarios, problems are taken on consideration and solved by means of analysis and visualization of relational networks.
The implementation is based on Java and Java-3D. The individual steps are described to product interactive two-dimensional and three-dimensional information visualizations and the different aspects of such an interactive analysis are highlighted.
At the end, the obtained results are combined and further work in this context is addressed.

2. STATE OF THE ART

Within the area of data visualization, a multiplicity of techniques is already present in the literature. In the 2d area the most popular techniques are the pie charts, bar charts, line diagrams and point diagrams. These methods result to be mostly inappropriate in order to represent large quantities of data. They neither can point out relations. Consequently, in order to highlight relative dependencies, adapted techniques must be used for the 2d representation as well as for the 3d representation of the data. Special layout techniques (see VISUALIZATION) offer great advantages to relate the data. For the representation of cyclic and periodic data the use of spirals is very useful (see Spirals). Within the area of information visualization, a possibility to represent large quantities of information on the display is realized by visualizing it in a 3d space. Thus utilizing in particular bar charts and line graphs, the third dimension is used for the representation of additional data in a common graph (see [CAMS99]). In a similar form several 2d visualization techniques were transferred into the 3d space as well. Moreover, new representational forms, based on coordinate graphics were developed. A significant example is the so-called Wormplots, which, functioning as an indirect visualization technique, allows a temporal periodical comparison between different populations (see [MARO97]).

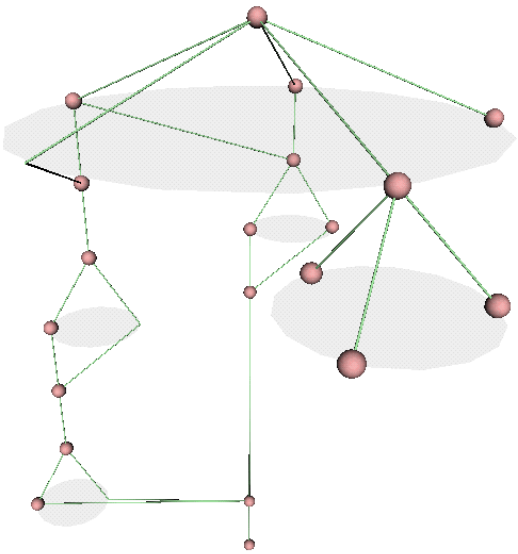

Figure 1 Graph structure in vrml

Besides the visualization of ordinal and quantitative data in 3d, also relational data have been successfully visualized in the three-dimensional space. Roberston et Al. presented a special kind of hierarchical structures´ visualization by utilizing the *Cone Trees*, as a result particular techniques are used for the explanation of hierarchic levels (see [ROMC91]). The offered examples prove that the inclusion of the third dimension can represent a valuable supplement of the current standard visualization techniques. Especially the number of information, which is at the same time representable, can be in this way increased. At the same time, navigation methods in the 3d space can be used as a natural metaphor to investigate the whole data record and not just a restricted range of it.

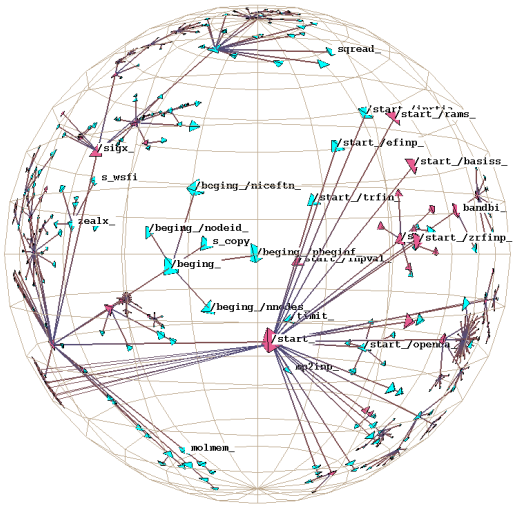

Figure 2 Graph structure of the Hyperbolic Viewer

The presented techniques use thereby specifically adapted navigation and interaction procedures. So it is recognized that there are so far no other interaction techniques (for the 3d navigation) able to offer an equal efficiency and generality. These have again to be evaluated in every case depending on the specific application.

