VISUAL AND COGNITIVE INTERPRETATION OF HETEROGENEOUS DATA

A.A. Zakharova [1,2], A.G. Podvesovskii [2], A.V. Shklyar [1,3]

[1] Keldysh Institute of Applied Mathematics RAS, Moscow, Russian Federation
[2] Bryansk State Technical University, Bryansk, Russian Federation – (zaa, apodv)@tu-bryansk.ru
[3] Tomsk Polytechnic University, Tomsk, Russian Federation – shklyarav@tpu.ru

Commission II, WG II/10

KEY WORDS: Visual Analytics, Cognitive Model, Fuzzy Cognitive Maps, Interpretation of Heterogeneous Data, Biomedical Data Models

ABSTRACT:

The paper identifies conditions that allow the use of visualization means as tools for practical study of heterogeneous data. Effectiveness requirements, the amount of input data or uncertainty of the overall goal of data analysis should be considered complicating factors for such a study. Development and practical application of visualization tools allow to overcome these factors as a result of using the advantages of visual perception for the source data interpretation. Use of cognitive maps has been proposed as a way to compensate for the subjective aspects of visual perception, as well as a tool for verifying the results of visual analysis. Combined use of visualization tools and cognitive maps forms cognitive interpretation technology, which allows solving the problems of research of empirical data belonging to specialized subject areas. An example application of this technology for processing biomedical data is considered.

1. INTERPRETATION OF HETEROGENEOUS DATA

As a result of the source data preprocessing, much additional information appears at the analyst's disposal (intermediate solutions, verification data, classification and clustering results, etc.). A volume of heterogeneous data is formed, that is a set of diverse data characterizing the object, process or system under study recorded on any medium. Interpretation of such data as a research problem is aimed at discovering dependencies of any type among the properties inherent in the objects under study.

The problem of heterogeneous data practical research is associated with a large amount of heterogeneous values and limited available resources (computational, temporal, human). Thus, heterogeneous data interpretation is a process of obtaining new information and is characterized by the amount of resources required. This condition is formed by the requirements for data interpretation tools.

2. BIOMEDICAL DATA ANALYSIS PROBLEMS

Practical biomedical data research is analysis of a large amount of multidimensional data containing a set of heterogeneous parameters having different types, level of validity, accuracy and origin. The main feature of traditional approaches is the simultaneous assessment of a large number of parameters characterizing one or another aspect of a biological object: its genome, genetic expression, protein composition, abundance of symbiotic microbial communities, metabolic composition, etc. The number of factors measured in a single object varies from hundreds to thousands, and the amount of raw information can be tens of gigabytes from one object. Meanwhile, for dependable establishing biological patterns within the framework of the experiment, it is necessary to investigate hundreds of such objects.

3. TECHNOLOGY OF INTERPRETATION

Data interpretation tools

Systematization of efforts to solve data analysis problems led to the development by the authors of a technology for heterogeneous data cognitive interpretation using visualization tools (Fig. 1). The developed technology involves implementation of a sequence of research stages aimed at creation and use of specialized research tools that meet the pre-formulated requirements for data analysis effectiveness. One of the key conditions for achieving this goal is the use of visualization tools to combine computational and cognitive resources available to the researcher.

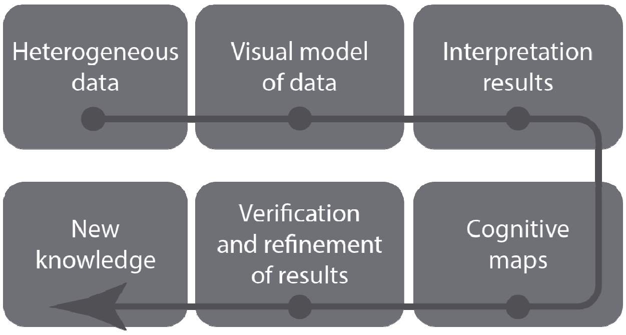


Fig. 1. Heterogeneous data interpretation technology

Visualization capabilities

The research tools are visual data models, i.e. software solutions allowing to obtain images that are a perceived representation of the source data. Visualization tools are a set of research tools, including visual models and ways of user interaction with them, which allow interpreting the source data (Zakharova et. al., 2017). The overall goal of the transition to the visual representation of these data is to use high-speed perception for searching and identifying features in the volume of the source data (heterogeneity, discontinuities, correlations,

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W12-243-2019 | © Authors 2019. CC BY 4.0 License.

symmetry, etc.), which indicate the existence of internal patterns in the data.

