Graph-Based Visual Analytics Tools for Digital Humanities Research *

Konstantin Ryabinin[1] [0000−0002−8353−7641],
Konstantin Belousov[1] [0000−0003−4447−1288], and
Svetlana Chuprina[1] [0000−0002−2103−3771]

Perm State University, Bukireva Str. 15, 614990, Perm, Russia
kostya.ryabinin@gmail.com, belousovki@gmail.com,
chuprinas@inbox.ru

Abstract. This paper is devoted to the development of the Web application for the visual analytics of the interconnected data within digital humanities research highly adaptable to the specifics of application domain and personal analytics preferences. The circular graph is proposed as a visual model to depict the interconnected data in a comprehensive way. The graph rendering software is organized according to the model-driven architecture utilizing ontology engineering methods and means, which ensure configuration flexibility and modification ease. The functioning scenarios of the application's visualization component can be changed without its source code modifications, just by editing the underlying ontology that describes data processing and rendering mechanisms. Extraction, transformation, loading and rendering of the data are configured in the intuitive way by data flow diagrams with the help of a high-level graphical editor. The described features are demonstrated on the real-world examples from the digital humanities application domain.

Keywords: Visual Analytics, Circular Graph, Data Filtering, Data Comparison, Ontology Engineering, Digital Humanities.

1 Introduction

Many tasks in digital humanities (DH) research involve the processing of the linked data, wherein the graph theory appears to be a powerful methodological and technological base for solving associated problems. Taking into account the specifics of DH, the considered data are normally quite big, but their handling requires human attendance and cannot be fully automated. One of the key means to help DH specialists to fulfill their everyday work is scientific visualization and visual analytics (VA) that allows to present related data in an observable interpretation-ready form. Our goal is to develop an ergonomic and flexible tool for graph-based visualization of interconnected data

* The reported study is supported by Ministry of Science and Higher Education of the Russian Federation, State Assignment No. FSNF-2020-0023.

that allows comprehensive VA within DH research. The new high-level component for circular graphs' visualization is presented to tackle data filtering problems and improve the cognitive power of visual analytics. Ontology-driven data extraction, transformation and loading (ETL) mechanism is proposed to enable obtaining the data from different sources and process them in a flexible way. The software developed is demonstrated by solving the problems from applied linguistics domain.

2 Background and Related Work

VA is no doubt a powerful methodology to conduct research in a field of DH, but, as indicated in [9], nowadays there is a noticeable talent gap between the VA scholars and digital humanists. While DH and VA have a huge potential of coevolution, the research results presented in the literature are typically valuable either only for DH, or only for VA, and rarely for both simultaneously [9]. This is because DH projects often lack researchers with deep computer science skills, and consequently have to rely on the existing general-purpose visualization tools, instead of driving the actual software development. But in that case, some tasks remain unsolved because of traditional software limitations [8,9]. W. Huang et al. tackle this problem by proposing a so-called user-centered approach to the process of visualization making (graph-based visualization in particular). This approach ensures the creation of cognitive graphics tools, which development comprises design and evaluation stages [8]. On the design stage, "the designer applies design principles and chooses the visualization best supporting perception and cognition", and on the evaluation stage "visualization is evaluated to understand how cognitive processes are affected" [8].

Similar, but slightly less formal approach is proposed by S. Jänicke, who describes an "ideal" VA+DH project as a close collaboration between the computer scientists and digital humanists, where each visualization feature proposed is immediately tested and validated in terms of its viability for DH research and then either approved for further development or rejected [9]. Working on our VA tools, we have chosen this exact strategy.

For graph visualization, the Gephi system is traditionally used [7]. Being feature-rich, this system, however, provides instruments for layout the graphs of free structure, while we found out, that sometimes the circular graphs [2] are more comprehensive by depicting data sets. Moreover, as stated in [15], it is often desirable to have the graph visualization tools in a Web application, without installing additional software.

An important point of graph visualization is the data preparation stage. To ensure the intuitive and flexible data preparation process we suggest to declare its steps by data flow diagrams (DFDs) [10]. A lot of popular visualization software use such an approach, for example, Blender, Maya, Substance Designer, etc., so it proved its efficiency in terms of data processing and rendering pipeline declaration.

