Large-Scale SPARQL Query Log Analysis

Angela Bonifati* Wim Martens† Thomas Timm†

*Lyon 1 University †University of Bayreuth

Motivation and Goals

Understand:

What is the structure of user-generated queries?
How are advanced features (e.g., property paths) used?

Investigate on a **large corpus** (180M queries):

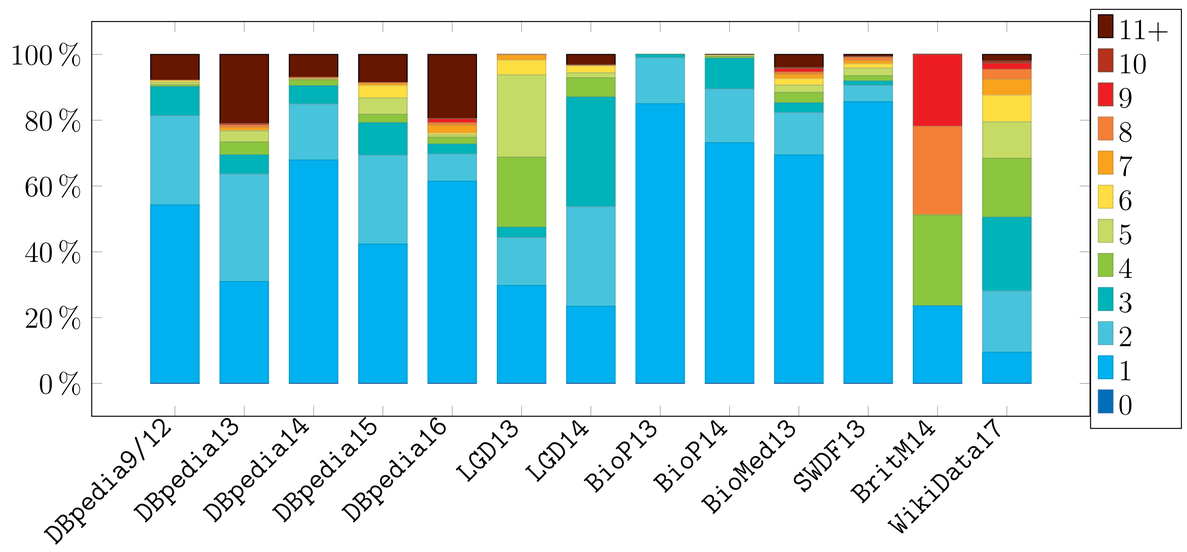
query size and **keyword usage**
shapes (and **treewidth**) of queries
inner **structure of property paths**
streaks, i.e. sequences of similar queries

The Investigated Query Logs

Source	*Total #Q*	*Valid #Q*	*Unique #Q*
DBpedia9/12	28,534,301	27,097,467	13,437,966
DBpedia13	5,243,853	4,819,837	2,628,005
DBpedia14	37,219,788	33,996,480	17,217,448
DBpedia15	43,478,986	42,709,778	13,253,845
DBpedia16	15,098,176	14,687,869	4,369,781
LGD13	1,841,880	1,513,868	357,842
LGD14	1,999,961	1,929,130	628,640
BioP13	4,627,271	4,624,430	687,773
BioP14	26,438,933	26,404,710	2,191,152
BioMed13	883,374	882,809	27,030
SWDF13	13,762,797	13,618,017	1,229,759
BritM14	1,523,827	1,513,534	135,112
WikiData17	309	308	308
Total	180,653,910	173,798,237	56,164,661

- **Diverse sources** (Biology, Museum, DBpedia, etc.) from **2009 to 2016**
- We focus on **unique and valid queries**

Size of Queries



Datasets	DBpedia9/12	DBpedia13	DBpedia14	DBpedia15	DBpedia16	LGD13	LGD14	BioP13	BioP14	BioMed13	SWDF13	BritM14	WikiData17
S/A	99.15%	91.88%	95.38%	93.05%	63.99%	29.01%	97.47%	100%	99.69%	12.87%	96.14%	98.64%	99.68%
$Avg_{\#T}$	2.38	3.98	2.09	2.94	3.78	3.19	2.65	1.16	1.42	2.44	1.51	5.47	3.94

- Size is measured by counting **nr. of triples per query**
- The **majority** of queries (over 90%) is **small** (<= 6 triples)
- Some datasets, e.g., BritM14, have different size distribution

