doi 10.18699/bgrs2024-10-20

Development of a hybrid tool system for semi-automatic ontology creation

Lebedev M.[1]*, Savostyanov A.[1, 2, 3]

[1] *Novosibirsk State University, Novosibirsk, Russia*

[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[3] *Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia*

* *ml220501@gmail.com*

Key words: ontologies; automation; tool; automatic ontology creation; mental illnesses; data structurization

Motivation and Aim: Nowadays, the excess of the information on the web hinders research and finding relevant information. For example, when searching for papers that include keywords “Depression” and “Genes” there are approximately 12,000 results. And when searching using keywords “Proteins” and “Neurotransmitter systems” about 137,000 results are given. That means that finding data connections is difficult if not impossible.

Ontologies aim to fix this problem by providing data systematization for different fields. For example, ontologies are widely used in medicine, one of them being ANDDigest that helps find genes associated with certain organisms and illnesses, along with a lot of other information [1–3]. However, creating anthologies is a time-consuming and difficult process that requires a lot of specific expert knowledge. This led to the development of tools that automate certain parts of ontology creation. The problem with automatic approach is that it is usually based on deduction, rather than induction, which can provide false information.

These negatives of both approaches (manual and automated) can be solved by creating a hybrid system utilizing automation algorithm and expert knowledge.

Methods and Algorithms: In this work a 4-level ontology system will be used [4]. The four levels are: ontological knowledge, general theoretical knowledge, empirical knowledge, probabilistic knowledge. Automated ontology creation steps correlate with ontology system levels: term extraction for ontological knowledge, link extraction for empirical knowledge and axiomatics extraction for theoretical knowledge.

After studying existing ontology creation tools, it is clear that there are two approaches: linguistic and statistical [5, 6].

Linguistic tools are most commonly used for text preprocessing, as well as parsing part-of-speech and syntactic information. Later this information can be used for term and link extraction.

Statistical tools are also used for term and link extraction, however they use very different principals. Using frequency and statistic information lets these tools become language independent.

Results: At this stage, a basis for tool creation has been developed. Firstly, it includes the tool’s structure: it will consist of several modules (axiomatics creation, text parsing, term extraction and link extraction) with the first one being an expert module, others being automatic and the latter two being expert-verified. Secondly, the requirements for the tool have been formulated: Technology stack – Python will be used for the tool and one of the more universal ontology systems will be used for automatic modules

(OntoGain or text2onto) as this provides the ability to use both statistical and linguistic algorithms without having to switch ecosystems; Integration – the tool must be compatible with modern existing ontology editors (like Protege) by utilizing commonly used formats (for example, omn or rdf); Simplicity – the tool must be usable by a person not fluent in computer science, which implies have an intuitive user interface; Discreteness – different modules need to be able to function on their own to save time and make testing and ontology creation more effective. Lastly, the list of classes that will be used in a final test ontology has been created. It was done by hand using Protege in collaboration with a field expert, however in the future, a special module will be developed to simplify this process.

Conclusion: Currently, all the necessary preparations for development have been made and the next step will involve integrating automation algorithms in a tool and testing it on scientific papers on mental illnesses and their studying and causes.

Funding: The study is supported within the budget project of the ICG SB RAS No. FWNR-2022-0020.).

References

1. Ivanov R., Kazantsev F., Zavarzin E. et al. ICBrainDB: An Integrated Database for Finding Associations between Genetic Factors and EEG Markers of Depressive Disorders. *J Pers Med*. 2022;12(1):53. doi 10.3390/jpm12010053
2. Ivanov R., Zamyatin V., Klimenko A., Matushkin Y., Savostyanov A., Lashin S. Reconstruction and Analysis of Gene Networks of Human Neurotransmitter Systems Reveal Genes with Contentious Manifestation for Anxiety, Depression, and Intellectual Disabilities. *Genes (Basel)*. 2019;10(9):699. doi 10.3390/genes10090699
3. Ivanisenko T.V., Saik O.V., Demenkov P.S. et al. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics*. 2020;21(Suppl. 11):228. doi 10.1186/s12859-020-03557-8
4. Naydanov C.A., Palchunov D.E., Sazonova P.A. Model-theoretic methods of integration of knowledge extracted from medical documents. *Vestnik NSU. Series: Information Technologies*. 2015;13(3):29-41 (in Russian)
5. Zagorulko Yu.A., Sidorova E.A., Zagorulko G.B., Akhmadeeva I.R., Sery A.S. Automation of the development of ontologies of scientific subject domains based on ontology design patterns. *Ontologiya proektirovaniya*. 2021;11(4):500-520. doi 10.18287/2223-9537-2021-11-4-500-520 (in Russian)
6. Asim M.N., Wasim M., Khan M.U.G., Mahmood W., Abbasi H.M. A survey of ontology learning techniques and applications. *Database (Oxford)*. 2018;2018:bay101. doi 10.1093/database/bay101