We use model-driven architecture based on the ontology engineering methods [6] to achieve the configuration flexibility and adaptation of the software to the specifics of the application domain without source code modification. We construct the ontologies within visual editor ONTOLIS [6].

Our previous research work was dedicated to the development of ontology-driven scientific visualization and VA system called SciVi[1] [13]. This system is portable across all the popular platforms (Windows, GNU/Linux, macOS, iOS, Android). It is organized as a client-server application, having both thick (native, written in C++ using Qt 5 framework) and thin (browser-based, written in TypeScript and JavaScript, utilizing HTML5 and CSS3) clients. The behavior of this system is fully controlled by underlying ontologies, which allow deep reconfigurations of SciVi, extension of its ETL and data rendering capabilities, whereby leveraging adaptation to the completely new visualization and analytics tasks without changing the source code of its core. Faced the problems in a field of DH during the case study of verbal and nonverbal behavior of social network users, we built the graph VA toolset upon the SciVi [12]. Tried out different graph layouts, we focused on the circular one because of its good readability [2].

We implemented a graph visualization SciVi component (called SciVi::CGraph) as a Web application in TypeScript utilizing PixiJS[2] rendering engine. The graph nodes are uniformly distributed on a circle and the edges are drawn as quadratic Bézier curves with the control point in the circle's center. Different slices of data can be displayed on the same graph using a scale of states that allows fast switching between them. Data slices can be organized in a hierarchy, therefore this scale supports multiple levels. Examples of different graphs can be found online: `https://graph.semograph.org/cgraph/`.

SciVi tools have been integrated into Semograph information system [4]. Semograph is aimed to solve different DH tasks involving methods of computational linguistics by supporting a wide range of operations on the textual content, including tagging, classification of terms, building semantic relations, etc. The integration with SciVi allowed to utilize advanced visualization features including the rendering of graphs.

3 ETL Mechanism

The conceptual scheme of the data processing within the SciVi system is shown in Fig. 1.

Currently, CSV format is used to transfer data from Semograph into SciVi, since export to this format is natively supported by Semograph. However, it is easy to switch to any other data representation since the ETL mechanism of SciVi is very flexible. This mechanism is implemented within the SciVi Data Processing Module and driven by the ontological knowledge base. Underlying ontologies describe different data formats and data interpretation rules, as well as available data preprocessing filters and data visualization techniques. Thanks to this, changing or extending these ontologies is enough to alter SciVi behavior adapting it to the new VA tasks. But the changing of ontologies requires knowledge engineering skills, thereby is unwanted for the end-users and is dedicated to the system administrator.

The end-users are provided with a more high-level steering instrument: Data Flow Editor. This SciVi module (based on the Rete[3] JavaScript framework) implements a

[1] `https://scivi.tools`

[2] `https://www.pixijs.com`

[3] `https://rete.js.org/`

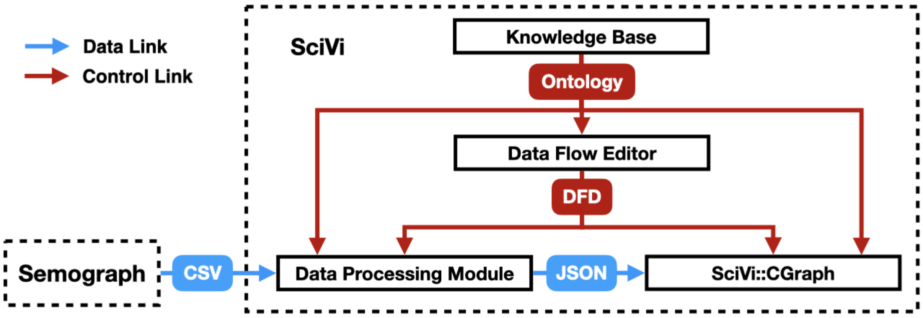


Fig. 1. Data processing pipeline within SciVi.

graphical user interface (GUI) to compose a data processing algorithm from the high-level building blocks utilizing DFDs. The example of DFD describing the extraction of data from an arbitrary CSV file is shown in Fig. 2.

