Reducing global consistency to local consistency in Ontology-based Data Access

Marco Console, Maurizio Lenzerini

Dipartimento di Ing. Informatica, Automatica e Gestionale "Antonio Ruberti"
SAPIENZA Università di Roma
Via Ariosto 25, I-00186 Roma, Italy
{console,lenzerini}@dis.uniroma1.it

1 Introduction

Ontology-based data access (OBDA) is a paradigm aiming at accessing and managing the data of an information system by means of an ontology [6]. An OBDA system is constituted by an OBDA specification, representing its intensional level, and one or more data sources, representing the extensional one. Depending on the relation the specification shares with the information system, we can divide OBDA systems into two main branches: (1) simple, if the information system is specifically designed to store the ontology instances, or (2) composite, if the information system is constituted by pre-existing data sources, that are not under the control of the OBDA modeler. In this paper we address the latter scenario, and assume that data sources are managed by a relational Data Base Management System (DBMS).

Most of the research on OBDA has concentrated on making query answering efficient. However, query answering is not the only service that an OBDA system must provide. Another crucial service is consistency checking. Current approaches to this problem involves executing expensive queries at run-time. Here, we address a fundamental problem for OBDA system: given an OBDA specification, can we avoid the consistency check on the whole OBDA system (global consistency check), and rely instead on the constraint checking carried out by the DBMS on the data source (local consistency checking)? If this is the case, whenever the DBMS accepts a database at the source, we know that its data are consistent with the OBDA system. In other words, we know that we can reduce global consistency to local consistency.

In the next sections, we present a formal framework for defining global and local consistency in OBDA systems, characterizing their relationship. We actually split this relationship in two parts, that we call protection and faithfulness. Intuitively, a source schema is faithful to an OBDA system if it does not block any data consistent with the ontology, and protects an OBDA system from inconsistency if its integrity constraints block every data that are in conflict with the ontology. By using these two notions, we present an algorithm for checking whether we can indeed reduce global consistency to local consistency in a relevant class of OBDA systems.

2 Ontology based data access

We consider *relational databases*, and refer the reader to [1] for a more detailed account of databases. A *schema* $\mathcal{S}$ is a pair $\langle \Sigma_{\mathcal{S}}, \mathcal{C}_{\mathcal{S}} \rangle$, where $\Sigma_{\mathcal{S}}$ is the alphabet of $\mathcal{S}$, and $\mathcal{C}_{\mathcal{S}}$ is the set of integrity constraints of $\mathcal{S}$, which are rules that each database conforming to the schema must obey. A *database* for $\mathcal{S}$, or simply a $\Sigma_{\mathcal{S}}$-database, is a finite set of ground atoms over the predicates in $\Sigma_{\mathcal{S}}$ and the constants in an alphabet Γ, subject to the unique name assumption. A $\Sigma_{\mathcal{S}}$-database D is *legal* for $\mathcal{S}$, written $D \models \mathcal{S}$, if satisfies all the integrity constraints in $\mathcal{C}_{\mathcal{S}}$, written $D \models \mathcal{C}_{\mathcal{S}}$.

An *ontology* is a conceptualization of a domain of interest expressed in terms of a formal language. Here, we consider logic-based languages, and, more specifically, Description Logics (DLs) [2].

An *OBDA specification* provides the characteristics of the three basic components of the system, as specified by the following definition.

Definition 1. *An* OBDA specification $\mathcal{B}$ *is a triple* $\langle \mathcal{T}, \mathcal{M}, \mathcal{S} \rangle$*, where*
- $\mathcal{T}$ *is a TBox, called the ontology of* $\mathcal{B}$*.*
- $\mathcal{S} = \langle \Sigma_{\mathcal{S}}, \mathcal{C}_{\mathcal{S}} \rangle$ *is a database schema, called the source schema of* $\mathcal{B}$*;*
- $\mathcal{M}$ *is a finite set of mapping assertions [4, 5] between* $\mathcal{S}$ *and* $\mathcal{T}$*, called the* mapping *of* $\mathcal{B}$*.*

Pairing an OBDA specification $\mathcal{B} = \langle \mathcal{T}, \mathcal{M}, \mathcal{S} \rangle$ with a $\Sigma_{\mathcal{S}}$-database D, we obtain an *OBDA system*. We define the semantics of an OBDA system by specifying which are the models of $\mathcal{B}$ relatively to D, denoted by $Mod_D(\mathcal{B})$.

