OntoDM: An Ontology of Data Mining

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

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

Abstract

Motivated by the need for unification of the field of data mining and the growing demand for formalized representation of outcomes of research, we address the task of constructing an ontology of data mining. The proposed ontology, named OntoDM, is based on a recent proposal of a general framework for data mining, and includes definitions of basic data mining entities, such as datatype and dataset, data mining task, data mining algorithm and components thereof (e.g., distance function), etc. It also allows for the definition of more complex entities, e.g., constraints in constraint-based data mining, sets of such constraints (inductive queries) and data mining scenarios (sequences of inductive queries). Unlike most existing approaches to constructing ontologies of data mining, OntoDM is a deep/heavy-weight ontology and follows best practices in ontology engineering, such as not allowing multiple inheritance of classes, using a predefined set of relations and using a top level ontology.

1 Introduction

Ontologies [9] are content theories about the classes of individuals, properties of individuals, and relations between individuals that are possible in a specified domain of knowledge. They define the terms for describing our knowledge about the domain. An ontology of a domain is beneficial in establishing a common (controlled) vocabulary for the describing the domain of interest. This is important for unification and sharing of knowledge about the domain and connecting with other domains.

While knowledge discovery in databases (KDD) and data mining have enjoyed great popularity and success in recent years, there is a distinct lack of a generally accepted framework that would describe and unify the area of data mining. The present lack of such a framework is perceived as an obstacle to the further development of the field. In [29], Yang and Wu collected the opinions of a number of outstanding data mining researchers about the most challenging problems in data mining research. Among the ten topics considered most important and worthy of further research, the development of an unifying framework for data mining is listed first.

Researchers in the field of data mining have tried to construct an ontology for data mining targeted to solve specific problems. Most of the developments are with the aim of automatic planning of data mining workflows [3, 30, 16]. Some of the developements are concerned with description of data mining services on the grid [7, 6].

The problems of these ontologies are that they are constructed with a specific task in mind and not to describe the complete domain of data mining. Almost all proposed ontologies, with the small exception of the work presented in [30], deal with propositional data mining algorithms and do not take into account the existence of data mining algorithms for mining structured data. Also, all of the approaches are superficial in sense that they look at data mining algorithms as black boxes, describing them only by their inputs and outputs, not trying to describe the basic components of the algorithms.

The engineering of ontologies is still a relatively new research field, and many of the steps in ontology design are manual, and can be considered as an art by itself. Even though there is no well-developed theory for ontology design, there exist good practices in ontology development that should be taken into consideration when designing an ontology of a domain. The proposals for ontology of data mining so far were not based on top-level ontology categories nor have used a predefined set of relations based on top-level ontology. Most of the semantic representations for data mining proposed so far are based on so called light-weight ontologies defining the semantics [17]. The reason is that the development of heavy-weight ontologies 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 and they still

978-0-7695-3503-6/08 $25.00 © 2008 IEEE
DOI 10.1109/ICDMW.2008.62

greatly facilitate computer applications, particularly search engines. In contrast to many other domains, data mining requires elaborate inference over data, and hence requires rigid heavy-weight ontologies to improve KDD process and support more intelligent data mining methods.

Another research topic in data mining, that was identified to be important in [29], is data mining for biological and environmental problems. By constructing an ontology of data mining we would be able to connect to biological and environmental domains where the degree of ontology development is really high. In biology domains ontologies have been used for different purposes [23]: as controlled vocabularies (Gene Ontology [1]), for representing encyclopedic knowledge (Foundation model of anatomy FMA [19]), as a specification of an information model (MAGE-OM, MAGE-ML, MGED ontology [2]), for specification of a data interchange format (BioPax [1]) and representation of semantics of data for information integration (TAMBIS [26]).

In this paper we propose an ontology of data mining that is based on the proposal for a general framework for data mining presented in [11]. Our ontology design takes into consideration the best practices in ontology engineering such as not allowing multiple inheritance of classes, using a predefined set of relations and using a top level ontology. 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.

