Fuzzy Systems and Data Mining VIII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220363

Ontology-Driven Visual Analytics Platform for Semantic Data Mining and Fuzzy Classification

Konstantin RYABININ [a,1], Roman CHUMAKOV [a], Konstantin BELOUSOV [a] and Mariia KOLESNIK [b]

[a] *Perm State University, Perm, Russia*
[b] *Perm Regional Museum, Perm, Russia*

Abstract. Visualization is claimed as one of the essential "V's" of Big Data since it allows presenting the data in a human-friendly way and is, therefore, a stepping-stone for the Big Data mining process. Visual analytics, in turn, ensures knowledge discovery out of the data through cognitive graphics and filtering capabilities. But to be efficient, visualization and analytics tools have to consider other Big Data "V's" by handling the large data volumes, keeping up with the data growth and changing velocity, and adapting to the variety of the data representation formats. We propose using ontology engineering methods to create a visual analytics platform controlled by an ontological knowledge base that describes supported data types, input formats, data filters, visual objects, and visualization algorithms, as well as available communication protocols and computing nodes, the platform modules can run on. This allows introducing new functions and distributed computation scenarios to the platform on the fly just by extending the underlying domain ontologies without changing the source code of the platform's core. The analytics flow inside this platform is described by task ontologies enabling semantic data mining process. As a result, seamless integration with different data sources is achieved, including plain files, databases, and even third-party soft- and hardware solvers. We demonstrate the viability of the approach proposed by solving several data mining and fuzzy classification problems, including the assessment of the citizens' regional identity according to the mental maps they draw and the reconstruction of ontogenesis of extinct synapsid *Titanophoneus potens* Efremov, 1938.

Keywords. Visual Analytics, Ontology Engineering, Semantic Data Mining, Fuzzy Classification, Sketch Maps, Paleontology

1. Introduction

One of the most significant challenges driving the Big Data initiatives is the data variety [1]. While other essential characteristics like volume and velocity can often be tackled by just increasing the computing power, handling the variety requires smart approaches to integrating different data sources, converting data formats, normalizing values, de-

[1]Corresponding Author: Konstantin Ryabinin, Perm State University, Bukireva Str. 15, 614068 Perm, Russia; E-mail: kostya.ryabinin@gmail.com.

feating corresponding uncertainties, etc. One of the possible ways to overcome the issue of variety lies in leveraging ontology engineering methods within the Big Data mining process. Being the formal models of domain-specific knowledge, ontologies can bridge the semantic gap between the actual data, the data mining (DM) tools, and the results of applying these tools [2]. Withal, the task ontologies provide a formalism to specify the DM flow by preserving the semantics of operations involved [2].

Besides volume, velocity, and variety, visualization is also recognized to be highly related to Big Data, because it is an essential part of the Big Data mining process. In this regard, ontologies can facilitate the assembling of individual visualization methods and tools into full-fledged visual analytics instruments [3].

Although ontologies are widely used in the Big Data mining process, still there is a lack of high-level self-service tools available for domain experts without requiring programming skills. To bridge this gap, we propose a unified architecture of an ontology-driven microservice-based client-server visual analytics platform and use it to implement a multi-purpose platform SciVi (`https://scivi.tools/`) capable of semantic DM. SciVi provides a high-level graphical user interface to declare DM and visual analytics pipelines in terms of data flows. The set of data processing operators for these pipelines is automatically generated by SciVi according to the ontological descriptions to meet the specifics of tasks being solved. Thanks to the declarative nature of SciVi extensibility, this platform is suitable not only for data analysts/scientists but also for researchers who develop or evaluate new DM algorithms.

In the present work, we discuss the SciVi platform architecture and demonstrate the viability of this platform by solving DM tasks from two very different application domains: Digital Humanities and Paleontology. The proposed DM pipelines involve fuzzy classification and visual analytics methods, which require adaptation and fine-tuning. To gain efficiency in this fine-tuning, we introduced new debug capabilities to the SciVi platform allowing us to monitor the intermediate results of the data processing pipeline.

