Fourth IEEE International Conference on eScience

Rule-based Classification Systems for Informatics

B. Krishnamurthy[a], T. Malik[b], S. Stamatis[a], V. Venkatasubramanian[a], J. Caruthers[a]
(a) Dept. of Chemical Engineering and (b) Cyber Center
Purdue University
West Lafayette, IN
bkrishna, tmalik,* sstamati, venkat, caruther@purdue.edu

1 Motivation

Classification of data is a central process in scientific analysis. For instance, in Chemistry, molecule classification is necessary to determine the family of molecules that participate in reaction [1]. In Biology, classification of proteins is central to develop an understanding of the molecular biology of an organism [6]. Traditionally, classification is performed by human experts who have the unique ability to recognize functional properties that are necessary and sufficient to place complex structures and phenomena into a particular class or group. However, classification is a time-consuming process and many academic institutions can no longer support large teams of scientists required for such activities [6]. Automated classication methods can quickly classify large volumes of data into detailed categories. Such methods when combined with expert knowledge of domain scientists can provide a complete classification system. In this paper, we use Chemistry as our domain to describe and provide a solution to the classification problem.

2 The Chemistry Classification Problem

In Chemistry, molecules are classified based on a variety of criteria. The most common method of classifying molecules is based on sub-structure search. For example, a molecule is classified as a styrene if the styrene skeleton is present in the structure. Another equally important method of classifying molecules is if molecules satisfy a set of rules on either the structure or its properties. For instance, the simplest way to detect a hydrocarbon is by looping over all atoms present to check if they are carbon or hydrogen. The corresponding sub-structure search will be complex and tedious; pattern 'C' or pattern 'H' are true and negate for all other atoms in the Periodic Table. Molecules are also classified based on their properties. For example, molecules more volatile than water are those that have their boiling point below $100°C$. Such a rule is not on the structure of the molecule but on its boiling point property. Thus classification in Chemistry is a combination of substructure search along with a set of rules on either the structure of the molecule or its properties.

Chemists also classify molecules based on role played by a molecule in a particular reaction. For instance, 1-hexene is a monomer in a polymerization reaction but is a fuel in a combustion reaction. Currently, we do not consider such molecules which require an experimental context for classification.

3 ChES: An Automated Classifier

We present ChES[4], an automated classifier for chemical data. ChES combines human understanding through an ontology and the diversity in classification types through a rule based system to classify complex molecular compounds. ChES is a fast and reproducible framework whose classification capabilities often times surpasses those of human experts. It is general in that when combined with a domain specific ontology and rules, it can be used to classify data for a scientific domain. ChES is the first system in ChemInformatics that seamlessly integrates chemistry knowledge with rule-based systems to produce effective classification. ChES has been developed as part of the SciAEther [4] project at Purdue. SciAEther [4] is building a computer-aided discovery environment that enables chemists to exploit their own and community data to rapidly and effectively discover knowledge.

4 Ontologies and Reasoning

The first step in classification of molecules is for scientists to share definition of various concepts. Definition of a concept can sometimes differ among research groups, which makes the classification results difficult to share. Ontologies provide a model for capturing the variety in concept definitions by capturing the human understanding of a domain within computer applications. In ChES, we have used Chemical Ontology (CO) [1], a well known ontology for chemistry to describe and share definitions.

Classification implies deducing a category from a set of constraints applied to ontological concepts and properties. Ontology languages (OWL) support expression of logical constraints on both concepts and properties. Reason-

978-0-7695-3535-7/08 $25.00 © 2008 IEEE
DOI 10.1109/eScience.2008.158

ers (or inference engines) such as Racer or FaCT can deduce classes from a set of conjunctive logical constraints only. However, in Chemistry, examples such hydrocarbons and halogens show that disjunctive constraints are common. Rule ML and SWRL foregoes computability and decidability in favor of allowing expressiveness of disjunctive constraints. However, its reasoners such as Bossam, Pellet are limited in functionality and have poor run time performance [3]. Infact scalable reasoning is still an open challenge [3].

5 ChES Architecture

We describe ChES, which uses CO ontology to capture class definitions but uses a rule-based system to allow for maximum expressiveness in terms of defining and reasoning on constraints. Rule-based systems such as Jess (Java expert system shell) [2] allow both logical predicates (conjunctive and disjunctive) and procedural functions. The Rete algorithm is at the center of rule based systems which allows efficient reasoning.

