Towards a semantic store of data mining models and experiments

Ilin Tolovski
Jožef Stefan International Postgraduate School & Jožef Stefan Institute
Ljubljana, Slovenia
ilin.tolovski@ijs.si

Sašo Džeroski
Jožef Stefan Institute & Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
saso.dzeroski@ijs.si

Panče Panov
Jožef Stefan Institute & Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
pance.panov@ijs.si

ABSTRACT

Semantic annotation provides machine readable structure to the stored data. We can use this structure to perform semantic querying, based on explicitly and implicitly derived information. In this paper, we focus on the approaches in semantic annotation, storage and querying in the context of data mining models and experiments. Having semantically annotated data mining models and experiments with terms from domain ontologies and vocabularies will enable researchers to verify, reproduce, and reuse the produced artefacts and with that improve the current research. Here, we first provide an overview of state-of-the-art approaches in the area of semantic web, data mining domain ontologies and vocabularies, experiment databases, representation of data mining models and experiments, and annotation frameworks. Next, we critically discuss the presented state-of-the-art. Furthermore, we sketch our proposal for an ontology-based system for semantic annotation, storage, and querying of data mining models and experiments. Finally, we conclude the paper with a summary and future work.

1. INTRODUCTION

Storing big amounts of data from a specific domain comes in hand with several challenges, one of them being to semantically represent and describe the stored data. Semantic representation enables us to infer new knowledge based on the one that we assert, i.e. the description and annotation of the data. This can be done by providing semantic annotations of the data with terms originating from a vocabulary or ontology describing the domain at hand. In computer and information science, ontology is a technical term denoting an artifact that is designed for a purpose, which is to enable the modeling of knowledge about some domain, real or imagined [15]. Ontologies provide more detailed description of a domain, first by organizing the classes into a taxonomy, and further on by defining relations between classes. With semantic annotation we attach meaning to the data, we can infer new knowledge, and perform queries on the data.

Data mining and machine learning experiments are conducted with faster pace than ever before, in various settings and domains. In the usual practice of conducting data mining experiments, almost none of the settings are recorded, nor the produced models are stored. These predicaments make for a research that is hard to verify, reproduce and upgrade. This is also in line with the FAIR (Findable, Accessible, Interoperable, Reusable) data principles, introduced by Wilkinson et al. [9]. Implementing these principles for the annotation, storing, and querying of data mining models and experiments will provide a solid ground for researchers interested in reproducing and reusing the results from the previous research on which they can build and improve.

In the literature, there exist some approaches that address some of these problems. In both ontology engineering and data mining community, there are approaches that aim towards describing the data mining domain, as described in Section 2. Furthermore, Vanschoren et al. [5] developed the OpenML system, a machine learning experiment database for storing various segments of a machine learning experiment such as datasets, flows (algorithms), runs, and completed tasks.

In other domains, such as life sciences, storing annotated data about experiments and their results is a common practice. This is mostly due to the fact that the experiments are more expensive to conduct, and require specific preparations. From the perspective of annotation frameworks, there are significant advances in these domains, such as The Center for Expanded Data Annotation and Retrieval (CEDAR) workbench [8] , and the OpenTox framework [11].

This paper is organized as follows. First, we make an overview of the state-of-the-art approaches in annotating, storing, and querying of models and experiments. Next, we critically assess these approaches and sketch a proposal for a system for annotating, storing and querying data mining models and experiments. Finally, we provide a summary and discuss the possible approaches for further work.

2. BACKGROUND AND RELATED WORK

The state-of-the-art in semantic annotation of data mining models and experiments provides very diverse research, ranging from domain-specific data mining ontologies, experiment databases, to new languages for deploying annotations in unified format. Here, we provide an introduction to the state-of-the-art in semantic web, ontologies and vocabularies, representations of data mining models and experiments, experiment databases, and annotation frameworks.

Semantic technologies. The Semantic Web is defined as an extension of the current web in which information is given well-defined meaning, enabling computers and people to work in cooperation [14]. The stack of technologies consists of multiple layers, however, in this paper we will focus on the ones essential for our scope of research. Resource Description Framework (RDF) represents a metadata data model for the Semantic Web, where the core unit of information is presented as a triple. A triple describes the subject by its relationship, which is what the predicate resembles, with the object. RDF files are stored in triple store (typically organized as relational or NoSQL databases [12]), on which we can perform semantic queries, by using querying languages such as SPARQL. Finally, ontology languages, such as Resource Description Framework Schema (RDFS) and Ontology Web Language (OWL), are formal languages used to construct ontologies. RDFS provides the basis for all ontology languages, defining basic constructs and relations, while OWL is far more expressive enabling us to define classes, properties, and instances.