2. Related Work

The concept of semantic DM was first introduced in 2009 [4] and 2010 [5] as an alloy of DM process and Semantic Web technologies to bring the machine-understandable meaning to the handled data. Since then, a lot of projects have adopted semantic DM to gain efficiency, context awareness, and flexibility of DM algorithms as reviewed by D. Dou et al. [2] and C. Sirichanya et al. [6]. P. K. Sinha et al. provided a systematic review of the most significant DM ontologies [7]. G. Amaral et al. summarized multiple benefits of ontologies to DM [8]. All the mentioned fundamental works and elaborate reviews form a consensus about high relevance, high demand, and wide prospects of semantic DM.

Another promising and demanded application of the principles declared by the Semantic Web is ontology-driven software development [9] that is a way to create adaptable software using ontology engineering methods and means. With the spreading of the microservice architectural style (MSA), research works arise devoted to the marriage of microservices and ontologies. For example, A. Versteden et al. "have shown that combining microservices with Semantic Technologies offers clean separation of concerns with nicely decoupled microservices" [10], ensuring "the services are talking about the same

content in the same way" [11]. G. Morais et al. provided the "first ontological formalization of MSA principles and anti-patterns concepts" [12]. A. Singer et al. proposed using lightweight ontologies within a system of microservices to represent descriptive information for "combining data from disparate sources and gathering new information" [13].

Aligning with the above works, we contribute to the semantic DM development by proposing the DM platforms' microservice-based architecture, in which both the DM pipeline and its underlying data processing operators are described by ontologies. Like in the state-of-the-art DM software like KNIME, Weka, RapidMiner, and Orange [14], we utilize data flow diagrams (DFDs) to visually represent DM pipelines enabling the users to compose them based on the data flow programming principles [15] within a high-level graphical user interface.

3. Multi-Purpose Ontology-Driven Data Flow Programming Platform

To achieve high flexibility and extensibility, SciVi adopts ontology-driven microservice architecture. The internal SciVi workflow is shown in Fig. 1.

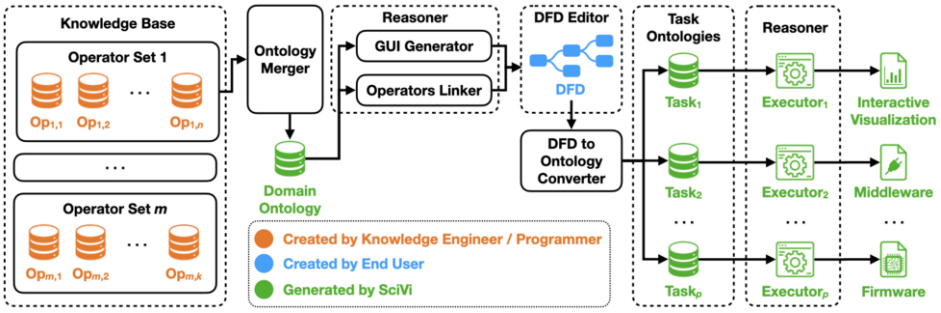

Figure 1. Internal SciVi workflow (arrows depict automatic transformations).

Each microservice is a software module denoted as an operator that is responsible for manipulating, visualizing, or analyzing the data. Each operator Op is formally specified by ontology adhering to the model $Op = \{I, S, O, \Phi\}$, where I is a set of typed inputs, S is a set of typed parameters (settings), O is a set of typed outputs, and Φ is a set of implementations. Each operators' implementation $\phi \in \Phi$ is tied to a particular computing resource, for example, to the thin client (in this case, the implementation is usually in JavaScript), to the server (usually in Python or in binary form), or to the specific computing device (usually in C/C++). Operators' ontologies are gathered into sets according to relevance to specific classes of tasks being solved by SciVi.

At startup, the user chooses the task class and the *Ontology Merger* module combines all the related operators' ontologies into the solid *Domain Ontology* that specifies, what functionality SciVi provides for solving the tasks of the chosen class. Traversing this ontology, *Reasoner* instructs *GUI Generator* to generate a graphical user interface (GUI) for each operator and *Operator Linker* to prepare an executable module for each operator's implementation. After that, the operators' palette is provided to the user and the user composes a DFD defining the desired data processing pipeline. In fact, this step involves visual programming within SciVi, but it does not require any hard skills from

the user. From their point of view, it is only a matter of chaining the high-level data processing stages and tuning these stages via the available settings using the GUI generated by SciVi. All the work related to type checking, type conversion, and operators' communication is fully automated by SciVi.

