LSQ: The Linked SPARQL Queries Dataset

Muhammad Saleem[1] Intizar Ali [2] Aidan Hogan [3]
Qaiser Mehmood[2] Axel Ngonga [1]

[1]AKSW, University of Leipzig, Germany

[2]INSIGHT, NUIG, Ireland

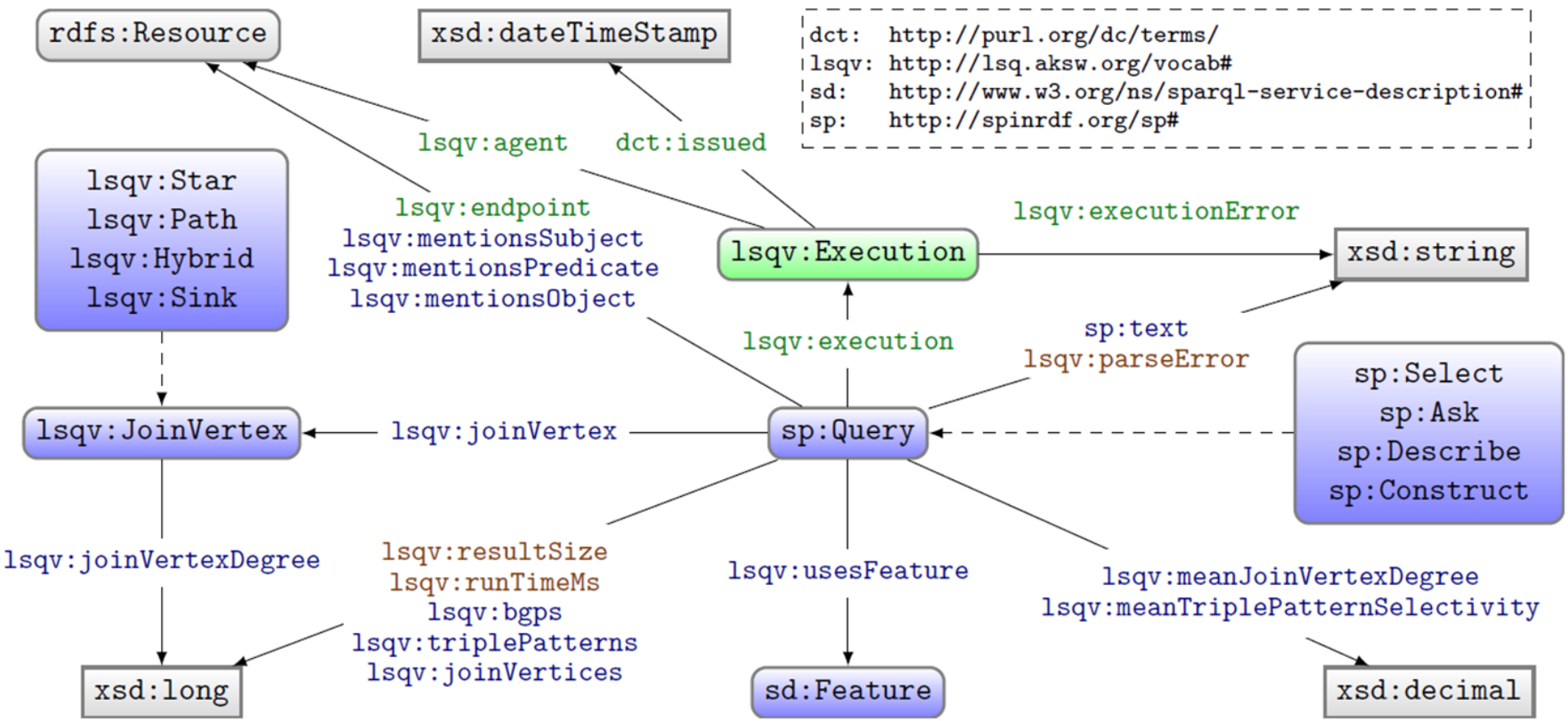
[3]DCC, Universidad de Chile

International Semantic Web Conference, Bethlehem, USA, 2015

LSQ: The Linked SPARQL Queries Dataset

- Linked Dataset of SPARQL queries extracted from endpoint logs
- DBpedia (232 million triples)
  - 30/04/2010–20/07/2010
- Linked Geo Data (1 billion triples)
  - 24/11/2010–06/07/2011
- Semantic Web Dog Food (300 thousand triples)
  - 16/05/2014–12/11/2014
- British Museum (1.4 million triples)
  - 08/11/2014–01/12/2014