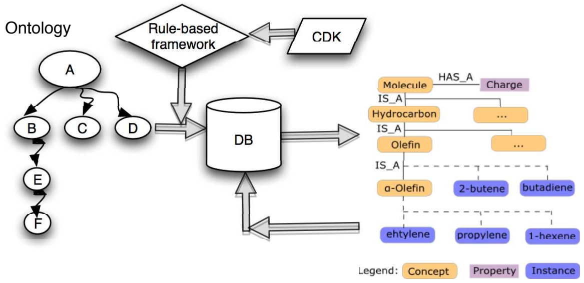


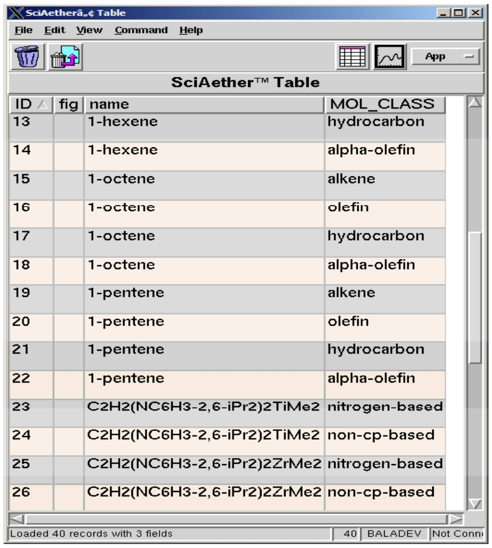
Figure 1. Architecture

Figure 1 shows the architecture of ChES. An ontology captures the relationships between various chemistry concepts such as Atom, Molecule and Bond. Instead of populating ABox assertions and TBox axioms, we import ontological concepts, relationships and instances into the rule-based framework. The rule declaration corresponds to ABox assertions and inferences classes as TBox axioms. The rule-based framework uses a Postgres database to store all its molecular instances. The resulting classes are also stored in the database and used to enhance the ontology. We use several functions in the Chemistry Development Kit (CDK) [5] library such as the UniversalIsomorphismTester to perform substructure search. The CDK is written in Java, which allows for efficient portability into the rule-based framework. We have created several chemistry-specific functions to enable rule specification and inferencing. These include rules for hydrocarbons, olefins, acids, etc. They are available through our Project website [4].

6 Classification Results

We have used 5000 molecules from PubChem to test our system as well as 192 molecules, which are related to single site polymerization in SciAEther. We have been able to correctly classify 95% of molecules, where the remaining 5% suffered from poor specification of rules. We used the incorrectly classified molecules to improve our rule specification. The correctness benchmark was manual inspection by chemists.

Figure 2 shows the classification results over some molecules selected from our test data on single site polymerization. The table shows all the classes of a specific molecule. The rules based approach is also flexible to allow for more complicated rules that chemists may not be able to quickly and consistently apply with visual inspection.



SciAether™ Table

ID	fig	name	MOL_CLASS
13		1-hexene	hydrocarbon
14		1-hexene	alpha-olefin
15		1-octene	alkene
16		1-octene	olefin
17		1-octene	hydrocarbon
18		1-octene	alpha-olefin
19		1-pentene	alkene
20		1-pentene	olefin
21		1-pentene	hydrocarbon
22		1-pentene	alpha-olefin
23		C2H2(NC6H3-2,6-iPr2)2TiMe2	nitrogen-based
24		C2H2(NC6H3-2,6-iPr2)2TiMe2	non-cp-based
25		C2H2(NC6H3-2,6-iPr2)2ZrMe2	nitrogen-based
26		C2H2(NC6H3-2,6-iPr2)2ZrMe2	non-cp-based

Loaded 40 records with 3 fields | 40 | BALADEV | Not Conn

Figure 2. Classification Results

References

[1] H. Feldman, M. Dumontier, S. Ling, N. Haider, and C. Hogue. CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. In *FEBS Letters*, volume 579, Aug 2005.

[2] JESS:http://herzberg.ca.sandia.gov/jess/.

[3] J. Lu, L. Ma, L. Zhang, J.-S. Brunner, C. Wang, Y. Pan, and Y. Yu. Sor: a practical system for ontology storage, reasoning and search. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 1402–1405. VLDB Endowment, 2007.

[4] SciAEther: www.SciAEther.com.

[5] C. Steinbeck, , Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The chemistry development kit (cdk): An open-source java library for chemo and bioinformatics. In *J. Chem. Inf. Comput. Sci.*, 2003.

[6] K. Wolstencroft, P. Lord, L. Tabernero, A. Brass, and R. Stevens. Protein classification using ontology classification. In *Bioinformatics*, volume 22, pages 530–538, 2005.