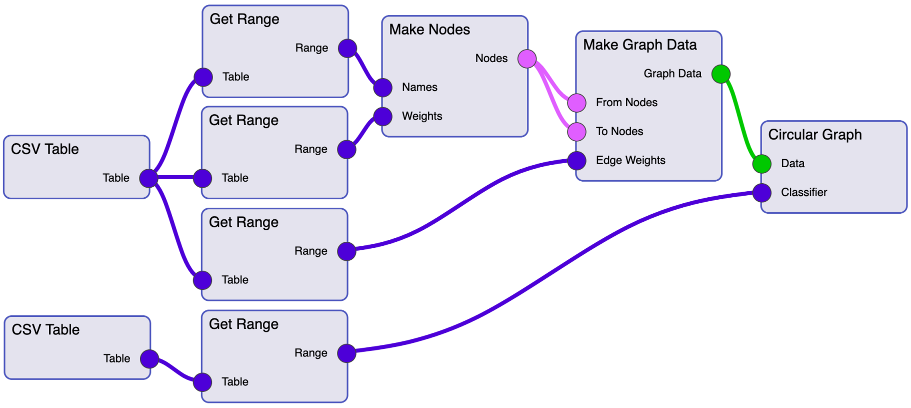


Fig. 2. DFD defining ETL and data visualization in SciVi.

Each node in the DFD represents a particular step in data obtaining, processing or visualization. For example, “CSV Table” defines file reading; “Get Range” allows to specify the subset of values within the CSV table; “Make Nodes” sets up the composition of the graph nodes internal representation; “Make Graph Data” corresponds to the stage of interconnecting the graph nodes with weighted edges; “Circular Graph” defines the data rendering using SciVi::CGraph visualization component. Links between DFD nodes depict the data flow and their color is bound to the type of transmitting data.

The set of available DFD nodes’ types correspond to the set of operations on the data available in SciVi. It is constructed automatically according to the underlying ontology and presented to the user as a toolbar palette. Each data processing operation has its own description that may be altered or extended to change the actual behavior of the entire system. For example, the ontology fragment describing “CSV Table” is shown in Fig. 3.

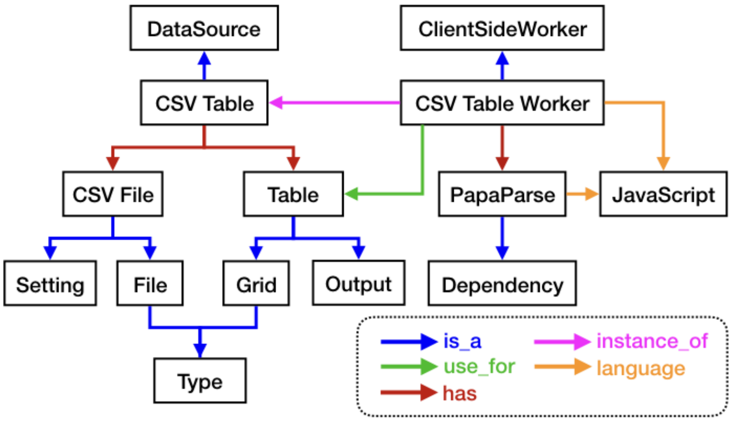


Fig. 3. Fragment of SciVi ontology describing CSV table reader.

It can be seen, that “CSV Table” node is treated as a data source, has CSV file as a setting parameter and table of values as an output. The actual implementation of this data reader is described by the “CSV Table Worker” concept in the ontology. This concept has an internal attribute (not drawn in the figure, since the figure shows concepts and relations only) with a link to the JavaScript code fragment that implements CSV reading with help of PapaParse[4] library. An important part of “CSV Table” description is “ClientSideWorker” concept. It identifies that the reading and parsing takes place within the browser (on the client side), without sending the data to the SciVi server. Although the SciVi architecture allows server-side processing, currently the amounts of data we faced in our tasks were small enough to be handled locally.

