COMP718: Ontologies and Knowledge Bases
Lecture 9: Ontology/Conceptual Model based Data Access

Maria Keet
email: keet@ukzn.ac.za
home: http://www.meteck.org

School of Mathematics, Statistics, and Computer Science
University of KwaZulu-Natal, South Africa

10 April 2012

Outline

1 OBDA Options

2 Some technical details
   - Introduction
   - The ontology language
   - The mapping layer
     - ‘Impedance’ mismatch
     - Mapping assertions
   - Query answering

An ontology with a very large ABox (intro last week)

- Scaling up to realistic size knowledge base handling large amounts of data
- To realise this, we need
  - A language of relatively low computational complexity
  - A way to store large amounts of data
  - Some mechanism to link up the previous two ingredients
  - Query (and reason over) the combination of the previous three
- Use the “Ontology-Based Data Access” (OBDA) approach
  - with the “ontology” in OBDA just a DL knowledge base
  - Most examples and use cases: the ‘ontology’ is a DL-formalised conceptual data model
- Example application with the WONDER system

⇒ What are the options to link an ontology to large amounts of data?
  - Two principal options (in KR view): query rewriting and data completion
  - Several implementation infrastructures: ‘external ABox’ most popular (realised with RDBMS or RDF Triple store)

⇒ What is there behind the scenes for the non-graphical OBDA-part in WONDER and the OBDA systems you set up in the lab?
OBDA options

- KR perspective (with OWA): query rewriting vs data completion
- DB perspective (with CWA): we probably won’t cover this in the lecture
- See slides obda-slides2012TomanCOMP718ukzn.pdf

Introduction

Definition (Ontology-Based Data Access system)

An OBDA system is a triple $O = \langle T, M, D \rangle$, where
- $T$ is a TBox
- $D$ is a relational database
- $M$ is a set of mapping assertions between $T$ and $D$

Note: this is for the current system, but one could conceive of a system that has an RDF triple store as $D$

Ontology-Based Data Access systems (static components)
- An ontology language
- A mapping language
- The data

Query answering in Ontology-Based Data Access systems
- Reasoning over the TBox
- Query rewriting
- Query unfolding
- Relational database technology

These slides are based on Calvanese’s MOSS’09 slides, which also will be made available

More precisely: “Option I, v1.0” mentioned in David Toman’s slides.
In addition to ABox scalability, there are other reasons to realise the ABox with $\mathcal{D}$:

- When we have no direct control over the data since it belongs to some external organization, which controls the access to it.
- When multiple data sources need to be accessed, such as in Information Integration.
- Deal with such a situation by keeping the data in the external (relational) storage, and performing query answering by leveraging the capabilities of the relational engine.
- New problems:
  - The so-called impedance mismatch between values in the relational database and the objects that the ABox expects.
  - How to link the TBox to the “ABox” that is realised as a $\mathcal{D}$?

### The DL-Lite family

- A family of DLs optimized according to the tradeoff between expressive power and complexity of query answering, with emphasis on data.
- Carefully designed to have nice computational properties for answering UCQs (i.e., computing certain answers):
  - The same complexity as relational databases.
  - Query answering can be delegated to a relational DB engine.
  - The DLs of the DL-Lite family are essentially the maximally expressive ontology languages enjoying these nice computational properties.
- Introduction of $\text{DL-Lite}_R$, a member of the DL-Lite family, essentially corresponds to OWL2 QL.

---

### DL-Lite$_R$ (compacter DL notation of OWL 2 QL)

**TBox assertions:**

- Concept inclusion assertions: $\text{Cl} \sqsubseteq \text{Cr}$, with:
  
  $\text{Cl} \rightarrow A \mid \exists Q$
  
  $\text{Cr} \rightarrow A \mid \exists Q \mid \neg A \mid \neg \exists Q$
  
  $Q \rightarrow P \mid P$

- Property inclusion assertions: $Q \sqsubseteq R$, with:
  
  $R \rightarrow Q \mid \neg Q$

**ABox assertions:** $A(c)$, $P(c_1, c_2)$, with $c_1$, $c_2$ constants

*Note:* DL-Lite$_R$ can be straightforwardly adapted to distinguish also between object and data properties (attributes).
### DL-Lite<sub>R</sub> (compacter DL notation of OWL 2 QL)

<table>
<thead>
<tr>
<th>ISA between classes</th>
<th>$A_1 \sqsubseteq A_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disjointness between classes</td>
<td>$A_1 \sqsubseteq \neg A_2$</td>
</tr>
<tr>
<td>Domain and range of properties</td>
<td>$\exists P \sqsubseteq A_1$, $\exists P^{-} \sqsubseteq A_2$</td>
</tr>
<tr>
<td>Mandatory participation ($\text{min card} = 1$)</td>
<td>$A_1 \sqsubseteq \exists P$, $A_2 \sqsubseteq \exists P^{-}$</td>
</tr>
<tr>
<td>ISA between properties</td>
<td>$Q_1 \sqsubseteq Q_2$</td>
</tr>
<tr>
<td>Disjointness between properties</td>
<td>$Q_1 \sqsubseteq \neg Q_2$</td>
</tr>
</tbody>
</table>