3. SYSTEM

As more and more corporate or institutional data are being computerized, intelligent and flexible methods for the analysis of such data are needed to enhance decision-support processes. Most of the potentially useful and valuable information in the data is usually hidden and not available to the user. This also happens for the most structured data in the growing number of databases that are used in different application areas. For instance Database Management Systems (DBMS) gave access to the stored data but do not provide any means to analyze the data.

Standard data analysis techniques (e.g. conventional numerical and statistical analysis) are not sufficient to solve the complex problem of '*discovering information*'. The reasons are the following:

- they have been developed just for standard applications intended for specific scientific or business applications

- they usually lack of appropriate methods that could deal with large quantities of data, complex data structures or heterogeneous data.
- they generally do not support the analysis of noisy, incomplete or erroneous data.

Information visualization techniques, which are commonly applied to large data structures of scientific data, show how efficiently graphical interactive methods for visual data analysis can be utilized in this context. This system represents an interactive tool exactly targeted for the application domains *'Data Mining'* and *'Knowledge Discovery in Databases'* (KDD).

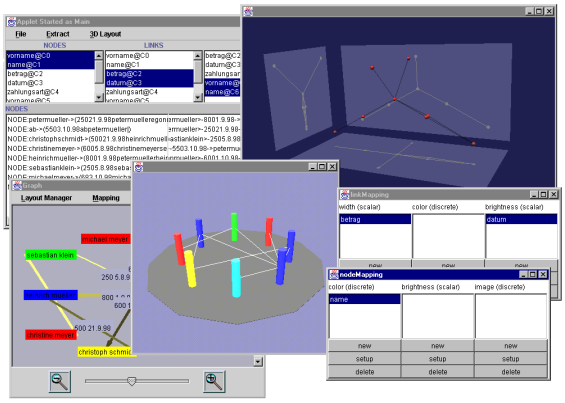

Figure 3 Screenshot of the Application

Providing selected information visualization techniques and flexible means to their combination, this application extends the repertoire of standard Data Mining tools. Since Knowledge and Information is usually the composite of not only a large number of data but also especially the relationships amongst themselves and between the data and a context, this system allows the interactive analysis of unknown data and the relations. A central element of the system is the interactive data selection panel, which allows for picking relevant parts of the data for the visualization and for interactively determining the objects and relations of observation. To visualize the relations, several layout strategies are available: i.e. a 2d hierarchical layout, a 2d circular layout, and a 2d and 3d constrained layout (that considers further visualization goals, such as limited edge crossings and context based neighborhood relations). Additional visualization techniques for qualitative and quantitative data extend the visualization spectrum in the system. In case of multi-parameter data, extra information can be displayed by mapping corresponding information on free visual variables such as color, transparency, size etc. Furthermore, conclusions and hypothesis regarding the data structure can be directly added to the data by defining new relations. Hereby the user can interactively create new visualizations and new views of the data and sift through the data to reveal discrepancies, to uncover trends, or to provide vital operational insights. This application is completely implemented in Java and therefore available on a large number of workstation systems.

VISUALIZATION OF ENTITY-RELATIONSHIP-DATA

The data are available as ASCII-tables. Different attributes (e.g.: name, first name, age, bank account, insurance...) can be selected, and their dependencies and connections be represented from a table. Hereby each object can be selected as a node and as a relationship between the nodes. The data type or the context, in which the data are observed, decides on the sort of representation. Both the 2d visualization as well as the 3d visualization structures are based on the dynamic extraction of temporary nodes and edges, selecting from objects attributes. Therefore, the resulting graphs can be in this way generated and visualized in the 2d and/or in the 3d space.

