Chemical Knowledge Representation with Description Graphs and Logic Programming

Despoina Magka
Dept. of Computer Science,
University of Oxford,
Parks Road, OX1 3QD, UK
desmag@cs.ox.ac.uk

Boris Motik
Dept. of Computer Science,
University of Oxford,
Parks Road, OX1 3QD, UK
boris.motik@cs.ox.ac.uk

Ian Horrocks
Dept. of Computer Science,
University of Oxford,
Parks Road, OX1 3QD, UK
ian.horrocks@cs.ox.ac.uk

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms**]: Representations

General Terms

Theory

1. MOTIVATION

OWL 2 is commonly used to represent objects with complex structure, such as complex assemblies in engineering applications, human anatomy or the structure of chemical molecules [2]. Towards that direction, the European Bioinformatics Institute (EBI) has developed the ChEBI ontology as a public dictionary of *molecular entities*, which is used to ensure interoperability of applications supporting tasks such as drug discovery. In order to automate the classification of molecules, ChEBI descriptions have been translated into OWL and then classified using state of the art Semantic Web reasoners. While this has uncovered numerous implicit subsumptions between ChEBI classes, the usefulness of the approach was somewhat limited by a fundamental inability of OWL 2 to correctly represent the structure of complex molecular entities. OWL 2 exhibits a so-called *tree-model* property, which prevents one from describing non-tree-like relationships using OWL 2 schema axioms. For example, OWL 2 axioms can state that butane molecules have four carbon atoms, but they cannot state that the four atoms in a cyclobutane molecule are arranged in a ring.

In our previous work, we considered a combination of OWL 2, rules and *description graphs* (DGs) [4]—a graphical notation for describing non-tree-like structures. As reported in [2], however, DGs solved only some of the problems related to the representation of structured objects, and our subsequent discussions with EBI researchers have revealed a number of drawbacks.

The most important weakness identified in [4] is the inability to describe structures defined by the *absence* of certain characteristics. For example, an inorganic molecule is commonly described as 'a molecule *not* containing a carbon atom', which can then be used to classify water as an inorganic molecule. Producing the required entailment using the approach from [4] is very cumbersome: apart from stating that 'each water molecule consists of one oxygen and two hydrogen atoms', one must additionally state that 'these three atoms are the only atoms in a water molecule' and that 'neither hydrogen nor oxygen atoms are carbon atoms'.

SWAT4LS-2011 December 7-9, 2011 London, UK

2. SUGGESTED FORMALISM

In response to this criticism, we suggest a radically different approach to modelling complex objects via an LP-based formalism that we call Description Graph Logic Programs (DGLP). At the syntactic level, our approach combines OWL 2 RL, Description Graphs (DGs) and rules. OWL 2 RL axioms are translated into logic programming (LP) rules. A DG abstracts the structure of a complex object by means of a directed labeled graph. For example, Figure 1 illustrates a DG that represents the structure of a cyclobutane molecule. DGs serve as syntactic constructs and can be translated into logic programs with function symbols. Additionally, the knowledge base may contain rules with negation as failure in their bodies. By interpreting negation under the closed-world assumption we are able to capture conditions based on the absence of information; we can thus express in a natural way chemical classes that are based on the exclusive composition of molecules. For instance, negation as failure enables us to represent hydrocarbons which are molecules consisting entirely of hydrogens and carbons.

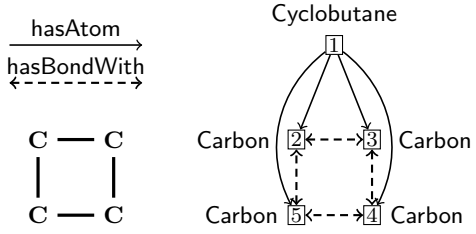

Figure 1: Chemical graph and DG for cyclobutane

Unfortunately, logic programs with function symbols can axiomatise infinite non-tree-like structures, so reasoning with DGLPs is trivially undecidable. Our goal, however, is not to model arbitrarily large structures, but to describe complex objects up to a certain level of granularity. For example, acetic acid has a carboxyl part, and carboxyl has a hydroxyl

No mol.	No rules	Loading time	T_1	T_2	T_3	T_4	T_5	Total time
10	1417	2.08	< 0.01	< 0.01	< 0.01	0.36	0.02	2.47
20	5584	8.35	< 0.01	< 0.01	0.02	2.07	0.21	10.66
30	8994	11.35	0.01	< 0.01	0.03	2.23	0.23	13.85
40	14146	16.14	0.01	< 0.01	0.04	2.58	0.29	19.06
50	21842	23.11	0.01	0.01	0.06	3.55	0.41	27.15
60	55602	168.71	0.04	0.02	0.51	109.88	21.68	300.84
70	77932	239.06	0.06	0.03	0.75	172.14	35.08	447.12

