

A Directed Hypergraph Formal Model for RDF



Student: M.Sc. Amadís Antonio Martínez-Morales ^{1,2}
Advisor: Dr. María-Esther Vidal ¹

¹ Universidad Simón Bolívar, Caracas, Venezuela

² Universidad de Carabobo, Valencia, Venezuela

Workshop on Semantic Web, Ontologies, and Databases

Universidad Simón Bolívar, Venezuela, February 12, 2008

Agenda

- Motivation
- Related Work
- Our Approach
- Initial Results
- Conclusions and Future Work



Motivation

- RDF is a W3C proposal to express metadata about resources in the Web
- An RDF management system requires support for two main tasks:
 1. Answering queries posed by users and software agents
 2. Semantic reasoning from the data to discover relationships between resources

Motivation

Related Work

Our Approach

Initial Results

Conclusions /
Future Work

Motivation

- The RDF data model allows several graph-based representations:
 - Labeled directed graphs
 - Undirected hypergraphs
 - Bipartite graphs
- Each one of these representations has its own limitations with respect to:
 - RDF data model expressive power
 - Support for the tasks of query answering and semantic reasoning

Related Work

(RDF Documents)

- An RDF document can be seen as an RDF graph:
 - nodes correspond to resources
 - arcs represent properties
- An RDF graph T is a set of RDF triples
- An RDF triple $(s,p,o) \in (U \cup B) \times U \times (U \cup B \cup L)$:
 - U : URI references
 - B : blank nodes
 - L : RDF literals
 - s is a subject
 - p is a predicate
 - o is an object

Related Work

(RDF Documents)

- Given an RDF graph T (without blank nodes):
 - Space complexity: $O(|T|)$
 - Time complexity of elemental query answering: $O(|T|^k)$
 - $|T|$ is the size of T
 - $1 \leq k \leq m$, where m is the number of subgoals in the query
- **Main idea:** preserve the coefficients and exponents as low as possible

Related Work

Labeled Directed Graph (LDG) model

Given an RDF graph T :

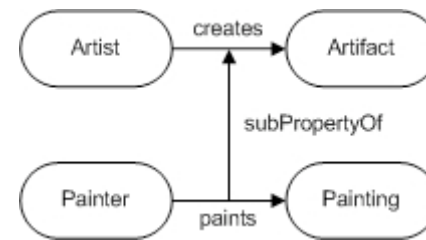
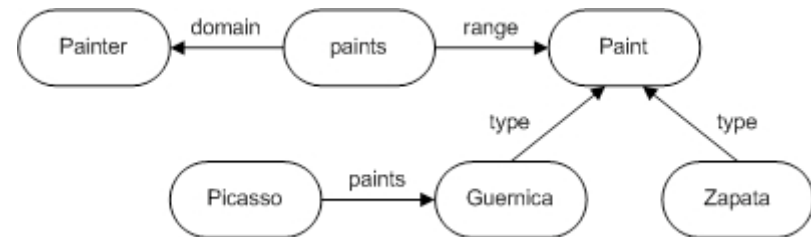
- $V \subseteq \text{sub}(T) \cup \text{obj}(T)$ is the set of nodes
- $E \subseteq \text{pred}(T)$ is the set of arcs
- RDF triples (s,p,o) are represented by labeled arcs $s \xrightarrow{p} o$
- Size of T : $|V| \leq 2|T|$ and $|E| = |T|$
- Space complexity: $O(|T|)$

Related Work

Labeled Directed Graph (LDG) model

This approach may violate some graph theory constraints:

- the intersection between the nodes and arcs labels must be empty
- the set of arcs must be a subset of the Cartesian product of the set of nodes
- It can not be considered a formal model for RDF

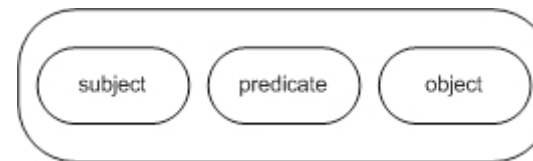


Related Work

Undirected Hypergraph (UH) model

Given an RDF graph T :