The rest of the paper is structured as follows: Section 2 provides the background for this work, in Section 3 we provide a detailed description of the OntoDM ontology and Section 4 discusses the possible uses and impacts of the ontology. In Section 5 we give a roadway for future research and development of the ontology.

2 Background

2.1 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 and scenarios, 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 a top 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 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 [5]. There are also standard initiatives to describe metabolomics experiments [20], proteomics experiments [27], etc. An ontology of data mining should follow this practice and define what is the minimum information required for the description of data mining investigations. For example, if a query uses a database, in order to be able to reproduce the results of the query, it is necessary to record not only what data it were used, but also access date, version of software etc.

2.2 Related work

In recent years, there is an increased need for formalized representations of the domain of data mining and formal representation of outcomes of research in general. In this sense, there exist proposals of ontologies of data mining. The research focus in most of the proposals are motivated by the need to have formalized description of data mining algorithms for constructing data mining workflows and description of the data mining services on the GRID. Other proposals deal with issues concerning the development of a framework for data mining. In this section we will briefly discuss the proposed approaches.

Data Mining Workflows In [3] the authors propose a prototype of an Intelligent Discovery Assistant (IDA) which provides users with systematic enumerations of valid data mining processes (sequences of data mining operators) and effective rankings of the processes by different criteria, in

[1] http://www.biopax.org/

order to facilitate the choice of data mining processes to execute to solve a concrete data mining task. This automated system takes the advantage of an explicit ontology of data mining operators (algorithms). The IDA determines the characteristics of the data and the desired mining result, and uses the ontology to search for and enumerate the data mining processes that are valid for producing the desired result and solving the data mining task, given the data. The ontology that is designed is a light-weight ontology that contains only a hierarchy of data mining operators divided into three main classes: preprocessing operators, induction algorithms and post processing operators. The leaves of the hierarchy are the actual operators. The operators are described by several properties: a specification of the conditions under which the operator can be applied, a specification of the operator's effect on the data mining process's state and the data, and estimations of the operator's effects when applied in a real situation. The ontology does not contain any information about the internal structure of the operators and the taxonomy is produced only according to the role that the operator has in the knowledge discovery process.

In [16] the authors build upon the work presented in [3] and propose an intelligent data mining assistant that combines planning and meta-learning for automatic design of data mining workflows. A knowledge driven planner relies on a knowledge discovery ontology (presented in the previous paragraph) to determine the valid set of operators for each step in the workflow. The probabilistic meta-learner is proposed for selecting the most appropriate operators by using relational similarity measures and kernel functions based on past data mining experiments.

The work in [30] also addresses the problem of semiautomatic design of workflows for complex knowledge discovery tasks. Similarly to the previous work [3], the idea is to automatically propose workflows for the given type of inputs and required outputs of the discovery process. This is done by defining a formal conceptualization of knowledge types and data mining algorithms in the form of an ontology, and a planning algorithm that accepts task descriptions expressed using the vocabulary of the ontology. The developed ontology in this case is also a light-weight ontology with its primary purpose to allow the planning algorithm to reason about which algorithms can be used to produce the results required by a specific data mining task. The authors propose two top classes of the ontology: <knowledge> and <algorithms>. The <knowledge> class captures the declarative elements of the knowledge discovery process. The <algorithms> class serves to define how pieces of knowledge are transformed into other pieces of knowledge. The ontology contains instances of several propositional algorithms and relational data mining algorithms.

Data mining services and resources on the GRID In [6] the authors introduce an ontology-based framework for automated construction of complex interactive data mining workflows as a means of improving productivity of Grid-enabled data systems. For this purpose they develop a data mining ontology which is based on concepts from industry standards like: predictive model mark-up language (PMML) [2], cross industry standard process for data mining (CRISP-DM) [10], WEKA [28] and Java data mining API [14]. The data mining ontology is a light-weight ontology built through the description of three essential classes of data mining components: DM-elements, DM-tasks and DM-services.

