THE VISUALIZATION METHOD PIPELINE FOR THE APPLICATION TO DYNAMIC DATA ANALYSIS

T. Galkin[1,a], D. Popov[2,b], V. Pilyugin[1,c], M. Grigorieva[3,d]

[1] National Research Nuclear University MEPhI, Moscow, Russia

[2] Skolkovo Institute of Science and Technology, Moscow Russia

[3] Lomonosov Moscow State University, Moscow, Russia

E-mail: [a] tpgalkin@mephi.ru, [b] dmitry.popov@skoltech.ru,

[c] vvpilyugin@mephi.ru, [d] Maria.Grigorieva@cern.ch

The new era of scientific research brings an enormous amount of data for scientists. These complex and multidimensional data structures are used for the verification of scientific hypothesis. Exploring such data by researchers requires the development of new technologies for its efficient processing, investigation and interpretation. Intellectual data analysis and statistical methods are rapidly developing, and this is where visualization methods are getting their place. This work describes mathematical basis of the developed visualization tool for the analysis of multidimensional dynamic data. This tool provides the pipeline of methods, which combined, allow to cope with a set of practical tasks (anomalies detection, cluster, trends and variation analysis) using visualization method. Authors provided mathematical models of geometrical operations under the data domain, algorithms for solving the mentioned classes of tasks and several use-cases with technological and economic data based on visualization method.

Keywords: visual analysis, dynamic data, time-variant data, multidimensional data, multivariate data, visualization, data analysis, multidimensional analysis.

1. Introduction

Intelligent computer algorithms are state-of-the-art of data analysis today. Artificial intelligence, machine learning and neural networks create the trend of discourse in the data science. However, at the same time some research point out the problem of understanding, interpretation and verification of the research results [1]. Various methods are available for these purposes. One of them is data visualization. Industry and science bring us tasks which include complex multidimensional data analysis, and data visualization can provide deep understanding of data based on its graphical representation. This paper describes the experience of applying the visualization method for multidimensional dynamic data analysis.

2. Background

The overview of visualization techniques for time-dependent multidimensional data can be found in [2]. The authors divide these techniques into static and dynamic. Moreover, they considered interactions with the visual representations. The recent overview [3] considers the different visualization techniques and data transformations. Data model for the visualization can be represented as the multidimensional Euclidian space and affine transformations within this space [4].

Various data structures from different research fields are successfully investigated by imaging, which provides analyst with the advanced and interactive means for data exploration. This paper describes the visualization pipeline, developed by the authors, based on 3D scatter plot diagram with colored distances between data objects in multidimensional space.

3. Visualization Method for the Dynamic Data

Dynamic multidimensional data is represented as a set of parameters of objects, changing in time. This data is stored as a set of numeric data values given for some periods of time.

3.1 Task formulation

Thus, the following task formulation is to solve:

Given:

Let *m* objects given, each of them is characterized by *n* parameters. The data is organized as a set of tables, such as follow:

Time = *j*	Parameter 1	Parameter 2	…	Parameter n
Object 1	x_{11}^{j}	x_{12}^{j}	…	x_{1n}^{j}
Object 2	x_{21}^{j}	x_{22}^{j}	…	x_{2n}^{j}
…	…	…	…	…
Object m	x_{m1}^{j}	x_{m2}^{j}	…	x_{mn}^{j}

This tabular data representation contains object parameter values at a specific point in time. Thus, table *j* is filled with parameter values for time t_j. It is stated that $t_1 < t_2 < \ldots < t_k$, and x_{il}^{j} – value of parameter *l* of object *i* in table *j* at the moment of time t_j ($l = (1,n)$, $i = (1,m)$, $j = (1,k)$).

To make the formulations easier, x_{il}^{j} for $l = (1,n)$ is called a *n-tuple* within the fixed *j*.

Task:

Find the subsets of similar objects, explore these subsets at each given point in time and make judgements about their behavior in time.

3.2 Data samples

In this research two data samples with dynamic data were used: technological and financial data.

Technological data sample were taken from Kaggle's dataset "CareerCon 2019 - Help Navigate Robots"[1]. It represents sensors data gathered while driving a small mobile robot over different floor surfaces: orientation, velocity, acceleration, etc. These data may help robots to recognize the floor surface. The dataset has ~4K objects and 128 time measurements. The total length of dataset is about 400K records. The visualization pipeline was applied to these data to visualize sensors data in order to explore the surface features.

