SPARQL Query Optimization Using Selectivity Estimation[1]

Abraham Bernstein, Markus Stocker, and Christoph Kiefer

DIS Dynamic and Distributed Information Systems

University of Zurich

Motivation

Performance of SPARQL queries not yet systematically investigated ➡ **OptARQ optimization suite for SPARQL**

Key features of OptARQ:

- static optimization for in-memory RDF graphs
- based on Jena ARQ framework
- join re-ordering strategies using selectivity estimation
- three approaches: **SEIOptARQ**, **QPIOptARQ**, and **HybridOptARQ**

OptARQ's Selectivity Estimation Approach

Selectivity of a triple pattern T (i.e., [ub:Student17 rdfs:label "Markus"])

$$sel(T) = sel(S) \times sel(P) \times sel(O)$$

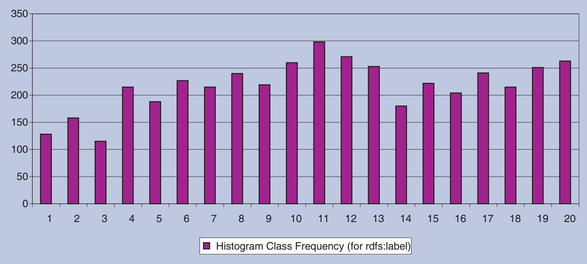
Selectivity of the subject, predicate, and object

$$sel(S) = \frac{1}{|R|}, \quad sel(P) = \frac{|T_P|}{|T|}, \quad sel(O) = \begin{cases} \frac{h_c(P,O_c)}{|T_P|}, & if\ P\ bound; \\ \sum_{P_i \in P} \frac{h_c(P_i,O_c)}{T_{P_i}}, & otherwise. \end{cases}$$

Example

R = number of resources = 500
$T_{rdfs:label}$ = number of triples matching rdfs:label = 500
T = number of triples = 1000

$$sel([ub:Student17\ rdfs:label\ "Markus"]) = \frac{1}{500} \times \frac{500}{1000} \times \frac{300}{500} = 0.0006$$



Optimization Rules

Rewrite filter variables

```
SELECT ?x ?y
WHERE
{ ?x rdf:type ub:Student .
  ?y rdf:type ub:Course .
  ?x ub:takesCourse ?y .
  ?s ub:teacherOf ?y .
  FILTER (?s = <http://Professor0>) .
  FILTER (?x != <http://Student17>)
}
```

➡

```
SELECT ?x ?y
WHERE
{ ?x rdf:type ub:Student .
  ?y rdf:type ub:Course .
  ?x ub:takesCourse ?y .
  <http://Professor0> ub:teacherOf ?y .
  FILTER (?x != <http://Student17>)
}
```

Move-up filter statements

```
SELECT ?x ?y
WHERE
{ ?x rdf:type ub:Student .
  ?y rdf:type ub:Course .
  ?x ub:takesCourse ?y .
  <http://Professor0> ub:teacherOf ?y .
  FILTER (?x != <http://Student17>)
}
```

➡

```
SELECT ?x ?y
WHERE
{ ?x rdf:type ub:Student .
  FILTER (?x != <http://Student17>) .
  ?y rdf:type ub:Course .
  ?x ub:takesCourse ?y .
  <http://Professor0> ub:teacherOf ?y
}
```

Re-order by SEI

```
SELECT ?x ?y
WHERE
{ ?x rdf:type ub:Student .
  FILTER (?x != <http://Student17>) .
  ?y rdf:type ub:Course .
  ?x ub:takesCourse ?y .
  <http://Professor0> ub:teacherOf ?y
}
```

➡

```
SELECT ?x ?y
WHERE
{ <http://Professor0> ub:teacherOf ?y .
  ?x rdf:type ub:Student .
  FILTER (?x != <http://Student17>) .
  ?y rdf:type ub:Course .
  ?x ub:takesCourse ?y
}
```

Re-order by QPI

```
SELECT ?x ?y
WHERE
{ ?x rdf:type ub:Student .
  FILTER (?x != <http://Student17>) .
  ?y rdf:type ub:Course .
  ?x ub:takesCourse ?y .
  <http://Professor0> ub:teacherOf ?y
}
```

➡

```
SELECT ?x ?y
WHERE
{ ?x ub:takesCourse ?y .
  <http://Professor0> ub:teacherOf ?y .
  ?x rdf:type ub:Student .
  FILTER (?x != <http://Student17>) .
  ?y rdf:type ub:Course
}
```