4 Visual Analytics Component

SciVi::CGraph VA component accepts the data in JSON representation. Once the user has created the DFD for the particular task and started the visualization, this component renders the graph and provides its own GUI allowing different interactions with that graph, including zooming, panning, nodes and edges selection, data filtering, etc. The most important distinctive features of SciVi::CGraph are described below.

4.1 Multilevel Ring Scale

In case, when a classification of graph nodes is defined, SciVi::CGraph draws a special ring scale around the graph to visually highlight the given nodes’ classes. The number

[4] https://www.papaparse.com

of rings in this scale is potentially unlimited, so the nodes' classifier can have multiple levels. A special tree view in a sidebar of the graph allows to explore the classifier and switch the visibility of nodes belonging to individual classes. Colors of the ring sectors, which depict the classes, can be assigned manually, but also set automatically based on the special heuristic algorithm that maps the classifier's hierarchy to the HSV color model in a way the neighbor ring sectors have distant colors to be visually distinguishable.

To evaluate different hypotheses, the user can change the order of scale rings by drag and drop, command the graph to sort the nodes accordingly and set the color of nodes to the color of any ring sector they belong to. These interactions help to find out, which order of hierarchy levels is the most meaningful one in terms of structuring the interconnected data.

Fig. 4 shows[5] the results of the correlation analysis of 38 topics extracted from 48 stories told by informants as self-presentation [11]. The sample of informants is balanced by sex, age, and education level. Graph nodes depict self-presentation texts, edge thickness represent correlations coefficients (all correlations are positive; all coefficients below 0.8 are filtered out). Social (education level: secondary, higher) and demographic (sex, age group) parameters are shown on the ring scale groping the nodes accordingly. The groups are nested according to the order of the rings.

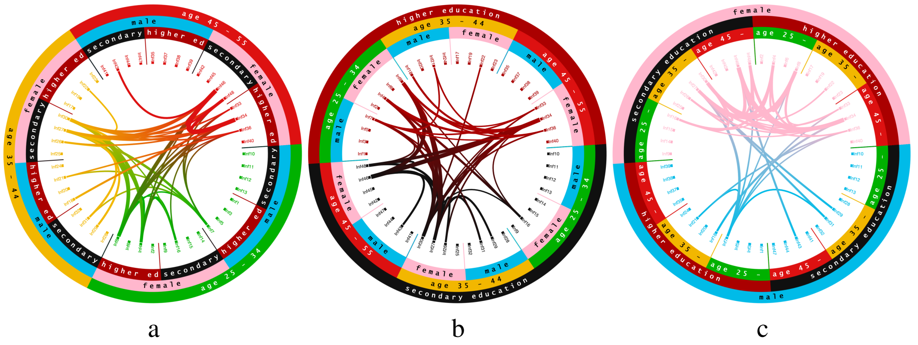


Fig. 4. Correlation of topics in self-presentations, grouped by age (a), education (b), and sex (c).

The aim is to find out, which parameter dominates by grouping the informants together. Related to DH it means to find, which social / demographic informant groups consolidate more by talking about themselves. Related to the graph theory it means to find, which layout of nodes provides their better clustering. The proposed mechanism of the ring scale reordering allows quick checking of different variants and inspecting them visually. While Fig. 4a (topmost grouping by age) and 4b (topmost grouping by education level) look messy, Fig. 4c reveals significant dense "community", corresponding to the stories told by females (at the same time, there is almost no correlation between

[5] The interactive graph is available online: `https://graph.semograph.org/cgraph/aboutmyself/index.html`