Let us define the visual data model as a visually perceptible image that correlates to this data according to some rule. The correlation rule is also referred to as a visualization metaphor (Zakharova et. al., 2013). Search for patterns in the data under study through the comprehension of the visual image becomes the task of visual analysis. As a result, the analysis process is completed after the visual model is perceived by the user as an image of the dependencies in the data found and understood by him.

The development of a visual model is an independent stage of technology and includes operations aimed at unifying input data format, adapting visual model controls to researcher's perception, choosing visual representation method appropriate to the interpretation objective, user qualification, available resources, etc. The advantage of this approach is the possibility of transition to data representation formally correct but different from the traditional one. As a result, cognitive interpretation of data becomes possible, which is difficult when using traditional data analysis techniques.

An essential aspect in the development and practical use of visualization tools is an active role of a host of subjective factors that manifest themselves in the interaction of the user and the visual image of the analyzed data (Chen, 2008). The presence of subjective factors can be both an advantage and a disadvantage of visualization tools; and the task of the developers is creating conditions to gain the maximum possible advantages (Bondarev et. al., 2016).

Due to the employment of visualization tools, processing of multidimensional data does not require conventional dimensional reduction methods. In addition, it becomes possible to visually monitor and interpret the intermediate results of the source data conversion and hence to form hypotheses about potential patterns in the data at the initial stages of research.

Another important feature of the proposed visual models is their interactivity. Unlike static visualization, such models are able to present the researcher with a multitude of images, corresponding both to individual data sets and various visualization metaphors. Comparing states of a controlled model corresponds to the mental hypothesis formulation. Each change to the model which coincides with the expectation becomes a confirmation. If the state does not meet a meaningful expectation, it attracts the user's attention and creates an event interpreted as a contradiction. This initiates formulation of a problem, resolution of which leads to the discovery of new information by the researcher. Thus, interactive features can serve as cognitive search tools.

Cognitive modelling capabilities

Within the framework of the proposed technology for interpreting heterogeneous data, a cognitive model is a tool for additional verification of hypotheses formed using a visual model. Such verification can be viewed, among other things, as a way to compensate for the above-mentioned subjective aspects of visual perception. Besides, due to its simulation modelling capabilities, the cognitive model can effectively complement the visual model, providing an opportunity to analyze effects of various control actions on the object, process or system under study, or find control actions that can bring this object (process, system) to the desired target state (Avdeeva et. al., 2016).

The cognitive model is based on formalization of cause-and-effect relations which occur between factors characterizing a system under study. The result of the formalization represents the system in the form of a cause-and-effect network, termed a cognitive map and having the following form:

$$G = < E, W >,$$

where $E = \{e_1, e_2, \ldots, e_K\}$ is a set of factors (also called concepts), W is a binary relation on the set E, which specifies a set of cause-and-effect relations between its elements.

Fuzzy logic is most commonly used as mathematical apparatus to represent and analyze cognitive models. There is a whole class of cognitive models based on different types of fuzzy cognitive maps (FCMs). One of FCM varieties well-proven in practical problems of analyzing and modeling semi-structured systems is Sylov's fuzzy cognitive maps (Isaev et al., 2017).

Problems solved by cognitive modeling can be divided into two groups:

- static (structure and target) analysis, which goals are finding the key factors influencing the targets most, identification of contradictions between the targets, feedback loops analysis, etc.;
- dynamic (scenario) analysis aimed at prediction of system states under various control actions and search for control solutions bringing the system to the target state.

4. BIOMEDICAL DATA MODELS

When creating visualization tools intended to study empirical data related to a particular field of knowledge, it is important to take into account a number of specific requirements to achieve the necessary research effectiveness. Definition of the subject area allows, for example, using specialized expressive means designed to represent such data. This provides a solution to the problem of over-informing and reduces the overall research time (Pirolli, 2007).

Use of visualization tools for interpretation of biomedical data can be exemplified by the solution of the problem of empirical data on patients' intestinal microbiota composition preliminary research, aimed at determining dependencies between individual parameters.

The source data for the test study have been sets of selected characteristics of patients' intestinal microbiota, supplemented by metadata, the need for which was determined at the stage of the preliminary study. The typical data dimension in the experimental study was 5–20 parameters. Meanwhile, data on groups of 200–250 people were analyzed simultaneously. The research was aimed at confirming the hypotheses about the relationship between the content of a number of bacteria (Lactobacillus, Cloacibacillus, Olsenella, Megasphaera, etc.) and the patient's diagnosis. In addition, the task of the visual research was to identify the relationships between microbiota characteristics, diseases, and patient's metadata, such as age, clinical record, and various social parameters.