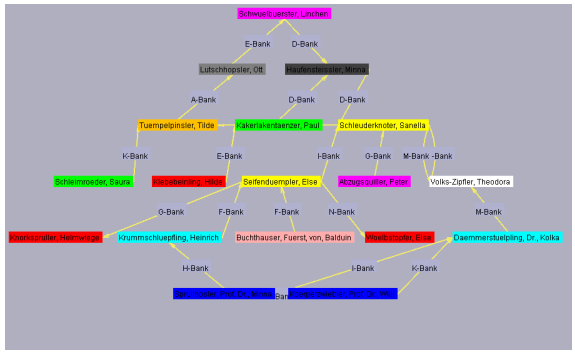

Figure 4 Representation of hierarchical relations.

The extraction of the temporary graphs is closely tied up as elementary constituent of the dynamic calculation of the data to the further visualization steps. Now, as a user chooses a specific layout, this calls (searching through the prepared data) the relative data that it requires. By using Mappers, the look of the entities´ attributes or relations are illustrated by relatively acting on the nodes´ or edges´ characteristics. To reach this scope, a Mapper offers several possibilities according to the context of visualization. The layout chart likewise depends on the context of visualization.

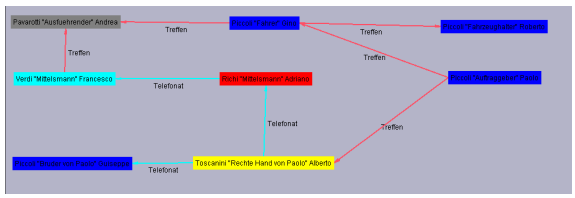

Figure 5: Example for Raster Layout

The system offers the following layout types:

- Circle Layout (see Figure 7)
- Hierarchical Layout (see Figure 4)
- Group separation Layout (see Figure 6)
- Raster Layout (see Figure 5)
- Spirals (see Spirals)

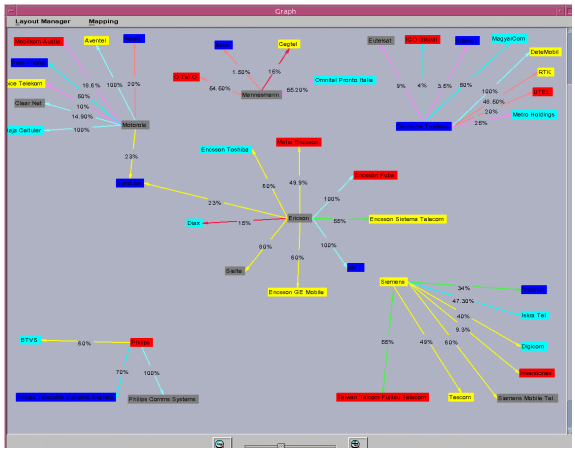

Figure 6 Subdivided occurring groups.

TRANSFER FROM 2D TO 3D REPRESENTATION BY THE EXAMPLE OF CIRCULAR CHART

In case of large data quantities, the circular chart offers a complete overview of all the data at one time. Therefore, the user gets a first impression of the special features of the data. The nodes are located on a circle. The edges run by the set, thus edge concentrations become clear. The user can now easily detect nodes with many edges, and move these out of the set (see Figure 7).

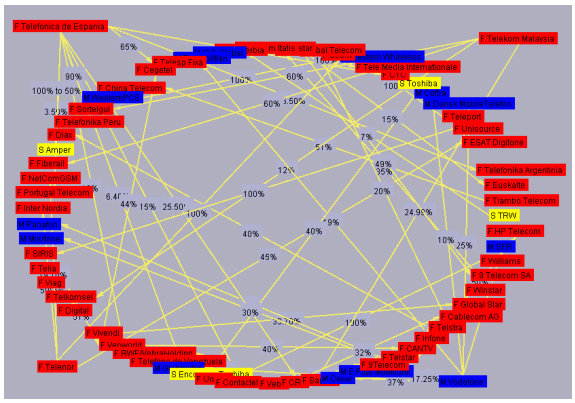

Figure 7 The Circle Layout offers the best overview with large data records.

