Benchmarking triple stores with biological data

Vladimir Mironov

NTNU, Trondheim, Norway

vladimir.mironov@bio.ntnu.no

SWAT4L
10 December 2010, Berlin, Germany

NTNU
Det skapande universitetet

Our Semantic Web Projects

Cell Cycle Ontology
http://www.semantic-systems-biology.org/cco
11,319,793 triples

BioGateway
http://www.semantic-systems-biology.org/biogateway
1,979,717,488 triples

NTNU
Det skapande universitetet

Cell Cycle Ontology graphs

cco_A_thaliana 356,903 triples
cco_A_thaliana_tc 469,484 triples

cco_S_pombe 406,131 triples
cco_S_pombe_tc 533,481 triples

cco_H_sapiens 836,622 triples
cco_H_sapiens_tc 1,076,760 triples

cco_S_cerevisiae 842,344 triples
cco_S_cerevisiae_tc 1,120,545 triples

cco 2,503,040 triples
cc_tc 3,170,556 triples

Cell Cycle Ontology queries

Retrieve the cell cycle proteins sharing the same function, location and processes
Retrieve functions, locations, processes of a specific cell cycle protein
Retrieve the cell cycle proteins that are'Breast cancer'-related as well as their interactions (if known)
Retrieve the cell cycle proteins that are a transformation of the given protein
Retrieve the core cell cycle proteins (IDs) participating in some known processes (in S pombe)
Retrieve the direct interactions in A thaliana and their participating cell cycle related proteins
Retrieve all the terms (protein, processes, etc) having 'cell cycle' as string in their names
Retrieve the core cell cycle proteins in S pombe that are located in the cell wall
Retrieve the core cell cycle protein names and their TAIR reference (AT code) in A thaliana
Retrieve the core cell cycle proteins located in the cytoplasm and having a hydrolysis-related function in A thaliana
Retrieve the name (label) of a given CCO id
Retrieve the neighbor terms of a given term
Retrieve the information of a given term
Retrieve the relation types used in CCO
Retrieve the children of a given term
Retrieve the parent terms of a given term
Retrieve the number of core cell cycle proteins in A thaliana
Retrieve terms based on specific key-words
Retrieve the number of cell cycle proteins in A thaliana (At ontology)
Retrieve the number of cell cycle genes in S pombe (Sp ontology)
Retrieve the number of cell cycle genes in S pombe (CCO ontology)
Retrieve all the protein-protein interactions in A thaliana
Retrieve the CCO id of a given specific term
Retrieve the CCO ids of a given term using regular expressions

Queries main features

	Simple Filters	More than 8 triple patterns	OPTIONAL operator	LIMIT modifier	ORDER BY modifier	DISTINCT modifier	REGEX operator	UNION operator	COUNT operator
Q1						x			
Q2								x	
Q3	x	x	x			x	x		
Q4									
Q5									
Q6									
Q7	x						x		
Q8						x			
Q9	x								
Q10	x	x				x	x		
Q11									
Q12								x	
Q13		x	x			x		x	
Q14					x				
Q15									
Q16									
Q17						x			x
Q18	x	x			x		x	x	
Q19						x		x	x
Q20						x			x
Q21						x			x
Q22									
Q23									
Q24	x			x			x		

Nature Precedings : doi:10.1038/npre.2010.5417.1 : Posted 19 Dec 2010

NTNU
Det skapande universitetet

Triple stores

- Virtuoso OpenSource 6.0.0
- Swift OWLIM 2.9.1
- 4Store 1.0.2
- Jena SDB 1.3.1
- Jena TDB 0.8.2

Benchmarking procedure

240 data points for each store:
- 10 graphs
- 24 queries

3 independent experiments for each store
- the store is cleared completely
- the graphs are loaded each time in the same order
- the queries are executed each time in the same order

For each store for the 3 experiments:
- the execution times averaged to produce 240 values
- relative standard errors were calculated to produce 240 values