Once the DFD is done, the *DFD to Ontology Converter* transforms it into the set of *Task Ontologies*, which specify the particular operations to be performed by particular computing resources SciVi can reach in the active network [16]. The specified operations are derived from the DFD and, if needed, automatically supplemented by communication actions to transmit data between the involved computing nodes. The reasoner traverses the *Task Ontologies* to spawn so-called *Executors*. They are software containers for each involved computing node to run the chain of appropriate operators within. Each *Executor* generates a specific result, for example, *Interactive Visualization* to graphically depict the processed data, *Middleware* to communicate with third-party software or hardware systems, or *Firmware* to configure devices within the ecosystem of the Internet of Things [17].

It is worth stressing that the SciVi platform has no default built-in DM functions. Instead, it provides bus mechanisms to seamlessly chain microservices (operators), which are easily added, extended, or modified just by extending the underlying ontologies with new knowledge. Thanks to this, SciVi tackles the Big Data variety issue by providing tools to adapt to multifarious data sources from different application domains. At the same time, SciVi enables materializing Schneiderman's Mantra of visual analytics [18] by maintaining operators for the overviewing, zooming, and filtering the data, as well as for querying the data details.

When a new microservice (operator) is added to SciVi, it sometimes requires testing and debugging. In this work, we introduce the special debug operators to SciVi, which allow inspecting the data being transmitted through the DFD. The user can link these operators to output sockets of other operators within the DFD and during the pipeline execution, "watches" appear on screen just like in integrated development environments. These "watches" are automatically updated to keep track of the transmitted values.

4. Use Case 1: Fuzzy Classification of Mental Maps

A mental map is a summarized spatial experience of a human that can tell a lot about their regional identity. When analyzed from a group of people, mental maps are a rich data source to study regional state, public mood, etc. within Digital Humanities research. In our previous work [19], we proposed a modification of the weighted fuzzy pattern matching algorithm (WFPM) [20] to reveal the regional identity of mental maps. In the present work, we improved this modified algorithm by taking into account a bigger data set of maps, which are drawn by informants all over Russia using our digital map editor Creative Maps Studio (`https://creativemaps.studio/`).

The following modifications and novel features are introduced:

1. The disjunctive form of WFPM was changed to the conjunctive form because it reduces the chance of false positives in our case.
2. It was noticed that some objects on the maps are depicted by the informants in approximately the same way, regardless of which region they are from. Therefore, a reduction factor is used for such objects to decrease their weight in classification.

3. It was found that informants often depict objects of their residence region in a distinctive (more verbose and precise) way so that these objects have a non-zero membership function only for this region. Therefore, the weight of corresponding objects is adaptively increased to stress their significance for classification.

The implementation of our WFPM version can be found in the SciVi open source repository: https://github.com/scivi-tools/scivi.web/tree/master/lib/wfpm/. Along with the code, all the essential formulas and explanations are provided. We evaluated this algorithm using 205 maps drawn by the informants from 7 known residence regions. 185 of them were used as patterns (training set), and 20 as a test set. The average precision of classification is 92%, which is fairly good. However, the precision is non-uniform across regions: the best precision hits 100%, the worst one is 80%. The factors affecting it are to be revealed during future work. For now, one of the obvious factors is geographical distance: the maps from regions, which are located near to each other, are less distinguishable than the ones from the regions located far apart. This demonstrates the spatial experience locality and seems to be logical.

The visualization of the WFPM-based classification results is shown in Fig. 2. The regions outlined on the geographical map correspond to the patterns assembled. Those fill colors depict the membership degrees of the selected map (see the slider in the bottom) according to the given color scale. For each classified map, this visualization helps the analyst to guess where the map's author resides. For example, the map with ID "110-171-1" was most likely drawn by a person from Bashkortostan (the region filled red).

