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.AlKhateeb@inrialpes.fr

RDF entailment [5], and by extension RDF queries, can be computed using a kind of graph homomorphism known as conceptual graphs projection [3]. Another approach, that has been successfully used in graph databases [6], is to use regular expressions to find paths in a graph (*i.e.*, given a directed labeled graph G and a regular expression E, find all pairs of nodes connected by a path such that the concatenation of the labels along the path belongs to the language generated by E, denoted by $L^*(E)$).

However, some queries that can be expressed in one approach cannot be expressed in the other. A query whose homomorphic image in the database is not a path cannot be expressed by a regular expression, while RDF semantics is not meant to express paths of unknown length.

To benefit from both approaches, we present an extension of RDF in which arcs can be labeled by regular expressions. We call this extension Path RDF or PRDF. We define the syntax and the semantics of this language. Then we have given a sound and complete inference mechanism for PRDF queries over RDF graphs, as well as for a particular case of PRDF query containment.

1 PRDF syntax

Here we consider a vocabulary V partitioned into a set U of URIrefs, a set L_p of plain literals, a set L_t of typed literals and a set B of blanks. The set $\mathcal{R}(V)$ of *regular expressions* over V contains URIrefs of V as *atomic expressions*, and the expressions inductively constructed using the usual operators $\cdot$, $|$, $+$, $*$, and !. A PRDF triple is a simple extension of RDF triples using regular expressions as predicates.

Definition 1 (PRDF Graphs). *Let V be an RDF vocabulary. A* PRDF triple *over V is an element of: $U \cup B \times \mathcal{R}(V) \times V$. A* PRDF graph *over V is a set of PRDF triples over V.*

2 PRDF semantics

A *PRDF interpretation* is an RDF interpretation [5]. However, an RDF interpretation must involve extra conditions to be a model for a PRDF graph. We define here the conditions required by two resources of an interpretation to satisfy a regular expression, effectively transposing classical path semantics within RDF's.

Definition 2 (Support of a regular expression). *Let $I = \langle IR, IP, I_{EXT}, I_S, I_L \rangle$ be a PRDF interpretation of the vocabulary V. A pair $\langle x, y\rangle$ of $IR \times IR$ supports a regular expression E of $\mathcal{R}(V)$ in I iff:*

- *If $E = \epsilon$, then $x = y$.*
- *If E is an URIref, then $\langle x, y\rangle \in I_{EXT}(I_S(E))$.*
- *If $E = E_1 \cdot E_2$, then there exists a resource z of IR such that $\langle x, z\rangle$ supports E_1 and $\langle z, y\rangle$ supports E_2 in I.*
- *Otherwise, there exists $m \in L^*(E)$ such that $\langle x, y\rangle$ supports m in I.*

Now we generalize the usual semantic conditions [5] characterizing the *models* of an RDF graph to: an interpretation I is a model of a PRDF graph G if there exists an extension I' of I to blanks of G such that for every triple $\langle s, E, o\rangle \in G$, $\langle I'(s), I'(o)\rangle$ supports E in I.

We note $G \models_{PRDF} Q$ (G *entails* Q) if every model of G is also a model of Q.

3 PRDF as a query language

We first consider PRDF graphs as queries over RDF graphs. The projection mechanism used to compute RDF entailments (based upon neighborhood) must be updated to take into account the complex paths of PRDF queries. We then use this updated mechanism for a particular case of PRDF query containment.

3.1 PRDF for querying RDF graphs

We define an inference mechanism that takes a PRDF graph as query and an RDF graph as a database. When using RDF graphs projection [3], two nodes that are linked by a predicate P must be mapped into nodes also linked by P. With PRDF graphs, two nodes linked by a regular expression E must be mapped into nodes linked by a path whose concatenation $E_1 \cdot \ldots \cdot E_k$ of labels is more specific than E, *i.e.*, $L^*(E_1 \cdot \ldots \cdot E_k) \subseteq L^*(E)$.