The representational form of a circular chart in 3 dimensions essentially corresponds to the same one in 2dimension (see e.g. [WEBL98]). This new view is realized by moving the previous 2d graph in the 3d space, i.e. positioning the 2d graph on a level, here represented by an underneath large disk (see Figure 9). As basis serves here the 2d Layouter already introduced, which was accordingly extended.

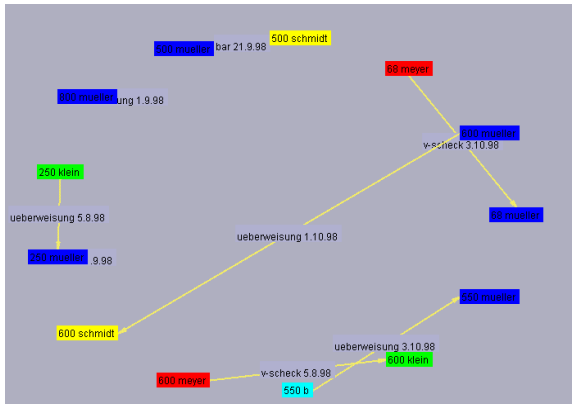

Figure 8 Circle Layout in 2d

Exactly as done for the 2d chart, here different Mappers again can illustrate the attributes. In particular the graphic attributes: node color, node form, edge color and edge thickness.

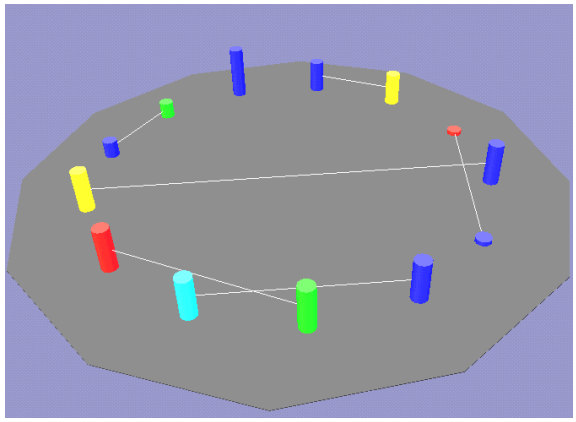

Figure 9 Circle Layout in 3d

In this representational form, figures results to be less suitable what concerns transparency and also texture mapping. Anyway a valuable additional attribute is the height of node items, which can be used to represent quantitative node attributes. Moreover, differently than what happens for the visual attribute "size" in the 2d representation, a variation of this item does not change the visibility of other items. So an advantage is represented by having a simpler layout algorithm that do not need to consider anymore complicate consideration of

variable node sizes. Also in the edges additional possibilities for attributes visualization are offered. Semantically different edges can be distinguished, besides using different colors, by having a different initial height, starting from a basic level or the base disk. The graph shows an example of a simple bank transaction as 3d chart (see Figure 9) opposite a 2d chart (see Figure 8) of the same data record.

SPIRALS

As previously introduced, the described system offers several layouts, in order to give the largest and best choice of possible visualization methods.
Every time an appropriate layout is selected according to the kinds of data that have to be visualized. “Spirals” is one of the available options and permits a different approach in visualizing information (see Figure 13).
The target is again the representation of a large amount of data.
In particular, this tool offers the possibility of giving a practical view of significant aspects of the data; nevertheless it permits a very compact representation of the data, which is obtained by utilizing all the space that is available on a display device.

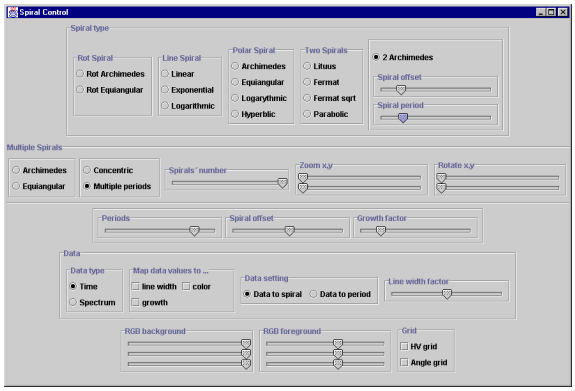

Figure 10: Control commands to vary the spiral characteristics.

