USAGE-CENTRIC BENCHMARKING OF RDF TRIPLE STORES

Mohamed Morsey, Jens Lehmann, Sören Auer and Axel-Cyrille Ngonga Ngomo

AKSW Research Group
University of Leipzig
Why Benchmarking?

- Highlight weak and strong points of systems
- Stimulate technical progress, make technology viable
- Part of the Scientific Method
- Track advancement in an area over time
- Make competing products comparable
History: (Relational) Databases

TPC price/perf trend 1990-2005: improved 58% per year
prices have declined 37%/y

TPC-A
price/perf trend
37% per year

TPC-C

TPC-C price/perf trend
37% per year

→ 15% algorithmic improvement per year

Goal: Devise a Benchmark Methodology which can be applied to RDF datasets

Apply it to a dataset in the LOD cloud
Goal: Devise a Benchmark Methodology which can be applied to RDF datasets

Apply it to a dataset in the LOD cloud
Goal: Devise a Benchmark Methodology which can be applied to RDF datasets

Apply it to a dataset in the LOD cloud
Benchmarking LOD Technology

- Goal: Devise a Benchmark Methodology which can be applied to RDF datasets
- Apply it to a dataset in the LOD cloud
Goal: Devise a Benchmark Methodology which can be applied to RDF datasets

Apply it to a dataset in the LOD cloud
Benchmarking and Reality

- Benchmarks for RDF triple stores are important
- They help to steer the direction of research
Benchmarks for RDF triple stores are important
They help to steer the direction of research

… but …
Benchmarks for RDF triple stores are important
They help to steer the direction of research

... but ...

Steering in the wrong Direction may not be helpful

Titanic 1912
Benchmarks for RDF triple stores are important
They help to steer the direction of research

... but ...

Steering in the wrong 
Direction may not be helpful

How can we ensure that a benchmark reflects reality?
Approach

Dataset

Query-Log

Triple Store

Benchmark Results
We started DBpedia with FUB and OpenLink in 2007

Extracts Information from Wikipedia and links it to other knowledge bases
- Freebase, OpenCyc, schema.org, UMBEL, GeoNames, Musicbrainz, CIA World Factbook, DBLP, Project Gutenberg, LinkedGeoData, DBtune Jamendo, Eurostat, Uniprot, Bio2RDF, US Census

Widely used:
- Applications: BBC, New York Times, Thomson Reuters, etc.
- 90+ languages
- > 1 billion RDF triples
- Community-maintained ontology
- 1000+ citations in research community
- >10 million queries per month against official triple store
## Motivation: Existing Benchmarks

<table>
<thead>
<tr>
<th></th>
<th>LUBM</th>
<th>SP²Bench</th>
<th>BSBM V3</th>
<th>DBPSB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test data</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td>Test queries</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td>Largest tested dataset</td>
<td>6.9M</td>
<td>25M</td>
<td>200M</td>
<td>300M</td>
</tr>
<tr>
<td>No. of queries</td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>20/25</td>
</tr>
<tr>
<td>Use case</td>
<td>Universities</td>
<td>Publications</td>
<td>E-commerce</td>
<td>Cross Domain</td>
</tr>
<tr>
<td>Classes</td>
<td>43</td>
<td>8</td>
<td>8</td>
<td>239</td>
</tr>
<tr>
<td>Properties</td>
<td>32</td>
<td>22</td>
<td>51</td>
<td>1200</td>
</tr>
</tbody>
</table>
Motivation: Existing Benchmarks

<table>
<thead>
<tr>
<th></th>
<th>LUBM</th>
<th>SP²Bench</th>
<th>BSBM V3</th>
<th>DBPSB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test data</strong></td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td><strong>Test queries</strong></td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td><strong>Largest tested dataset</strong></td>
<td>6.9M</td>
<td>25M</td>
<td>200M</td>
<td>300M</td>
</tr>
<tr>
<td><strong>No. of queries</strong></td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>20/25</td>
</tr>
<tr>
<td><strong>Use case</strong></td>
<td>Universities</td>
<td>Publications</td>
<td>E-commerce</td>
<td>Cross Domain</td>
</tr>
<tr>
<td><strong>Classes</strong></td>
<td>43</td>
<td>8</td>
<td>8</td>
<td>239</td>
</tr>
<tr>
<td><strong>Properties</strong></td>
<td>32</td>
<td>22</td>
<td>51</td>
<td>1200</td>
</tr>
</tbody>
</table>
### Motivation: Existing Benchmarks

<table>
<thead>
<tr>
<th></th>
<th>LUBM</th>
<th>SP²Bench</th>
<th>BSBM V3</th>
<th>DBPSB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test data</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td>Test queries</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td>Largest tested dataset</td>
<td>6.9M</td>
<td>25M</td>
<td>200M</td>
<td>300M</td>
</tr>
<tr>
<td>No. of queries</td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>20/25</td>
</tr>
<tr>
<td>Use case</td>
<td>Universities</td>
<td>Publications</td>
<td>E-commerce</td>
<td>Cross Domain</td>
</tr>
<tr>
<td>Classes</td>
<td>43</td>
<td>8</td>
<td>8</td>
<td>239</td>
</tr>
<tr>
<td>Properties</td>
<td>32</td>
<td>22</td>
<td>51</td>
<td>1200</td>
</tr>
</tbody>
</table>
## Motivation: Existing Benchmarks

