Learning Semantic Web rules within the framework of $\mathcal{SHIQ}$+log

Francesca A. Lisi
Dipartimento di Informatica
Università degli Studi di Bari
Via E. Orabona 4, 70125 Bari, Italy
lisi@di.uniba.it

Floriana Esposito
Dipartimento di Informatica
Università degli Studi di Bari
Via E. Orabona 4, 70125 Bari, Italy
esposito@di.uniba.it

ABSTRACT

In this paper we face the problem of learning Semantic Web rules within a decidable instantiation of the $\mathcal{DL}$+log framework which integrates the DL $\mathcal{SHIQ}$ and positive DATALOG. To solve the problem, we resort to the methodological apparatus of Inductive Logic Programming.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Learning

Keywords

Semantic Web rules, Inductive Logic Programming

1. INTRODUCTION

Among the many recent KR proposals for Semantic Web rules, $\mathcal{DL}$+log [9] is a powerful framework for the tight integration of Description Logics (DLs) [1] and disjunctive DATALOG with negation (DATALOG$^{\neg\vee}$) [3]. More precisely, it extends a DL KB with *weakly-safe* DATALOG$^{\neg\vee}$ rules. Note that the condition of weak safeness allows to overcome the main representational limits of the approaches based on the DL-safeness condition, e.g. the possibility of expressing conjunctive queries (CQ) and unions of conjunctive queries (UCQ), by keeping the integration scheme still decidable. In particular, the decidability of reasoning in $\mathcal{DL}$+log depends on the decidability of Boolean CQ/UCQ containment in the $\mathcal{DL}$ chosen to instantiate the framework. As far as we know, the most powerful decidable instantiation of $\mathcal{DL}$+log is currently obtained by choosing $\mathcal{SHIQ}$ [4] as $\mathcal{DL}$.

Acquiring and maintaining Semantic Web rules is very demanding and can be automated though partially by applying Machine Learning (ML) algorithms based on induction. In this paper we face the problem of learning Semantic Web rules by adopting $\mathcal{SHIQ}$+log restricted to positive DATALOG [2] as KR framework and Inductive Logic Programming (ILP) [8] as ML approach. Since ILP has been historically concerned with rule induction from examples within the KR framework of Horn Clausal Logic (HCL), we reformulate core ILP ingredients to tackle with the hybrid DL-CL representation and reasoning of $\mathcal{SHIQ}$+log.

2. INDUCING $\mathcal{SHIQ}$+LOG RULES

We assume that the **data** are represented as a $\mathcal{SHIQ}$+log KB $\mathcal{B}$ where the intensional part $\mathcal{K}$ (i.e., the TBox $\mathcal{T}$ plus the set Π_R of rules) plays the role of *background theory* and the extensional part (i.e., the ABox $\mathcal{A}$ plus the set Π_F of facts) contribute to the definition of *observations*. As an example, suppose we have a $\mathcal{SHIQ}$+log KB (adapted from [9]) consisting of the following intensional knowledge $\mathcal{K}$:

$[A1]$ $\texttt{RICH}\sqcap\texttt{UNMARRIED} \sqsubseteq \exists\ \texttt{WANTS-TO-MARRY}^{-}.\top$
$[R1]$ $\texttt{RICH(X)} \leftarrow \texttt{famous(X)}, \texttt{scientist(X,us)}$

and the following extensional knowledge $\mathcal{F}$:

```
UNMARRIED(Mary)
UNMARRIED(Joe)
famous(Mary)
famous(Paul)
famous(Joe)
scientist(Mary,us)
scientist(Paul,us)
scientist(Joe,it)
```

that can be split into $\mathcal{F}_{\texttt{Joe}} = \{\texttt{UNMARRIED(Joe)}, \texttt{famous(Joe)}, \texttt{scientist(Joe,it)}\}$, $\mathcal{F}_{\texttt{Mary}} = \{\texttt{UNMARRIED(Mary)}, \texttt{famous(Mary)}, \texttt{scientist(Mary,us)}\}$, and $\mathcal{F}_{\texttt{Paul}} = \{\texttt{famous(Paul)}, \texttt{scientist(Paul,us)}\}$.

