Integration-Oriented Ontology

Sergi Nadal, Alberto Abelló

Synonyms

Ontology-Based Data Access

Definition

The purpose of an integration-oriented ontology is to provide a conceptualization of a domain of interest for automating the data integration of an evolving and heterogeneous set of sources using Semantic Web technologies. It links domain concepts to each of the underlying data sources via schema mappings. Data analysts, who are domain experts but not necessarily have technical data management skills, pose ontology-mediated queries over the conceptualization, which are automatically translated to the appropriate query language for the sources at hand. Following well stablished rules when designing schema mappings allows to automate the process of query rewriting and execution.

Overview

Information integration, or data integration, has been an active problem of study for decades. Shortly, it consists on given a single query involving several data sources get a single answer.

Semantic Web technologies are well-suited to implement such data integration settings, where given the simplicity and flexibility of ontologies, they constitute an ideal tool to define a unified interface (i.e., *global* vocabulary or schema) for such heterogeneous environments. Indeed, the goal of an ontology is precisely to conceptualize the knowledge of a domain, see Gruber (2009).

Ontologies are structured into two levels, the TBox and the ABox, where the former represents general properties of concepts and roles (namely Terminology), and the latter represents instances of such concepts and roles (namely Assertions). Such knowledge is commonly represented in terms of the Resource Description Framework (RDF) in the form of triples *subject-predicate-object* (see Wood et al (2014)), which enables to automate its processing, and thus opens the door to exchange such information on the Web as Linked Data, see Bizer et al (2009). Therefore, a vast number of ontologies, or vocabularies, have been proposed to achieve common consensus when sharing data, such as the RDF Data Cube Vocabulary or the Data Catalog Vocabulary.

Data providers make available their datasets via SPARQL endpoints, that implement a protocol where given a SPARQL query, a set of triples is obtained, properly annotated with respect to an ontology. Hence, the data analysts' queries must separately fetch the sets of triples of interest from each endpoint and integrate the results. Furthermore, such settings assume that both TBox and ABox are materialized in the ontology. Figure 1 depicts the Semantic Web architecture, where the data analyst is directly responsible of firstly decomposing the query and issuing the different pieces into several homogeneous SPARQL endpoints, and afterwards collecting all the results and composing the single required answer.

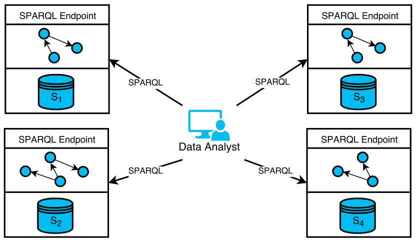

Fig. 1 Example of query execution in the Semantic Web

The alternative architecture for data integration is that of a federated schema, that serves as a unique point of entry, which is responsible of mediating the queries to the respective sources. Such federated information systems are commonly formalized as a triple $\langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, such that $\mathcal{G}$ is the global schema, $\mathcal{S}$ is the source schema and $\mathcal{M}$ are the schema mappings between $\mathcal{G}$ and $\mathcal{S}$, see Lenzerini (2002). With such definition two major problems arise, namely how to design mappings capable of expressing complex relationships between $\mathcal{G}$ and $\mathcal{S}$, and how to translate and compose queries posed over $\mathcal{G}$ into queries over $\mathcal{S}$. The former led to the two fundamental approaches for schema mappings, *global-as-view* (GAV) and *local-as-view* (LAV), while the latter has generically been formulated as the problem of rewriting queries using views, see Halevy (2001) for a survey.