Financial data sample represents data of banking system. This data was obtained from the open sources[2] and describes 81 banks for a period of 13 month by a set features like sales profit, deposits, overdue debt and others. The main idea of the exploration of these data is to uncover suspicious, anomalous banks visually, and detect a point in time or a period when the anomalous behavior takes place.

3.3 Task solving method

For solving the data analysis task, the scientific visualization method was used [5].

Both data samples were represented as a set of dynamic objects. Each object with all the corresponding features at each point in time is stored in a table. A set of tables for all objects at different points in time form a preprocessed dataset for loading into the visualization application.

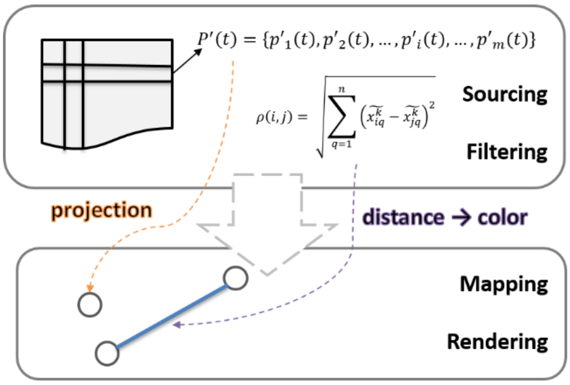
$P'(t) = \{p'_1(t), p'_2(t), \dots, p'_i(t), \dots, p'_m(t)\}$
$\rho(i,j) = \sqrt{\sum_{q=1}^{n} \left(\widetilde{x_{iq}^{k}} - \widetilde{x_{jq}^{k}}\right)^2}$
Sourcing
Filtering
projection
distance → color
Mapping
Rendering

Figure 1. The visualization pipeline

The visual analysis of data has two stages. The first stage is the visualization itself: data tables are transformed into geometrical objects on screen. This transformation implies four steps: sourcing (obtaining the data from the source), filtering (getting the data ready for the application), mapping (corresponding geometric objects placement on the scene) and rendering (making the resulting picture of the scene). After the visualization is ready, the second stage of the analysis - the interpretation of images is performed by the analyst. Parameters of each step of the visualization pipeline may be changed by the analyst interactively in order to generate another visualization sample. This makes the process of data analysis iterative and interactive.

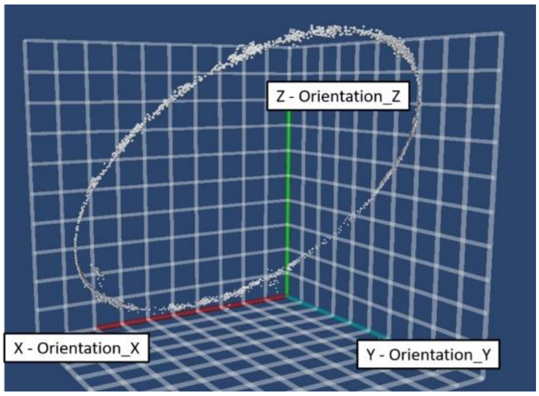

Figure 2. Technological data visualization

3.3.1 Visualization pipeline

Figure 1 shows the transformation between the data table and the visualization. Each line in the table represents the data object. Features of objects are transformed into multidimensional coordinates. The objects are then projected into spheres.

The application backend calculates distances between all pairs of objects in multidimensional space, and display it as segments between corresponding pairs

[1] https://www.kaggle.com/c/career-con-2019/data
[2] https://www.banki.ru/

of spheres. Segments are colored from blue to red, depending on how close the objects are in the original multidimensional space. This allows to observe similarity of objects in multidimensional space looking at the 3D visualization.

4. Implementation and Application

The algorithms of the visualization method pipeline were implemented in C# programming language, using Unity graphics engine.

4.1 Technological data sample visualization

At the figure 2, X Y and Z values are the orientation parameter, in degrees. The spheres in the picture form a circle. That is the key point of visual analysis. Human experts are good at interpretation of graphical images, which is hard to be programmed automatically. Moving the time slider allows to observe changes of parameters values in time and can be useful in the detection of specific points of time when anomalous behavior takes place.

One more thing that can be visually discovered is presented in the figure 3. An analysist found a cluster of spheres, and this cluster does not change in time. They are marked with red color at the picture. Further investigation showed that these spheres correspond to the specific type of floor surfaces. Therefore, such visualization is useful for the problem of clusterization.

4.2 Economic data sample visualization

The figure 4 shows that the spheres lay on a plane, which noticeably rotates over the time. The analyst may observe visually the direction and the velocity of movement of some financial parameters, making conclusions about common financial situation for banks. Also, such visualization allows to catch anomalous banks, which features are changing in time along other trajectories.