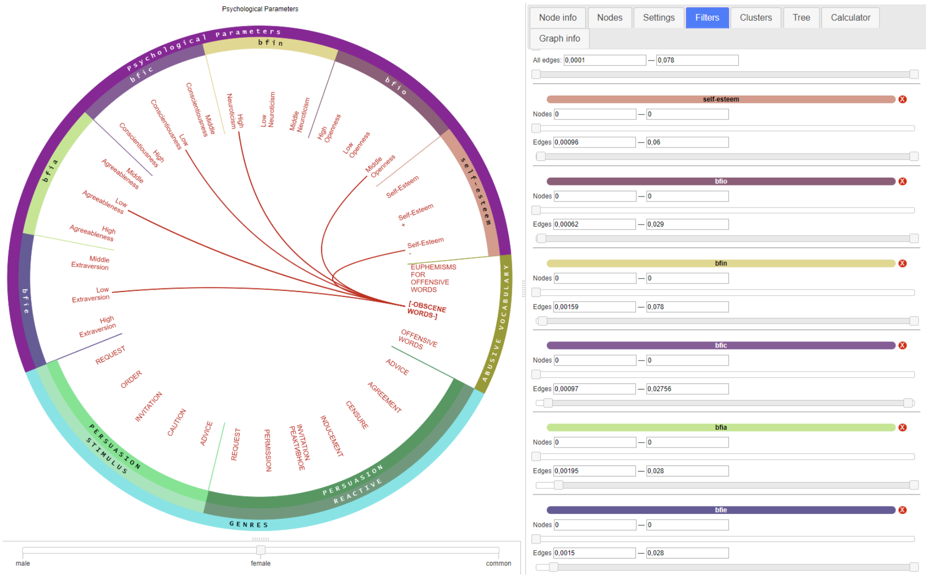
males' stories). Further interpretation of this material is outside of this paper's scope, but the corresponding milestone of related DH research is considered to be reached. It is worth noting, that it took less than a minute to find this solution using SciVi::CGraph.

4.2 Equalizing Filter

Sometimes the noisy data on a single graph may have a non-uniform distribution of the noise strength. In this case, filtering the entire data set with the single threshold appears to be meaningless and threshold adaptivity is required. We often face this problem in multipartite graphs comprising interconnected data of different nature, or data, which parts were differently preprocessed. To tackle this problem, we propose a so-called equalizing filter that can have individual parameters for selected groups of nodes and edges (resembling the sound equalizer that can differently affect selected parts of the spectrum).

Currently, the equalizing filter within SciVi::CGraph operates as a set of range-based cutoff functions tied to the ring scale. By default, there is one global cutoff function (affecting the entire graph), but, if needed, the user can add auxiliary local ones for any sector of the ring scale. If a node or an edge is affected by multiple cutoff functions (global one and multiple local ones according to the hierarchy of the ring scale), their ranges are intersected to build the resulting filter. Node or edge is filtered out if its weight lays outside the functions' range intersection.

The practical use case of the equalizing filter is demonstrated in Fig. 5.



All edges: 0,0001 — 0,078

self-esteem
Nodes 0 — 0
Edges 0,00096 — 0,06

bfio
Nodes 0 — 0
Edges 0,00062 — 0,029

bfin
Nodes 0 — 0
Edges 0,00159 — 0,078

bfic
Nodes 0 — 0
Edges 0,00097 — 0,02756

bfia
Nodes 0 — 0
Edges 0,00195 — 0,028

bfie
Nodes 0 — 0
Edges 0,0015 — 0,028

Fig. 5. Relationships between the verbal behavior of social network users and their psychological characteristics.

This figure represents[6] the filtered data of the relationships between the verbal behavior of social network users (SNUs) and their psychological characteristics. The psychological parameters are obtained by two questionaries (personality features and self-esteem) [14] fulfilled by the sample of SNUs. The verbal behavior is revealed with the help of the linguistic analysis from the comments written by these users in social networks. The filtering is individual for each psychological parameter because each of them has its own statistical features (minimum, maximum, average, standard deviation). This approach allows leaving only the dominant indicators for each psychological parameter. Fig. 5 demonstrates, that after equalizing the indicators, it can be revealed that the SNUs of the female gender, who use obscene words in the public social network space, are characterized by low self-esteem, low conscientiousness, low agreeableness, high neuroticism, middle openness, and low extraversion.

4.3 Graph State Calculator

To visually compare the structure of data slices displayed in the graph, we implemented a special graph state calculator. It allows to perform a sequence of basic set operations on the graph states: union, intersection, difference, and symmetric difference.