Definition 3 (PRDF-projection). *Let Q and G be two PRDF graphs over V. A* PRDF-projection *from Q into G is a mapping π from the nodes of Q into the nodes of G such that:*

- *for every node x of Q, $label(\pi(x)) \leq label(x)$;*

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

- *for every arc a of Q whose ends (denoted by $\gamma(a)$) are $\langle x, y\rangle$, there exist arcs $a_1, \ldots, a_k$ of G with $\gamma(a_1) = \langle \pi(x), n_1\rangle$, $\gamma(a_2) = \langle n_1, n_2\rangle$, $\ldots$, $\gamma(a_k) = \langle n_{k-1}, \pi(y)\rangle$ such that $L^*(label(a_1)\cdot \ldots \cdot label(a_k)) \subseteq L^*(label(a))$.*

Note that $\leq$ is the smallest preorder on V such that blanks are more general than the other elements.

PRDF projection of a PRDF graph Q into an RDF graph G can be computed using standard projection techniques by initially computing all pairs of nodes of G that satisfy regular expressions of Q. This initial treatment can use the techniques in [1; 6].

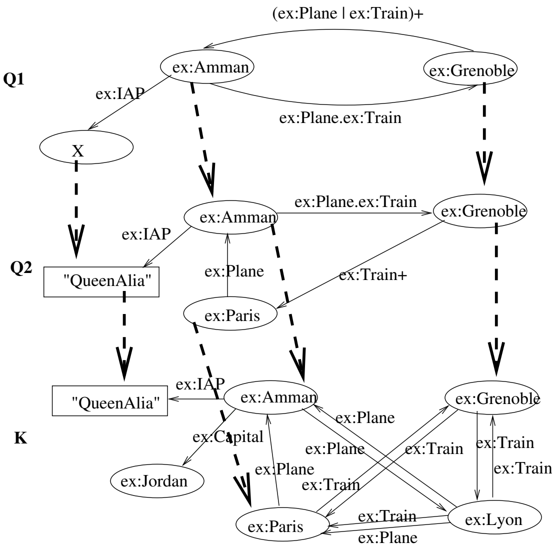

Figure 1: A PRDF projection.

A PRDF projection from the PRDF graph Q_2 to the RDF graph G is represented in dashed line in Fig. 1. Note that the two nodes of the triple ⟨ex:Amman, ex:Plane.ex:Train, ex:Grenoble⟩ of Q_2 are projected into two nodes that are not neighbors in G, but are connected by a path.

The following theorem [2] expresses the soundness and completeness of PRDF projection with respect to our semantics.

Theorem 1. *Let G be an RDF graph over V and Q be a PRDF graph over V. Then $G \models_{PRDF} Q$ iff there is a PRDF projection from Q into G.*

3.2 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 sound for computing PRDF queries containment, but it is not complete in the general case [2].

Solving the general problem in a sound and complete way, *i.e.*, containment of *union of conjunctive 2-way regular path queries (UC2RPQs),* is known to be EXSPACE-complete [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-labeled arcs are not blank nodes.

Definition 4 (Anchored PRDF graph). *A PRDF triple $\langle s, E, o\rangle$ over V is* anchored *if E is an atomic expression (i.e., an URIref) or if neither s nor o are blanks. A PRDF graph is* anchored *if all its triples are anchored.*

We proved [2] that query containment of an anchored PRDF graph into a PRDF graph can be computed by PRDF projection.

Theorem 2. *Let Q_1 and Q_2 be two PRDF graphs over V such that Q_1 is an* anchored PRDF graph. *Then $Q_2 \models_{PRDF} Q_1$ iff there is a PRDF projection from Q_1 into Q_2.*

For instance, consider the two PRDF graphs Q_1 and Q_2 of the Fig. 1. There exists a PRDF projection from Q_1 into Q_2, and Q_1 is anchored, therefore, according to the theorem 2, Q_2 entails Q_1 or Q_2 is contained in Q_1.

4 Conclusion

For querying RDF graphs we introduced graphs labeled by regular expressions as a query language. We found that 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. Both problems are thus NP-complete.

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

References

[1] Serge Abiteboul and Victor 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, 2005.

[3] Jean-François Baget. RDF entailment as graph homomorphism. In *4th ISWC2005*, to appear.

[4] Diego Calvanese, Giuseppe De Giacomo, 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] Patrick Hayes. RDF semantics. W3C Recommendation, February 2004.

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