

OntoDM: Towards an Ontology of Data Mining Investigations (Extended Abstract)

Panče Panov¹, Larisa N. Soldatova², Sašo Džeroski¹

¹ Department of Knowledge Technologies, Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
{Pance.Panov,Saso.Dzeroski}@ijs.si

² Computer Science Department, Aberystwyth University
Penglais, Aberystwyth, SY23 3DB, Wales, UK
lss@aber.ac.uk

1 Introduction

When one sets out to construct an ontology then, what one is doing is designing a representational artifact that is intended to represent the universals and relations amongst universals that exist, either in a given domain of reality (e.g data mining domain) or across such domains. The engineering of ontologies is still a relatively new research field and some of the steps in ontology design remain manual and can be considered as an art by itself. Recently there was a significant progress in automatic ontology learning [9], application of text mining [12], and ontology mappings [8]. However the construction of a good quality ontology with the use of automatic and even semi-automatic techniques still requires manual definition of the key upper level entities of the domain of interest. Good practices in ontology development are: following an upper level ontology as a template, the use of formally defined relations between the entities, not allowing multiple inheritances [15].

Researchers in the field of data mining have tried to construct ontologies describing data mining entities that were targeted to solve specific problems. Most of the developments are with the aim of automatic planning of data mining workflows [1, 17, 7, 5]. Some of the developments are concerned with description of data mining services on the GRID [4, 3]. The proposals for ontology of data mining so far were not based on upper level categories nor have used a predefined set of relations based on upper level ontology. Most of the semantic representations for data mining proposed so far are based on so called light-weight ontologies defining the semantics [10]. The reason is that the development of a heavy-weight ontology is difficult and time consuming. Light-weight ontologies are often shallow, without rigid relations between the defined entities, but they are relatively easy to develop by semi/automatic methods and they still greatly facilitate computer applications.. In contrast to many other domains, data mining requires elaborate inference over its entities, and hence requires rigid heavy-weight ontologies to improve KDD (Knowledge Discovery in Databases) process and to support more intelligent data mining methods.

In this work we propose an ontology of data mining that is based on the proposal for a general framework for data mining presented in [6]. Our ontology design takes into consideration the best practices in ontology engineering. We use an upper level ontology BFO (Basic Formal Ontology)³ to define the top level classes, the OBO Relational Ontology (RO)⁴ to define the semantics of the relationships between the DM entities, and provide is-a completeness and single is-a inheritance for all DM entities. We also developed our ontology in the most general fashion in order to be able to represent the complex entities in data mining that are becoming more and more popular research areas such as mining structured data and constraint-based mining.

In previous work [11] we presented an initial version of OntoDM sufficient for the representation of data mining tasks and complex data types. The version described in the current paper has been sufficiently updated compare to the previous version by the description data mining algorithms, scenarios and aligning the OntoDM top level structure with ontology of biomedical investigations (OBI)⁵.

2 Motivation

The motivation for developing an ontology of data mining is multi-fold. Firstly, as it was mentioned in the introduction, the area of data mining is developing rapidly and one of the most challenging problems deals with developing a general framework for data mining, mining structured data and data mining of biological and environmental data. By developing an ontology of data mining we are taking one step towards solving this problem. The ontology would define what are the basic entities in data mining: data types, data mining tasks, generalizations, algorithms, components of algorithms, constraints, etc. The ontology also defines relations between the entities. When the basic entities are defined, we can build upon them and define more complex entities, like data mining queries, scenarios and experiments, that are necessary in data mining applications.

Secondly, there exist several proposals for ontology of data mining but all of them are light-weight ontologies aimed at solving a particular problem in data mining, are of a limited scope and highly use-dependent. Data mining is a domain that needs a heavy-weight ontology where much attention is paid to the rigorous meaning of each class, semantically rigorous relations between classes and compliance to an upper level ontology and the domains of application (e.g., biology, environmental sciences).