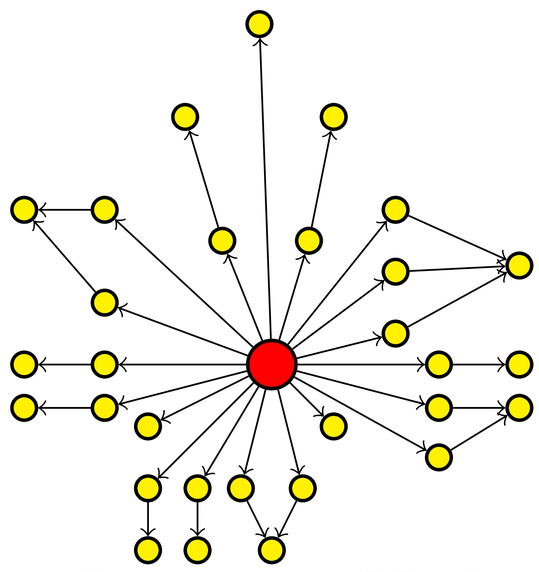
Keyword Usage

Element	*Absolute*	*Relative*	*Element*	*Absolute*	*Relative*
Select	49,409,913	87.97%	Distinct	12,198,198	21.72%
Ask	2,789,420	4.97%	Limit	9,545,249	17.00%
Describe	2,578,311	4.49%	Offset	3,455,500	6.15%
Construct	1,386,908	2.47%	Order By	1,159,231	2.06%
Filter	22,547,561	40.15%	Count	320,035	0.57%
And	15,863,942	28.25%	Max	3,660	0.01%
Union	10,465,706	18.63%	Min	3,632	0.01%
Opt	9,106,419	16.21%	Avg	263	< 0.01%
Graph	1,519,899	2.71%	Sum	68	< 0.01%
Not Exists	926,849	1.65%	Group By	168,444	0.30%
Minus	766,380	1.36%	Having	12,276	0.02%

- **Select** queries are frequently occurring
- **Aggregates** are used surprisingly sparsely

Shape Analysis for Conjuctive Queries

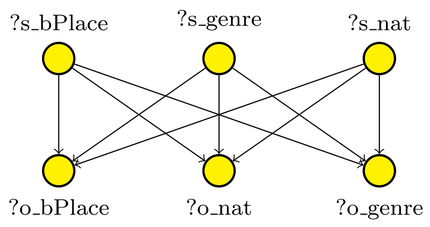
Shape	*#Queries*	*Relative %*
single edge	12,273,871	77.98%
chain	15,561,944	98.87%
chain set	15,570,042	98.93%
star	147,457	0.94%
tree	15,723,163	99.90%
forest	15,731,535	99.95%
cycle	4,550	0.03%
flower	15,730,043	99.94%
flower set	15,738,439	100.00%
treewidth ≤ 2	15,739,056	100.00%
treewidth $= 3$	1	0.00%
total	15,739,057	100.00%



A **flower query** in DBpedia

- Our shape classification achieves **full coverage** of all queries
- Low number of **cyclic queries** in relation to entire corpus

Treewidth Analysis and Complex Queries



A query with treewidth=3

- **Treewidth** generalizes graph acyclicity
- Forests (and all subclasses thereof) have **treewidth=1**
- Cycles, flowers, and flower sets have **treewidth=2**
- We found queries with **treewidth up to 3**

Property Paths (Navigational Regular Expressions)

Expression Type	ℓ	*Relative*	*Expression Type*	*Relative*
$(a_1 + \cdots + a_\ell)^*$	2–4	29.10%	$a^*b?$	< 0.01%
$.$		25.48%	abc^*	< 0.01%
a^*		19.66%	$(ab^*) + c$	< 0.01%
$a_1 \cdots a_\ell$	2–6	8.66%	$.^*$	< 0.01%
a^*b		7.73%	$(a_1 + a_2)?$	< 0.01%
$(a_1 + \cdots + a_\ell)$	1–6	6.61%	$.?$	< 0.01%
$(a_1 + \cdots + a_\ell)^+$	1–2	1.54%	$a^* + b$	< 0.01%
$a_1?a_2? \cdots a_\ell?$	1–5	1.15%	$a + b^+$	< 0.01%
$a(b_1 + b_2)?$		0.01%	$a^+ + b^+$	< 0.01%
$a_1a_2? \cdots a_\ell?$	2–3	0.01%	$(ab)^*$	< 0.01%
$A_1 \cdots A_\ell$	2–6	< 0.01%		

247,404 expressions

Notation:
. is a wildcard

A_i abbreviate $(a_1 + \cdots + a_n)$

- **Very few different types of expressions**
- **Structurally simple**:
  Transitive closures are used over (disjunctions of) symbols

Streaks – Query Evolution over Time

Length	*DBP14*	*DBP15*	*DBP16*
1–10	42,272	167,292	199,375
11–20	3,732	24,001	37,402
21–30	2,425	4,813	17,749
31–40	884	667	5,849
41–50	283	162	1,998
51–60	88	40	711
61–70	26	8	357
71–80	15	4	129
81–90	5	1	54
91–100	4	0	27
>100	5	0	24

- **Streaks**:
  Sequences of **similar queries within close distance** in the logs
- Results from logs over 3 days, on non-deduplicated logs

Conclusion

- Queries are often **structurally simple**
  (e.g., small treewidth, simple property paths)
- Query logs exhibit **streaks**,
  which motivates **multi-query optimization**
- We saw **significant differences** between
  different types of logs
- Demo tool: https://github.com/PoDMR/darql
  Fast interactive, visual exploration of queries