The **language $\mathcal{L}$ of hypotheses** allows for the generation of $\mathcal{SHIQ}$+log rules of the form

$$p(\vec{X}) \leftarrow r_1(\vec{Y_1}), \ldots, r_m(\vec{Y_m}), s_1(\vec{Z_1}), \ldots, s_k(\vec{Z_k})$$

where $m \geq 0$, $k \geq 0$, $p(\vec{X})$ is an atom built out of either a $\mathcal{SHIQ}$-predicate or a DATALOG-predicate, each $r_j(\vec{Y_j})$ is an atom with a DATALOG-predicate, and each $s_l(\vec{Y_l})$ is an atom with a $\mathcal{SHIQ}$-predicate. Note that p represents the target predicate, denoted as c if p is a DATALOG-predicate and as C if p is a $\mathcal{SHIQ}$-predicate. The former case aims at inducing $c(\vec{X}) \leftarrow$ rules that will enrich the DATALOG part of the KB. E.g., suppose that the DATALOG-predicate happy is the target concept and the building blocks for the language $\mathcal{L}^{\mathtt{happy}}$ are in the set $\{\texttt{famous/1}, \texttt{RICH/1}, \texttt{WANTS-TO-MARRY/2}, \texttt{LIKES/2}\}$. The following rules

$H_1^{\mathtt{happy}}$ $\texttt{happy(X)} \leftarrow \texttt{RICH(X)}$
$H_2^{\mathtt{happy}}$ $\texttt{happy(X)} \leftarrow \texttt{famous(X)}$
$H_3^{\mathtt{happy}}$ $\texttt{happy(X)} \leftarrow \texttt{famous(X)}, \texttt{WANTS-TO-MARRY(Y,X)}$

belonging to $\mathcal{L}^{\mathtt{happy}}$ can be considered hypotheses for happy. Note that $H_3^{\mathtt{happy}}$ is weakly-safe. The latter case aims at inducing $C(\vec{X}) \leftarrow$ rules that will extend the DL part (i.e., the input ontology). E.g., suppose that the target concept is the DL-predicate LONER. If $\mathcal{L}^{\mathtt{LONER}}$ is defined over $\{\mathtt{famous/1}, \mathtt{scientist/2}, \mathtt{UNMARRIED/1}\}$, then the rules

$$\begin{array}{ll}
H_1^{\mathtt{LONER}} & \mathtt{LONER(X)} \leftarrow \mathtt{scientist(X,Y)} \\
H_2^{\mathtt{LONER}} & \mathtt{LONER(X)} \leftarrow \mathtt{scientist(X,Y)}, \mathtt{UNMARRIED(X)} \\
H_3^{\mathtt{LONER}} & \mathtt{LONER(X)} \leftarrow \mathtt{scientist(X,Y)}, \mathtt{famous(X)}
\end{array}$$

belong to $\mathcal{L}^{\mathtt{LONER}}$ and represent hypotheses for LONER.