Fig. 6 demonstrates[7] the states of “Moscow” geoconcept. In this research, under the term “geoconcept” we understand a set of collective opinions about a geographical object. These opinions can be revealed from the associations people come up with [16]. Graph nodes represent the semantic categories of associations (extracted according to the special classifier within Semograph system), edges identify the co-presence of linked categories in the analyzed associates (derived from a group of informants). The actual structure of geoconcept presented as a set of association categories depends on the region. In this experiment we collected 3 datasets: in Perm (Fig. 6a), Biysk (Fig. 6b) and Orenburg. The state scale (drawn below the graph) provides quick navigation between these data sets and makes it possible to visually compare them. However, to make this comparison more elaborated and meaningful, set operations can help. As an example, Fig. 6c shows the intersection of Perm and Biysk data sets, allowing to view their common parts.

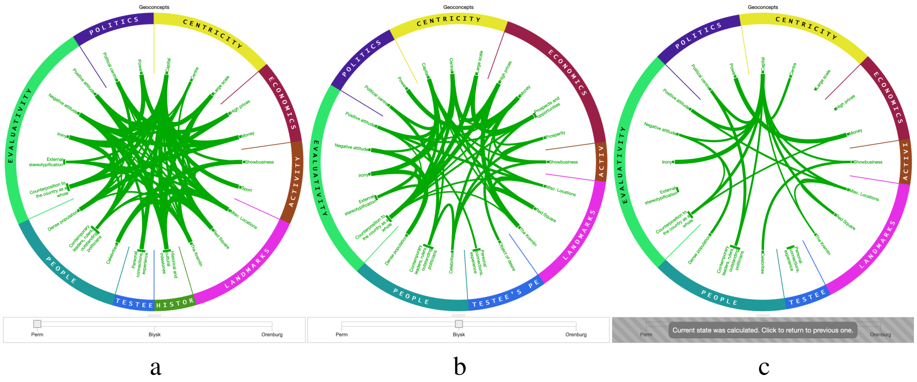
Thus, the graph state calculator provides a good basis for conducting comparative DH studies and facilitates the process of interpreting research results.

5 Conclusion

Thanks to the features discussed, SciVi::CGraph allows advanced interactive VA of interconnected data in DH. According to the feedback from the DH researchers of Perm State University, this tool outperforms traditional graph analysis software like Gephi in the tasks, which require special analytics features. Like SciVi VA system,

[6] The interactive graph is available online: `https://graph.semograph.org/cgraph/psycho_reduced/index.html`

[7] The interactive graph is available online: `https://graph.semograph.org/cgraph/geoconcepts_reduced/index.html`



Current state was calculated. Click to return to previous one.

Fig. 6. States of "Moscow" geoconcept as viewed in Perm (a) and Biysk (b) along with their intersection (c).

SciVi::CGraph is an open-source project licensed under the terms of GPLv3: `https://github.com/scivi-tools/scivi.graph`.

SciVi::CGraph is being iteratively developed in close collaboration with DH specialists and each new feature is immediately evaluated in real-world research projects (in exact accordance with the cooperation model described in [9]).

For example, SciVi::CGraph was used by exploring the egocentric field of speaker in the macedonian language [5], in the study of social network users' speech within the research project of Perm State University supported by Ministry of Education and Science of the Russian Federation, state assignment No.34.1505.2017/4.6 [3], and by the semiotic analysis of geomental maps [16]. Also, SciVi::CGraph was utilized in the Sirius education center within the project "Images of Large Russian Cities in the Linguistic Consciousness of Senior Schoolchildren" [1].

Taking into account the needs of conducted DH research, we plan to extend our scientific visualization system SciVi with new feature-rich visualization components for free structure graphs, graphs with volumetric 3D layout and graphs pinned to geographic maps.

We would like to thank Alexey Gorodilov, Elena Erofeeva and Ekaterina Khudyakova for fruitful discussions on the papers topic.