We say in advance that the only disadvantage that we register in using such a representation could be given by the lost of one coordinate (i.e. the information that is registered in the y coordinate in a typical Cartesian system). That means that, with the intent of achieving our targets in optimizing the visualization of some kinds of information, arises the risk of losing the perception of the dimension quality of the data in favor of the data quantity.
Nevertheless, we explain how we can cope with this problem and resolve it.
This can be realized by means of the following methods:

- introducing a chart where the significant information can be visualized.
- registering the entity of data by using radials axis. In this case the resolution that can characterize the information is related to the step that regulates the distance between the windings of the spiral. As a result, we can use zoom tools, in case we need to diminish this step to increase the length of the spiraling axes and so the number of information.
- transferring the whole spiraling system in a three-dimensional environment, so that we save an axis where it is again possible to register the quantity information related to our data.

We clarify now when and how this layout brings considerable advantages. A spiraling system can be used in various cases of interest, especially in the following ones:

- Great quantity of general data
- Serial data
- Periodical/cyclic data
- Multivariate/multidimensional data

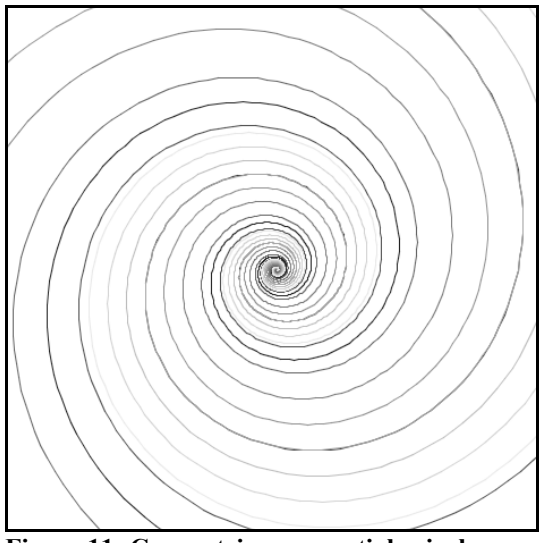

Figure 11: Concentric exponential spirals.

We observe how this kind of representation is able to offer several advantages in each of the four fields cited before.
As previously explained, the need of visualizing large amounts of data becomes day by day more important. For this reason, a spiraling system offers

the most compact representation, as the polar system is characterized by having a spiraling axis around the origin. Thus, the information can be visualized in the whole available graph, instead of just laying in a Cartesian coordinate system.

Consequently also large databases, scientific data, etc. can be efficiently stored in such a system.
This also can permit (by means of distance functions) to represent the entities depending on their distance to a reference point (typically the spiral origin) and to define regions of interest.
It is anyway observing the other three categories, that the Spiral layout is able to offer the greatest advantages.
Principally in these cases in fact, a spiraling system can be adopted to usefully highlight existing relations among the data.
This, besides being useful, also offers an easy overview and recognition of information to the user.

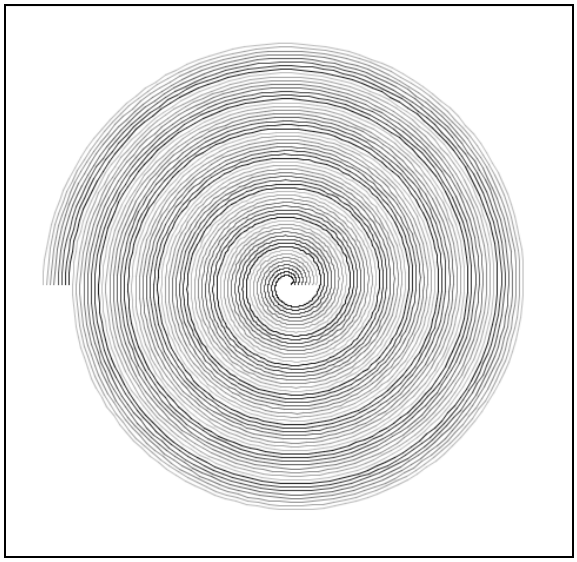

Figure 12: Multiple Archimedean spiral.