In the context of GRID programming in [7] the authors propose a design and implementation of an ontology of data mining. The motivation for building the ontology comes from the context of the author's work in Knowledge grid [8]. The main goals of the ontology are to allow the semantic search of data mining software and other data mining resources and to assist the user by suggesting the software to use on the basis of the user's requirements and needs. The proposed DAMON (DAta Mining ONtology) light weight ontology is built through a characterization of data mining software that is available. The top level classes of the ontology are as follows: <task>[3],<method>, <algorithm>, <software>, <suite>, <data source> and <human interaction>. They have been identified by using the following criteria: the data mining task performed by the software, type of methodologies that the software uses in the data mining process, the type of data sources the software works on and the degree of required interaction with the user. The top classes are extended in *is_a* hierarchies and the relations between them are expressed using limited set of non-standard relations.

Data Mining Framework On a seminar entitled Data Mining: The Next Generation [18] held in 2005, one of the discussion topics was compositionality of data mining operators. The participants of the seminar proposed to describe data mining operations in terms of their signatures, that is, in terms of the domain and range of the functions that they are computing. In order to structure the large number and variety of data processing and data mining operations, they proposed to organize the signatures into hierarchies. The purpose of hierarchies is to order the operators conceptually and also to get a better understanding of the commonalities and differences between them. In the higher level of the hierarchy, the signatures are described by general terms,

[2] `http://www.dmg.org/`

[3] In this paper we will denote the ontology class names with the notation <class_name> and the names of relations will be in italic font e.g. *is_a*. This notation, along with the names of classes and relations, is compliant with the naming convention presented in [21].

like patterns or models, and in the lower levels of hierarchy the signatures are specialized for certain types of patterns or models. The operators can be organized in several ways: according to the generic basic operations, according to the type of data or pattern domain or according to the type of operation itself. Although this proposal was made some time ago, we are not aware that these hierarchies were developed.

Description of scientific investigations There exist several formalisms for description of scientific investigations and outcomes of research. In this part we will focus on two proposals that are relevant for describing data mining investigations: OBI (Ontology for Biomedical Investigations)[4] and EXPO [25].

OBI aims to provide a standard for the representation of biological and biomedical investigations. It employs rigid logic and semantics, it uses a top level ontology BFO (Basic Formal Ontology)[5] and OBO RO (Relational Ontology)[6] to define the top classes and a set of relations. OBI defines occurrences (processes) and continuances (materials, instruments, qualities, roles, functions) relevant to biomedical domains. OBI is fully compliant with the existing formalisms in biomedical domains.

A generic ontology of experiments EXPO tries to define principal entities for the representation of scientific investigations. It uses SUMO[7] as a top level ontology and a minimized set of relations (*is-a*, *part-of* and *attribute-of*) in order to provide compliance with the existing formalisms. EXPO defines types of investigations: <computational investigation>, <physical investigation> and their principal components: <investigator>, <method>, <result>, <conclusion>.

3 The Proposed Solution: OntoDM

Our ontology of data mining (OntoDM) aims to provide a structured vocabulary of entities for the description of the domain of data mining. In the OntoDM ontology we consider a data mining investigation as a type of a scientific investigation and follow the philosophy of OBI and EXPO, extending their top level classes by data mining specific classes. In this way OntoDM will be designed with a sound theoretical foundation, will be compliant with other domains and will be 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 editor[8]. It consists of three main components: classes, a hierarchical structure (*is_a* relations) of classes and relations (other than *is_a* relations) between classes.

The version of the ontology presented in this paper is available online at: `http://kt.ijs.si/panovp/OntoDM/`.

3.1 Design Principles

OntoDM aims to follow the OBO Foundry principles[9] in ontology engineering that are wide spread in the biomedical domains. The main OBO Foundry principles say 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", "it is orthogonal to OBO ontologies" and "it follows a naming convention" [21].

The OntoDM ontology defines around 100 classes. All of the classes are extensions of top level classes that correspond and can be easily mapped to OBI and EXPO. The top level classes are as follows: <informational_entity>, <aggregate>, <procedure>, <process>, <quality>, <representation> and <role>.