- Each RDF triple $t = (s,p,o) \in T$ is a hyperedge
- Each element of t (subject s , predicate p , and object o) is a node



Related Work

Undirected Hypergraph (UH) model

Given an RDF graph T :

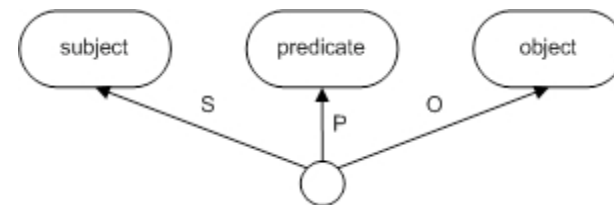
- Size of T : $|V| = |\text{univ}(T)|$ and $|E| = |T|$
- Space complexity: $O(\max(|\text{univ}(T)|, |T|))$
- UH are a generalization of undirected graphs, losing the notion of direction, which impacts the task of semantic reasoning
- It may not be easy to graphically represent large RDF graphs

Related Work

Bipartite Graph (BG) model

Given an RDF graph T :

- There can be two types of nodes in V :
 - Statement nodes St (one for each RDF triple $(s,p,o) \in T$)
 - Value nodes Val (one for each element $x \in \text{univ}(T)$)
- Arcs in E relate statement and value nodes



Related Work

Bipartite Graph (BG) model

Given an RDF graph T :

- Size of T : $|V| = |\text{univ}(T)| + |T|$ and $|E| = 3|T|$
- Space complexity: $O(\max(|\text{univ}(T)|, |T|))$
- BG satisfy the requirement of a formal representation for RDF
- Reification, entailment, and semantic reasoning have not been addressed yet

Motivation

Related Work

Our Approach

Initial Results

Conclusions /
Future Work

Our Approach

(Goals)

A Directed Hypergraph (DH) Model for RDF to:

- Provide a formal representation of RDF to reduce space and time complexity
- Represent and manage RDF documents efficiently
- Implement efficient query evaluation algorithms

Motivation

Related Work

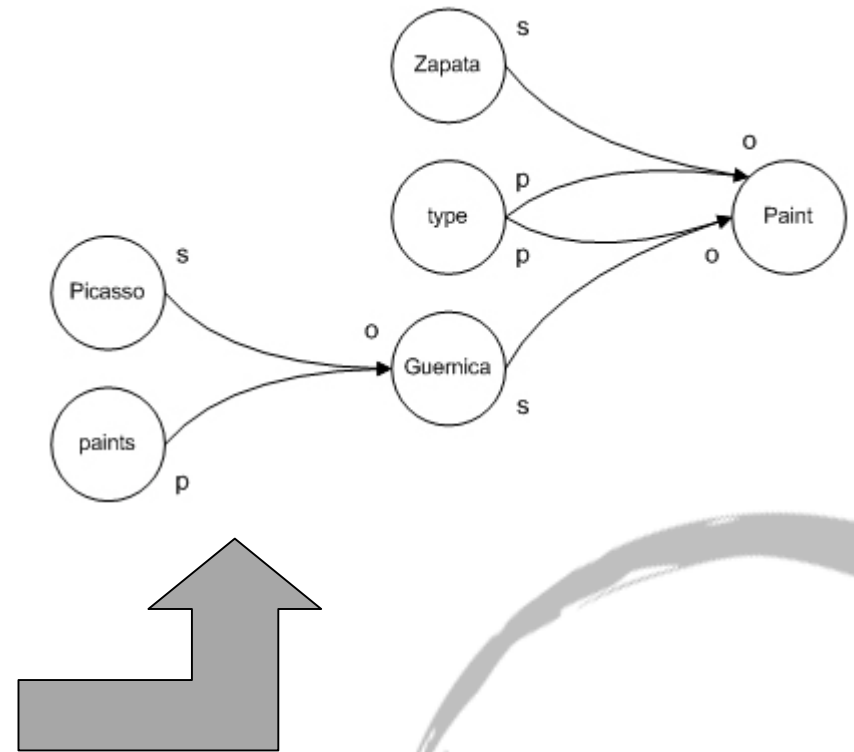
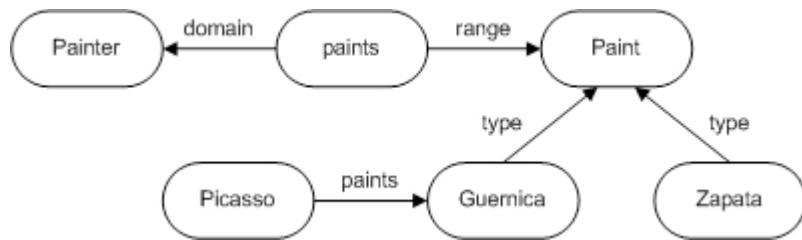
Our Approach