Consequently, Serial data, together with Periodical data, are the most important group (see [WEAM01]).
Almost every kind of data is characterized by having temporal information, thus a time dimension has to be considered.
Also cyclic behavior can be often found in nature (heartbeats, audio signals, chemistry, etc.) but also simply the seasons in a year or repetitive dates in a week).
The high dimensionality of the variables on which the data depend are also an important aspect to consider: for example the human heart beat is related to the age, sport activity, time in the day, weather conditions, etc.
As a result, this information can be now concentrated in a limited space, not utilizing many graphs anymore, but using simple and efficient methods (e.g. concentric spirals, see Figure 11, Figure 12).

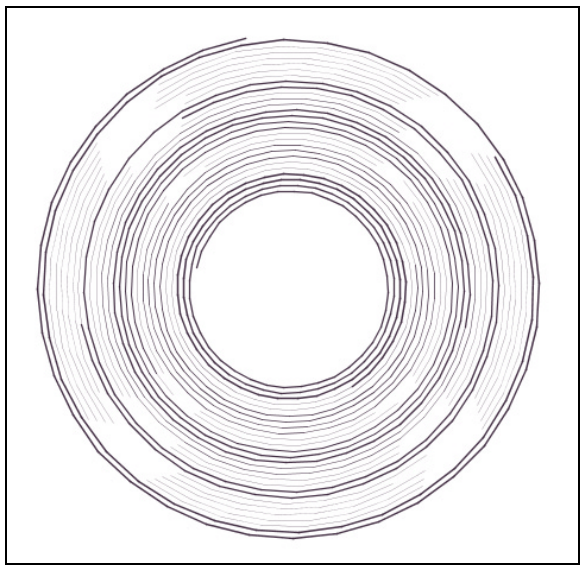

Figure 13: Representing data on the spiral.

An interesting aspect is that data often belong to more than one category, as generally they are characterized by having several relations and by being depending on many factors.
In conclusion, this tool is thought to permit improved data visualization, not only with regard to the amounts of information that can be displayed, but also highlighting relations existing between the entities. With the help of colors, different kind of spirals and interactive tools a user can easily recognize and analyze relevant data characteristics (see Figure 10).
We think that this can offer a valid method for information visualization and analysis. Consequently, as a future work, we consider to improve it by giving again more generality, analyzing the most possible different data categories (in astronomy, science, economy, etc.).

3D VISUALIZATION OF ENTITY-RELATIONSHIP-DATA

The idea of the technique for the graphs visualization in the three-dimensional space, used here, is based on the following request: different data with similar characteristics should be detected fast and surely. Spatial proximity seems to be a promising concept for the user, in order to arrange data similarity. By similarity we understand here the distance of nodes, which represents the selected relations in the graph. If two nodes of the graph are directly connected, then the distance of 1 is assigned to them. If two nodes are indirectly, i.e. connected by a third node

together, then they have the distance of 2. Generally the unitary distance names the minimum number of edges in a path between two nodes in a graph. The scenario can also be generalized, occupying the edges with 'weights'. The distance of two nodes corresponds then to the total sum of weights on an optimal path (a path, on which the total sum of the edge weights is insignificant).
We give now a formal explanation of the described concept: we define a graph $G(V,E)$, where $V = (v0, v1,...,vn)$ is the node quantity and E the edge quantity.