Finally, there is a growing demand for formalized semantic representations of research results in all areas of science. Knowledge discovery and data mining applications are struggling with vast volumes of data and knowledge repositories of different not standardized formats describing research findings. Biology is leading the way in developing standards for recording and representation of

³ BFO: <http://www.ifomis.org/bfo>

⁴ RO: <http://www.obofoundry.org/ro/>

⁵ OBI: <http://obi-ontology.org/>

scientific data. For example already more than 50 journals require compliance of papers reporting microarray experiments to MIAME (the Minimum Information About a Microarray Experiment) standard. An ontology of data mining should follow this practice and define what is the minimum information required for the description of a data mining investigation.

3 OntoDM Design and Description

Our ontology of data mining (OntoDM) aims to provide a structured vocabulary of entities sufficient for the description of data mining scenarios and workflows. OntoDM aims to follow the OBO Foundry principles⁶ in ontology engineering that are widely accepted in the biomedical domains. The main OBO Foundry principles state that "the ontology is open and available to be used by all", "is in a common formal language", "includes textual definition of all terms", "uses relations which are unambiguously defined", "is orthogonal to OBO ontologies" and "follows a naming convention" [14]. In this way, OntoDM will be built on a sound theoretical foundation, will be compliant with other (e.g., biological) domains and can be widely re-usable. Our ontology intends to be compatible with other formalisms, to share and reuse already formalized knowledge.

OntoDM is expressed in OWL-DL and is being developed using the Protege ontology editor⁷. It consists of three main components: classes, a hierarchical structure (*is-a* relations) of classes and relations (other than *is-a* relations) between instances. All three major components are described in the following subsections. Availability: OntoDM is online at: <http://kt.ijs.si/panovp/OntoDM/>.

OntoDM is based on the proposal of a general framework for data mining by Džeroski [6]. The framework proposes a set of basic entities of data mining. The basic entities identified are the following (please consult [6] for a detailed description of the entities):

- **dataset**, which consists of **data items**;
- **datatype**, which can be **primitive** (**nominal**, **boolean**, **numeric**), or **structured** (**set**, **sequence**, **tree**, **graph**);
- **data mining task**, which includes **predictive modeling**, **pattern discovery**, **clustering** and **probability distribution estimation**;
- **generalization**, the output of a data mining algorithm, which can be: **predictive model**, **pattern**, **clustering**, **probability distribution**;
- **data mining algorithm**, which solves a data mining task and produces generalizations from a dataset and includes components of algorithms such as: **distance function**, **kernel function**, **refinement operator**;
- **function**, which can be: an **aggregation function**, **prototype function**, **evaluation function**, **cost function** etc;

⁶ OBO Foundry: http://ontoworld.org/wiki/OBO_foundry

⁷ Protege: <http://protege.stanford.edu>

- **constraint**, which include **evaluation** and **language constraint** (**hard constraint**, **soft constraint**, **optimization constraint**) and
- **data mining scenarios**, related to **queries** and **inductive queries**.

The entities listed above are used to describe different dimensions of data mining. These are all orthogonal dimensions and different combinations among these should be facilitated. Through combination of these basic entities, one should be able to describe most of the diversity present in data mining approaches today. One should be also able to derive new data mining approaches and insights. The identification of the basic entities in data mining is a key point in the development of a data mining query language, which would support the design and implementation of data mining algorithms, as well as their composition into knowledge discovery scenarios relevant for concrete applications. While the above basic entities were identified in the proposed framework, an ontological approach is still needed, so that all the relations between the entities could be formally defined and expressed in a formal language.

4 Conclusion and Further Work

In this work we present a proposal for an ontology of data mining. Unlike most existing approaches to constructing ontologies of data mining, our ontology *OntoDM* is a deep/heavy-weight ontology. It also follows best practices in ontology engineering, such as not allowing multiple inheritance of classes, using formally defined set of relations and using an upper level ontology.