Bird's eye view

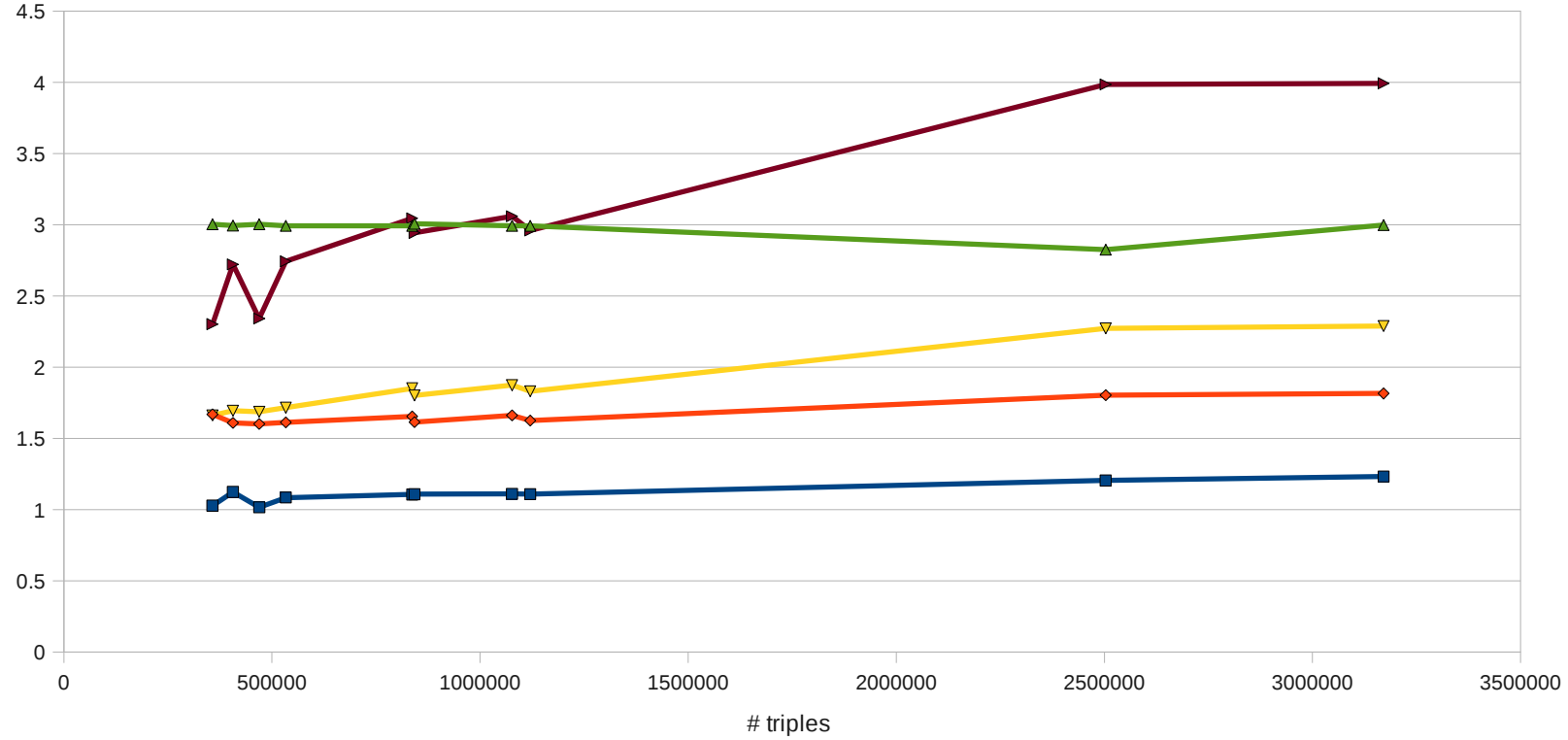
Store	Total time	RSE
Virtuoso	203.939	0.053
Jena SDB	730.492	0.020
Jena TDB	1445.572	0.235
OWLIM	14257.964	0.156
4Store	47566.530	0.097