*Note: DL-Lite<sub>R</sub> cannot capture completeness of a hierarchy. This would require disjunction (i.e., OR).*

*Note 2: DL-Lite<sub>R</sub> cannot capture functionality on roles ($\text{max card} = 1$)

---

### Relational database as ABox

- Sources store data, which is constituted by values taken from concrete domains, such as strings, integers, codes, ...
- Instances of concepts and relations in an ontology are (abstract) objects
- Solution:
  - Specify how to construct from the data values in the relational sources the (abstract) objects that populate the ABox of the ontology
  - Embed this specification in the mappings between the data sources and the ontology
- Use a *virtual ABox*, where the objects are not materialized

---

### Solution to the impedance mismatch

- Define a mapping language that allows for specifying how to transform data into abstract objects, where
  - Each mapping assertion maps a query that retrieves values from a data source to a set of atoms specified over the ontology
- Basic idea: use Skolem functions in the atoms over the ontology to "generate" the objects from the data values
- Semantics of mappings:
  - Objects are denoted by terms (of exactly one level of nesting)
  - Different terms denote different objects (i.e., we make the unique name assumption on terms)

---

### Example

- Actual data is stored in a DB:
  - An employee is identified by her SSN.
  - A project is identified by its name.

  $D_1[\text{SSN: String, PrName: String}]
  \begin{align*}
  \text{Employees and projects they work for} \\
  \text{Employee's code with salary} \\
  \text{Employee's Code with SSN} \\
  \end{align*}$

- Intuitively:
  - An employee should be created from her SSN: $\text{pers(SSN)}$
  - A project should be created from its name: $\text{proj(PrName)}$

```java
Employee
empCode: Integer, salary: Integer

1..x

worksFor

1..x

Project

projectName: String
```

...
Associate objects in the ontology to data in the tables

- Introduce an alphabet $\Lambda$ of function symbols, each with an associated arity.
- Use value constants from an alphabet $\Gamma_V$ to denote values.
- Use object terms instead of object constants to denote objects: an object term has the form $f(d_1, \ldots, d_n)$ with $f \in \Lambda$, and each $d_i$ is a value constant in $\Gamma_V$.

**Example**

- If a person is identified by her SSN, we can introduce a function symbol $\text{pers}/1$. If $\text{NRM18JUL18}$ is a SSN, then $\text{pers(NRM18JUL18)}$ denotes a person.
- If a person is identified by her name and dateOfBirth, we can introduce a function symbol $\text{pers}/2$. Then $\text{pers(Mandela, 18/07/18)}$ denotes a person.

Mapping assertions, formally

- Mapping assertions are used to extract the data from the DB to populate the ontology.
- Use of variable terms, which are like object terms, but with variables instead of values as arguments of the functions.

**Definition (Mapping assertion between a database and a TBox)**

A mapping assertion between a database $\mathcal{D}$ and a TBox $\mathcal{T}$ has the form

$$\Phi \rightsquigarrow \Psi$$

where

- $\Phi$ is an arbitrary SQL query of arity $n > 0$ over $\mathcal{D}$;
- $\Psi$ is a conjunctive query over $\mathcal{T}$ of arity $n' > 0$ without non-distinguished variables, possibly involving variable terms.

**Definition (Mapping assertion in $\mathcal{M}$ in an OBDA system)**

A mapping assertion between a database $\mathcal{D}$ and a TBox $\mathcal{T}$ in $\mathcal{M}$ has the form

$$\Phi(\vec{x}) \rightsquigarrow \Psi(\vec{z}, \vec{y})$$

where

- $\Phi$ is an arbitrary SQL query of arity $n > 0$ over $\mathcal{D}$;
- $\Psi$ is a conjunctive query over $\mathcal{T}$ of arity $n' > 0$ without non-distinguished variables;
- $\vec{x}, \vec{y}$ are variables with $\vec{y} \subseteq \vec{x}$;
- $\vec{z}$ are variable terms of the form $f(\vec{z})$, with $f \in \Lambda$ and $\vec{z} \subseteq \vec{x}$. 
Semantics of mappings

Intuitively: \( I \) satisfies \( \Phi \rightsquigarrow \Psi \) with respect to \( D \) if all facts obtained by evaluating \( \Phi \) over \( D \) and then propagating answers to \( \Psi \), hold in \( I \).

**Definition (Satisfaction of a mapping assertion with respect to a database)**

An interpretation \( I \) satisfies a mapping assertion \( \Phi(\vec{x}) \rightsquigarrow \Psi(\vec{\imath}, \vec{\jmath}) \) in \( M \) with respect to a database \( D \), if for each tuple of values \( \vec{v} \in \text{Eval}(\Phi, D) \), and for each ground atom in \( \Psi[\vec{x}/\vec{v}] \), we have that:

- If the ground atom is \( A(s) \), then \( s^I \in A^I \);
- If the ground atom is \( P(s_1, s_2) \), then \( (s_1^I, s_2^I) \in P^I \).