Still, providing an integrated view over a heterogeneous set of data sources is a challenging problem, commonly referred as the data variety challenge, that traditional data integration techniques fail to address in Big Data contexts given the heterogeneity of formats in data sources, see Horrocks et al (2016). Current attemps, namely *ontology-*

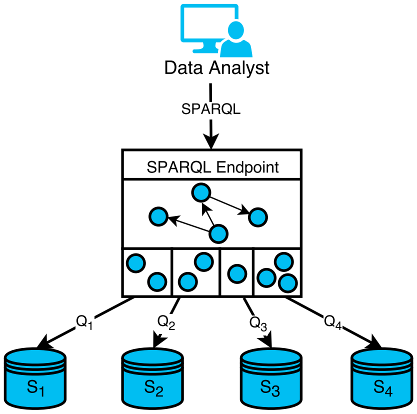

Fig. 2 Example of query execution in integration-oriented ontologies

based data access (OBDA), adopt ontologies to represent $\mathcal{G}$, as depicted in Figure 2. In such case, unlike in the Semantic Web, the TBox resides in the ontology as sets of triples, but the ABox in the data sources, in its source format. The most prominent OBDA approaches are based on generic reasoning in description logics (DLs) for query rewriting (see Poggi et al (2008)). They present data analysts with a virtual knowledge graph (i.e., $\mathcal{G}$), corresponding to the TBox of the domain of interest, in an *OWL2 QL* ontology, see Grau et al (2012). Such TBox is built upon the *DL-Lite* family of DLs, which allows to represent conceptual models with polynomial cost for reasoning in the TBox, see Calvanese et al (2007). This rewritings remain tractable as schema mappings follow a GAV approach, that characterize concepts in $\mathcal{G}$ in terms of queries over $\mathcal{S}$, providing simplicity in the query answering methods, consisting just of unfolding the queries to express them in terms of the sources. More precisely, given a SPARQL query over $\mathcal{G}$, it is rewritten by generic reasoning mechanisms into a first-order logic expression, and further translated into a union of conjunctive queries over $\mathcal{S}$. Nevertheless, despite the flexibility on querying, the management of the sources is still a problem (magnified in Big Data settings), as the variability in their content and structure would potentially entail reconsidering all existing mappings (a well known drawback in GAV).

Alternatively to generic reasoning-based approaches, we have vocabulary-based approaches, which are not necessarily limited by the expressiveness of a concrete DL, as they do not rely on generic reasoning algorithms. In such settings, tailored metadata models for specific integration tasks have been proposed, focusing on linking data sources by means of simple annotations with external vocabularies, see Varga et al (2014). With such simplistic approach, it is possible to define LAV mappings, that characterize elements of the source schemata in terms of a query over the common ontology (facilitating the management of evolution in the sources). However, in the case of LAV, the trade-off comes at the expense of query answering, which now becomes a computationally complex task. Different proposals of specific algorithms exist in the literature for query rewriting under LAV mappings, such as the Bucket algorithm (see Levy et al (1996)), or MiniCon (see Pottinger and Halevy (2001)). In the context of integration-oriented ontologies, new specific algorithms must be devised that leverage the information of the metamodel and the annotations to automatically rewrite queries over $\mathcal{G}$ in terms of $\mathcal{S}$.

Key Research Findings

In the line of OBDA, *DL-Lite* set the cornerstone for research, and many extensions have been proposed in order to increase the expressivity of the ontology while maintaining the cost of reasoning for query answering tractable. Some examples of such approaches are the families of description logics $\mathcal{ELHI}$ (see Pérez-Urbina et al (2009)), *Datalog*$^{\pm}$ (see Gottlob et al (2011)) or $\mathcal{ALC}$ (see Feier et al (2017)). Attention has also been paid to change management in this context, and the definition of a logic able to represent temporal changes in the ontology, see Lutz et al (2008). Relevant examples include *TQL* that provides a temporal extension of *OWL2 QL* and enables temporal conceptual modeling (see Artale et al (2013)), and a logic that delves on how to provide such temporal aspects for specific attributes (see Keet and Ongoma (2015)).