More details

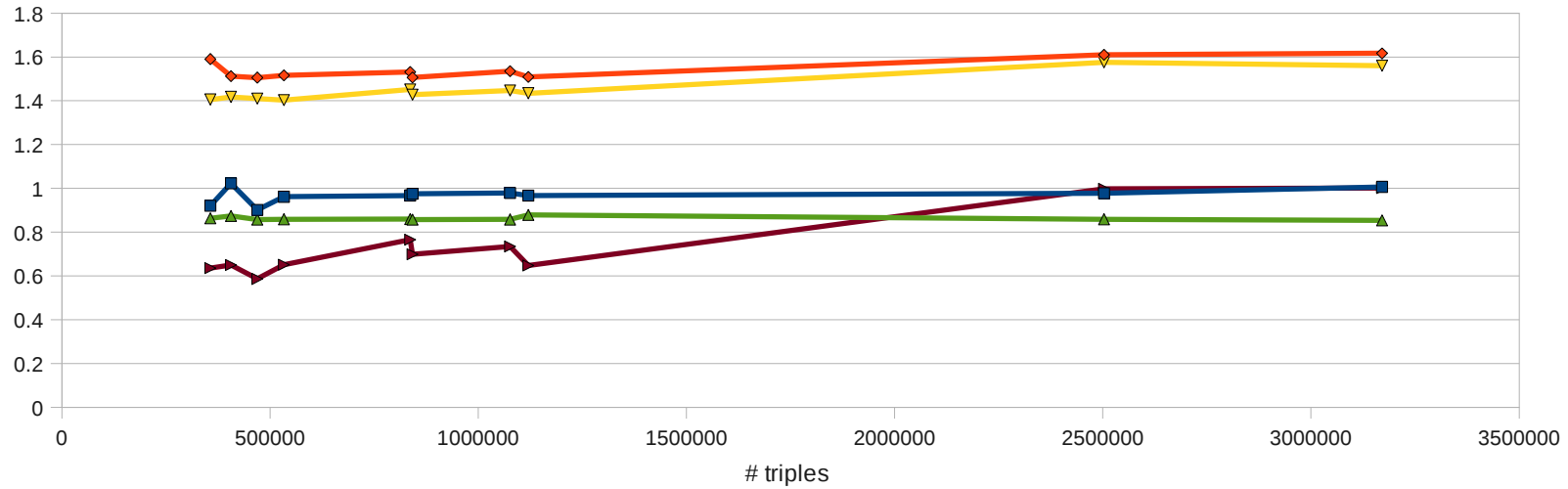
Query	Virtuoso	Jena SDB	Jena TDB	4Store	OWLIM	Avg
Q5	2.639	13.446	11.000	1.526	0.408	5.804
Q23	5.630	13.343	10.454	1.343	0.009	6.156
Q11	5.343	13.339	10.703	1.419	0.011	6.163
Q16	5.617	13.345	10.825	1.346	0.009	6.228
Q15	6.163	13.342	10.544	1.390	0.018	6.291
Q22	5.170	13.709	10.981	1.428	0.173	6.292
Q4	5.916	13.348	10.773	1.539	0.017	6.319
Q8	7.094	13.336	10.449	1.577	0.049	6.501
Q12	7.198	13.373	10.731	1.400	0.030	6.546
Q2	7.281	13.337	10.768	1.438	0.052	6.575
Q6	4.054	14.523	10.573	1.779	2.020	6.590
Q19	2.065	13.390	9.795	1.326		6.644
Q9	5.820	13.711	10.699	2.133	1.067	6.686
Q21	3.379	13.335	9.818	1.316		6.962
Q10	4.679	13.757	11.676	2.529	4.664	7.461
Q17	5.648	13.390	10.119	1.350		7.627
Q20	6.110	13.387	10.686	1.315		7.875
Q1	1.897	18.064	14.258	1.647	8.024	8.778
Q13	1.658	52.545	14.156	1.569	0.034	13.992
Q24	2.813	24.719	38.619	14.366	27.242	21.552
Q7	2.617	26.519	39.248	14.110	28.996	22.298
Q14	5.775	13.338	46.433	1.401	91.894	31.768
Q3	3.358	30.476	27.049	3.654	1121.702	237.248
Q18	22.840	76.013	493.596	24999.569	8325.734	6783.550

Nature Precedings : doi:10.1038/npre.2010.5417.1 : Posted 19 Dec 2010

Scalability



Scalability



Conclusions

- Virtuoso is a safe choice
- Avoid Jena
- OWLIM and 4Store have unacceptably slow responses for some queries
- Consider OWLIM for particularly large stores
- Avoid 4Store in case of very large stores

Acknowledgements

My thanks

- to my co-authors for their contribution,
- to FUGE Midt Norge for financial support,
- to all of you for your attention.

NTNU
Det skapande universitetet

NTNU
Det skapande universitetet