References

1. Linguists of Perm University Became Curators of Scientific Research at the Educational Center "Sirius". http://www.psu.ru/ (2019), `http://www.psu.ru/news/lingvisty-permskogo-universiteta-stali-kuratorami-nauchnogo-issledovaniya-v-obrazovatelnom-tsentre-sirius`, last accessed 12 Aug 2020
2. Ageev, A.: A Triangle-Free Circle Graph with Chromatic Number 5. Discrete Mathematics **152**, 295–298 (1996). https://doi.org/10.1016/0012-365X(95)00349-2
3. Belousov, K., Erofeeva, E., Baranov, D., Zelyanskaya, N., Shchebetenko, S.: The Multi-Parameter Analysis of Linguistic Data in the Information System Semograf (On the Example

of the Study of Social Network Users' Speech). Tomsk State University Journal of Philology pp. 6–29 (2020). https://doi.org/10.17223/19986645/64/1

4. Belousov, K., Erofeeva, E., Leshchenko, Y., Baranov, D.: "Semograph" Information System as a Framework for Network-Based Science and Education. Smart Innovation, Systems and Technologies **75**, 263–272 (2017). https://doi.org/10.1007/978-3-319-59451-4_26
5. Boronnikova, N., Taleski, A., Belousov, K., Ryabinin, K.: Visual Representation of the Egocentric Field of the Speaker in the Macedonian Language (Experimental Study). Perm University Herald. Russian and Foreign Philology **10**, 13–27 (2018). https://doi.org/10.17072/2037-6681-2018-3-13-27
6. Chuprina, S., Nasraoui, O.: Using Ontology-based Adaptable Scientific Visualization and Cognitive Graphics Tools to Transform Traditional Information Systems into Intelligent Systems. Scientific Visualization **8**, 23–44 (2016)
7. Grandjean, M.: Introduction to Network Visualization with GEPHI. http://www.martingrandjean.ch (2013), `http://www.martingrandjean.ch/introduction-to-network-visualization-gephi/`, last accessed 12 Aug 2020
8. Huang, W., Luo, J., Bednarz, T., Duh, H.: Making Graph Visualization a User-Centered Process. Journal of Visual Languages and Computing **48**, 1–8 (2018). https://doi.org/10.1016/j.jvlc.2018.07.001
9. Jänicke, S.: Valuable Research for Visualization and Digital Humanities: A Balancing Act. In: Workshop on Visualization for the Digital Humanities, IEEE VIS 2016, Baltimore, Maryland, USA (2016)
10. Lee, B., Hurson, A.: Issues in Dataflow Computing. Advances in Computers **37**, 285–333 (1993). https://doi.org/10.1016/S0065-2458(08)60407-6
11. Pavlova, D., Garanovich, M.: Variability of Semantic Structure of Oral Spontaneous Monologues "About Myself" depending on "Gender" Factor. Nauchnyi dialog pp. 107–122 (2019). https://doi.org/10.24224/2227-1295-2019-5-107-122
12. Ryabinin, K., Belousov, K., Chuprina, S., Shchebetenko, S., Permyakov, S.: Visual Analytics Tools for Systematic Exploration of Multi-Parameter Data of Social Web-Based Service Users. Scientific Visualization **10**, 82–99 (2018). https://doi.org/10.26583/sv.10.4.07
13. Ryabinin, K., Chuprina, S.: High-Level Toolset for Comprehensive Visual Data Analysis and Model Validation. Procedia Computer Science **108**, 2090–2099 (2017). https://doi.org/10.1016/j.procs.2017.05.050
14. Shchebetenko, S.: Reflexive Characteristic Adaptations Explain Sex Differences in the Big Five: But not in Neuroticism. Personality and Individual Differences **111**, 153–156 (2017). https://doi.org/10.1016/j.paid.2017.02.013
15. Tominski, C., Abello, J., Schumann, H.: CGV – An Interactive Graph Visualization System. Computers & Graphics **33**, 660–678 (2009). https://doi.org/10.1016/j.cag.2009.06.002
16. Zelyanskaya, N., Belousov, K., Ichkineeva, D.: Naive Geography and Geopolitical Semiotics: the Semiotic Analysis of Geomental Maps of Russians. Semiotica **2017**, 235–253 (2017). https://doi.org/10.1515/sem-2016-0231