Ontologies & vocabularies. Currently, there are several ontologies that describe the data mining domain. These include the OntoDM ontology [16], DMOP ontology [7], Expose [4], KDDOnto [1], and KD ontology [10]. MEX [2] is an interoperable vocabulary for annotating data mining models and experiments with metadata. In addition there have been developments in formalism for representing scientific experiments in general, such as the EXPO ontology [6].

Representation of models. With the constant development of new environments for developing data mining software, it is necessary to have a unified representation of the constructed data mining models and the conducted experiments. The first open standard was the Predictive Model Markup Language (PMML). For a period of time it provided transparent and intuitive representation of data mining models and experiments. However, due to the fast growth in the development of new data mining methods, PMML was unable to follow the pace and extend its more and more complicated specification. Its successor, the Portable Format for Analytics (PFA), was developed having the PMML's drawbacks as guidelines for improvement.

Experiment and model databases. Storing already conducted experiments in a well structured and transparent manner is essential for researchers to have available, verifiable, and reproducible results. An experiment database is designed to store large number of experiments, with detailed information on their environmental setup, the datasets, algorithms and their parameter settings, evaluation procedure, and the obtained results [3]. The state-of-the-art in storing setups and results is abundant with approaches and solutions in different domains. For example, OpenML[1] is the biggest machine learning repository of data mining datasets, tasks, flows, and runs, the BioModels[2] repository stores more than 8000 experiments and models from the domains of systems biology, and ModelDB[3] is an online repository for storing computational neuroscience models.

Annotation frameworks. When it comes to frameworks for (semi) automatically or manually annotating data, there are several solutions that exist outside of the data mining domain, which provide innovative approaches and good foundation for development in the direction of creating a software to enable ontology-based semantic annotation of models and experiments, their storage and querying. The CEDAR Workbench [13] provides an intuitive interface for creating templates and metadata annotation with concepts defined in the ontologies available at BioPortal[4]. On the other hand, OpenTox [11] represents domain specific framework that provides unified representation of the predictive modelling in the domain of toxicology.

3. CRITICAL ASSESSMENT

In this section, we will critically assess the presented state-of-the-art in Section 2 in the context of semantic annotation, storage and querying of data mining models and experiments.

The state-of-the-art in *ontology design* for data mining provides well documented research with various ontologies that thoroughly describe the domain from different aspects and can be used in various applications. OntoDM provides unified framework of top level data mining entities. Building on this, it describes the domain in great detail, containing definitions for each part of the data mining process. Because of the wide reach, it lacks a particular use case scenario. On the other hand, this same property makes this ontology suitable for wide range of applications where there is a need of describing a part of the domain.

Ontologies like EXPO and Exposé have a essential meaning in the research since the first one describes a very wide and important interdisciplinary domain, while the latter uses it as a base for defining a specific sub-domain. DMOP ontology describes the process of algorithm and model selection in the context of semantic meta mining. Both the KD ontology and KDDOnto describe the knowledge discovery process in the context of constructing knowledge discovery workflows. They differ mainly in the key concepts on which they were built. At the same time, the MEX vocabulary provides a lightweight framework for automating the metadata generation. Since it is tied with Java environment, it provides a library which only uses the MEX API and can also be implemented in other programming languages.

All in all, the current state of the art in ontologies for data mining provides a good foundation for development of applications which will be based on one or several of these ontologies. Given the wide of coverage they can be easily be combined in a manner to suit the application at hand.

In the area of *descriptive languages for data mining models and experiments*, one can see the path of progress in research. PMML was the first, ground-breaking, XML-based descriptive language. However, with the expansion of the data mining domain, several weaknesses of PMML emerged. The language was not extensible, users could not create chains of models, and it was not compatible with the distributed data processing platforms. Therefore, the same community started working on a new, extensible, portable

[1] https://www.openml.org/

[2] http://www.ebi.ac.uk/biomodels/

[3] https://senselab.med.yale.edu/modeldb/

[4] https://bioportal.bioontology.org/

language. Since its inception, the PFA format was intended to fill the small gaps that PMML had. Made up of analytic primitives, implemented in Python and Scala, it provides the users with more customizable framework, where they can create custom models, model chains, and implement them in a parallelized setting.

Storing and annotating experiments is of great significance in multiple scenarios. First, in domains where conducting the experiment is not a trivial task, i.e. the physical or financial conditions challenge the process, there needs to be a database where the setup and the findings of the experiment will be saved. For example, in BioModels.net we have two groups of experiments: Manually curated with structured metadata, and experiments without structure. The main drawback with this type of storage is the need for manual curation of the metadata. It is repetitive, time consuming task for which there is a strong need to be automated.