Regarding vocabulary-based approaches purely in the context of data integration, R2RML has been proposed as a standard vocabulary to define mappings from relational databases to RDF (see Cyganiak et al (2012)), and different techniques ad-hoc appeared to automatically extract such mappings, see Jiménez-Ruiz et al (2015).

Examples of Application

Prominent results of generic reasoning OBDA systems are Ontop (see Calvanese et al (2017)), Grind (see Hansen and Lutz (2017)) or Clipper (see Eiter et al (2012)). Ontop provides an end-to-end solution for OBDA, which is integrated with existing Semantic Web tools such as Protégé or Sesame. It assumes a *DL-Lite* ontology for $\mathcal{G}$ and while processing an ontology-mediated query it applies several optimization steps. Oppositely, Grind assumes that $\mathcal{G}$ is formulated using the $\mathcal{EL}$ DL, which has more expressivity than *DL-Lite*, but does not guarantee that a first-order logic rewriting exists for the queries. Clipper, relies on Horn-$\mathcal{SHIQ}$ ontologies, a DL that extends *DL-Lite* and $\mathcal{EL}$. All works assume relational sources, and translate first-order logic expressions into SQL queries. Nonetheless, recent works present approaches to adopt NOSQL stores as underlying sources, such as MongoDB, see Botoeva et al (2016).

Regarding vocabulary-based approaches, we can find an RDF metamodel to be instantiated into $\mathcal{G}$ and $\mathcal{S}$ for the purpose of accommodating schema evolution in the sources, and use RDF named graphs to implement LAV mappings, see Nadal et al (2017). Such approach simplifies the definition of LAV mappings, for non-technical users, which has been commonly characterized as a difficult task. Query answering, however, is restricted to only traversals over $\mathcal{G}$. Another proposal is GEMMS, a metadata management system for data lakes, that automatically extracts metadata from the sources to annotate them with respect to the proposed metamodel, see Quix et al (2016).

Future Directions for Research

Despite that much work has been done in the foundations of OBDA and description logics, there are two factors that complicate the problem in the context of Big Data, namely size and lack of control over the data sources. Given this lack of control and the heterogeneous nature of data sources, it is needed to further study the kind of mappings and query languages that can be devised for different data models. Nevertheless, this is not enough, new governance mechanisms must be put in place giving rise to Semantic Data Lakes, understood as huge repositories of disparate files that require advanced techniques to annotate their content and efficiently facilitate querying. Related to this efficiency, more attention must be paid to query optimization and execution, specially when combining independent sources and data streams. Finally, from another point of view, given that the goal of integration-oriented ontologies is to enable access to non-technical users, it is necessary to empower them by providing clear and traceable query plans that, for example, trace provenance and data quality metrics throughout the process.

References

Artale A, Kontchakov R, Wolter F, Zakharyaschev M (2013) Temporal description logic for ontology-based data access. In: IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013, pp 711–717

Bizer C, Heath T, Berners-Lee T (2009) Linked data - the story so far. International Journal on Semantic Web Information Systems 5(3):1–22

Botoeva E, Calvanese D, Cogrel B, Rezk M, Xiao G (2016) OBDA beyond relational dbs: A study for mongodb. In: 29th International Workshop on Description Logics

Calvanese D, De Giacomo G, Lembo D, Lenzerini M, Rosati R (2007) Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. J Autom Reasoning 39(3):385–429

Calvanese D, Cogrel B, Komla-Ebri S, Kontchakov R, Lanti D, Rezk M, Rodriguez-Muro M, Xiao G (2017) Ontop: Answering SPARQL queries over relational databases. Semantic Web 8(3):471–487

Cyganiak R, Das S, Sundara S (2012) R2RML: RDB to RDF mapping language. W3C recommendation, W3C, http://www.w3.org/TR/2012/REC-r2rml-20120927/

Eiter T, Ortiz M, Simkus M, Tran T, Xiao G (2012) Query rewriting for horn-shiq plus rules. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.