Apart form the well-defined and known foundation relations *is_a* and *part_of*, OntoDM includes the relations from the OBO Relational ontology (RO) [22] *has_participant* and *has_agent*, the relations *has_input* and *has_output* that were recently introduced into OBI, the relations *has_role* and *has_quality* that are used in OBI, and relations *has_representation* and *has_information*, defined by Soldatova et.al [24].

3.2 Description of OntoDM

Basic entities. OntoDM is based on the proposal of a general framework for data mining by Džeroski[11]. The framework proposes a set of basic entities of data mining. The basic entities identified are:

- dataset;
- datatype (primitive and structured);
- data mining tasks (predictive modeling, pattern discovery, clustering, probability distribution estimation);
- generalizations (predictive model, pattern, a clustering, probability distribution);
- data mining algorithms;
- components of data mining algorithms (distance functions, kernel functions, features) and

[4] OBI: `http://obi.sourceforge.net/`
[5] BFO: `http://www.ifomis.org/bfo`
[6] RO: `http://www.obofoundry.org/ro/`
[7] SUMO: `http://www.ontologyportal.org/`
[8] Protege: `http://protege.stanford.edu`
[9] OBO Foundry: `http://ontoworld.org/wiki/OBO_foundry`

- constraints (evaluation and language).

The entities listed above are used to describe different dimensions of data mining. These are 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. The above entities were identified in the proposed framework, however an ontology approach is needed, so that all the relations between entities could be properly identified and expressed in a formal language. In this section we will focus on two basic entities in our ontology, <dataset> and <data_mining_task>. These two entities will be described in detail. Other basic entities will be briefly described by emphasizing their role in data mining and showing the basic relations between them.

Dataset Data is the most basic entity in data mining. Most typically, data is encountered in the form of a <dataset>. A data mining algorithm takes as input a dataset. An individual <data_example> in the dataset has its own structure, e.g., consists of values for several attributes. The attributes may be of different datatype and can take values from different ranges. It is usually assumed that all data items are of the same type and share the same structure.

The class <data_type> has two subclasses: <primitive_datatype> and <stuctured_datatype> (see Figure 1). Primitive data types e.g. real, integer, boolean, discrete, are usually taken as a starting point; and more complex (structured) data types are built by using a <datatype_constructor> which contains information on how the structured data type is constructed and what primitive datatypes are used. Structured datatypes include: tuples, sets, sequences, graphs etc. It is of crucial importance to be able to deal with structured data, as these are attracting an increasing amount of attention within data mining.

In OntoDM, we represent the class <dataset> as an extension of the top level class <aggregate> as presented in Figure 1. A dataset has a structure and has data examples that belong to it. This is represented by two properties: *part_of* <data_example> and *has_information* <datset_structure>.

The class <dataset_structure> gives information about the characteristics of a dataset (e.g. number of data examples and number of attributes). This is represented by two dataset properties: *has_number* <number_of_data_examples> and *has_number* <number_of_attributes>. Dataset properties relate dataset instances to concrete properties, in this case class of integer numbers.

The attributes are regarded as qualities of a dataset so the relation to the class <dataset> is expressed via the property *has_quality* <attribute>. Every attribute, in a given data example, has a specific value so we express this with the relation *has_information* <attribute_value>. The class <attribute> expresses the datatype of the attribute and this is done by defining a property *has_information* with the class <datatype>.

Because a <datatype> can have complex structure, we enable our ontology to deal with structured data. The <structured_datatype> class is connected with the <datatype_constructor> via the relation *has_information*. As attributes can have different roles in the dataset, we defined a class <attribute_role>. By using this design schema, we have the same treatment of primitive and structured data types that can appear as attributes in a dataset.