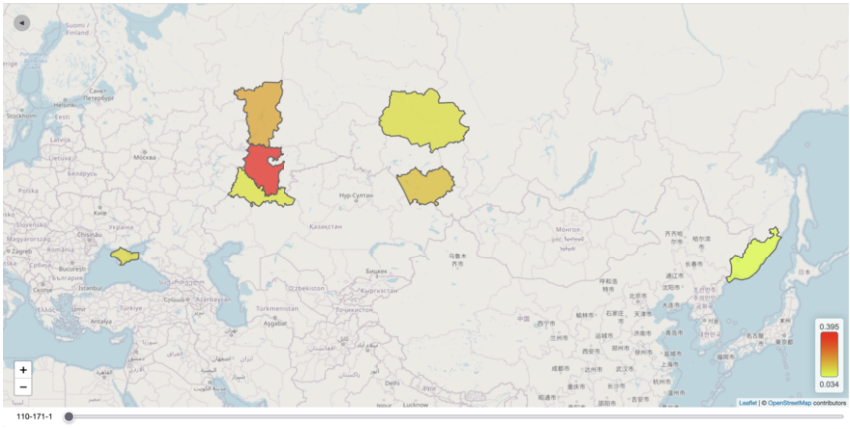

Figure 2. Matching the mental map against patterns (interactive view within SciVi is available online https://scivi.semograph.com/?preset=mmapsClass.json&start=true).

5. Use Case 2: Reconstruction of *Titanophoneus* Ontogenesis

The study of extinct animals' life cycle is a big problem in paleontology due to the lack of data, especially when ages like the Permian period are considered. Nevertheless, there are sometimes fossils available, representing different ontogenetic stages of the same or the related species. Having two or more of such fossils, one can try to model the entire ontogenetic process.

In our previous work, we reconstructed skull 3D models of young and adult *Titanophoneus potens* Efremov, 1938 individuals based on real fossils. Then we used SciVi to create a cyber-physical exhibit for a paleontological museum [21]. This exhibit demonstrates the ontogenesis of the *Titanophoneus* as a linear morphing from young to adult. In this work, we set up an experiment based on visual analytics, under which we extrapolate the ontogenesis beyond the known fiducial stages. The results are shown in Fig. 3. The variable t stands for interpolation/extrapolation factor: for $t \in [0;1]$, interpolation from young ($t = 0$) to adult ($t = 1$) individuals is rendered, and beyond this segment extrapolation takes place.

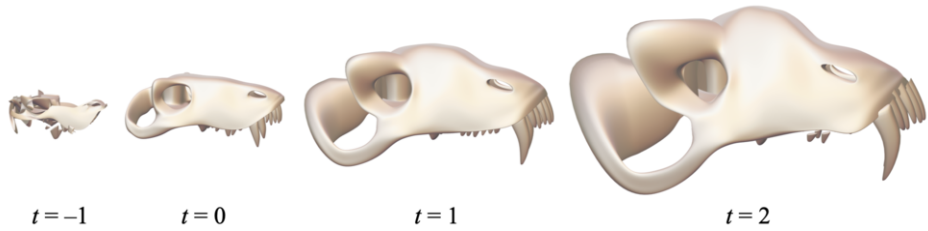

Figure 3. Extrapolating the *Titanophoneus* ontogenesis (interactive view within SciVi is available online https://scivi.semograph.com/paleo?preset=titanophone.json&start=true).

As seen in the figure, the model of "elderly" *Titanophoneus* ($t = 2$) looks pretty natural. Moreover, it resembles another synapsid from the same family (Anteosauridae), *Anteosaurus magnificus* Watson, 1921. This resemblance is weak evidence that the linear extrapolation is correct and *Titanophoneus potens* and *Anteosaurus magnificus* may have similar growth curves. In contrast, the model of "juvenal" *Titanophoneus* ($t = -1$) collapses. This phenomenon indicates that the growth curve of *Titanophoneus* is non-linear in the juvenile segment.

6. Conclusion

In this paper, we discuss the architecture and implementation of an ontology-driven microservice-based visual analytics platform SciVi that is suitable for solving different DM tasks from various application domains. To enable debugging of individual microservices and the entire pipeline, we introduce special operators that help to monitor the intermediate data flowing through the pipeline.

We demonstrate the viability of SciVi by solving the task of mental maps fuzzy classification revealing the regional identity of informants who have drawn the mental maps and the task of *Titanophoneus potens* ontogenesis reconstruction revealing non-linear nature of its aging changes. These are just two among multiple SciVi use cases, but the others are beyond this papers' scope due to the volume limitations. Currently, the SciVi platform is about to grow in four major directions: fostering the tools of advanced visual analytics (VASciVi Workbench), enabling experiments in virtual reality (VRSciVi Workbench), enabling brain-to-computer interfaces (NeuroSciVi Workbench), and supporting ontology-driven Edge Computing (EdgeSciVi Workbench). Each workbench reuses the main principles of SciVi and brings its own array of microservices, which act as operators joined into sets covering particular task classes.