In the domain of neuroscience, ModelDB provides an online service for storing and searching computational neuroscience models. In this database, alongside the files that constitute the models, researchers also need to upload the code that defines the complete specification of the attributes of the biological system represented in the model, together with files that describe the purpose and application of the model. Therefore, researchers can search the database for models with specific applications describing biological systems.

OpenML provides a good framework for storing and annotating data mining datasets, experimental setups and runs, as well as algorithms. One particular drawback of OpenML is that it does not store the actual models that are produced from each experimental run, and one can not query the models. Furthermore, it's founded on relational-database which can not provide execution of semantic queries.

All in all, these three examples show significant advances in storing and annotating models and experiments. However, there is also a significant room for improvement in the direction of storing the models and experiments into NoSQL databases that are better suited for this task.

Finally, in the context of annotation tools the CEDAR Workbench and the OpenTox Framework provide a good insight in annotation frameworks. CEDAR enables the user to execute the annotations in modular manner by creating templates and adding elements to them. After curating the annotations, they can export the schemas either in JSON, JSON-LD, or RDF file. OpenTox [11] is also based on ontology terms and represents a complete framework that describes the predictive process in toxicology, starting with toxicity structures and ending with the predictive modelling.

4. A PROPOSAL FOR SEMANTIC STORE OF MODELS AND EXPERIMENTS

After analysing the previous and current research, we can conclude that despite the great achievements, there is a wide area for improvement in which we will contribute in the upcoming period by developing an ontology-based framework for storage and annotation of data mining models and experiments. In order to annotate a data mining experiment, we need to have complete information about the conditions in which that experiment was conducted. Namely, we need to have an annotated dataset, annotation of the algorithm and its parameter settings for the specific run of the experiment. Since one experiment usually consists of multiple algorithm runs we annotate each run separately, as well as each of the results from each of them. For annotating the results, we use the definitions of the performance metrics formalized in the data mining ontologies. A sketched example of the proposed solution is shown in Figure 1.

The proposed system for ontology-based annotation, storage, and querying of data mining experiments and models will consist of several components. The users will interact with the system through an user interface enabling them to run experiments on a data mining software, which will export models and experiment setups to a semantic annotation engine. For example, for testing purposes we plan to use CLUS[5] software for predictive clustering and structured output prediction, which generates different types of models and addresses different data mining tasks.

In the semantic annotation engine, the data mining models and experiments will be annotated with terms from the extended OntoDM ontology and then stored in a database. Once stored, the users will be able to semantically query the models and experiments in order to infer new knowledge. This will be done through a querying engine based on the SPARQL language, accessible through a user interface.

In order to perform annotation, we will extend the existing OntoDM ontology by adding a number of new terms, linking it to other domain ontologies, such as Exposé and EXPO. Linking OntoDM to these ontologies will extend the domain of OntoDM towards connecting the data mining entities that it already covers with new entities that describe the experimental setup and principles. With this we will obtain a schema for annotation of data mining models and experiments. The schema will then be used to annotate the data mining models and experiments through a semantic annotation engine. The engine will have to read the models and experiments from a data mining software system, annotate them with terms from developed schema and produce an RDF representation of the annotated data.

Furthermore, the RDF graphs will be stored in a triple store database. Since the data mining models and experiments differ a lot in their structure, we have yet to decide on the type of database in which we will store them. The data stored in this way is set for performing semantic queries on top of it. Therefore, we will develop a SPARQL-based querying enigne so the users can perform predefined or custom semantic queries on top of the storage base.

Finally, the format of the results is another point where we need to decide whether the results will be presented as RDF graphs, or in a different format (such as JSON) that is easier to interpret. This software package along with the storage will then be added as a module to the CLUS software, developed at the Department of Knowledge Technologies.

