An Approach to Probabilistic Data Integration for the Semantic Web *

Andrea Calì [1] and Thomas Lukasiewicz [2, 3]

[1] Facoltà di Scienze e Tecnologie Informatiche, Libera Università di Bolzano
Piazza Domenicani 3, I-39100 Bolzano, Italy
cali@inf.unibz.it
[2] Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza"
Via Salaria 113, I-00198 Rome, Italy
lukasiewicz@dis.uniroma1.it

Abstract. In previous work, we have introduced probabilistic description logic programs for the Semantic Web, which combine description logics, normal programs under the answer set (resp., well-founded) semantics, and probabilistic uncertainty. In this paper, we continue this line of research. We propose an approach to probabilistic data integration for the Semantic Web that is based on probabilistic description logic programs, where probabilistic uncertainty is used to handle inconsistencies between different data sources. It is inspired by recent works on probabilistic data integration in the database and web community [5,2].

1 Overview

Towards sophisticated reasoning capabilities for the Semantic Web, the work [4] has introduced *probabilistic description logic programs* (or *probabilistic dl-programs*), which combine description logics, normal programs under the answer set (resp., well-founded) semantics, and probabilistic uncertainty. Probabilistic dl-programs are an expressive formalism, which generalizes Poole's ICL [6], which in turn generalizes (amongst others) influence diagrams, Bayesian networks, Markov decision processes, normal form games, and structural causal models. Intuitively, a probabilistic dl-program consists of (i) a description logic knowledge base L, (ii) a normal program P involving queries to L [1], and (iii) a probability distribution on a set of total choices. It represents a set of probability distributions on a set of first-order interpretations. Instead of querying L in P, a variant of dl-programs also allows for using L to constrain the terms in P (which may e.g. be used to resolve naming inconsistencies between different data sources).

In this paper, we describe how probabilistic dl-programs can be used for modeling data integration systems with probabilities. A data integration system [3], in its most general form, is a triple $\langle \mathsf{G}, \mathsf{S}, \mathsf{M} \rangle$, where (i) G is the *global* or *mediated schema*, representing the domain of interest of the system, (ii) S is the *source schema*, representing the data sources that take part in the system, and (iii) M is a *mapping* that establishes a relation between the source schema and the global schema. There exist different approaches to the specification of the mapping, which is a crucial aspect in a data integration system. A common issue in data integration is the fact that data may be inconsistent and/or

[3] Alternate address: Institut für Informationssysteme, Technische Universität Wien, Favoritenstraße 9-11, A-1040 Vienna, Austria; e-mail: lukasiewicz@kr.tuwien.ac.at.

* This work was partially supported by the DFG through a Heisenberg Professorship.

redundant relative to the global schema G, which in general incorporates constraints expressed as rules. In other words, the same information may come from different data sources, with different degrees of certainty, which we model by means of rules in probabilistic dl-programs (or *dl-rules*). More formally, we partition the vocabulary Φ into the pairwise disjoint sets Φ_{G}, Φ_{S}, and Φ_c: the symbols in Φ_{G} are of arity at least 1, and represent the (virtual) global predicates; the symbols in Φ_{S} are of arity at least 1, and represent source predicates; the symbols in Φ_c are constants. The mapping M between Φ_{G} and Φ_{S} is then specified by *mapping dl-rules*, which have only predicates in Φ_{S} and Φ_c in the body, and only predicates in Φ_{G} and Φ_c in the head. A probabilistic dl-program modeling a data integration system may have: (i) *source dl-rules* (over Φ_{S} and Φ_c): they express properties and constraints of the data sources; (ii) *global dl-rules* (over Φ_{G} and Φ_c): they express properties and constraints on the global schema, which enhance its expressiveness to better fit the application domain; global dl-rules cannot comprise ground facts; in fact, such facts specify the contents of data sources, while the global schema (at least in the "traditional" data integration setting) must remain strictly virtual; (iii) *mapping dl-rules* as specified above. To summarize, G consists of Φ_{G} and the global dl-rules, S consists of Φ_{S} and the source dl-rules, and M consists of the mapping dl-rules. The fact that the mapping dl-rules are probabilistic allows for a high flexibility in the treatment of the uncertainty that is present when pieces of data come from heterogeneous sources whose informative content in general partially overlaps.

2 Example

Consider a typical rule-based data integration, where the global predicate $buy(C, X)$ is derived from either the source predicate $s_1(C, X, Y)$ or the source predicates $s_2(C, D)$ and $s_3(D, X)$. Moreover, suppose that C resp. X are restricted to customers resp. products from a description logic knowledge base L, and that there may be inconsistencies between the two different ways of deriving $buy(C, X)$. To consistently integrate them, we assign to each derivation a total choice from $\{a, \overline{a}\}$ along with user-defined probabilities that depend on the reliability of the derivations (e.g., $\mu(a) = 0.7$ and $\mu(\overline{a}) = 0.3$). The following dl-rules then realize the probabilistic data integration:

$$buy(C, X) \leftarrow s_1(C, X, Y), DL[Customer](C), DL[Product](X), a\,;$$
$$buy(C, X) \leftarrow s_2(C, D), s_3(D, X), DL[Customer](C), DL[Product](X), \overline{a}\,.$$

So, every fact that holds by the first resp. second dl-rule has the probability 0.7 resp. 0.3, while every fact that holds by both dl-rules has the probability 1. Note that in addition to being inconsistent, two data sources may also be independent from each other.

References

1. T. Eiter, T. Lukasiewicz, R. Schindlauer, and H. Tompits. Combining answer set programming with description logics for the Semantic Web. In *Proc. KR-2004*, pp. 141–151.
2. A. Halevy, M. Franklin, and D. Maier. Principles of dataspace systems. In *Proc. PODS-2006*.
3. M. Lenzerini. Data integration: A theoretical perspective. In *Proc. PODS-2002*, pp. 233–246.
4. T. Lukasiewicz. Probabilistic description logic programs. In *Proc. ECSQARU-2005*, pp. 737–749. Extended version in *Int. J. Approx. Reasoning*, in press.
5. M. van Keulen, A. de Keijzer, and W. Alink. A probabilistic XML approach to data integration. In *Proc. ICDE-2005*, pp. 459–470.
6. D. Poole. The independent choice logic for modelling multiple agents under uncertainty. *Artif. Intell.*, 94(1–2):7–56, 1997.