An **observation** $o_i \in O$ is represented as a couple $(p(\vec{a_i}), \mathcal{F}_i)$ where $\mathcal{F}_i$ is a set containing ground facts concerning the individual $\vec{a_i}$. We assume $\mathcal{K} \cap O = \emptyset$. We say that $H \in \mathcal{L}$ *covers* $o_i \in O$ w.r.t. $\mathcal{K}$ iff $\mathcal{K} \cup \mathcal{F}_i \cup H \models p(\vec{a_i})$. Note that the coverage test can be reduced to query answering in $\mathcal{SHIQ}$+log KBs which in its turn can be reformulated as a satisfiability problem of the KB. E.g., the hypothesis $H_3^{\mathtt{happy}}$ covers the observation $o_{\mathtt{Mary}} = (\mathtt{happy(Mary)}, \mathcal{F}_{\mathtt{Mary}})$ because $\mathcal{K} \cup \mathcal{F}_{\mathtt{Mary}} \cup H_3^{\mathtt{happy}} \models \mathtt{happy(Mary)}$. Conversely it does not cover the observations $o_{\mathtt{Joe}} = (\mathtt{happy(Joe)}, \mathcal{F}_{\mathtt{Joe}})$ and $o_{\mathtt{Paul}} = (\mathtt{happy(Paul)}, \mathcal{F}_{\mathtt{Paul}})$. It can be proved that $H_1^{\mathtt{happy}}$ covers $o_{\mathtt{Mary}}$ and $o_{\mathtt{Paul}}$, while $H_2^{\mathtt{happy}}$ all the three observations. With reference to $\mathcal{L}^{\mathtt{LONER}}$, the hypotheses $H_1^{\mathtt{LONER}}$ and $H_3^{\mathtt{LONER}}$ cover the observations $o_{\mathtt{Mary}} = (\mathtt{LONER(Mary)}, \mathcal{F}_{\mathtt{Mary}})$, $o_{\mathtt{Paul}} = (\mathtt{LONER(Paul)}, \mathcal{F}_{\mathtt{Paul}})$ and $o_{\mathtt{Joe}} = (\mathtt{LONER(Joe)}, \mathcal{F}_{\mathtt{Joe}})$. Conversely, $H_2^{\mathtt{LONER}}$ covers only $o_{\mathtt{Mary}}$ and $o_{\mathtt{Joe}}$.

The **generality order** for $\mathcal{L}$ is based on a relation of subsumption, named $\mathcal{K}$*-subsumption* and denoted as $\succeq_{\mathcal{K}}$, between $\mathcal{SHIQ}$+log rules which is defined as follows. Let $H_1, H_2 \in \mathcal{L}$ be two hypotheses standardized apart, $\mathcal{K}$ a background theory, and σ a Skolem substitution[1] for H_2 with respect to $\{H_1\} \cup \mathcal{K}$. We say that $H_1 \succeq_{\mathcal{K}} H_2$ iff there exists a ground substitution θ for H_1 such that (i) $head(H_1)\theta = head(H_2)\sigma$ and (ii) $\mathcal{K} \cup body(H_2)\sigma \models body(H_1)\theta$. Note that condition (ii) is a variant of the Boolean CQ/UCQ containment problem because $body(H_2)\sigma$ and $body(H_1)\theta$ are both Boolean CQs. The difference between (ii) and the original formulation of the problem is that $\mathcal{K}$ encompasses not only a TBox but also a set of rules. Nonetheless this variant can be reduced to the satisfiability problem for finite $\mathcal{SHIQ}$+log KBs. Indeed the skolemization of $body(H_2)$ allows to reduce the Boolean CQ/UCQ containment problem to a CQ answering problem. Due to the aforementioned link between CQ answering and satisfiability, checking (ii) can be reformulated as proving that the KB $(\mathcal{T}, \Pi_R \cup body(H_2)\sigma \cup \{\leftarrow body(H_1)\theta\})$ is unsatisfiable. Once reformulated this way, (ii) can be solved by applying the algorithm NMSAT-$\mathcal{DL}$+log. Thus, a procedure for testing $\succeq_{\mathcal{K}}$ can be built on top of the reasoning mechanisms for $\mathcal{SHIQ}$+log. E.g., it can be checked that $H_1^{\mathtt{happy}} \not\succeq_{\mathcal{K}} H_2^{\mathtt{happy}}$ and $H_2^{\mathtt{happy}} \not\succeq_{\mathcal{K}} H_1^{\mathtt{happy}}$, i.e. the two hypotheses are incomparable with respect to $\mathcal{K}$-subsumption, and that $H_1^{\mathtt{LONER}} \succeq_{\mathcal{K}} H_2^{\mathtt{LONER}}$ but not viceversa. It can be proved that $\succeq_{\mathcal{K}}$ is a decidable quasi-order (i.e. it is a reflexive and transitive relation) for $\mathcal{SHIQ}$+log rules.