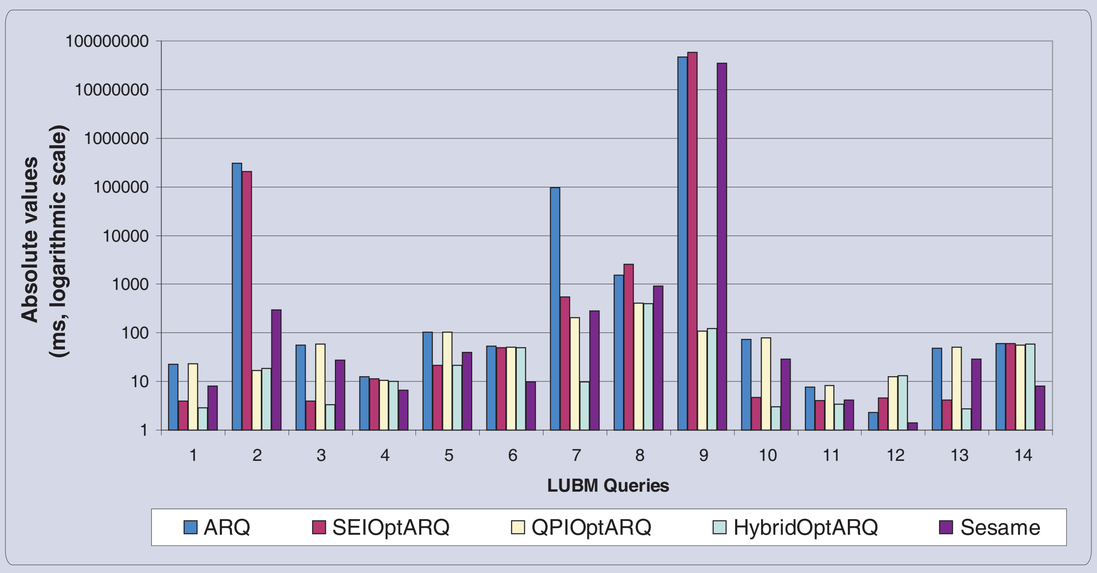
Hybrid SEI-QPI

```
SELECT ?x ?y
WHERE
{ ?x rdf:type ub:Student .
  FILTER (?x != <http://Student17>) .
  ?y rdf:type ub:Course .
  ?x ub:takesCourse ?y .
  <http://Professor0> ub:teacherOf ?y
}
```

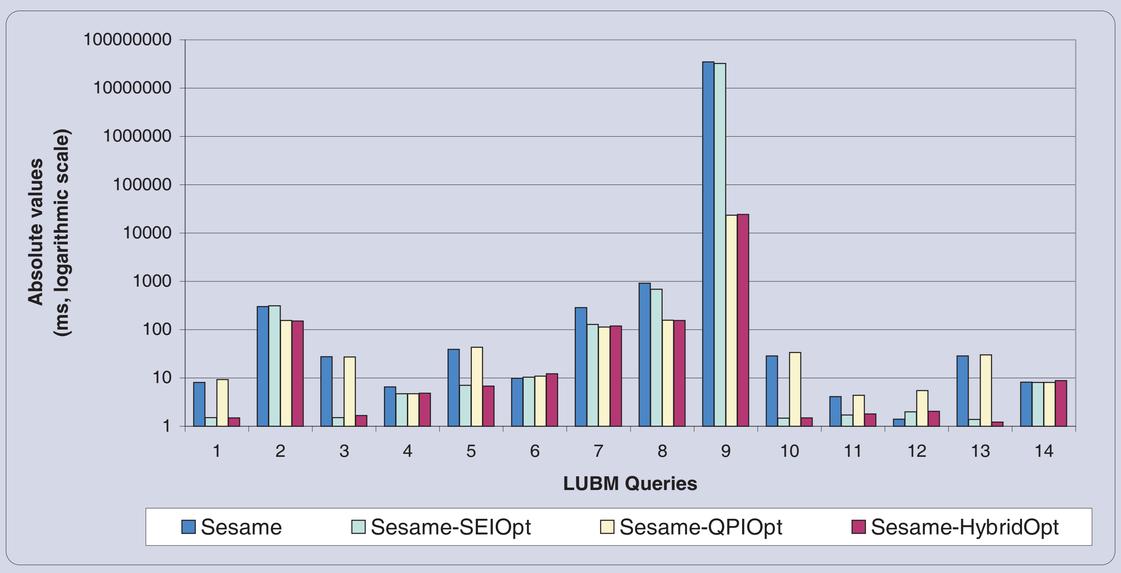
➡

```
SELECT ?x ?y
WHERE
{ <http://Professor0> ub:teacherOf ?y .
  ?x ub:takesCourse ?y .
  ?x rdf:type ub:Student .
  FILTER (?x != <http://Student17>) .
  ?y rdf:type ub:Course
}
```

