A Directed Hypergraph Model for RDF

Amadís Antonio Martínez Morales[1] and María Esther Vidal Serodio[2]

[1] Universidad de Carabobo, Venezuela, E-mail: aamartin@uc.edu.ve
[2] Universidad Simón Bolívar, Venezuela, E-mail: mvidal@ldc.usb.ve

RDF is a proposal of the W3C to express metadata about resources in the Web. The RDF data model allows several representations, each one with its own limitations at expressive power and support for the tasks of query answering and semantic reasoning. In this paper, we present a directed hypergraph model for RDF to represent RDF documents efficiently, overcoming those limitations.

1 Related Work

An RDF document can be viewed as a graph: nodes are resources and arcs are properties. Formally [4], suppose there are three infinite sets: $\mathcal{U}$ (URIs), $\mathcal{B} = \{b_j : j \in \mathbb{N}\}$ (blank nodes), and $\mathcal{L}$ (literals). $(s, p, o) \in (\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L})$ is an RDF triple, s is called the subject, p the predicate, and o the object. An RDF graph T is a set of RDF triples. The universe of T, $univ(T)$, is the set of elements of $\mathcal{U} \cup \mathcal{B} \cup \mathcal{L}$ that occur in the triples of T. $sub(T)$ (resp. $pred(T)$, $obj(T)$) is the set of all elements in $univ(T)$ that appear as subjects (resp. predicates, objects) in T. RDF graphs allow several representations: labeled directed graphs (LDG) [4,7], undirected hypergraphs (UH) [5], and bipartite graphs (BG) [5,6].

In the LDG model, given an RDF graph T, nodes in V are elements of $sub(T) \cup obj(T)$, and arcs in E are elements of $pred(T)$ [4,7]. Each $(s, p, o) \in T$ is represented by a labeled arc, $s \xrightarrow{p} o$. The number of nodes and arcs for LDG representation is $|V| \leq 2|T|$ and $|E| = |T|$ [5]. This approach may violate some of the graph theory constraints. Thus, while LDG model is the most widely used representation, it can not be considered a formal model for RDF [1].

In the UH model, given an RDF graph T, each $t = (s, p, o) \in T$ is a hyperedge in E and each element of t (subject s, predicate p, and object o) is a node in V. The number of nodes and hyperedges for UH representation is $|V| = |univ(T)|$ and $|E| = |T|$ [5]. However, UH represent RDF documents as a generalization of undirected graphs, losing the notion of direction in RDF graphs, which impacts the task of semantic reasoning. Besides, it can be hard to graphically represent large RDF graphs, like the museum example [2].

In the BG model, given an RDF graph T, there are two types of nodes in V: statement nodes St (one for each $(s, p, o) \in T$) and value nodes Val (one for each $x \in univ(T)$). Arcs in E relate statement and value nodes as follows: Each $t \in St$ has three outcoming arcs that point to the corresponding node for the subject, predicate, or object of the triple represented by t. The number of nodes and arcs for BG representation is $|St| = |T|$, $|Val| = |univ(T)|$, and $|E| = 3|T|$ [5,6]. While BG satisfy the requirement of a formal model for RDF, issues such as reification, entailment and reasoning have not been addressed yet [1].

2 Proposed Solution

Directed hypergraphs (DH) have been used as a modeling tool to represent concepts and structures in many application areas: formal languages, relational databases, production and manufacturing systems, public transportation systems, between others [3]. RDF DH are formally defined as follows:

Definition 1. *Let T be an RDF graph. The RDF directed hypergraph representing T is a tuple $\mathcal{H}(T) = (W, E, \rho)$ such that: $W = \{w : w \in univ(T)\}$ is the set of nodes, $E = \{e_i : 1 \leq i \leq |T|\}$ is the set of hyperarcs, and $\rho : W \times E \rightarrow \{s, p, o\}$ is the role function of nodes w.r.t. hyperarcs. Let $t \in T$ be an RDF triple, $e \in E$ an hyperarc, and $w \in W$ a node such that $w \in head(e) \cup tail(e)$. Then the following must hold: (i) $(\rho(w, e) = s) \Leftrightarrow (w \in tail(e)) \wedge (w \in sub(\{t\}))$, (ii) $(\rho(w, e) = p) \Leftrightarrow (w \in tail(e)) \wedge (w \in pred(\{t\}))$, and (iii) $(\rho(w, e) = o) \Leftrightarrow (w \in head(e)) \wedge (w \in obj(\{t\}))$*

The information is only stored in the nodes, while hyperarcs preserve the role of nodes and the notion of direction in RDF graphs. Thus, the space complexity of our approach must be smaller than the complexity of representations in section 1. The number of nodes and hyperarcs for DH representation is $|W| = |univ(T)|$ and $|E| = |T|$. Besides, concepts and algorithms of hypergraph theory could be used to manipulate RDF graphs under this representation.

3 Preliminary Experimental Results

Labeled directed graph (LDG) and directed hypergraph (DH) representations were studied over twenty synthetic simple RDF documents, randomly generated using an uniform distribution. Around 33% of the resources simultaneously played the role of subject, predicate, or object. Document sizes were increased, ranging from 1000 to 100000 triples. We used two metrics in this study: (a) the space in memory required to store information, measured as the number of nodes and arcs and (b) the number of comparisons required to answer an elemental query. In both cases, the LDG approach showed a trend of linear dependence on the size of the document, while DH exhibited a more independent behavior.

4 Conclusions and Future Plan

We proposed a directed hypergraph model for RDF. Initial results make us believe that our approach scales better than existing representations. In the future, we propose to: (1) extend this representation for RDFS graphs, (2) develop query evaluation algorithms for conjunctive and SPARQL queries, (3) study the impact of this model on the tasks of query answering and semantic reasoning, and (4) conduct empirical studies to analyze the goodness of our approach.

References

1. F. Dau, “RDF as Graph-Based, Diagrammatic Logic”, *Proceedings of ISMIS’06*, pages 332–337, 2006.
2. FORTH Institute of Computer Science (2003), “RQL v2.1 User Manual”. [Online] `http://139.91.183.30:9090/RDF/RQL/Manual.html`
3. G. Gallo, G. Longo, S. Pallottino, and S. Nguyen, “Directed Hypergraphs and Applications”, *Discrete Applied Mathematics*, Vol. 42, No. 2, pages 177–201, 1993.
4. C. Gutierrez, C. A. Hurtado, and A. O. Mendelzon, “Foundations of Semantic Web Databases”, *Proceedings ACM Symposium on PODS*, pages 95–106, 2004.
5. J. Hayes, *A Graph Model for RDF*, Diploma Thesis, Technische Universitt Darmstadt / Universidad de Chile, 2004.
6. J. Hayes and C. Gutierrez, “Bipartite Graphs as Intermediate Model for RDF”, *Proceedings of the ISWC’04*, pages 47–61, 2004.
7. G. Klyne and J. J. Carroll (2004), “Resource Description Framework (RDF): Concepts and Abstract Syntax”, W3C Recommendation. [Online] `http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/`