\( \text{Eval}(\Phi, D) \) denotes the result of evaluating \( \Phi \) over \( D \), \( \Psi[\vec{x}/\vec{v}] \) denotes \( \Psi \) where each \( x_i \) is substituted with \( v_i \).

---

Query answering

**Two approaches for query answering over \( O \)**

- **Bottom-up approach:**
  - Explicitly construct an ABox \( A_M, D \) using \( D \) and \( M \), and compute the certain answers over \( \langle T, A_M, D \rangle \).
  - Conceptually simpler, but less efficient (PTime in the data).
- **Top-down approach**
  - Unfold the query w.r.t. \( M \) and generate a query over \( D \).
  - Is more sophisticated, but also more efficient
- OBDA with QUONTO/Quest uses the top-down approach
Top-down approach to query answering, intuition

- **Reformulation:** compute the perfect reformulation (rewriting), $q_{pr} = \text{PerfectRef}(q, T)$, of the original query $q$ using the inclusion assertions of the TBox $T$ so that we have a UCQ.

- **Unfolding:** compute a new query $q_{unf}$ from $q_{pr}$ by using the (split version of) the mappings in $\mathcal{M}$
  - Each atom in $q_{pr}$ that unifies with an atom in $\Psi$ is substituted with the corresponding query $\Phi$ over the database
  - The unfolded query is such that $\text{Eval}(q_{unf}, D) = \text{Eval}(q_{pr}, A_{\mathcal{M}, D})$

- **Evaluation:** delegate the evaluation of $q_{unf}$ to the relational DBMS managing $D$

More examples, rewriting rules and algorithm are described on pp290-297 of the MOSS’09 slides, and more details on unfolding are on pp248-251 of the MOSS’09 slides.

Example

Consider the query $q(x) \leftarrow \text{worksFor}(x, y)$
the perfect rewriting is

$$r_{q,T} = q(x) \leftarrow \text{worksFor}(x, y)$$

$$q(x) \leftarrow \text{Employee}(x)$$
To compute $\text{unfold}(r_{q,T})$, we first split $M$ as follows (always possible, since queries in the right-hand side of assertions in $M$ are without non-distinguished variables):

$M_{1,1}$: $\text{SELECT SSN, PrName} \quad \leadsto \quad \text{Employee(pers(SSN))}$
FROM $D_1$

$M_{1,2}$: $\text{SELECT SSN, PrName} \quad \leadsto \quad \text{Project(proj(PrName))}$
FROM $D_1$

$M_{1,3}$: $\text{SELECT SSN, PrName} \quad \leadsto \quad \text{projectName(proj(PrName), PrName)}$
FROM $D_1$

$M_{1,4}$: $\text{SELECT SSN, PrName} \quad \leadsto \quad \text{workFor(pers(SSN), proj(PrName))}$
FROM $D_1$

$M_{2,1}$: $\text{SELECT SSN, Salary} \quad \leadsto \quad \text{Employee(pers(SSN))}$
FROM $D_2$, $D_3$

$M_{2,2}$: $\text{SELECT SSN, Salary} \quad \leadsto \quad \text{salary(pers(SSN), Salary)}$
FROM $D_2$, $D_3$

Then, we unify each atom of the query

\[
r_{q,T} = \begin{cases} q(x) \quad & \text{worksFor}(x, y) \\ q(x) \quad & \text{Employee}(x) \end{cases}
\]

with the right-hand side of the assertion in the split mapping, and substitute such atom with the left-hand side of the mapping

$q(\text{pers(SSN)}) \quad \leadsto \quad \text{SELECT SSN, PrName}$
FROM $D_1$

$q(\text{pers(SSN)}) \quad \leadsto \quad \text{SELECT SSN, Salary}$
FROM $D_2$, $D_3$

The construction of object terms can be pushed into the SQL query, by resorting to SQL functions to manipulate strings (e.g., string concat).

To generate an SQL query, one can follow different strategies:

- Substitute each view predicate in the unfolded queries with the corresponding SQL query over the source:
  - joins are performed on the DB attributes
  - does not generate doubly nested queries
  - the number of unfolded queries may be exponential

- Construct for each atom in the original query a new view.
  This view takes the union of all SQL queries corresponding to the view predicates, and constructs also the Skolem terms
  - avoids exponential blow-up of the resulting query, since the union (of the queries coming from multiple mappings) is done before the joins
  - joins are performed on Skolem terms
  - generates doubly nested queries

Which method is better, depends on various parameters.
1 OBDA Options

2 Some technical details
   - Introduction
   - The ontology language
   - The mapping layer
     - 'Impedance' mismatch
     - Mapping assertions
   - Query answering