[5] http://sourceforge.net/projects/clus

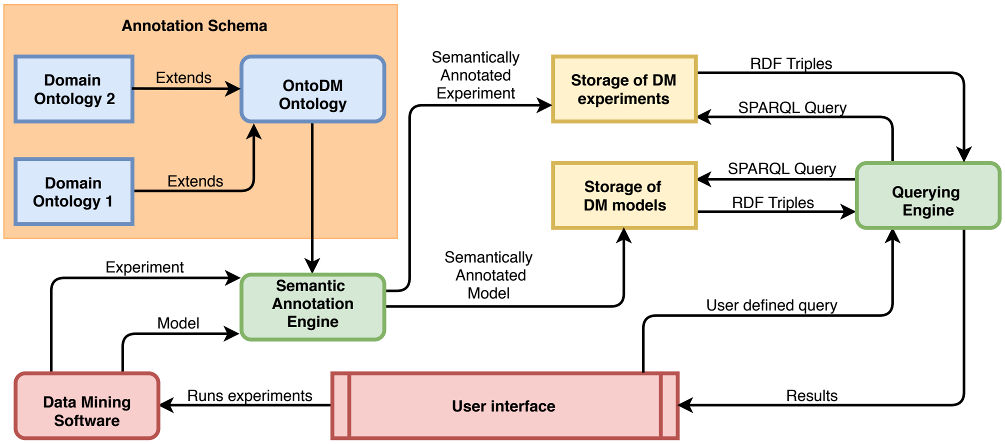

Figure 1. Schema of the proposed solution

5. CONCLUSION & FURTHER WORK

In this paper, we presented the state-of-the-art in annotation, storage and querying in the light of designing a semantic store of data mining models and experiments. We first gave an overview of semantic web technologies, such as RDF, SPARQL, RDFS, and OWL that provide a complete foundation for annotation and querying of data.

Furthermore, we critically reviewed the state-of-the-art ontologies and vocabularies for describing the domain of data mining provide detailed description of the domain of data mining and machine learning (OntoDM, Expose, KD Ontology, DMOP and KDDOnto, MEX). Next, we focused on experiment databases as repositories where the experiment datasets, setups, algorithm parameter settings, and the results are available for the performed experiments in various domains. Furthermore, we saw that annotation frameworks provide environments for (semi) automatically or manually annotating data, by discussing two frameworks from the domains of biomedicine and toxicology in order to analyze best practices present in those domains.

Finally, given the performed analysis of the state-of-the-art, we outlined our proposal for an ontology-based framework for annotation, storage, and querying of data mining models and experiments. The proposed framework consists of an annotation schema, a semantic annotation engine, and storage for data mining models and experiments with a querying engine, all of which will be controlled from an user interface. It will allow users to semantically query their data mining models and experiments in order to infer new knowledge.

In the future, we plan to adapt this framework for the needs of research groups or companies that conduct high volume of data mining experiments, enabling them to obtain a queryable knowledge base consisting of annotated metadadata for all experiments and produced models. This will enable them to reuse existing models on new data for testing purposes, infer knowledge based on past experimental results, all while saving time and computational resources.

Acknowledgements

The authors would like to acknowledge the support of the Slovenian Research Agency through the projects J2-9230, N2-0056 and L2-7509 and the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia through its scholarship program.

6. REFERENCES

[1] Claudia Diamantini et al. KDDOnto: An ontology for discovery and composition of kdd algorithms. *Towards Service-Oriented Knowledge Discovery (SoKD'09)*, pages 13–24, 2009.

[2] Diego Esteves et al. MEX Vocabulary: a lightweight interchange format for machine learning experiments. In *Proceedings of the 11th International Conference on Semantic Systems*, pages 169–176. ACM, 2015.

[3] Hendrick Blockheel et al. Experiment databases: Towards an improved experimental methodology in machine learning. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 6–17. Springer, 2007.

[4] Joaqin Vanschoren et al. Exposé: An ontology for data mining experiments. In *Towards service-oriented knowledge discovery (SoKD-2010)*, pages 31–46, 2010.

[5] Joaqin Vanschoren et al. Taking machine learning research online with OpenML. In *Proceedings of the 4th International Conference on Big Data, Streams and Heterogeneous Source Mining*, pages 1–4. JMLR. org, 2015.

[6] Larisa N Soldatova et al. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3(11):795–803, 2006.

[7] Maria C Keet et al. The Data Mining OPtimization Ontology. *Web Semantics: Science, Services and Agents on the World Wide Web*, 32:43–53, 2015.

[8] Mark A Musen et al. The Center for Expanded Data Annotation and Retrieval. *Journal of the American Medical Informatics Association*, 22:1148–1152, 2015.

[9] Mark D. Wilkinson et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 2016.

[10] Monika Záková et al. Automating knowledge discovery workflow composition through ontology-based planning. *IEEE Transactions on Automation Science and Engineering*, 8:253–264, 2011.

[11] Olga Tcheremenskaia et al. OpenTox predictive toxicology framework: toxicological ontology and semantic media wiki-based openToxipedia. In *Journal of biomedical semantics*, page S7, 2012.

[12] Olivier Curé et al. *RDF database systems: triples storage and SPARQL query processing*. Morgan Kaufmann, 2014.

[13] Rafael S Gonçalves et al. The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that Describe Scientific Experiments. In *International Semantic Web Conference*, pages 103–110. Springer, 2017.

[14] Tim Berners-Lee et al. The semantic web. *Scientific American*, 284:34–43, 2001.

[15] Tom Gruber. Ontology. *Encyclopedia of database systems*, pages 1963–1965, 2009.

[16] Panče Panov. *A Modular Ontology of Data Mining*. PhD thesis, Jožef Stefan IPS, Ljubljana, Slovenia, 2012.