Definition 2. *Let* $\mathcal{B} = \langle \mathcal{T}, \mathcal{M}, \mathcal{S} \rangle$ *be an OBDA specification, and let* D *be a* $\Sigma_{\mathcal{S}}$*-database. Then* $Mod_D(\mathcal{B}) = \{ \mathcal{I} \mid \mathcal{I} \models \mathcal{T},\ (D, \mathcal{I}) \models \mathcal{M},$ and $D \models \mathcal{C}_{\mathcal{S}} \}$.

Checking whether an OBDA system, constituted by $\mathcal{B}$ and D, is satisfiable amounts to checking whether $Mod_D(\mathcal{B}) \neq \emptyset$. In practice, the system is managed by suitable software components, including a database management system ensuring that $D \models \mathcal{C}_{\mathcal{S}}$.

3 Framework for global and local consistency

We begin our analysis of global and local consistency with the formal definition of these two notions.

Definition 3. *Let* $\mathcal{B} = \langle \mathcal{T}, \mathcal{M}, \langle \Sigma_{\mathcal{S}}, \mathcal{C}_{\mathcal{S}} \rangle \rangle$ *be an OBDA specification, and let* D *be a* $\Sigma_{\mathcal{S}}$*-database. Then the OBDA system constituted by* $\mathcal{B}$ *and* D *is said to be* locally consistent *if* $D \models \mathcal{C}_{\mathcal{S}}$*, whereas is said to be* globally consistent *if* $Mod_D(\langle \mathcal{T}, \mathcal{M}, \langle \Sigma_{\mathcal{S}}, \emptyset \rangle \rangle) \neq \emptyset$*,*

The above definition captures the idea that, while the domain ontology $\mathcal{T}$ forms the intensional level of the whole system, the database D together with $\mathcal{M}$ determines its extensional level. The schema $\mathcal{S}$ is simply the structure designed for accommodating the data stored at the source, but it does not really contribute to the semantics of the OBDA system. So global consistency is indeed different from checking the satisfiability of the

whole $\mathcal{B}$, while local consistency merely means that the database D is legal with respect to the source schema.

Further, global consistency of $\mathcal{B}$ and D can be reduced to local consistency exactly when, for all $\Sigma_{\mathcal{S}}$-databases D, $Mod_D(\langle \mathcal{T}, \mathcal{M}, \langle \Sigma_{\mathcal{S}}, \emptyset \rangle \rangle) \neq \emptyset$ is equivalent to $D \models \mathcal{C}_{\mathcal{S}}$. We actually split this notion in two parts, corresponding to the two parts of the equivalence, and we call such parts protection and faithfulness, respectively.

Definition 4. *Let $\mathcal{B} = \langle \mathcal{T}, \mathcal{M}, \mathcal{S} \rangle$ be an OBDA specification, where $\mathcal{S} = \langle \Sigma_{\mathcal{S}}, \mathcal{C}_S \rangle$. Then, $\mathcal{S}$ is said to protect $\mathcal{T}$ and $\mathcal{M}$ from inconsistency if for all $\Sigma_{\mathcal{S}}$-database D such that $Mod_D(\langle \mathcal{T}, \mathcal{M}, \langle \Sigma_{\mathcal{S}}, \emptyset \rangle \rangle) = \emptyset$, we have that $D \not\models \mathcal{C}_{\mathcal{S}}$.*

Intuitively, the schema $\mathcal{S}$ protects $\mathcal{B}$ from inconsistency whenever its constraints block every database which would break global consistency.