3. CONCLUSIONS AND FUTURE WORK

Our proposal for learning Semantic Web rules adopts a decidable KR framework, $\mathcal{SHIQ}$+log, that is the most powerful among the ones currently available for the tight integration of DLs and CLs. We would like to emphasize that the results of this paper are valid for any other decidable instantiation of $\mathcal{DL}$+log with positive DATALOG. More details of our proposal can be found in [6]. Compared with related works on learning hybrid DL-CL rules [10, 5], it relies on a more expressive DL (i.e., $\mathcal{SHIQ}$), thus getting closer to the actual DLs underlying OWL. Also it can learn rules with a DL-predicate in the head, thus affecting the input ontology. For the future we plan to define ILP algorithms starting from the ingredients identified in this paper. Also we would like to extend the proposal to $\mathcal{SHIQ}$+log with DATALOG$^{\neg\vee}$. Finally, from the application side, we will consider some of the use cases for Semantic Web rules. A preliminary study of an application to ontology evolution can be found in [7].

4. REFERENCES

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press, 2003.

[2] S. Ceri, G. Gottlob, and L. Tanca. *Logic Programming and Databases.* Springer, 1990.

[3] T. Eiter, G. Gottlob, and H. Mannila. Disjunctive DATALOG. *ACM Transactions on Database Systems*, 22(3):364–418, 1997.

[4] I. Horrocks, U. Sattler, and S. Tobies. Practical reasoning for very expressive description logics. *Logic Journal of the IGPL*, 8(3):239–263, 2000.

[5] F. Lisi. Building Rules on Top of Ontologies for the Semantic Web with Inductive Logic Programming. *Theory and Practice of Logic Programming*, 8(03):271–300, 2008.

[6] F. Lisi and F. Esposito. Foundations of Onto-Relational Learning. In F. Železný and N. Lavrač, editors, *Inductive Logic Programming*, volume 5194 of *Lecture Notes in Artificial Intelligence*, pages 158–175. Springer, 2008.

[7] F. Lisi and F. Esposito. Supporting the Evolution of $\mathcal{SHIQ}$ Ontologies with Inductive Logic Programming - A Preliminary Study. In *Proc. Int. Workshop on Ontology Dynamics (IWOD-08)*, 2008.

[8] S. Nienhuys-Cheng and R. de Wolf. *Foundations of Inductive Logic Programming*, volume 1228 of *Lecture Notes in Artificial Intelligence*. Springer, 1997.

[9] R. Rosati. $\mathcal{DL}$+log: Tight integration of description logics and disjunctive datalog. In P. Doherty, J. Mylopoulos, and C. Welty, editors, *Proc. of 10th Int. Conf. on Principles of Knowledge Representation and Reasoning*, pages 68–78. AAAI Press, 2006.

[10] C. Rouveirol and V. Ventos. Towards Learning in CARIN-$\mathcal{ALN}$. In J. Cussens and A. Frisch, editors, *Inductive Logic Programming*, volume 1866 of *Lecture Notes in Artificial Intelligence*, pages 191–208. Springer, 2000.

[1] Let $\mathcal{B}$ be a clausal theory and H be a clause. Let $X_1, \ldots, X_n$ be all the variables appearing in H, and $a_1, \ldots, a_n$ be distinct constants (individuals) not appearing in $\mathcal{B}$ or H. Then the substitution $\{X_1/a_1, \ldots, X_n/a_n\}$ is called a *Skolem substitution* for H w.r.t. $\mathcal{B}$.