The ontology *OntoDM* as presented here is in its early stages of development and hence much work remains to be done. We first need to populate the proposed classes of data mining entities with individuals, identify shortcomings of our ontology in the process and refine the structure of *OntoDM* as needed. While the current version of *OntoDM* is expressed in OWL-DL, the next level of development would require it to be translated into first-order logic and extended with axioms.

Formalizing the knowledge about the domain of data mining and building of a heavy weight ontology of data mining is a time and resource consuming task and should be a community effort. That is why one of the aims of our work is also to invite researchers from the area of data mining to contribute to the ontology by suggesting improvements in the definitions of the entities and by using the knowledge in the ontology in their applications. Our goal is to have a mature ontology of data mining that is sufficient and expressive enough to describe the current trends in data mining. This would be also a helpful step in developing standards for data mining and would lead towards an ontology of data mining investigations.

References

1. Abraham Bernstein, Foster Provost, and Shawndra Hill. Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. *IEEE Trans. on Knowl. and Data Eng.*, 17(4):503–518, 2005.
2. Peter Brezany, Ivan Janciak, and A Min Tjoa. *Data Mining with Ontologies: Implementations, Findings and Frameworks*, chapter Ontology-Based Construction of Grid Data Mining Workflows. IGI Global, 2007.
3. Mario Cannataro and Carmela Comito. A data mining ontology for grid programming. In *Proceedings of the 1st Int. Workshop on Semantics in Peer-to-Peer and Grid Computing (SemPGrid2003)*, pages 113–134, 2003.
4. Claudia Diamantini and Domenico Potena. Semantic annotation and services for kdd tools sharing and reuse. In *ICDMW '08: Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*, pages 761–770, Washington, DC, USA, 2008. IEEE Computer Society.
5. Sašo Džeroski. Towards a general framework for data mining. In Saso Džeroski and Jan Struyf, editors, *KDID*, volume 4747 of *Lecture Notes in Computer Science*, pages 259–300. Springer, 2006.
6. Alexandros Kalousis, Abraham Bernstein, and Melanie Hilario. Meta-learning with kernels and similarity functions for planning of data mining workflows. In Pavel Brazdil, Abraham Bernstein, and Larry Hunter, editors, *Proceedings of the Second Planning to Learn Workshop (PlanLearn) at the ICML/COLT/UAI 2008*, pages 23–28, 2008.
7. Lord Ph. Pocock M. Lister, A. and A. Wipat. Annotation of sbml models through rule-based semantic integration. In *Proceedings of Bio-ontologies SIG/ ISMB 2009*, 2009.
8. Evguenia Malaia. *Engineering ontology: domain acquisition methodology and practice*. VDM Verlag Dr. Muller, Saarbrucken, 2009.
9. Riichiro Mizoguchi. Tutorial on ontological engineering - part 3: Advanced course of ontological engineering. *New Generation Comput.*, 22(2), 2004.
10. Panče Panov, Sašo Džeroski, and Larisa Soldatova. OntoDM: An ontology of data mining. In *ICDMW '08: Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*, pages 752–760, Washington, DC, USA, 2008. IEEE Computer Society.
11. Philpp Cimiano (Eds.) Paul Buitelaar. *Ontology learning and population: bridging the gap between text and knowledge*. IOS Press. The Netherlands., 2008.
12. D. Schober, W. Kusnierczyk, S. E Lewis, and J. Lomax. Towards naming conventions for use in controlled vocabulary and ontology engineering. In *Proceedings of BioOntologies SIG, ISMB 2007*, pages 29–32, 2007.
13. Larisa N. Soldatova and Ross D. King. Are the current ontologies in biology good ontologies? *Nature Biotechnology*, 23(9):1095–1098.
14. Monika Zakova, Petr Kremen, Filip Zelezny, and Nada Lavrac. Planning to learn with a knowledge discovery ontology. In Pavel Brazdil, Abraham Bernstein, and Larry Hunter, editors, *Proceedings of the Second Planning to Learn Workshop (PlanLearn) at the ICML/COLT/UAI 2008*, pages 29–34, 2008.