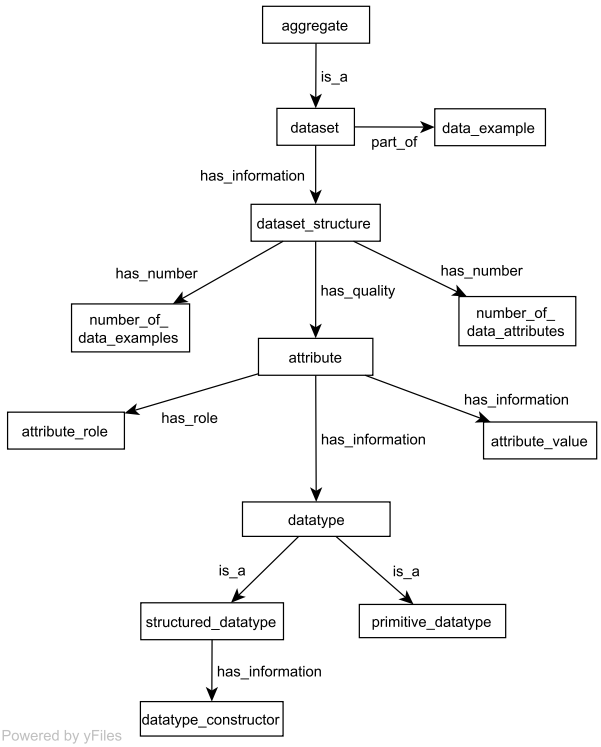

Figure 1. A representation of datasets in OntoDM

In Figure 2, the previously described classes and relations are illustrated on a real world dataset [12] by defining concrete instances of the classes. The authors [12], want

to classify diterpenes (chemical compounds) based on the NMR spectrum. As can be observed from the figure, the expression formalism of our ontology allows us to specify that the diterpenes dataset has a structure consisting of two attributes: 'spectrum' and 'compound class'. The attribute 'spectrum' is structured, while the attribute 'compound class' is a primitive attribute. In this formalism, we describe also the roles of each of the attributes. For the case of the data mining task, which is in this case predictive modeling, the role of the attribute 'spectrum' is descriptive and the role of the attribute 'compound class' is target.

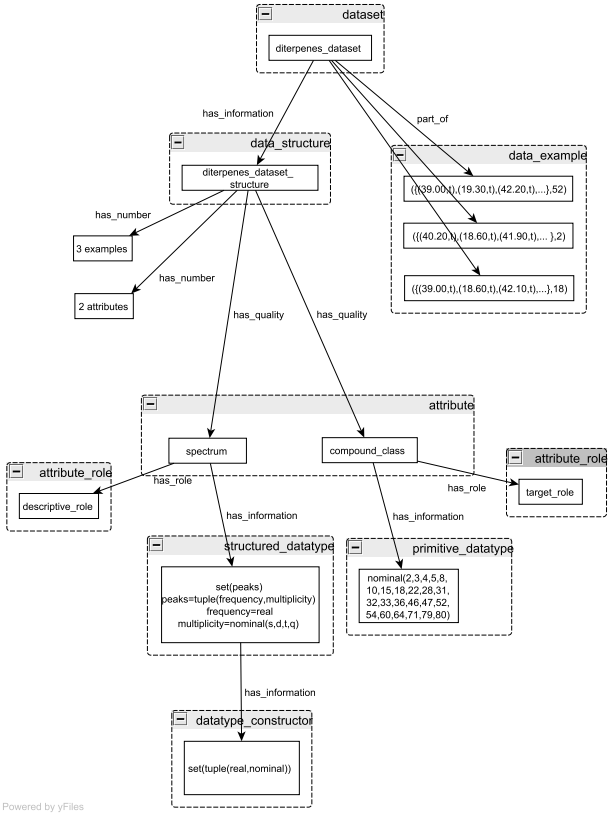

Figure 2. Example of representation of the Diterpenes dataset [12]

Data Mining Task The general task of data mining is to produce a some type of generalization from a given dataset. A plethora of data mining tasks have been considered in the literature. The basic data mining tasks that have been identified in [11] are as follows: estimation of the (joint) probability distribution, learning a (probabilistic) predictive model, clustering and pattern discovery. These basic data mining tasks are represented in OntoDM as separate classes via *is_a* relation (See figure 3).