Initial Results

Conclusions /
Future Work

Our Approach (Example)

DH for an RDF graph



Motivation

Related Work

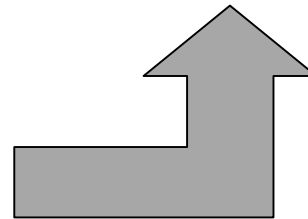
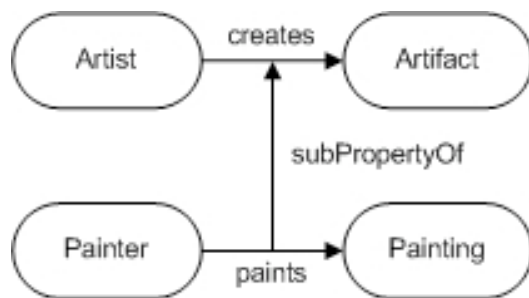
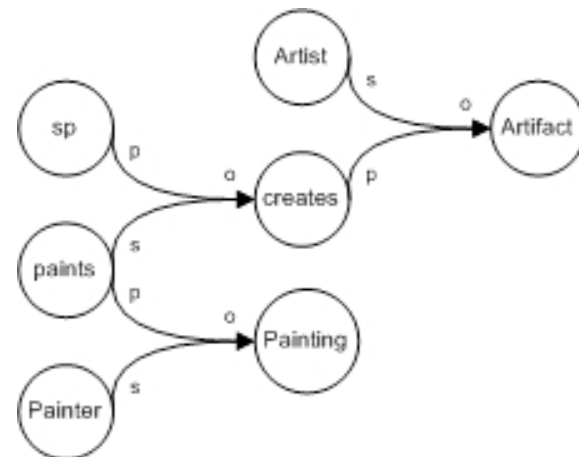
Our Approach

Initial Results

Conclusions /
Future Work

Our Approach (Example)

DH for an RDF graph



Motivation

Related Work

Our Approach

Initial Results

Conclusions /
Future Work

Our Approach

Given an RDF graph T , $H(T) = (W, E, \rho)$ is the RDF Directed Hypergraph representing T :

- Nodes: $W = \{w : w \in \text{univ}(T)\}$
- Hyperarcs: $E = \{e_i : 1 \leq i \leq |T|\}$
- $\rho : W \times E \rightarrow \{s, p, o\}$ is the role function of nodes w.r.t. hyperarcs

Let $t \in T$, $e \in E$, and $w \in W$ such that $w \in \text{head}(e) \cup \text{tail}(e)$. Then:

- $(\rho(w, e) = s) \Leftrightarrow w \in \text{tail}(e) \wedge w \in \text{sub}(\{t\})$
- $(\rho(w, e) = p) \Leftrightarrow w \in \text{tail}(e) \wedge w \in \text{pred}(\{t\})$
- $(\rho(w, e) = o) \Leftrightarrow w \in \text{head}(e) \wedge w \in \text{obj}(\{t\})$
 - $\text{head}(e)$ is the set of incoming nodes of e
 - $\text{tail}(e)$ is the set of outgoing nodes of e

Our Approach

Given an RDF graph T :