LSQ Data Model



LSQ Statistics

High-level analysis of the queries and query executions in the LSQ dataset for each log (QE = Query Executions, UQ = Unique Queries, PE = Parse Errors, RE = Runtime Error, ZR = Zero Results, SEL = SELECT, CON = CONSTRUCT, DES = DESCRIBE; percentages are with respect to UQ)

Dataset	QE №	UQ №	PE №	RE №	ZR №	SEL %	CON %	DES %	ASK %
DBpedia	1,728,041	1,208,789	426,425	69,523	176,257	94.6	0.9	0.1	4.4
LGD	1,656,254	311,126	13,546	50,059	143,574	89.3	2.3	0.0	8.4
SWDF	1,411,483	99,165	13,645	475	25,674	68.8	0.0	31.1	0.1
BM	879,426	129,989	100,916	0	29,073	100	0.0	0.0	0.0
Overall	5,675,204	1,749,069	554,532	120,057	374,578	91.6	1.2	2.3	4.9

LSQ Statistics

Percentage of unique queries containing different types of joins (a query may contain multiple join types)

Dataset	Star %	Path %	Hybrid %	Sink %	No Join %
DBpedia	38.58	8.60	6.79	6.31	61.23
LGD	28.18	9.46	7.57	1.24	72.00
SWDF	10.70	11.25	4.01	0.93	84.25
BM	0.00	0.00	0.00	0.00	100.00
Overall	33.05	8.79	6.62	4.51	66.51

LSQ Statistics

Comparison of the mean values of different query features across all query logs (RS = Result Size, TPs = Triple Patterns, JVs = Join Vertices, MJVD = Mean Join Vertex Degree, MTPS = Mean Triple Pattern Selectivity)

Dataset	RS	BGPs	TPs	JVs	MJVD	MTPS	Runtime (ms)
DBpedia	87.57	1.81	2.22	0.40	0.78	0.002	20.26
LGD	161.90	1.75	2.16	0.37	0.75	0.030	32.28
SWDF	19.65	2.57	2.94	0.26	0.35	0.025	11.98
BM	0.00	1.00	1.00	0.00	0.00	0.000	6.78
Overall	122.45	1.74	2.04	0.24	0.45	0.013	26.40

LSQ Statistics

Percentage of queries using various specific SPARQL features

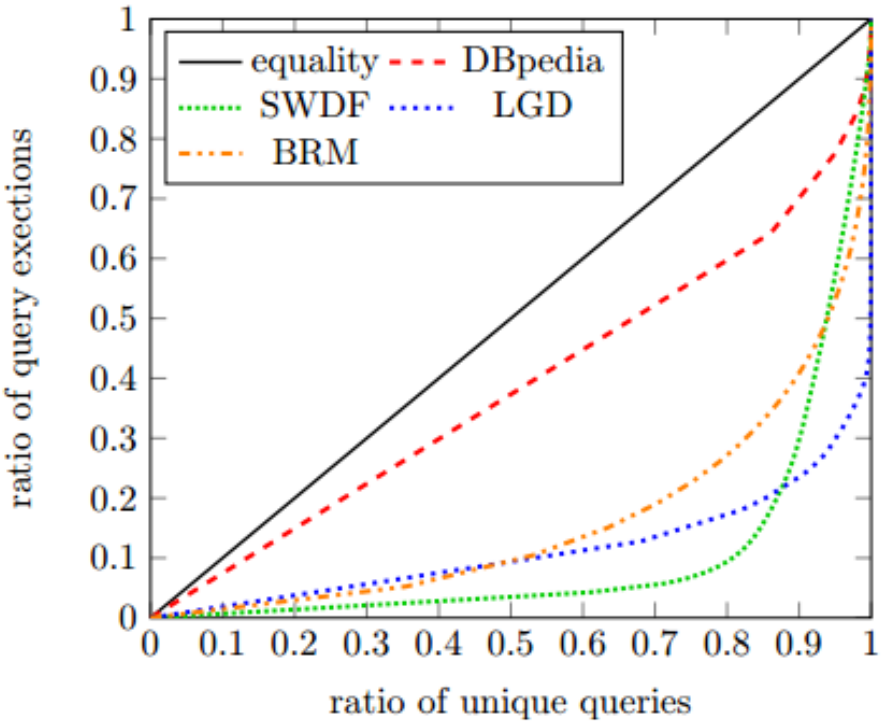
Dataset	UNION	OPTIONAL	DISTINCT	FILTER	REGEX	SERVICE	Sub-Query
DBpedia	4.42	36.20	18.44	23.47	2.90	0.0005	0.00
LGD	9.65	25.10	22.25	31.10	1.25	0.0000	0.01
SWDF	32.71	25.32	45.40	0.95	0.06	0.0012	0.02
BM	0.00	0.00	100.00	0.00	0.00	0.0000	0.00
Overall	7.64	31.78	23.30	23.19	2.22	0.0004	0.01