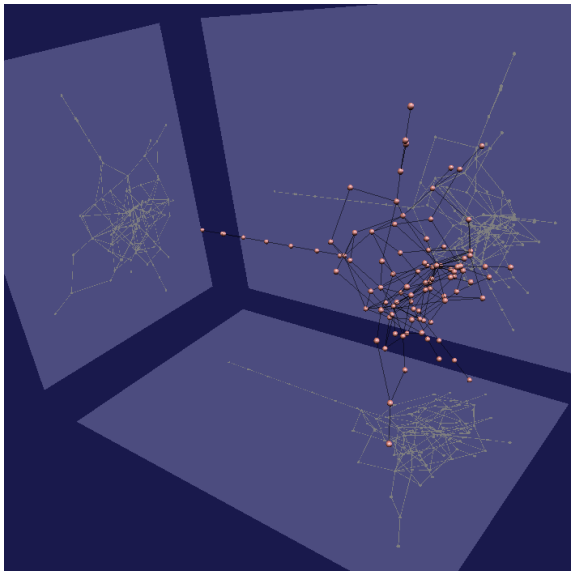

Figure 14: 3d View of Entity-Relationship-Data

Every node is identified by a natural number, while an edge result defined by pairs of natural numbers. Additionally, the edges can contain a real weight. Target of the 3d graph visualization is to find proper positions for the nodes, so that the distances between two nodes correspond to a distance in the graph. At first we need the distance between every two node. These distances are entered into a spacer matrix A:
$A = (a_{ij})$, $a_{ij} = d(v_i,v_j)$.
Let consider the algorithm of Kruskal (see e.g. [Pana99]): it is used for the determination of a minimum stretching tree or in case of unweighted graphs with simple width search. Utilizing this algorithm, a complete column of the matrix is obtained. The problem consists now on determining the nodes´ positions, starting from the given spacer matrix. First it is marked that the distances of the nodes in the three-dimensional space can only approximate the distances in the graph.
Now, starting from the given spacer matrix, it is necessary to find a good approximation of three-dimensional coordinates. We avail ourselves of a multivariate statistics´ technique: the Multi-dimensionally scaling. By means of it, the data dimensionality is reduced, in order to keep the distances in the original dimension, if possible according to the maximal value.
So, exist methods, which do not need the coordinates in the original dimension at all. Everything that is needed is a matrix of the distances of the data values. Then, from this matrix, concrete coordinates can be got and used in a k-dimensional space (for k smaller than half the matrix dimension). The basic idea is to obtain (from the spacer matrix A) an interior product matrix B of the original coordinates $X = [x_1, x_{2, ...} x_n] \rightarrow (B = X X^T)$. Using the Euclidean standard on $x_{i,}$ shaping effects are achieved:

$$b_{r,s} = \frac{1}{2} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} a_{i,s}^2 + \frac{1}{n} \sum_{j=1}^{n} a_{r,j}^2 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j}^2 - a_{r,s}^2 \right)$$

Thus, the interior product matrix B is defined. As the interior product matrix B is symmetrically and positively semi definite, it results to have n positive eigenvalues. The matrix of the eigenvalues E is sorted along the diagonal according to its quantity $e1,...,ep,...,en$. If V furthermore is the matrix of the accordingly sorted eigenvectors, then is $B = V E V^T$. An approximation B' using the positive or a section of the positive eigenvalues (in a matrix E') and the appropriate eigenvectors (V') results in:
$B' = V' E' V' T = V ' E'^{1/2} E'^{1/2} V'^T$.
Being $B = X X^T$, the approximation is achieved in a small dimension p as $X ' = V ' E'^{1/2}$.

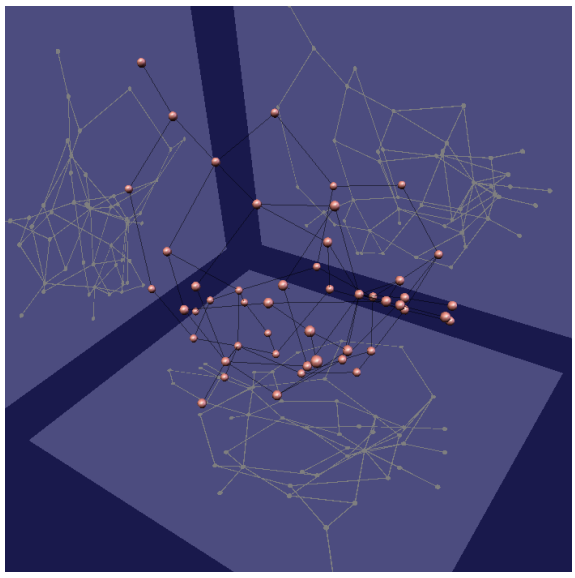

Figure 15: This Figure shows an example one in this form visualized relations network.