The developed visual model is a set of visual images of individual properties that are present in the heterogeneous data volume (Fig. 2). Each such image is a horizontal plane, which contains data corresponding to an individual property of the objects under study. The plane is defined by two orthogonal scales. The main scale corresponds to the selected property and is scaled in the model space in order to provide the most complete presentation of information to the user. The additional scale is used to display the comparison parameter selected as the basis for the research metaphor.

Arrangement of property images relative to each other is used as an independent scale, reflecting the degree of connectivity between one previously selected property and all other data. Points of different planes displaying properties of an object are combined into a single visual image of this object, which is prepared for further analysis.

The visual model management interface is designed to perform data filtering operations, select visual presentation methods and control the interpretation process (Fig. 2). The interpretation results have confirmed a number of pre-formed hypotheses and made possible the transition from the visual to the formal model of relationships between the parameters. To verify these hypotheses, cognitive maps can be used.

When constructing a cognitive map, selected properties act as concepts. Cause-and-effect relations between concepts are set on the basis of hypotheses about the relationship between properties; methods of structural and parametric identification of cognitive maps are used. An example of a cognitive map built in IGLA cognitive modeling support system is shown in Fig. 3.

Methods of structure and target and scenario analyses can be applied to a constructed cognitive map. Structure and target analysis allows to evaluate the reliability of hypotheses about the existing relationships between properties, as well as to identify the crucial properties that have a significant impact on the system or process under study, or, conversely, are subject to significant influence from the system. Within the framework of the scenario analysis, simulation modelling of various control actions is carried out in order to assess their consequences or to search for actions ensuring specified target states. This allows analyzing identified patterns over time and also forming additional hypotheses.