T_1: hydrocarbons, T_2: inorganic molecules, T_3: molecules with exactly two carbons, T_4: molecules with a four-membered ring, T_5:molecules with a benzene ring

Table 1: Times for deciding chemical classes of ChEBI molecules

part, but hydroxyl does not have an acetic acid part (see Figure 2). To address undecidability, several syntactic acyclicity conditions have been suggested such as weak acyclicity [1] or super-weak acyclicity [3]; these conditions ensure termination of the reasoning algorithms by examining the syntactic structure of the program's rules. However, these conditions rule out some very simple and intuitive logic programs, such as the one that represents the acetic acid structure. As a remedy, we present a novel *semantic acyclicity* condition to ensure decidability. In particular, we require the modeller to specify a (possibly empty) ordering on DGs that, intuitively, describes which DGs are allowed to imply existence of other DGs. Using a suitable test, one can then check whether implications between DGs are acyclic and hence whether DGs describe structures of bounded size only. We show that reasoning for semantically acyclic DGLPs with *stratified* negation is decidable. The resulting semantic acyclicity condition allows for the modelling of naturally-arising molecular structures, such as acetic acid, that would be precluded by existing syntax-based acyclicity conditions [1, 3].

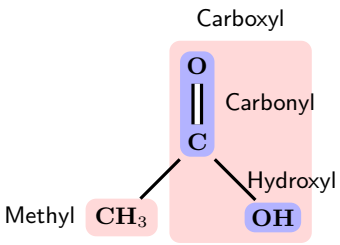

Figure 2: The chemical graph of acetic acid

We thus obtain a novel expressive and decidable formalism that is well suited for modelling objects with complex structure, such as chemical molecules, and which can enable automatic classification of chemical entities based on descriptions of their properties and structure. Full definitions and proofs of all technical results can be found in the technical report available online.[1]

3. IMPLEMENTATION

In order to test the feasibility of our approach in practice, we have developed a prototypical implementation based on the XSB system.[2] The objective of our implementation was to gain some insight into the capabilities and scalability of our approach by performing experiments using test data extracted from the ChEBI ontology.

The chemical classes that we modelled appear in the legend of Table 1. We tested the scalability of our prototype by evaluating logic programs ranging in size from 10 to 70 molecules. For each LP of different size we executed five tests ($\mathsf{T}_1 - \mathsf{T}_5$); each test corresponds to a chemical class and determines which molecules are classified under this class. Table 1 exhibits CPU timings for loading the LP rules to XSB and for executing the tests.

All the logic programs were found to be acyclic. Additionally, we verified that the subsumptions were computed as expected: for instance, according to the answers acetylene has exactly two carbons, cyclobutane has a four-membered ring and dinitrogen is inorganic. None of these inferences can be derived using the approach from [4] due to the lack of negation as failure or using OWL due to the tree-model property. Finally, we observe that all tests were accomplished in a reasonable amount of time: no test required more than a few minutes. Given the prototypical character of our application, we consider these results to be encouraging and we take them as evidence of practical feasibility of our approach. Additionally, performance could be greatly improved by e.g. loading the LP rules that represent *one molecule only* and then testing for its subsumers. As a consequence, we are optimistic that the current implementation may serve as a basis for building a fully-scalable chemical classification system in the future.

[1] `http://www.cs.ox.ac.uk/isg/people/despoina.magka`
[2] `http://xsb.sourceforge.net/`

4. ACKNOWLEDGMENTS

This work was supported by the EU FP7 project SEALS and by the EPSRC projects ConDOR, ExODA and LogMap.

5. REFERENCES

[1] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data exchange: Semantics and Query Answering. In *ICDT*, pages 207–224, 2003.

[2] J. Hastings, M. Dumontier, D. Hull, M. Horridge, C. Steinbeck, U. Sattler, R. Stevens, T. Hörne, and K. Britz. Representing Chemicals using OWL, Description Graphs and Rules. In *OWLED*, 2010.

[3] B. Marnette. Generalized Schema-Mappings: from Termination to Tractability. In *PODS*, 2009.

[4] B. Motik, B. C. Grau, I. Horrocks, and U. Sattler. Representing Ontologies Using Description Logics, Description Graphs, and Rules. *Artif. Int.*, 173:1275–1309, 2009.