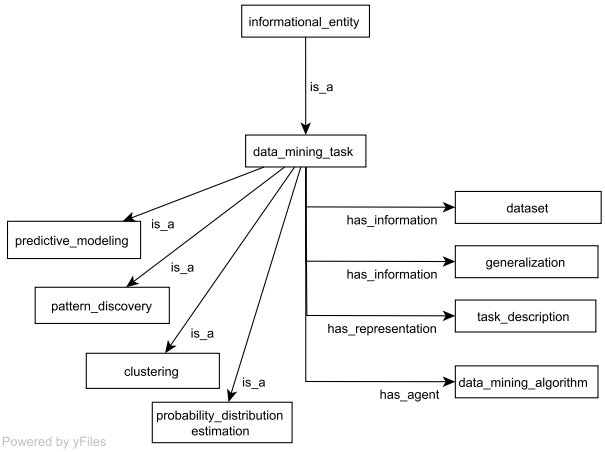

Figure 3. A representation of a data mining task in OntoDM

In our ontology the class <data_mining_task>, as presented in Figure 3 is an extension of top level class <informational_entity>. A data mining task has the property *has_representation* <task_description>. By this relation, we allow that a data mining task has a description that can be in the form of text or some other format. For a data mining task to be completely defined we need relations with the classes <dataset> and <generalization>. These relations are defined via the property *has_information*. In this case this means that a dataset and a generalization provide information for the data mining task. In the definition of a data mining task, we also need a active entity that transforms an data mining task description into action. This is expressed with the relation *has_agent* with the class <data_mining_algorithm>, which has an active role in this case.

Generalizations A Generalization is directly related to the data mining tasks and an output of a data mining algorithm. The class <generalization> is one of the fundamental classes in our ontology. The basic generalization types are: probability distributions, (probabilistic) predictive models, clusterings and patterns. These are defined as subclasses of generalization. All of the different types of generalizations are defined on a given type of data, except for predictive models, which are defined on a pair of data types.

Generalizations inherently have a dual nature. They can be seen as functions that take as input data points and map them to: probabilities, boolean values, class predictions or cluster assignments. On the other hand they can be treated as data structures and as such represented, stored and ma-

nipulated. This dual aspect is also represented in OntoDM.

Data Mining Algorithms Data mining algorithms are the core active elements in the data mining process. This is one of the reasons why it is important to identify the basic components thereof. The basic components of data mining algorithms are: distance functions, features, kernel functions and generality/refinement operators. The key notions in data mining have been studied extensively and are reasonably well understood for primitive data types. The basic idea of the unified approach in mining structured data as proposed in [11], is to derive the basic components of the algorithms for a complex data type (built through using type constructors) from information about the structure of the type and the basic components of the primitive data types. By introducing the basic components of data mining algorithms, we can also introduce the notion of a generic data mining algorithm that would be parameterized with selected basic components. In OntoDM, a data mining algorithm is represented both with defining the inputs and outputs of the algorithm, and by its internal structure with the basic components.

Constraints If we recall briefly from the previous paragraphs, data mining is concerned with finding generalizations that are valid in a given dataset. A generalization is said to be valid if it satisfies a given set of constraints. This is one of the important facts why constrains are regarded as basic entity in OntoDM. The constraints that are considered here depend heavily on the data mining task at hand. Given that generalizations have dual nature, i.e., have both a data and a function aspect, we can have constraints on each of these aspects: language constraints and evaluation constraints.

Language constraints concern the data part of the generalization. They can define a subclass or a sub-language of the class of generalizations or they may involve a language cost function on the data part of generalization. Evaluation constraints concern the function aspect of generalizations. They are usually boolean functions involving evaluation functions and comparing them to constant threshold. Evaluation function measure the validity of a generalization on a given dataset.