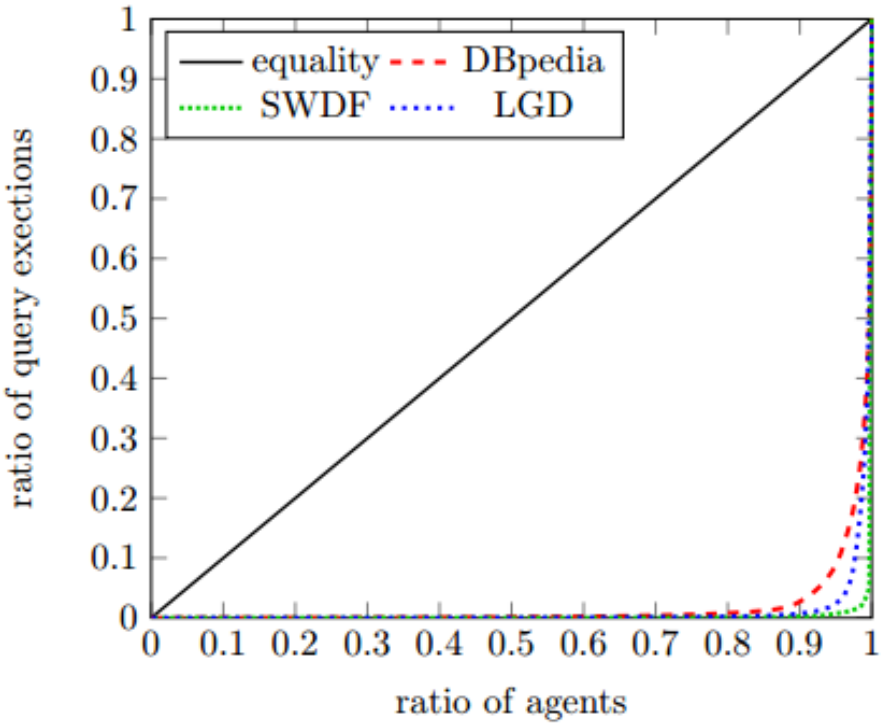
LSQ Statistics

Percentage of queries using various classes of features

Dataset	Solution Mod.	Aggregates	(¬)Exists	Binding	Graph
DBpedia	1.036	0.001	0.001	0.000	0.002
LGD	60.443	0.007	0.000	0.000	0.000
SWDF	33.265	2.405	0.001	0.008	0.001
BM	0.000	0.000	0.000	0.000	0.000
Overall	18.117	0.174	0.001	0.001	0.001

LSQ Statistics





90% of the agents issues fewer than 3% queries

Use-Cases

- Custom Benchmarks
- SPARQL Adoption
- Caching
- Usability
- Meta-Querying

Conclusion and Future Work

- First Linked Dataset of real-world SPARQL queries
- 5.7 million query executions, 73 million triples
- 90% of the agents issues fewer than 3% queries
- LSQ is available from (`http://aksw.github.io/LSQ/`)
- Add more logs, e.g., Bioportal, Strabon
- Update current logs (esp. DBpedia)
- Link to the benchmark generation framework FEASIBLE (`http://feasible.aksw.org/`)