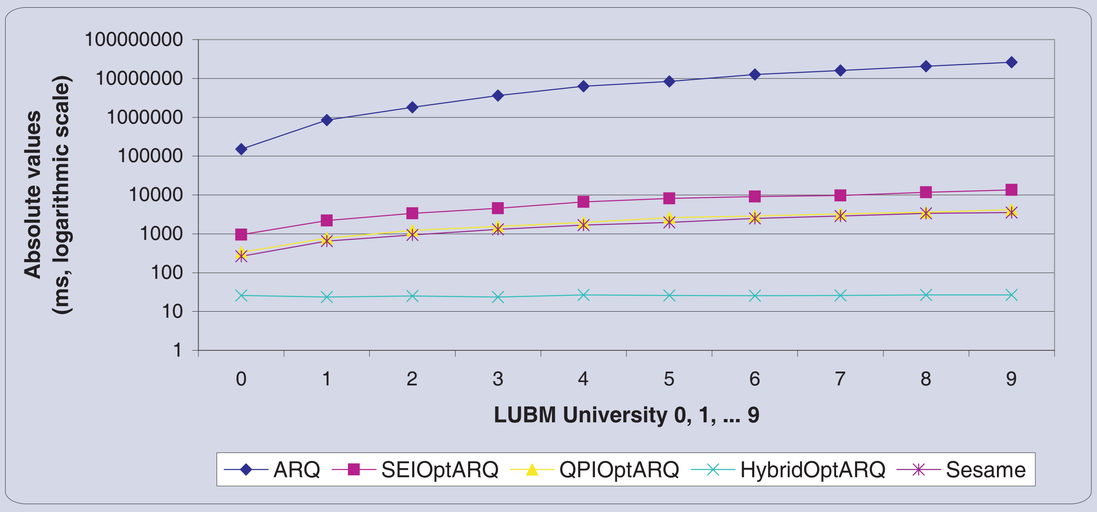
Evaluation for In-Memory RDF Graphs



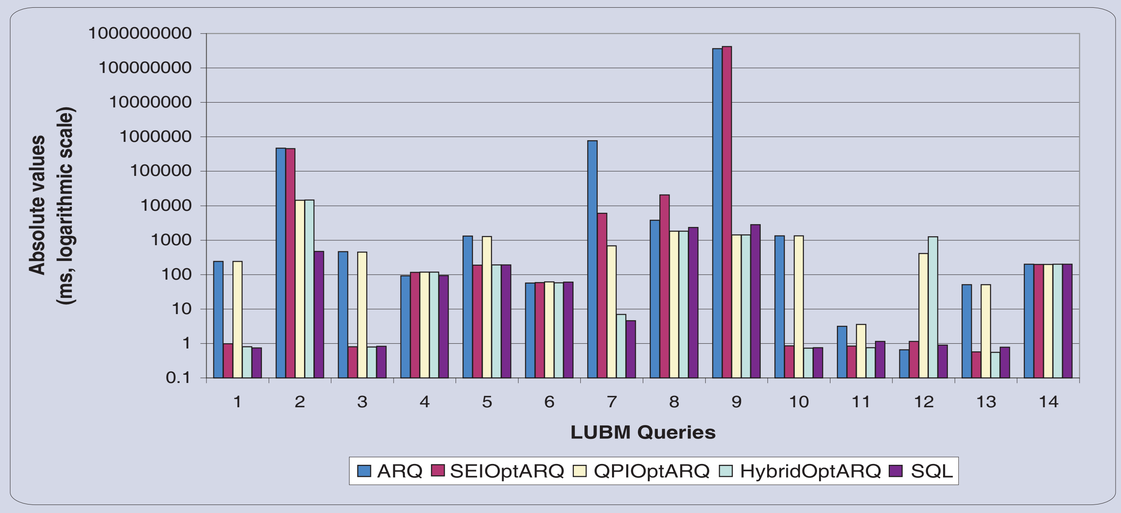
LUBM in-memory evaluation for ARQ, Sesame and the SEI/QPI/Hybrid optimization approaches.



LUBM in-memory evaluation for Sesame using the SEI/QPI/Hybrid optimized queries.



In-memory performance of LUBM query 7 for different ontology sizes.



LUBM on-disk evaluation using the SQL version of the SPARQL queries.

[1] **OptARQ: A SPARQL Optimization Approach based on Triple Pattern Selectivity Estimation.**
Abraham Bernstein, Christoph Kiefer, and Markus Stocker. Technical Report IFI-2007.02, Department of Informatics, University of Zurich, 2007.