Definition 5. *Let $\mathcal{B} = \langle \mathcal{T}, \mathcal{M}, \mathcal{S} \rangle$ be an OBDA specification, where $\mathcal{S} = \langle \Sigma_{\mathcal{S}}, \mathcal{C}_S \rangle$. Then, $\mathcal{S}$ is said to be faithful to $\mathcal{T}$ and $\mathcal{M}$ in $\mathcal{B}$ if for all $\Sigma_{\mathcal{S}}$-database D such that $Mod_D(\langle \mathcal{T}, \mathcal{M}, \langle \Sigma_{\mathcal{S}}, \emptyset \rangle \rangle) \neq \emptyset$, we have that $D \models \mathcal{C}_{\mathcal{S}}$.*

Intuitively, the schema $\mathcal{S}$ is faithful to $\mathcal{B}$ if it does not constrain the source in such a way to filter out data that would not cause the OBDA system to fall into inconsistency.

Theorem 1. *Let $\mathcal{B} = \langle \mathcal{T}, \mathcal{M}, \langle \Sigma_{\mathcal{S}}, \mathcal{C}_S \rangle \rangle$ be an OBDA specification. Then $\mathcal{S}$ is faithful to $\mathcal{B}$ if and only if $Mod_D(\langle \mathcal{T}, \mathcal{M}, \langle \Sigma_{\mathcal{S}}, \mathcal{C}_S \rangle \rangle) = Mod_D(\langle \mathcal{T}, \mathcal{M}, \langle \Sigma_{\mathcal{S}}, \emptyset \rangle \rangle)$.*

The two notions of protection and faithfulness give raise to two decision problems, namely check whether $\mathcal{S}$ protects $\mathcal{B}$ (Protection) and check whether $\mathcal{S}$ is faithful to $\mathcal{B}$ (Faithfulness).

4 Results

Unfortunately, even for OBDA specifications having decidable query answering procedures, the decision problems associated to protection and faithfulness are both undecidable. In recent studies, we discovered cases in which an algorithm for solving those problems actually exists. In particular, in one relevant scenario, we restricted the TBox to be expressed in the *DL-Lite*$_{\mathcal{R}}$ fragment (see [3, 7]), the mapping language to be GLAV-based, (see [4, 5]), with both the head and the body of each mapping assertion being conjunctive queries, and the source schemata to be expressed in terms of the relational model with key, foreign key and denial constraints. Note that this combination of languages allows us to capture a large amount of real world scenarios.

Relying on the finite controllability of query answering under keys and foreign keys (see [8]), we were able to prove the following.

Theorem 2. *Protection can be solved in PTIME with respect to $\mathcal{T}$ and $\mathcal{M}$, and in NP with respect to $\mathcal{S}$.*

Theorem 3. *Faithfulness can be solved in PTIME with respect to $\mathcal{S}$ and $\mathcal{T}$, and in NP with respect to $\mathcal{M}$.*

We plan to continue our investigation at considering the case of OBDA systems where the source schema contains constraints that do not fall into the class of constraints studied here, or where the DLs used for expressing the ontology goes beyond *DL-Lite*$_{\mathcal{R}}$.

Acknowledgements: Work partially supported by the EU under FP7, project Optique (Scalable End-user Access to Big Data), grant n. FP7-318338.

References

1. Abiteboul, S., Hull, R., Vianu, V.: Foundations of Databases. Addison Wesley Publ. Co. (1995)
2. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press (2010), paperback edition
3. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: *DL-Lite*: Tractable description logics for ontologies. In: Proc. of AAAI 2005. pp. 602–607 (2005)
4. Halevy, A.Y.: Answering queries using views: A survey. VLDB Journal 10(4), 270–294 (2001)
5. Lenzerini, M.: Data integration: A theoretical perspective. In: Proc. of PODS 2002. pp. 233–246 (2002)
6. Lenzerini, M.: Ontology-based data management. In: Proc. of CIKM 2011. pp. 5–6 (2011)
7. Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.: Linking data to ontologies. J. on Data Semantics X, 133–173 (2008)
8. Rosati, R.: On the decidability and finite controllability of query processing in databases with incomplete information. In: Proc. of PODS 2006. pp. 356–365 (2006)