4.3 Other applications

This application was also tested on metadata from ATLAS Grid Information System[3] as shown on figure 5. The visualization shows the appearing of computing queues, and the duration of these queues in time, and can be used to observe some specific tendencies, which may be unobvious without graphic representation.

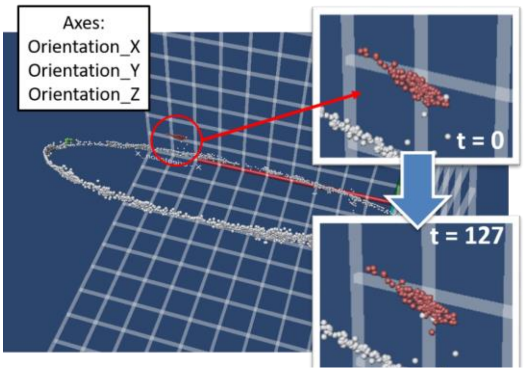

Figure 3. Cluster evolution in time

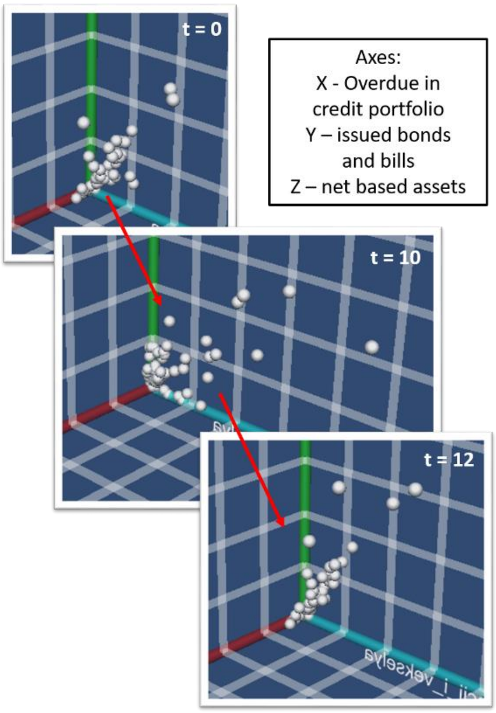

Figure 4. Economic data visualization

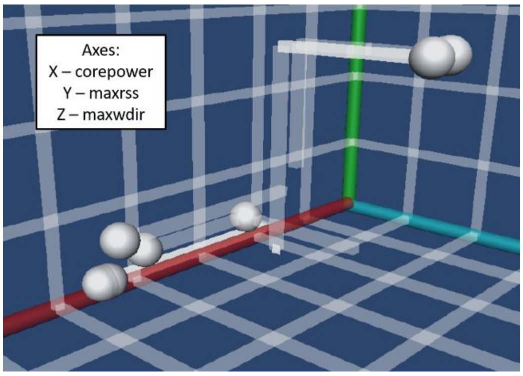

Figure 5. ATLAS Grid Information System metadata visualization

[3] http://atlas-agis.cern.ch/agis/

5. Conclusion

An interactive and iterative method of data analysis and the application for the dynamic data analysis using the visualization pipeline was developed. The method considers dynamic objects as time-dependent points of the Eucledian space. The visualization system utilizes a 3D scatter plot diagram with colored distances in the multidimensional space.

The developed visualization application was tested on different data samples, showing the applicability wide variety of domains.

Further research will be focused on an adaptation of the developed software for more complex tasks of cluster analysis and searching for correlations within the data.

Acknowledgement

This work has been supported by the RSF grant No. 18-71-10003.

References

[1] J. Thomas, K. Cook, "Illuminating the Path: A Research and Development Agenda for Visual Analytics", IEEE Press.

[2] W. Muller and H. Schumann, "Visualization methods for time-dependent data - an overview," Proceedings of the 2003 Winter Simulation Conference, 2003., New Orleans, LA, USA, vol.1., pp. 737-745, 2003.

[3] W. Cui, "Visual Analytics: A Comprehensive Overview", IEEE, vol. 7, p. 81555 - 81573, DOI: 10.1109/ACCESS.2019.2923736.

[4] I. Milman, A. Pasko, V. Pilyugin, "Survey of approaches to multidimensional data geometrization in the analysis using computer visualization", Scientific Visualization, vol. 7, no. 2, pp. 21-37, 2015.

[5] V. Pilyugin, E. Malikova, A. Pasko and V. Adzhiev, "Scientific visualization as method of scientific data analysis," Scientific Visualization, pp. 56-70, 2012.