Exploiting Paths for Entity Search in RDF Graphs
Minsuk Kahng and Sang-goo Lee
Department of Computer Science and Engineering
Seoul National University, South Korea

Abstract
- Propose an entity retrieval model for RDF data.
- Aim to capture indirect relationships between nodes in the RDF graph by considering paths as fields.

Problem
- Retrieve entities using (semi-)structured RDF data.
  • Called ‘Semantic search’ or ‘Ad-hoc object retrieval’.

Data: What is RDF?
- Semantic Web (standard) data model
- Consists of triples.
- Triples form a graph.

Observation
- Most existing models assume the descriptions of an entity exist only at directly linked nodes (distance=1).
  • Ex. For ‘<../movie/72>’, nodes like ‘"James Cameron’’ are NOT considered.
  • But even if two nodes are not directly linked, they are somewhat related to each other.

Motivation
- We aim to capture indirect relationships btw nodes.
- We assume the descriptions of an entity exist at any (literal) node that is reachable from the resource node.
- Each path from E to L_i is considered as a field.

Proposed Model
- Simulates the generation process of query Q by following paths from a resource node E to several related literal nodes L_i.

P(E|Q) \propto P(Q|E)P(E) = \prod_{j=1}^{Q} P(q_i|E) - \sum_{j=1}^{m} P(q_i|L_j)P(L_j|E)

Finally,

P(E|Q) = P(E) \prod_{j=1}^{Q} P(L_j|E)P(q_i|L_j)

Bayes's Rule
Term independence
Through Literals L_j related to E

Resource Prior
- Like doc prior
- Use 'q' of literals

Path Importance
- Path: seq. of pred.
- How to determine?

Literal LM
- Lang Model for each literal
- Generation prob. of query term given L_j

P(q_i|L_j) = \frac{f(q_i, L_j)}{|L_j| + \mu}

The weights are determined by defining the importance of predicates, then aggregated (above table). Otherwise, it can be learned.

Evaluation
Setup
- Followed the standard evaluation framework used in SemSearch Challenge 2010.
- Data: BTC collection (886M triples, 175M resources, 296M literals)
- Query set: 92 entity queries from Yahoo query logs
- Relevance Judgments: obtained from SemSearch10
- Baselines approaches:
  • Pseudo-document (plain text)
  • Attributes (directly linked literals) as fields (uni. & diff. weights)

Results
- Only few long paths improve performance (e.g. sameAs).
- More research needed to develop ways to learn path weights.
- We added extra judgments due to insufficient relevance judgments.
- We look forward to further research in our path-based approach.
  • See our paper “Ranking Objects by Following Paths in E-R Graphs” in Ph.D workshop at CIKM 2011.