- Each node corresponds with an element $w \in \text{univ}(T)$
- The hyperarcs only preserve the role of each node and the concept of direction
- This approach must require less amount of memory than other representations
- Size of T : $|V| = |\text{univ}(T)|$ and $|E| = |T|$
- Space complexity: $O(\max(|\text{univ}(T)|, |T|))$
- DH defines implicit position-based indexes for an RDF document, which can support efficient evaluation of queries over the document

Motivation

Related Work

Our Approach

Initial Results

Conclusions /
Future Work

Initial Results

1. A directed hypergraph model for RDF as a proposal to represent RDF documents efficiently
2. An analysis of the expressive power of this representation with respect to the RDF data model
3. A formal study of the space complexity of this representation to storage the information
4. An empirical study of the impact of this approach on the task of query answering

Motivation

Related Work

Our Approach

Initial Results

Conclusions /
Future Work

Initial Experimental Results

- Labeled Directed Graph (LDG), Bipartite Graph (BG), and Directed Hypergraph (DH) representations were studied empirically
- We tested our prototype using the Lehigh University Benchmark (LUBM)
- The LUBM features an ontology for the university domain and synthetic data scalable to an arbitrary size
- The ontology used in the benchmark is called Univ-Bench
- Univ-Bench describes universities and departments, and the activities that occur at them

Motivation

Related Work

Our Approach

Initial Results

Conclusions /
Future Work

Initial Experimental Results

- To identify each dataset, we use the following notation: LUBM(N, S), the dataset that contains N universities beginning at University0 and is generated using a seed value of S
- We have created five sets of test data:
 - LUBM(1, 0)
 - LUBM(5, 0)
 - LUBM(10, 0)
 - LUBM(20, 0)
 - LUBM(50, 0)

which contain files for 1, 5, 10, 20, and 50 universities respectively, the largest one having over 6.800.000 triples in total

Initial Experimental Results

- Three metrics were addressed to accomplish this preliminary experimental study:
 1. Load Time
 2. Memory Size
 3. Query Response Time
- We have exhaustive results of the first two metrics

