An Approach to Probabilistic Data Integration for the Semantic Web

Andrea Calì [1] Thomas Lukasiewicz [2 3]

[1] Facoltà di Scienze e Tecnologie Informatiche
Libera Università di Bolzano, Italy

[2] Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza", Italy

[3] Institut für Informationssysteme
Technische Universität Wien, Austria

Outline

- Probabilistic Description Logic Programs
- Probabilistic Data Integration

Probabilistic description logic programs (Lukasiewicz, ECSQARU-05, URSW-05, IJAR-07):

- Combine DLs, normal programs under the answer set resp. well-founded semantics, and probabilistic uncertainty.
- Generalize Poole's ICL, which in turn generalizes influence diagrams, Bayesian networks, MDPs, normal form games, and structural causal models.
- Consist of (i) a description logic knowledge base L, (ii) a normal program P involving queries to L, and (iii) a probability distribution on a set of total choices.
- Represent 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.

$PC \sqcup Camera \sqsubseteq Electronics$; $PC \sqcap Camera \sqsubseteq \bot$;
$Book \sqcup Electronics \sqsubseteq Product$; $Book \sqcap Electronics \sqsubseteq \bot$;
$Textbook \sqsubseteq Book$;

$Product \sqsubseteq \geq 1\, related$;
$\geq 1\, related \sqcup \geq 1\, related^{-} \sqsubseteq Product$;

$Textbook(tb\_ai)$; $Textbook(tb\_lp)$;
$PC(pc\_ibm)$; $PC(pc\_hp)$;

$related(tb\_ai, tb\_lp)$; $related(pc\_ibm, pc\_hp)$;
$provides(ibm, pc\_ibm)$; $provides(hp, pc\_hp)$.

- Description logic knowledge base L (as above).
- Finite set L of dl-rules:

  $avoid(X) \leftarrow DL[Camera](X), not\ offer(X), avoid\_pos;$
  $offer(X) \leftarrow DL[PC \uplus pc; Electronics](X), not\ brand\_new(X), offer\_pos;$
  $buy(C, X) \leftarrow needs(C, X), view(X), not avoid(X), v\_buy\_pos;$
  $buy(C, X) \leftarrow needs(C, X), buy(C, Y), also\_buy(Y, X), a\_buy\_pos.$
- Probabilities associated with all atomic choices:

  $\{avoid\_pos : .9, avoid\_neg : .1\},\ \{offer\_pos : .9, offer\_neg : .1\},$
  $\{v\_buy\_pos : .7, v\_buy\_neg : .3\},\ \{a\_buy\_pos : .7, a\_buy\_neg : .3\}.$

$\Rightarrow$ Probabilities associated with all total choices:

$\{avoid\_pos, offer\_pos, v\_buy\_pos, a\_buy\_pos\} : .9 \times .9 \times .7 \times .7, \ldots$

$\Rightarrow$ Probability intervals associated with all conditional events:

Probabilistic query: $?(buy(c, x) \mid needs(c, x) \wedge buy(c, y) \wedge also\_buy(y, x) \wedge view(x) \wedge not avoid(x))[L, U]$

A data integration system is a triple $\langle G, S, M \rangle$, where

- G is the global or mediated schema, representing the domain of interest of the system,
- S is the source schema, representing the data sources that take part in the system, and
- M is a mapping that establishes a relation between the source schema and the global schema.

The vocabulary Φ is partitioned into:

- Φ_G: its symbols are of arity at least 1, and represent the (virtual) global predicates;
- Φ_S: its symbols are of arity at least 1, and represent the source predicates;
- Φ_c: its symbols are constants.

A probabilistic dl-program modeling a data integration system:

- source dl-rules (over Φ_S and Φ_c): they express properties and constraints of the data sources;
- 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;
- mapping dl-rules: they specify the mapping M between Φ_G and Φ_S; they have only predicates in Φ_S and Φ_c in the body, and only predicates in Φ_G and Φ_c in the head.

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 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 $s_3(D, X)$.

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:

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

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.