Complex Path Queries for RDF Graphs*

Faisal Q. Alkhateeb, Jean-François Baget, Jérôme Euzenat
INRIA Rhône-Alpes,
655 avenue de l'Europe
38330 Montbonnot Saint-Martin, France
Faisal.Al-Khateeb@inrialpes.fr

Inferences in RDF can be computed by a sort of graph homomorphism, known as conceptual graph projection [3]. It is possible to find the results of a query in the form of RDF graph in computing the set of projections of this query in the databases (semantic web).

Another approach, that has been successfully used in graph databases [5], is to use regular expressions for finding regular simple paths in a graph (i.e. given a directed labeled graph G and a regular expression R, find all pairs of nodes connected by a simple path such that the concatenation of the labels along the path satisfies R).

However, both approaches do not find the same set of answers (see some examples in [2]). In this paper we present an extension of RDF in which the arcs can be labeled by regular expressions over the set of URIrefs, we call this extension Path RDF or PRDF. We define the syntax and the semantics of this extension. Then we give an inference mechanism, for PRDF over RDF graphs, which is sound and complete. We also prove the completeness of a possible case for computing the containment of complex path queries.

1 Path RDF: Syntax and semantics

In this section, we define the syntax and the semantics of PRDF: an extension of RDF accepting paths.

1.1 PRDF syntax

Let U denote the set of URIrefs of an RDF vocabulary, L_p denote the set of plain literals, L_t the set of typed literals and B the set of blanks. A PRDF triple is a simple extension of RDF triples by considering path expressions as predicates.

Definition 1 (PRDF Triple). *Let $C \subseteq \{.,|,+,*,!\}$ be a set of operations, V an RDF vocabulary and $\mathcal{L}(C,V)$ the set of regular expressions over $\langle C,V\rangle$. A* PRDF triple *over $\langle C,V\rangle$ is an element of: $(U \cup B) \cap V \times \mathcal{L}(C,V) \times V$.*

A PRDF graph is a set of such PRDF triples.

1.2 PRDF semantics

A PRDF interpretation is an RDF interpretation. However, an RDF interpretation must involve extra conditions to be a model for a PRDF graph. So, we define a support that gives the semantics of a regular expression in an RDF interpretation: this semantics is the transposition of the classical path semantics within RDF's.

Definition 2 (Support of triple in a PRDF interpretation). *Let $C \subseteq \{.,+,*,|,!\}$ be a set of operations, V an RDF vocabulary, $I = \langle IR, IP, I_{EXT}, I_S, I_L\rangle$ a PRDF interpretation of V, B a set of blanks and $I' : V \cup B \rightarrow IR \cup IP$ an extension of I to blanks. An RDF triple $\langle x,E,y\rangle \in G$ is supported in I by I' iff:*

- *If $E = \epsilon$, then $I'(x) = I'(y)$.*
- *If $E = u$, then $\langle I'(x), I'(y)\rangle \in I_{EXT}(I'(u))$.*
- *If $E = E'.E''$, then $\exists z \in B; \langle x,E',z\rangle$ and $\langle z,E'',y\rangle$ are supported by I'.*
- *Else then $\exists m \in L^*(E); \langle x,m,y\rangle$ is supported by I'.*

As usual, an interpretation I is a model of G if there exists an extension I' of I to blanks of G such that every triple $T = \langle s,E,o\rangle \in G$ is supported by I'.

2 PRDF as a query language

We first consider PRDF as a query language to make inferences over RDF graphs. The classical definition of RDF consequence as well as that of projection [3] do not work anymore with PRDF graphs because they cannot deal with path on arcs (they miss extra arcs and nodes). So, we define a projection mechanism [2] that extends the classical projection [3] and the basic querying mechanism [1]. Then we consider the containment of PRDF queries.

2.1 PRDF for querying RDF graphs

We define an inference mechanism that take a PRDF graph as query and an RDF graph as a database. Though the basic querying mechanism [1] searches for all pairs of objects that are connected by a path conforming to a regular expression, our inference mechanism searches all paths between the images of the projection of each two nodes connected by an arc labeled by a regular expression of a PRDF query in the database (semantic web) conforming that regular expression.

Definition 3 (PRDF-projection). *Let $C \subseteq \{.,+,*,|,!\}$ be a set of operations, V an RDF vocabulary, H a PRDF graph over $\langle C,V\rangle$ or an RDF graph over V and G an RDF graph*

*This work has been partially supported by the Knowledge Web European network of excellence (IST-2004-507482)

over V or a PRDF graph over $\langle C, V\rangle$. A PRDF-projection from H to G is an application $\pi : N(H) \rightarrow N(G)$ such that:

- $\forall x \in N(H)$, $\lambda(\pi(x)) \leq \lambda(x)$*;*
- $\forall a \in A(H)$ *with* $\gamma(a) = \langle x, y\rangle$, $\exists a_1, ..., a_k \in A(G)$ *with* $\gamma(a_1) = \langle \pi(x), n_1\rangle$, $\gamma(a_2) = \langle n_1, n_2\rangle$,..., $\gamma(a_k) = \langle n_{k-1}, \pi(y)\rangle$ *such that* $L^*(\lambda(a_1).\lambda(a_2)...\lambda(a_k)) \subseteq L^*(\lambda(a))$.