The use of the largest positive eigenvalues supplies the best approximation of the distances in lower dimension. In the case, the 3d graph visualization is related to the 3 largest eigenvalues.

This concept is to be used also for the 2d graph visualization. Anyway the appropriate adjustment of the layout algorithm in 2d is still pending.
To get a better perceptibility of the spatial arrangement in this representational form, additionally shadow-walls are used. They represent the appropriate projection of the graph into the 2d space. Detailed information about individual nodes is callable by user interactions. By clicking on a node or an edge, the relative information is visualized in an information border.
This representational form can be further optimized by different techniques.

- The potentiality of extra mono ocular visual factors can be used in this context.
- In order to improve depth perception, the use of depth sharpness and area selection are here offered for visualization.
- Further interaction techniques can complete the representational form, and particularly the interaction with the node that is based on the special representation on the 2d shadow walls, which provides additional 2d information.
- Also *Brushing* techniques are meaningful, as they permit a selection and a *Highlighting* of data elements in the 3d space.
- Furthermore a simultaneous selection of the appropriate items into the active 2d representational forms is possible.

4. CONCLUSION

In this report the fundamental aspects of information visualization were presented and discussed.
In the context of the analysis of large quantities of data, promising results were presented. They increase the packing density of visual representation and supply natural and intuitive metaphors for navigation in these quantities of data.
Thereby, these are able to complete the conventional representation methods.
The 3d representation, based on the presented algorithm, represents a new technique for the presentation of multivariate relational data.
It appears as generally applicable to the recognition of structures and symmetries in the relations' networks, in the most flexible way.
Complex relations networks are representable in connection to additional possibilities, so that the information is mapped on attributes such as color and texture. The visualization of cyclic data on a spiral enables faster highlights and symmetries that result being easier to detect.

5. FUTURE WORK

Further work concerns the analysis and qualitative evaluation of 2d and 3d visualization techniques and their application type for information visualization. The most interesting possibilities are situated in the area of visualization with spirals. Also the advancement of fundamental 3d visualization, navigation and interaction techniques is intended for the improved use of 3d techniques in information visualization.

6. BIBLIOGRAPHY

[CAMS99] Card, Stuart K.; Mackinlay, Jock D.; Shneiderman, B.: Readings in Information Visualization, Morgan Kaufman Pub., San Francisco, 1999

[HEDM98] Herman, I.; Delest, M.; Melancon, G.: Tree Visualization and Navigation Clues for Information Visualization, Computer Graphics Forum, Vol. 17, No. 2, Juni 1998, pp. 153-165

[MARO97] Matthews, Geoffrey; Roze, Mike: Wormplots, Computer Graphics & Applications, Vol 17, No. 6, Nov./Dez. 1997, pp 17-20

[PANA99] Panagiotis, Papaioannou: Kruskal's Algorithm, http://students.ceid.upatras.gr/~papagel/project/kruskal.htm, 1999

[ROMC91] Robertson, G.G., Mackinlay, J.D., und Card, S.K.: *Cone Trees: Animated 3D* Visualizations *of Hierarchical Information*, Proc. of ACM Conf. on Human Factors in Computing Systems, 1991, pp. 189-193

[WEBL98] Westphal, Christopher; Blaxton, Teresa: Data Mining Solutions – Methods and Tools for Solving Real-World Problems, John-Wiley & Sons, Inc., New York/Chichester1998

[WEAM01] Marc Weber, Marc Alexa, Wolfgang Mueller, Visualizing time-series on spirals. Accepted for publication in Proceedings of InfoVis 2001, 2001