УДК 004.822

Формальные способы проверки качества данных

С.А. Шабалов[1]

[1]Московский государственный университет им. М.В. Ломоносова

Постановка задачи

Дано: онтологическая модель описания данных о теплофизическиз свойствах веществ.

Задача: исследовать спецификации языков OWL/OWL 2 на предмет возможности описания различных видов ограничений на данные, особенно те, которые невозможно формализовать при помощи реляционной модели, такие как монотонность и уникальность объектов, связанных одним отношением со сторонним объектом, а также, если это возможно, реализовать их на примере предметной области теплофизических свойств веществ.

Онтология

Онтология (в информатике) - это попытка всеобъемлющей и детальной формализации некоторой области знаний с помощью концептуальной схемы. Обычно такая схема состоит из структуры данных, содержащей все релевантные классы объектов, их связи и правила (теоремы, ограничения), принятые в этой области. Онтологии используются в информатике как форма представления знаний о реальном мире или его части. Основные сферы применения — это моделирование бизнес процессов, семантическая паутина (англ. Semantic Web), искусственный интеллект.

Спецификации языка описания

Существует несколько основных подможеств языков описания OWL/OWL2, определенные их спецификациями.

OWL Lite: призван ограничить выразительность языка и повысить скорость алгоритмов. Отсутствует в OWL 2.

OWL DL: обеспечивает максимальную выразительность при сохранении полноты вычислений (все логические заключения будут гарантированно вычислимыми) и разрешаемости (все вычисления завершатся за определенное время). Ограничения: класс может быть подклассом многих классов, но не может сам быть экземпляром другого класса.

OWL Full: предназначен для пользователей, которым нужна максимальная выразительность и синтаксическая свобода RDF без гарантий вычисления. Например, в OWL Full класс может рассматриваться одновременно как собрание индивидов и как один индивид в своём собственном значении. OWL Full позволяет строить такие онтологии, которые расширяют состав предопределённого (RDF или OWL) словаря. Маловероятно, что какое-либо программное обеспечение будет в состоянии осуществлять полную поддержку каждой особенности OWL Full.

В настоящий момент актуальной считается вторая версия OWL, в которой добавлено значительное количество новых возможностей по формализации данных и т.д.

Предметная область

Теплофизические свойства веществ – набор естественнонаучных данных по физическим, химическим, эксплуатационным свойствам. Такие данные широко представлены в печатных пособиях и мировых БД для широчайшего круга веществ: чистых и растворов, органических и неорганических, наноструктур и материалов, характеризуемых технологией изготовления.

Для данной предметной области были выделены основные сущности и разработана модель ее описания.

Pure_Chemical_Substance: перечень/словарь химических веществ.

Physical_Quantity: перечень/словарь физических величин.

State: перечень/словарь агрегатных состояний.

Crystal_structure: перечень/словарь кристаллических систем, к которым может быть отнесено вещество в определенном состоянии.

Magnetic_structure: перечень/словарь магнитных состояний, которыми может обладать вещество.

Dimension: перечень/словарь физических размерностей.

Uncertainty_type: перечень/словарь неопределенностей физических измерений.

Substance_in_State: перечень химических веществ для которых осмыслены, с точки зрения предметной области, измерения в заданном состоянии.

Point_Of_Measure: массивы фактографических данных по с численными значениями свойств из наборов данных.

Measurement_Uncertainty: связанные с конкретными точками измерений неопределенности, в свою очередь характеризующиеся типом и значением.

Data_Set: сущность, относительно которой группируются численные измерения согласованные относительно характера значений (экспериментальные, справочные и.т.д.) и значений наборов констант их сопровождающих, дополнительное ограничение: одно вещество в одном состоянии.

Data_Source: сущность, содержащая информацию об источнике данных (например библиографическую ссылку и.т.д.).

Function_Definition: сущность, определяющая понятие функции.

Domain_Of_Function_Definition: сущность хранит информацию об областях определения функций определенных в Function_Definition.

Control_Functions: функции, с помощью которых можно проверить согласованность измерений относительно физических законов.

Ограничения

Для данной предметной области были определены ограничения, не все из которых возможно формализовать при помощи только лишь реляционной модели. К таким относятся:

1. Ограничение на соблюдение характера монотонности физической величины-функции. Пример: значение изобарной теплоемкости при фиксированном давлении растет с увеличением температуры.

2. Корректность связи физических величин и размерностей. Пример: температуру нельзя измерить в атмосферах и.т.п.

3. Определение диапазонов применимости свойств, например есть свойства, которые имеют смысл только в состоянии идеальный газ, или же наоборот неприменимы к какому-то набору состояний.

Подобные свойства необходимо реализовать при помощи возможностей OWL/OWL2 или доказать невозможность этого.

Литература

1. OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004. https://www.w3.org/TR/2004/REC-owl-ref-20040210/

2. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition). W3C Recommendation 11 December 2012. https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/

3. XML Schema Datatypes in RDF and OWL. W3C Working Group Note 14 March 2006. https://www.w3.org/TR/2006/NOTE-swbp-xsch-datatypes-20060314/