Interest-Based RDF Update Propagation

Kemele M. Endris, Sidra Faisal, Fabrizio Orlandi, Sören Auer, Simon Scerri

EIS - Enterprise Information Systems
University of Bonn & Fraunhofer IAIS
Motivation: Problems accessing remote SPARQL endpoints
Motivation: Problems accessing remote SPARQL endpoints

• Challenges:
  • Availability is not guaranteed
  • Low performance due to high load and traffic
  • Restriction on the query forms and number of results
Motivation: Problems accessing remote SPARQL endpoints

• Challenges:
  • Availability is not guaranteed
  • Low performance due to high load and traffic
  • Restriction on the query forms and number of results

• Possible solution:
  • Setup a local replica of source datasets
Motivation: Syncing using a naïve approach does not scale

• Setting up local replica:
  • Requires manual infrastructure setup and maintenance
  • Requires full data loading
  • Data Freshness not guaranteed

• Approach to sync with source (Naïve):
  1. Fully-loading releases of dataset dumps in intervals
     • Replace old versions of a dataset with new ones
  2. Continuous synchronization
     • Propagate all changes, including irrelevant ones
Definition: Changeset - $\Delta$

- Delta of a dataset $\mathcal{V}$ within two time points: $t_1$ and $t_0$ ($t_1 > t_0$)

$$\Delta(\mathcal{V}_{t_1}) = <\mathcal{D}_{t_1-t_0}, \mathcal{A}_{t_1-t_0}>$$

- Two components:
  - Set of removed triples
    $$\mathcal{D}_{t_1-t_0} = \mathcal{V}_{t_0} \setminus \mathcal{V}_{t_1}$$
  - Set of added triples
    $$\mathcal{A}_{t_1-t_0} = \mathcal{V}_{t_1} \setminus \mathcal{V}_{t_0}$$
Dataset Mirror Tool: Naïve approach

• A typical dataset mirror tool applies a changeset on a replica
  • It removes the set of deleted triples and insert the set of added triples
    • e.g., DBpedia Live Mirror tool

\[
\mathcal{V}_{t_1} = (\mathcal{V}_{t_0} \setminus D_{t_1-t_0}) \cup \mathcal{A}_{t_1-t_0}
\]

removed Triples:
- dbr:Marcel dbp:goals 1.
- dbr:Marcel dbo:team dbr:FNFT.
- dbr:Tim02 foaf:name "Tim Berners-Lee".
- dbr:Cristiano_Ronaldo dbo:goals 205.

added Triples:
- dbr:Cristiano_Ronaldo dbo:goals 230.
- dbr:Barack_Obama foaf:name "Barack Obama".
- dbr:Rio_Ferdinand foaf:Person.
- dbr:Rio_Ferdinand dbo:Athlete.
- dbr:Arvid_Smit foaf:Person.
- dbr:Arvid_Smit dbo:Athlete.

Our Approach: Interest-based update propagation

• Setup a slice (subset) of a dataset
• Interest-based update propagation
  • Interest expressions using SPARQL basic graph patterns
  • Interest expressions are evaluated on deltas (changesets) of the source dataset
  • Only triples that match interest expression shipped to the replica dataset

iRap: Interest-based RDF update propagation framework
Interest-based update propagation

Source

$V_{t_1}$

Changeset $\Delta(V_{t_1})$

$D_{t_1-t_0}$

$A_{t_1-t_0}$

Interest evaluation

Interest ($i_g$)

Candidate Generation

Candidate Assertion

Candidate Generation

Candidate Assertion

Potentially Interesting dataset

Potentially interesting changeset $\Delta(p_{t_1})$

Interesting Changeset $\Delta(\tau_{t_1})$

$\tau_{t_1-t_0} \cup \tau'_{t_1-t_0}$

$\alpha_{t_1-t_0}$

Replica $\tau_{t_0}$

Interest expression

• Based on SPARQL graph pattern
  \( i_g = \langle \tau, b, op \rangle \)

• Composed of:
  • \( b \) - BGP – basic graph pattern
  • \( op \) - OGP – optional graph pattern
  • \( g \) - Source dataset URI – where changesets are downloaded from
  • \( \tau \) - Replica (target) dataset URI – where interesting changes are propagated to

CONSTRUCT WHERE {
  ?athlete a dbo:Athlete .
  OPTIONAL {
  }
}
Interest evaluation (1)

• Two steps:
  I. **Interest candidate generation**
     • Performs matching between the interest expression, $i_g$, and a changeset, $\Delta(V_{t1})$
     • Generates a set of candidate triples
       • e.g., `{ dbr:Cristiano_Ronaldo dbp:goals 205. }`
  II. **Interest candidate assertion**
     • Evaluates candidate triples combined with the interest query on the target dataset
       • e.g.,
         ```query
         SELECT * WHERE {
             ?a a dbo:Athlete .
             ?a dbp:goals ?goals.
         } VALUES (?a ?goals){ (dbr:Cristiano_Ronaldo 205) }
         ```
     • Performs matching between the *assertion* query with the *target* dataset
Interest evaluation (2)

- Execution of the two steps on a changeset results in a set of:
  - Interesting removed (and added) triples
  - Potentially removed (and added) interesting triples
  - Uninteresting triples
Remember?!

- A typical dataset mirror tool

\[ \mathcal{V}_{t_1} = (\mathcal{V}_{t_0} \setminus \mathcal{D}_{t_1-t_0}) \cup \mathcal{A}_{t_1-t_0} \]

**Removed Triples:**
- `dbr:Marcel dbp:goals 1 .
- `dbr:Marcel dbo:team dbr:FNFT`.
- `dbr:Tim%02 foaf:name "Tim Berners-Lee" .
- `dbr:Cristiano_Ronaldo dbo:goals 205`.