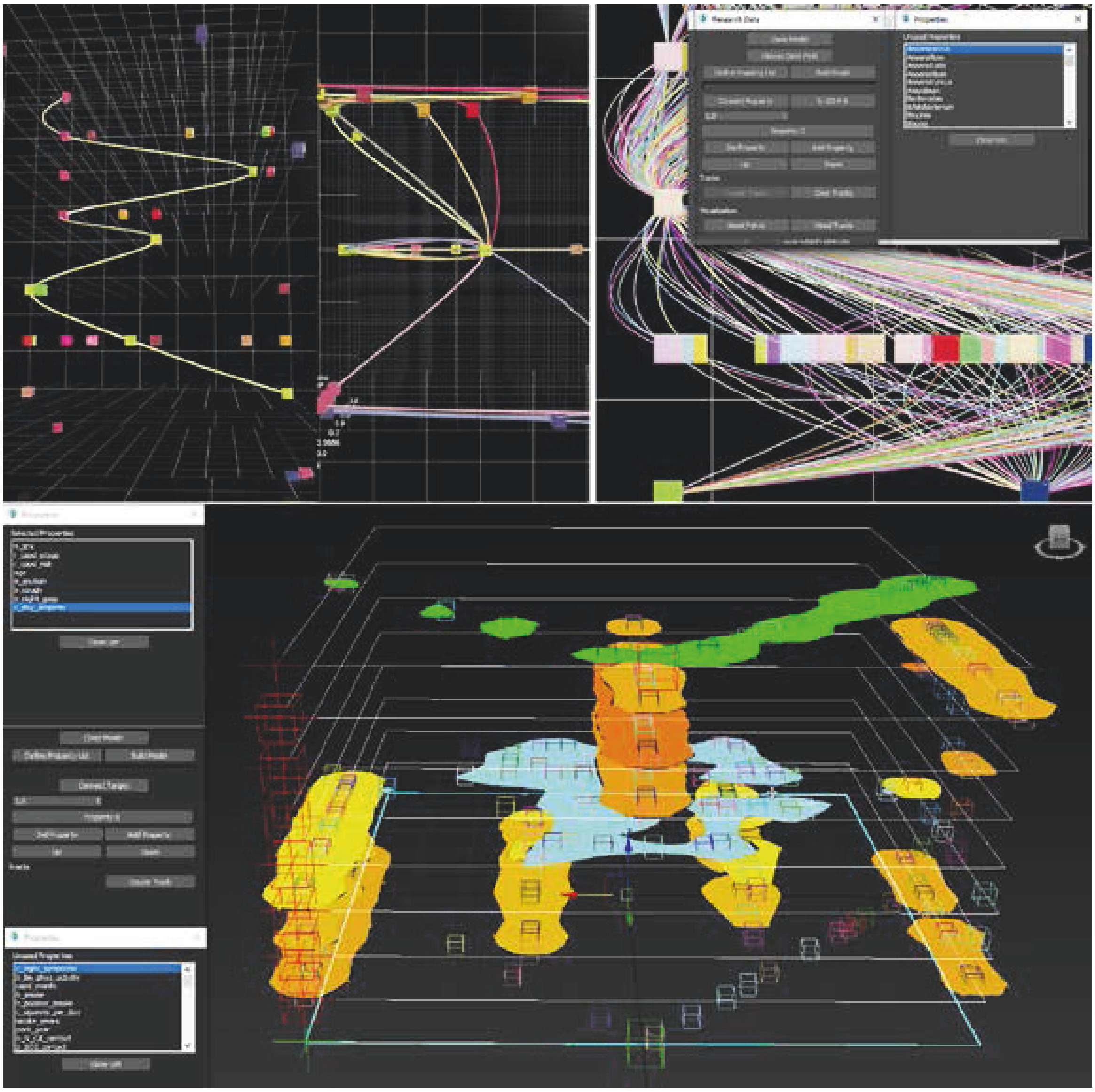
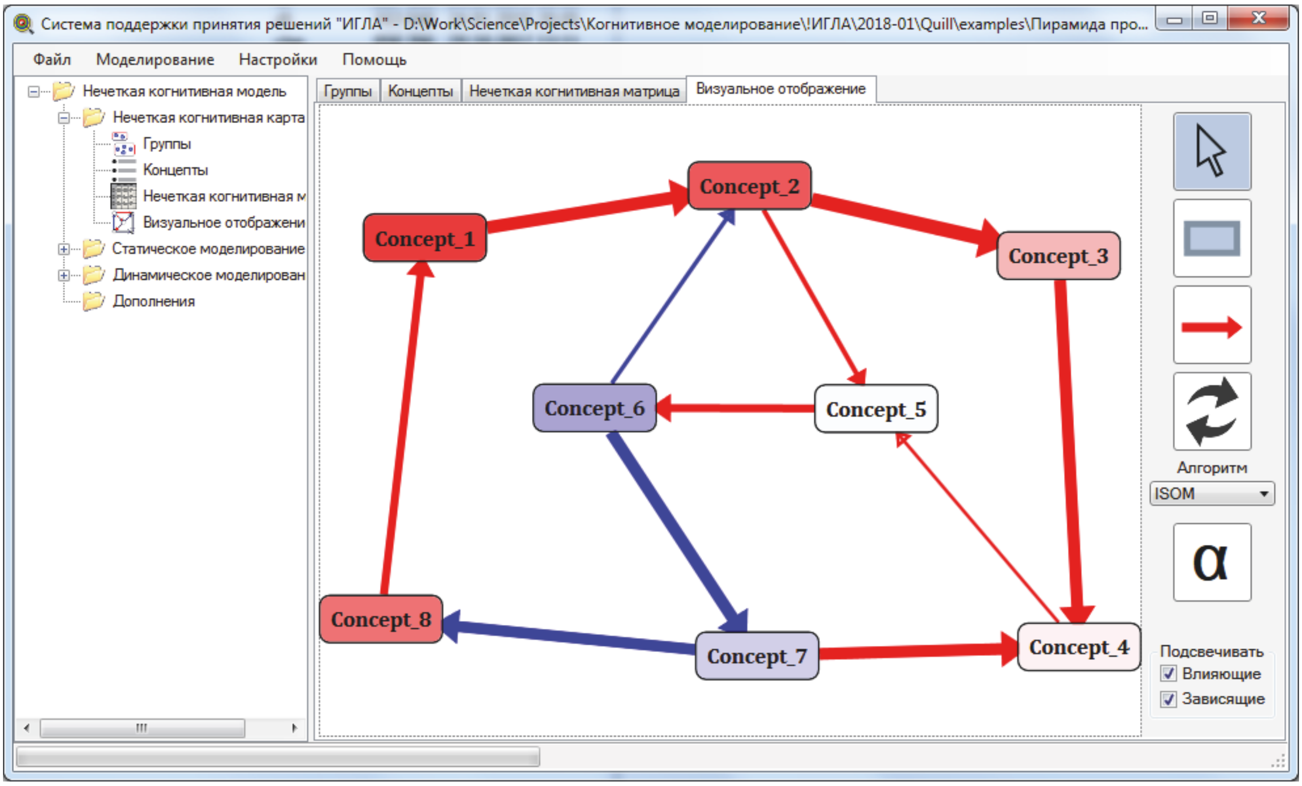