Depending on the output, language and evaluation constraints can be defined as: boolean constraints, optimization constraints and soft constraints. Boolean constraints are obtained by imposing a threshold on the value of a function. This can be a threshold on the language cost function or on an evaluation function. Boolean constraints are either satisfied or not. On the other hand, optimization constraints ask for generalizations that have a maximal/minimal value for a given cost or evaluation function. If we define language and evaluation constraints as boolean functions, we view them as hard constraints. The fact that constraints actually define what patterns are valid or interesting in data mining, and that interestingness is not a dichotomy, has lead to introduction of soft constraints.

In the current version of the ontology we have represented all of the described basic data mining entities. In the process of building the ontology we have also identified and defined a large number of supporting entities which are necessary for describing the domain of data mining. The task that will follow in the further development of the ontology will be to revise and refine the entities by looking at concrete instances and trying to describe them with the proposed formalism.

4 Discussion

Our proposal for an ontology of data mining includes descriptions of basic data mining entities. These basic entities can be used to define more complex entities that are of importance especially in applications of data mining.

The concept of an inductive database [15] employs a database perspective on knowledge discovery, where the knowledge discovery process is composed of query sessions. In this case ordinary queries can be used to access and manipulate the data, while inductive queries (data mining queries) can be used to generate (mine), manipulate and apply generalizations. This is why it is important to represent the complex entity query in our ontology, and this is possible because all of the basic entities of data mining have been identified and represented.

Real life applications of data mining typically require interactive sessions and involve formulation of a complex sequence of inter-related inductive queries, which we call a KDD scenario [4]. KDD scenarios can be described at different level of detail and precision and can serve multiple purposes. At the most detailed level of description, KDD scenarios can serve to document the exact sequence of data mining operations undertaken by a human analyst on a specific task. This would facilitate, for example, the repetition of the entire sequence after an erroneous data entry has been corrected in the source data. At higher level of abstraction, the scenarios enable the re-use of already performed analyses,e.g., on a new dataset of the same type. The explicit storage and manipulation of scenarios would greatly facilitate the KDD process in whole, reduce human effort and thus alleviate a major bottleneck in applying KDD in practice. Our proposed ontology can be used for formalizing and describing KDD scenarios.

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 be a helpful step in developing standards for data mining.

5 Summary and Future work

In this paper 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 a predefined set of relations and using a top level ontology.

OntoDM is based on a recent proposal of a general framework for data mining, and includes definitions of basic data mining entities, such as datatype and dataset, data mining task, data mining algorithm and components thereof (e.g., distance function), etc. It also allows for the definition of more complex entities, e.g., constraints in constraint-based data mining, sets of such constraints (inductive queries) and data mining scenarios (sequences of inductive queries). OntoDM is general-purpose and has not been designed with a specific use in mind: Rather, it can be used to support a number of relevant activities, such as describing data mining services and resources, data mining experiments/investigations, as well as data mining scenarios/workflows.

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, 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: This is needed to suport reasoning about OntoDM entities (e.g., about roles, which have a crucial meaning in OntoDM). Finally, we need to transform our current effort of developing OntoDM into a collaborative community effort.

Acknowlegements

This work is supported by the IST-FP6-516169 project Inductive Queries for Mining Patterns and Models (IQ).

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.

[2] C. A. Ball and A. Brazma. MGED standards: work in progress. *Omics : a journal of integrative biology*, 10(2):138–144, 2006.

[3] A. Bernstein, F. Provost, and S. 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.

[4] J.-F. Boulicaut, M. Klemettinen, and H. Mannila. Modeling KDD processes within the inductive database framework. In *Data Warehousing and Knowledge Discovery*, pages 293–302, 1999.

[5] A. Brazma et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, 29:365–371, December 2001.

[6] P. Brezany, I. Janciak, and A. M. Tjoa. *Data Mining with Ontologies: Implementations, Findings and Frameworks*, chapter Ontology-Based Construction of Grid Data Mining Workflows. IGI Global, 2007.

[7] M. Cannataro and C. Comito. A data mining ontology for grid programming. In *Proceedings of the 1st Int. Workshop on Semantics in Peer-to-Peer and Grid Computing (SemP-Grid2003)*, pages 113–134, 2003.