Feier C, Kuusisto A, Lutz C (2017) Rewritability in monadic disjunctive datalog, mmsnp, and expressive description logics (invited talk). In: 20th International Conference on Database Theory, ICDT 2017, March 21-24, 2017, Venice, Italy, pp 1:1–1:17

Gottlob G, Orsi G, Pieris A (2011) Ontological queries: Rewriting and optimization. In: Proceedings of the 27th International Conference on Data Engineering, ICDE 2011, April 11-16, 2011, Hannover, Germany, pp 2–13

Grau BC, Fokoue A, Motik B, Wu Z, Horrocks I (2012) OWL 2 web ontology language profiles (second edition). W3C recommendation, W3C, http://www.w3.org/TR/2012/REC-owl2-profiles-20121211

Gruber T (2009) Ontology. In: Encyclopedia of Database Systems, pp 1963–1965

Halevy AY (2001) Answering queries using views: A survey. VLDB Journal 10(4):270–294

Hansen P, Lutz C (2017) Computing fo-rewritings in *EL* in practice: From atomic to conjunctive queries. In: 16th International Semantic Web Conference (ISWC), pp 347–363

Horrocks I, Giese M, Kharlamov E, Waaler A (2016) Using semantic technology to tame the data variety challenge. IEEE Internet Computing 20(6):62–66

Jiménez-Ruiz E, Kharlamov E, Zheleznyakov D, Horrocks I, Pinkel C, Skjæveland MG, Thorstensen E, Mora J (2015) Bootox: Practical mapping of rdbs to OWL 2. In: The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II, pp 113–132

Keet CM, Ongoma EAN (2015) Temporal attributes: Status and subsumption. In: 11th Asia-Pacific Conference on Conceptual Modelling, APCCM 2015, Sydney, Australia, January 2015, pp 61–70

Lenzerini M (2002) Data integration: A theoretical perspective. In: 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), pp 233–246

Levy AY, Rajaraman A, Ordille JJ (1996) Querying heterogeneous information sources using source descriptions. In: 22th International Conference on Very Large Data Bases (VLDB), pp 251–262

Lutz C, Wolter F, Zakharyaschev M (2008) Temporal description logics: A survey. In: 15th International Symposium on Temporal Representation and Reasoning, TIME 2008, Université du Québec à Montréal, Canada, 16-18 June 2008, pp 3–14

Nadal S, Romero O, Abelló A, Vassiliadis P, Vansummeren S (2017) An integration-oriented ontology to govern evolution in big data ecosystems. Information Systems XXX(YYY):ZZZ–ZZZ, (to appear)

Pérez-Urbina H, Motik B, Horrocks I (2009) A comparison of query rewriting techniques for dl-lite. In: Proceedings of the 22nd International Workshop on Description Logics (DL 2009), Oxford, UK, July 27-30, 2009

Poggi A, Lembo D, Calvanese D, De Giacomo G, Lenzerini M, Rosati R (2008) Linking data to ontologies. Journal on Data Semantics 10:133–173

Pottinger R, Halevy AY (2001) Minicon: A scalable algorithm for answering queries using views. VLDB Journal 10(2-3):182–198

Quix C, Hai R, Vatov I (2016) GEMMS: A generic and extensible metadata management system for data lakes. In: 28th International Conference on Advanced Information Systems Engineering (CAiSE), pp 129–136, CAiSE Forum

Varga J, Romero O, Pedersen TB, Thomsen C (2014) Towards next generation BI systems: The analytical metadata challenge. In: 16th International Conference on Data Warehousing and Knowledge Discovery (DaWaK), pp 89–101

Wood D, Cyganiak R, Lanthaler M (2014) RDF 1.1 concepts and abstract syntax. W3C recommendation, W3C, http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225

Cross-References

From the ToC under section *Big Data Integration*:

- Data Integration

- Knowledge Bases in the Big Data Era
- Semantics for Big Data Integration
- Schema Mapping