Fig. 2. States of a visual model and its properties management interface


Система поддержки принятия решений "ИГЛА" - D:\Work\Science\Projects\Когнитивное моделирование\!ИГЛА\2018-01\Quill\examples\Пирамида про...
Файл
Моделирование
Настройки
Помощь
Нечеткая когнитивная модель
Нечеткая когнитивная карта
Группы
Концепты
Нечеткая когнитивная м
Визуальное отображени
Статическое моделирование
Динамическое моделирован
Дополнения
Группы
Концепты
Нечеткая когнитивная матрица
Визуальное отображение
Concept_1
Concept_2
Concept_3
Concept_6
Concept_5
Concept_8
Concept_7
Concept_4
α
Подсвечивать

Алгоритм ISOM

☑ Влияющие

☑ Зависящие

Fig. 3. Example of a fuzzy cognitive map

5. EXPERIMENTAL EVALUATION AND ASSESSMENT OF TECHNOLOGY

To organize research of biomedical data, performance criteria have been identified, achievement of which was considered one of the goals of the technology being developed. Firstly, completion of the study of the source data set is believed to be a necessary result. Therefore, the performance criterion of a visualization tool is research speed preservation for various tasks. In other words, time required to interpret sets of different source data should be predictable and, preferably, manageable.

Secondly, a common feature of the tasks for the solution of which visualization tools are developed is their interdisciplinary nature. Thus, one of the performance criteria of the proposed technology is ensuring operational interaction of specialists with different levels of prior knowledge and specialization when forming hypotheses aimed at interpreting data. In particular, it is necessary to reduce unreasonable information load on the user, which prevents cognitive interpretation of data.

To determine the degree of dependence of visual research duration on the source data sets, interpretation time measurement was carried out in solving a series of test problems using the developed visualization tools. The difference in the tasks of one series was the use of various data sets while maintaining the type of the visualization task and the research objective. The results obtained (Fig. 4) allowed to draw conclusions about the close performance of visual analytics tools used to solve different tasks. Thus, the requirement of visual research duration predictability within the framework of the proposed technology is fulfilled.

An additional result achieved at the stage of test solving within the framework of the cognitive interpretation technology being developed was the confirmation of the hypothesis of visualization tools semiotic meaning. Experimental performance measurement of a visualization tool developed for one task in solving another task, differing in the subject area or the research objective formulation, has been carried out. After some initial stage of solution, the duration of which is individual and related to the visual perception capabilities, differences in the visualization tool performance for different tasks became negligible. This is indicative of the transition to the use of visualization tools as a formal language system.

6. CONCLUSION

The paper proposes a technology for interpreting heterogeneous data which is based on the combined use of visual and cognitive models and allows solving the problem of studying multidimensional heterogeneous data related to specialized subject areas. The technology was applied to analyze and interpret biomedical data by constructing and analyzing data visual images with the subsequent verification of the formed hypotheses using cognitive models. This made it possible to significantly reduce the stages of research related to the selection of interrelated parameters and properties in the data volume. Also an important advantage of the technology proposed is the possibility of participation in the problem solution of researchers qualified in the subject area which serves as the source of data.