[8] M. Cannataro and D. Talia. The knowledge grid. *Commun. ACM*, 46(1):89–93, 2003.

[9] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins. What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 14(1):20–26, 1999.

[10] P. Chapman, R. Kerber, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth. The CRISP-DM process model. Discussion Paper, March 1999. http://www.crisp-dm.org.

[11] S. Džeroski. Towards a general framework for data mining. In S. Dzeroski and J. Struyf, editors, *KDID*, volume 4747 of *Lecture Notes in Computer Science*, pages 259–300. Springer, 2006.

[12] S. Džeroski, S. Schulze-Kremer, K. R. Heidtke, K. Siems, and D. Wettschereck. Applying ILP to diterpene structure elucidation from ^{13}c NMR spectra. In S. Muggleton, editor, *Inductive Logic Programming Workshop*, volume 1314 of *Lecture Notes in Computer Science*, pages 41–54. Springer, 1996.

[13] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with DOLCE, 2002.

[14] M. F. Hornick, E. Marcadé, and S. Venkayala. *Java Data Mining: Strategy, Standard, and Practice: A Practical Guide for architecture, design, and implementation (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2006.

[15] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. *Comm. Of The Acm*, 39:58–64, 1996.

[16] A. Kalousis, A. Bernstein, and M. Hilario. Meta-learning with kernels and similarity functions for planning of data mining workflows. In P. Brazdil, A. Bernstein, and L. Hunter, editors, *Proceedings of the Second Planning to Learn Workshop (PlanLearn) at the ICML/COLT/UAI 2008*, pages 23–28, 2008.

[17] R. Mizoguchi. Tutorial on ontological engineering - part 3: Advanced course of ontological engineering. *New Generation Comput.*, 22(2), 2004.

[18] R. Ramakrishnan, R. Agrawal, J.-C. Freytag, T. Bollinger, C. W. Clifton, S. Dzeroski, J. Hipp, D. Keim, S. Kramer, H.-P. Kriegel, U. Leser, B. Liu, H. Mannila, R. Meo, S. Morishita, R. Ng, J. Pei, P. Raghavan, M. Spiliopoulou, J. Srivastava, and V. Torra. Data mining: The next generation. In R. Agrawal, J. C. Freytag, and R. Ramakrishnan, editors, *Perspectives Workshop: Data Mining: The Next Generation*, number 04292 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2005. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

[19] C. Rosse and J. L. V. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *J. of Biomedical Informatics*, 36(6):478–500, December 2003.

[20] S.-A. Sansone at al. Metabolomics standards initiative - ontology working group. work in progress. *Metabolomics*, 3(3):249–256, 2007.

[21] 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.

[22] B. Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5), 2005.

[23] B. Smith and N. Shah. Ontologies for biomedicine - how to make them and use them. Tutorial notes at ISMB/ECCB 2007, 2007.

[24] L. N. Soldatova, W. Aubrey, R. D. King, and A. Clare. The exact description of biomedical protocols. *Bioinformatics*, 24(13), 2008.

[25] L. N. Soldatova and R. D. King. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3(11):795–803, 2006.

[26] R. D. Stevens, P. G. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. Paton, C. A. Goble, and A. Brass. Tambis: Transparent access to multiple bioinformatics information sources. *Bioinformatics*, 16:200–0, 2000.

[27] C. F. Taylor at al. The minimum information about a proteomics experiment (miape). *Nature Biotechnology*, (25):887 – 893, 2007.

[28] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.

[29] Q. Yang and X. Wu. 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making*, 5(4):597–604, 2006.

[30] M. Žáková, P. Kremen, F. Železný, and N. Lavrač. Planning to learn with a knowledge discovery ontology. In P. Brazdil, A. Bernstein, and L. Hunter, editors, *Proceedings of the Second Planning to Learn Workshop (PlanLearn) at the ICML/COLT/UAI 2008*, pages 29–34, 2008.