Acknowledgments

This work was supported by Russian Science Foundation (grant number 20-18-00336).

References

[1] Bean R. Variety, Not Volume, Is Driving Big Data Initiatives; 2016. Visited on 28 Feb 2022. MIT Sloan Management Review. Available from: https://sloanreview.mit.edu/article/variety-not-volume-is-driving-big-data-initiatives/.

[2] Dou D, Wang H, Liu H. Semantic Data Mining: A Survey of Ontology-Based Approaches. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015); 2015. p. 244–251.

[3] Ryabinin K, Chuprina S. Development of Ontology-Based Multiplatform Adaptive Scientific Visualization System. Journal of Computational Science. 2015;10:370–381.

[4] Novak PK, Vavpetič A, Trajkovski I, Lavrac N. Towards Semantic Data Mining with g-SEGS. In: Proceedings of the 11th International Multiconference Information Society; 2009. .

[5] Liu H. Towards Semantic Data Mining. In: Proceedings of the 9th International Semantic Web Conference (ISWC 2010); 2010. Available from: http://iswc2010.semanticweb.org/pdf/448.pdf.

[6] Sirichanya C, Kraisak K. Semantic Data Mining in the Information Age: A Systematic Review. International Journal of Intelligent Systems. 2021;36(8):3880–3916.

[7] Sinha PK, Gajbe SB, Debnath S, Sahoo S, Chakraborty K, Mahato SS. A Review of Data Mining Ontologies. Data Technologies and Applications. 2021.

[8] Amaral G, Baião F, Guizzardi G. Foundational Ontologies, Ontology-Driven Conceptual Modeling, and Their Multiple Benefits to Data Mining. WIREs Data Mining and Knowledge Discovery. 2021;11(4).

[9] Pan JZ, Staab S, Assmann U, Ebert J, Zhao Y. Ontology-Driven Software Development. Springer; 2013.

[10] Versteden A, Pauwels E, Papantoniou A. An Ecosystem of User-Facing Microservices Supported by Semantic Models. In: Joint Proceedings of the 5th International Workshop on Using the Web in the Age of Data (USEWOD '15) and the 2nd International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES '15). vol. 1362 of CEUR Workshop Proceedings. CEUR-WS.org; 2015. .

[11] Versteden A, Pauwels E. State-of-the-Art Web Applications Using Microservices and Linked Data. In: Proceedings of the 4th Workshop on Services and Applications over Linked APIs and Data. vol. 1629 of CEUR Workshop Proceedings. CEUR-WS.org; 2016. .

[12] Morais G, Adda M. OMSAC – Ontology of Microservices Architecture Concepts. In: 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON); 2020. p. 293–301.

[13] Singer A, Wenzel K, Müller J, Langer T, Putz M. Ontology-Based Microservices Architecture; 2018. p. 508–511.

[14] Naik A, Samant L. Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. Procedia Computer Science. 2016;85:662–668.

[15] Sousa T. Dataflow Programming: Concept, Languages and Applications. In: Doctoral Symposium on Informatics Engineering. vol. 130; 2012. .

[16] Ryabinin K, Chuprina S, Labutin I. Tackling IoT Interoperability Problems with Ontology-Driven Smart Approach. Lecture Notes in Networks and Systems. 2021;342:77–91.

[17] Ryabinin K, Chuprina S. Ontology-Driven Edge Computing. Lecture Notes in Computer Science. 2020;12143:312–325.

[18] Shneiderman B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: Proceedings 1996 IEEE Symposium on Visual Languages; 1996. p. 336–343.

[19] Ryabinin K, Belousov K, Chumakov R. Ontology-Driven Data Mining Platform for Fuzzy Classification of Mental Maps. Frontiers in Artificial Intelligence and Applications. 2021;340:363–370.

[20] Dubois D, Prade H, Testemale C. Weighted Fuzzy Pattern Matching. Fuzzy Sets and Systems. 1988;28(3):313–331.

[21] Ryabinin K, Kolesnik M, Akhtamzyan A, Sudarikova E. Cyber-Physical Museum Exhibits Based on Additive Technologies, Tangible Interfaces and Scientific Visualization. Scientific Visualization. 2019;11(4):27–42.