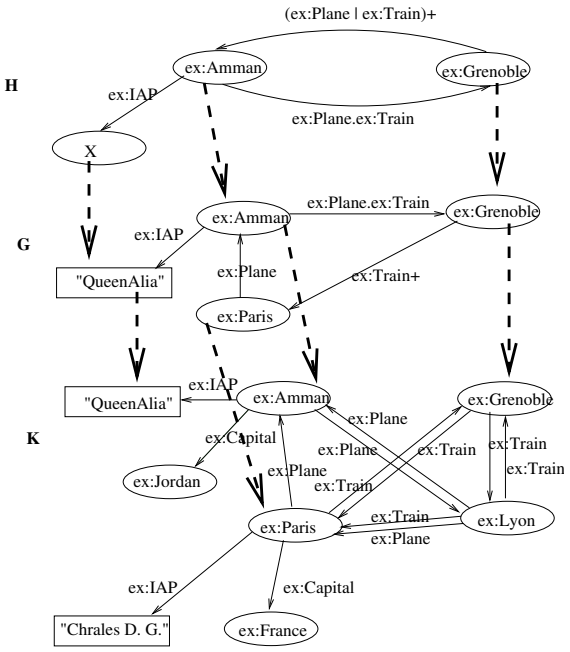

Figure 1: A PRDF projection.

A PRDF projection from a PRDF graph to an RDF graph is represented in dashed line of the Fig. 1. In this projection, it must that the nodes of G project in nodes of K which are more specific. And for each two nodes of G connected with an arc labeled with a regular expression, there is at least a path between the images of these two nodes in the RDF graph. We have proved the soundness and completeness (Theorem 1) of the projection [2].

Theorem 1. *Let $C \subseteq \{., +, *, |, !\}$ be a set of operations and V an RDF vocabulary. Let G an RDF graph over V and H a PRDF graph over $\langle C, V\rangle$. Then $G \models_{PRDF} H$ iff there is a PRDF projection from H to G.*

2.2 PRDF Query containment

Query containment (or entailment between PRDF queries) consists of checking whether or not one query yields a subset of the results of another one. It can be very useful when, for instance, one wants to use queries as indexes over a set of graphs.

PRDF projection is, in general, sound for calculating the containment of two PRDF queries. It is not complete in the general case.

Solving the general problem in a sound and complete way tends to be between EXSPACE and 2EXPTIME [4]. However, we exhibited several restrictions of the problem [2] which have lower complexity.

One of these restrictions involves anchored PRDF graphs, i.e., PRDF graphs in which the extremities of path-labelled graphs are not blank nodes.

Definition 4 (Anchored PRDF graph). *Let $C \subseteq \{., +, *, |, !\}$ be a set of operations, V an RDF vocabulary and H a PRDF graph over $\langle C, V\rangle$. Then H is said* anchored *if $\forall a \in A(H)$ with $\gamma(a) = \langle s, o\rangle$, $\lambda(a) = E$ and E is a regular expression not atomic*[1]*. Then s and o must not be blanks.*

We proved that query containment of an anchored PRDF graphs into a PRDF graph can be computed by projection.

Theorem 2. *Let $C \subseteq \{., +, *, |, !\}$ be a set of operations and V an RDF vocabulary. Let G and H two PRDF graphs over $\langle C, V\rangle$ such that H is an* anchored PRDF graph*. Then $G \models_{PRDF} H$ iff there is a PRDF projection from H to G.*

For instance, consider the two PRDF graphs H and G of the Fig. 1. There exists a PRDF projection from H to G as illustrated in dashed line. Therefore, according to the theorem 2, G entails H or G is contained in H.

3 Conclusion

For querying RDF graphs we introduced graphs labelled by regular expressions as a query language. We found that classical graph projection techniques are sound and complete for querying RDF and that in the case of anchored PRDF graphs, query containment can even be decided by projection.

We plan to investigate how far this query language can be extended by preserving good computational properties.

References

[1] S. Abiteboul and V. Vianu. Regular path queries with constraints. *Journal of Computer and System Sciences*, 58:428–452, 1999.

[2] Faisal Q. Alkhateeb. Graphe à chemins: Graphe RDF/RDFS étiquetés par des expressions algébriques. Master's thesis, Université de Joseph Fourier. INRIA Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot Saint-Martin, France, 2005.

[3] Jean-François Baget. RDF entailmentb as graph homomorphism. In *4th ISWC*, 2005 to appear.

[4] Diego Calvanese, De Giacomo Giuseppe, and Moshe Y. Vardi. Decidable containment of recursive queries. In *Proc. of the 9th Int. Conf. on Database Theory (ICDT 2003)*, volume 2572 of *Lecture Notes in Computer Science*, pages 330–345. Springer, 2003.

[5] Alberto Mendelzon and Peter Wood. Finding regular simple paths in graph databases. *SIAM Journal on Computing*, 24(6):1235–1258, 1995.

[1] A regular expression is said atomic if it is an element of the vocabulary.