Motivation

Related Work

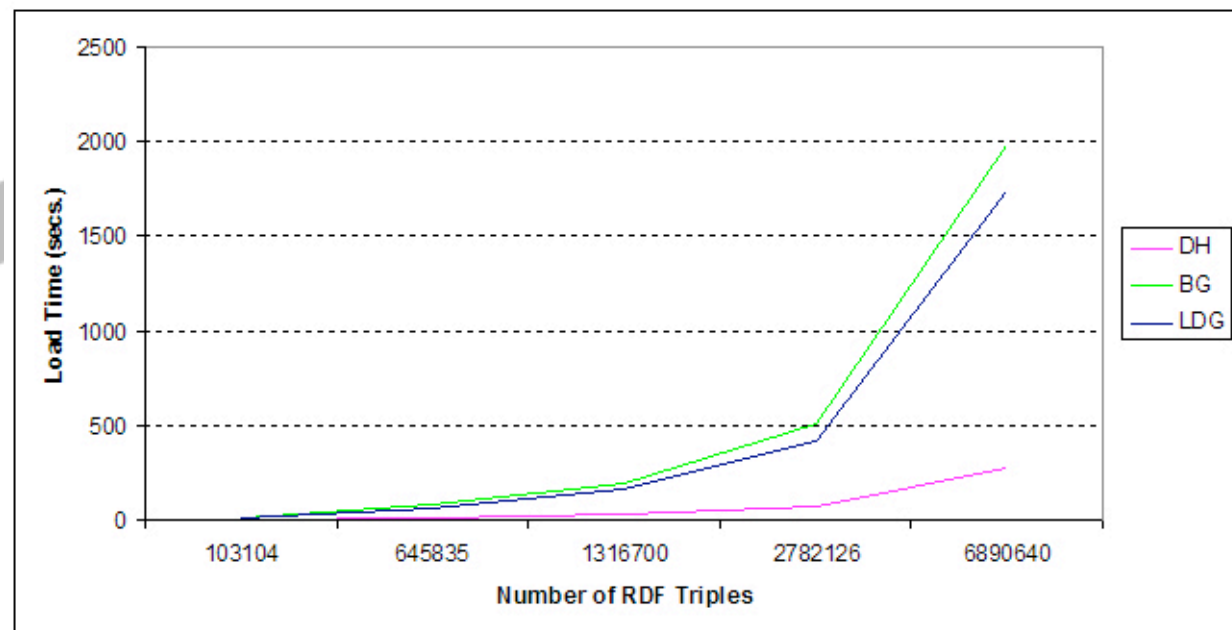
Our Approach

Initial Results

Conclusions /
Future Work

Initial Experimental Results

The load time of all the approaches increases as the number of triples in the documents. DH load time is better than LDG and BG load time; both approaches have similar behavior



Load Time

Motivation

Related Work

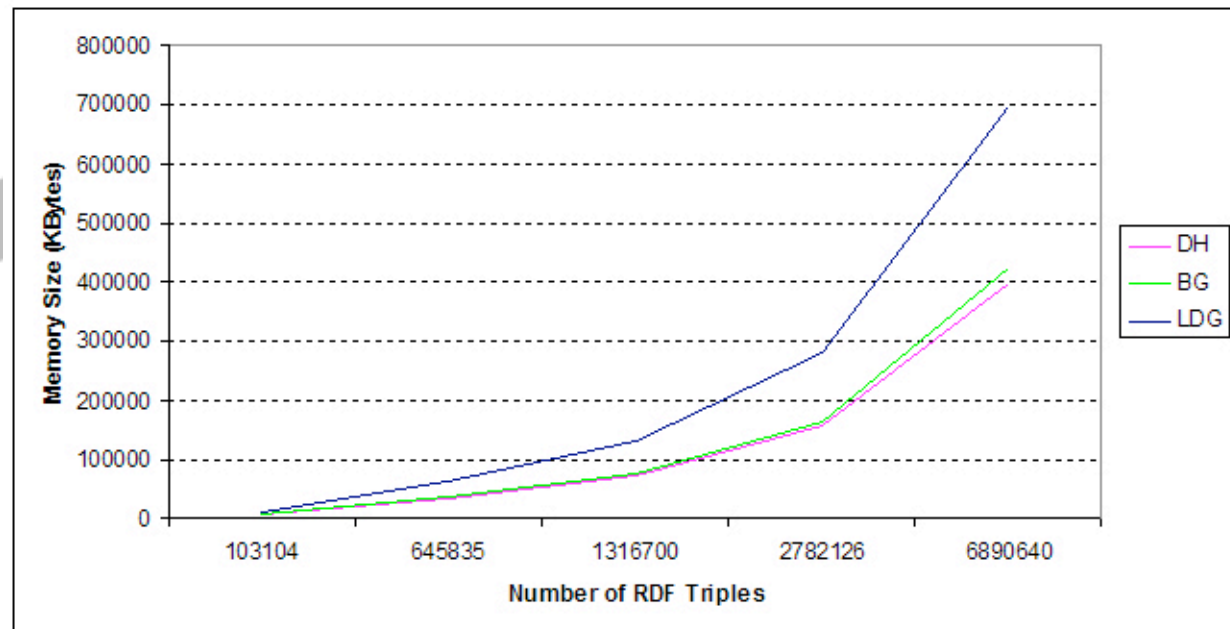
Our Approach

Initial Results

Conclusions /
Future Work

Initial Experimental Results

Memory size of all approaches depend on the size of the document, DH and BG have similar behavior



Memory Size

Motivation

Related Work

Our Approach

Initial Results

Conclusions /
Future Work

Conclusions / Future Work

- Initial results make believe that this approach scales better than existing representations to manage large RDF documents
- Future work:
 - Develop query evaluation algorithms for conjunctive and SPARQL queries
 - Extend this representation for RDFS graphs
 - Study the impact of this model on issues like blank nodes, reification, entailment, and on the tasks of query answering and semantic reasoning
 - Conduct empirical studies to analyze the goodness of this model

Motivation

Related Work

Our Approach

Initial Results

**Conclusions /
Future Work**