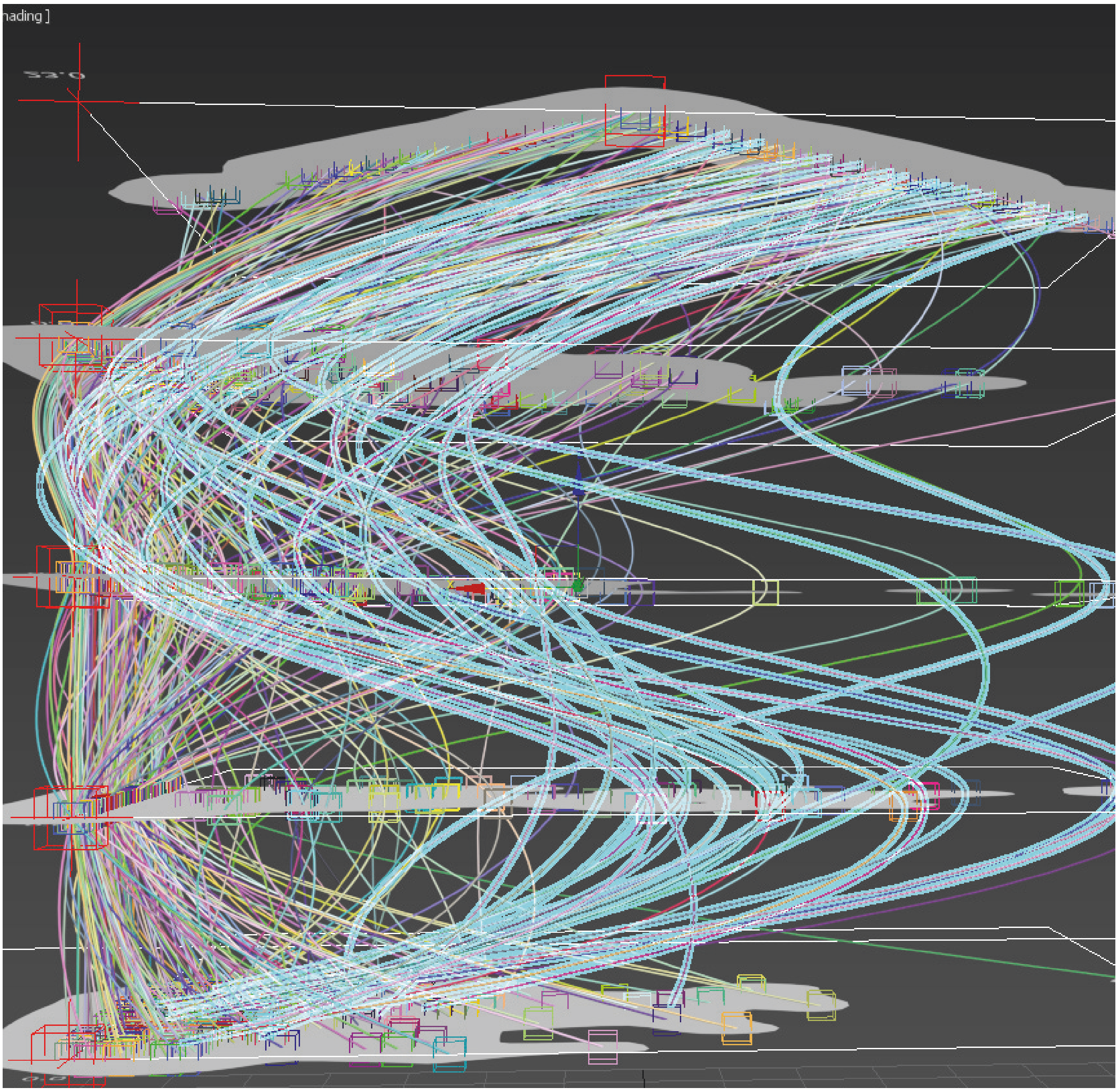


Fig. 4. Strongly linked data visualization example

ACKNOWLEDGEMENTS

This research was supported by Russian Science Foundation, project 18-11-00215.

REFERENCES

Avdeeva, Z., Raikov, A., Ermakov, A., 2016. Big Data Refining on the Base of Cognitive Modeling. In: *IFAC-PapersOnLine*, Vol. 49, Issue 32, pp. 147-152, https://www.doi.org/10.1016/j.ifacol.2016.12.205

Bondarev, A.E., Galaktionov, V.A., 2016. Multidimensional Data Analysis and Visualization for Time-Dependent CFD Problems. *Programming and Computer Software*, 41 (5), pp. 247–252.

Chen, C., 2008. An Information-Theoretic View of Visual Analytics. *IEEE Computer Graphics and Applications*. , 28 (1), pp. 18-23.

Isaev, R.A., Podvesovskii A.G., 2017. Generalized Model of Pulse Process for Dynamic Analysis of Sylov's Fuzzy Cognitive Maps. In: *CEUR Workshop Proceedings of the Mathematical Modeling Session at the International Conference Information Technology and Nanotechnology (MM-ITNT 2017)*, Vol. 1904, pp. 57-63, http://ceur-ws.org/Vol-1904/paper11.pdf

Pirolli, P.L., 2007. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press.

Zakharova, A.A., Shklyar, A.V., 2013. Visualization Metaphors. *Scientific Visualization*, 5 (2), pp. 16-24.

Zakharova, A.A., Vekhter, E.V., Shklyar, A.V., 2017. Methods of Solving Problems of Data Analysis Using Analytical Visual Models. *Scientific Visualization*, 9 (4), pp. 78-88.

Revised May 2019