Scalable Multi-Query Optimization for SPARQL

Wangchao Le, Anastasios Kementsietsidis, Songyun Duan, Feifei Li

Motivation 1: Access Control for RDF data

Motivation 2: Web Data Integration on Query Endpoints

Finding Maximal Common Connected Substructures for SPARQL MQO

Theorem 1 Given two graphs, finding the maximal common connected subgraphs amounts to finding the maximal common connected induced subgraphs in their linegraphs.

Theorem 2 Given two graphs, finding the maximal common connected induced subgraphs amounts to finding the maximal cliques with strong covering trees in their product graph.

*Challenges: (I) deal with hundreds of graphs in one shot; (II) blend selectivity into the structure-based MQO.

Build & Optimize Linegraphs and Product graph

- Prune by missed matched constants.
- Mask every linegraph edge e with \(\cap_{u\in e} [\tau_u] \cap [\tau_v] = \emptyset\)

- Example query

Clique, Maximal Common Induced Subgraph and Strong Covering Tree

Blend Cost into MOO

Cost(Q) = \(\min|\text{sel}(t)|\) + \(\Delta\) Q is a Type 1 query, \(t \in \text{GP}\)

Q is a Type 2 query, \(t \in \text{GP}\)

Observation: >90% of query evaluation time for our MOO is on evaluating the common structure (do it once for all queries), resulting in less time in evaluating the non-common substructures; while the pure structure-based MOO (MQO-S) is sensitive to the variances of common substructures rewritten, leading to more overhead in evaluating the non-common substructures.

Experiments on varying selectivity

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Default</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of queries</td>
<td>(Q)</td>
<td>100</td>
<td>50 to 500</td>
</tr>
<tr>
<td>Query size (num of tripl. patterns)</td>
<td></td>
<td>6</td>
<td>5 to 10</td>
</tr>
<tr>
<td>Number of seed queries</td>
<td></td>
<td>3</td>
<td>1 to 10</td>
</tr>
<tr>
<td>Size of seed queries</td>
<td></td>
<td>3</td>
<td>1 to 10</td>
</tr>
<tr>
<td>Max selectivity of patterns in Q</td>
<td></td>
<td>0.01</td>
<td>0.1% to 40%</td>
</tr>
<tr>
<td>Max selectivity of patterns in (Q)</td>
<td></td>
<td>random</td>
<td>6 to 10</td>
</tr>
</tbody>
</table>

MQO-S: MOO based on structure. MOO: cost-based MOO. No-MOO: No MOO.