**Added Triples:**
- `dbr:Cristiano_Ronaldo dbo:goals 230 .
- `dbr:Barack_Obama foaf:name "Barack Obama" .
- `dbr:Rio_Ferdinand a foaf:Person .
- `dbr:Rio_Ferdinand a dbo:Athlete .
- `dbr:Arvid_Smit a dbo:Athlete .

Remember?!

- A typical dataset mirror tool

\[ V_{t_1} = (V_{t_0} \setminus D_{t_1 \rightarrow t_0}) \cup A_{t_1 \rightarrow t_0} \]

**Removed Triples:**
- `dbr:Marcel dbp:goals 1 .`
- `dbr:Marcel dbo:team dbr:FNFT .`
- `dbr:Tim%02 foaf:name "Tim Berners-Lee" .`
- `dbr:Cristiano_Ronaldo dbo:goals 205 .`

**Added Triples:**
- `dbr:Cristiano_Ronaldo dbo:goals 230 .`
- `dbr:Barack_Obama foaf:name "Barack Obama" .`
- `dbr:Rio_Ferdinand a foaf:Person .`
- `dbr:Rio_Ferdinand a dbo:Athlete .`
- `dbr:Arvid_Smit a dbo:Athlete .`

**Interest query:**

```
CONSTRUCT WHERE {
  ?athlete a dbo:Athlete .
  OPTIONAL {
  }
}
```

Interest-based update propagation

- Interesting Changeset $\Delta(\tau_{t_1})$:
  - Interesting removed triples
  - Interesting added triples

### Interesting Removed Triples:

- dbr:Marcel dbp:goals 1.
- dbr:Cristiano_Ronaldo dbo:goals 205.

### Interesting Added Triples:

- dbr:Cristiano_Ronaldo dbo:goals 230.
- dbr:Rio_Ferdinand a dbo:Athlete.

### Athlete Dataset $\tau_{t_0}$:

- dbr:Marcel a dbo:Athlete.
- dbr:Marcel dbp:goals 1.
- dbr:Cristiano_Ronaldo a dbo:Athlete.
- dbr:Cristiano_Ronaldo dbp:goals 205.
- dbr:Cristiano_Ronaldo foaf:homepage "http://cristianoronaldo.com".

Interest-based update propagation

• Interesting Changeset Δ(τ_{t1}):
  • Interesting removed triples
  • Interesting added triples

• t1):
  • Potentially interesting removed triples
  • Potentially interesting added triples

---

Interesting Removed Triples:

dbr:Marcel dbp:goals 1 .
dbr:Cristiano_Ronaldo dbo:goals 205 .

Interesting Added Triples:

dbr:Cristiano_Ronaldo dbo:goals 230 .
dbr:Rio_Ferdinand a dbo:Athlete .
dbr:Rio_Ferdinand dbp:goals 7 .

Potentially Interesting Triples:

dbr:Barack_Obama foaf:homepage "http://www.barackobama.com/".
dbr:Arvid_Smit a dbo:Athlete .
dbr:Marcel a dbo:Athlete .

Athlete Dataset τ_{t1}:

dbr:Cristiano_Ronaldo a dbo:Athlete .
dbr:Cristiano_Ronaldo dbo:goals 230 .
dbr:Cristiano_Ronaldo foaf:homepage "http://cristianoronaldo.com".
dbr:Rio_Ferdinand a dbo:Athlete .
dbr:Rio_Ferdinand dbp:goals 7 .

iRap Framework

• Reference implementation of our approach
• Implemented using Java and Jena
• Open-source
  • http://eis.iai.uni-bonn.de/Projects/iRap
Experimental setting

• DBpedia 2014 – as a source dataset
  • 12,057 Changesets
    from Oct 01 – 15, 2014

• Two target datasets
  1) Football dataset
    • Slice of DBpedia: 265K triples
  2) Location dataset
    • Complete DBpedia 2014 dataset

• Two interest expressions:
  1) //Football dataset interest
     CONSTRUCT WHERE {
       ?footballer a dbo:SoccerPlayer .
       ?footballer foaf:name ?name .
       ?team rdfs:label ?teamName .
     }

  2) //Location dataset interest
     CONSTRUCT WHERE {
       ?location a ?type .
       ?location wgs:lat ?lat .
       OPTIONAL {
         ?location dcterms:subject ?subject
       }
     }

Results: Football dataset

Results: Football dataset

- Interesting triples:
  - 0.38% of the total removed triples
  - 0.34% of the total added triples
Results: Football dataset growth
Results: Location dataset
Results: Location dataset

- Interesting triples:
  - 4.38% of the total removed triples
  - 1.81% of the total added triples
Results: Location dataset growth

Conclusion

• A novel approach for interest-based RDF update propagation and detailed formalizations

• Our evaluation shows that our method can significantly cut down the size of the data updates

• Summary of results from 12,057 changesets of DBpedia 2014

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Interesting removed triples</th>
<th>Interesting added triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Football dataset</td>
<td>0.38%</td>
<td>0.34%</td>
</tr>
<tr>
<td>Location dataset</td>
<td>4.38%</td>
<td>1.81%</td>
</tr>
</tbody>
</table>
Thank you for your attention!

Questions?

• More on our Website: [http://eis.iai.uni-bonn.de/Projects/iRap](http://eis.iai.uni-bonn.de/Projects/iRap)

• Contact:
  Google group: [irap-ld@googlegroups.com](mailto:irap-ld@googlegroups.com)
  Twitter: @KemeleM - @BadmotorF