<table>
<thead>
<tr>
<th></th>
<th>LUBM</th>
<th>SP²Bench</th>
<th>BSBM V3</th>
<th>DBPSB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test data</strong></td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td><strong>Test queries</strong></td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td><strong>Largest tested dataset</strong></td>
<td>6.9M</td>
<td>25M</td>
<td>200M</td>
<td>300M</td>
</tr>
<tr>
<td><strong>No. of queries</strong></td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>20/25</td>
</tr>
<tr>
<td><strong>Use case</strong></td>
<td>Universities</td>
<td>Publications</td>
<td>E-commerce</td>
<td>Cross Domain</td>
</tr>
<tr>
<td><strong>Classes</strong></td>
<td>43</td>
<td>8</td>
<td>8</td>
<td>239</td>
</tr>
<tr>
<td><strong>Properties</strong></td>
<td>32</td>
<td>22</td>
<td>51</td>
<td>1200</td>
</tr>
</tbody>
</table>
### Motivation: Existing Benchmarks

<table>
<thead>
<tr>
<th></th>
<th>LUBM</th>
<th>SP²Bench</th>
<th>BSBM V3</th>
<th>DBPSB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test data</strong></td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td><strong>Test queries</strong></td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td><strong>Largest tested dataset</strong></td>
<td>6.9M</td>
<td>25M</td>
<td>200M</td>
<td>300M</td>
</tr>
<tr>
<td><strong>No. of queries</strong></td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>20/25</td>
</tr>
<tr>
<td><strong>Use case</strong></td>
<td>Universities</td>
<td>Publications</td>
<td>E-commerce</td>
<td>Cross Domain</td>
</tr>
<tr>
<td><strong>Classes</strong></td>
<td>43</td>
<td>8</td>
<td>8</td>
<td>239</td>
</tr>
<tr>
<td><strong>Properties</strong></td>
<td>32</td>
<td>22</td>
<td>51</td>
<td>1200</td>
</tr>
</tbody>
</table>
**Motivation: Existing Benchmarks**

<table>
<thead>
<tr>
<th></th>
<th>LUBM</th>
<th>SP²Bench</th>
<th>BSBM V3</th>
<th>DBPSB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test data</strong></td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td><strong>Test queries</strong></td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td><strong>Largest tested dataset</strong></td>
<td>6.9M</td>
<td>25M</td>
<td>200M</td>
<td>300M</td>
</tr>
<tr>
<td><strong>No. of queries</strong></td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>20/25</td>
</tr>
<tr>
<td><strong>Use case</strong></td>
<td>Universities</td>
<td>Publications</td>
<td>E-commerce</td>
<td>Cross Domain</td>
</tr>
<tr>
<td><strong>Classes</strong></td>
<td>43</td>
<td>8</td>
<td>8</td>
<td>239</td>
</tr>
<tr>
<td><strong>Properties</strong></td>
<td>32</td>
<td>22</td>
<td>51</td>
<td>1200</td>
</tr>
</tbody>
</table>

*August 26, 2012*
## Motivation: Existing Benchmarks

<table>
<thead>
<tr>
<th></th>
<th>LUBM</th>
<th>(SP^2\text{Bench})</th>
<th>BSBM V3</th>
<th>DBPSB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test data</strong></td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td><strong>Test queries</strong></td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td><strong>Largest tested dataset</strong></td>
<td>6.9M</td>
<td>25M</td>
<td>200M</td>
<td>300M</td>
</tr>
<tr>
<td><strong>No. of queries</strong></td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>20/25</td>
</tr>
<tr>
<td><strong>Use case</strong></td>
<td>Universities</td>
<td>Publications</td>
<td>E-commerce</td>
<td>Cross Domain</td>
</tr>
<tr>
<td><strong>Classes</strong></td>
<td>43</td>
<td>8</td>
<td>8</td>
<td>239</td>
</tr>
<tr>
<td><strong>Properties</strong></td>
<td>32</td>
<td>22</td>
<td>51</td>
<td>1200</td>
</tr>
</tbody>
</table>
Query List Generation

- Query Log
  - Preprocessing
  - Clustering
- Clean Queries
- Query Clusters
  - Feature Selection
- Query List
Query List Generation

- Query Log
  - Preprocessing

- Clean Queries
  - Clustering

- Query Clusters
  - Feature Selection

- Query List
Preprocessing

Variable Unification

Discard Uncommon Queries

Common Keyword Removal
Preprocessing

Variable Unification

Discard Uncommon Queries

Common Keyword Removal

SELECT ?o WHERE {
  ?s foaf:homepage ?o .
  ?s rdf:type <http://rdfs.org/void#Dataset>. }

Morsey, Lehmann, Auer, Ngonga: Usage-Centric Benchmarking of RDF Triple Stores

August 26, 2012
Preprocessing

- Variable Unification
- Discard Uncommon Queries
- Common Keyword Removal

```
SELECT ?var0 WHERE {
  ?var1 foaf:homepage ?var0 .
  ?var1 rdf:type <http://rdfs.org/void#Dataset> .
}
```
Preprocessing

Variable Unification

Discard Uncommon Queries

Common Keyword Removal

Query Frequency > 10 → keep

SELECT ?var0 WHERE {
  ?var1 foaf:homepage ?var0 .
  ?var1 rdf:type <http://rdfs.org/void#Dataset> .
}
Preprocessing

Variable Unification

Discard Uncommon Queries

Common Keyword Removal

?var0 { ?var1 foaf:homepage ?var0 .
?var1 rdf:type <http://rdfs.org/void#Dataset>. }
Query List Generation

- Query Log
  - Preprocessing

- Clean Queries
  - Clustering

- Query Clusters
  - Feature Selection

- Query List
Clustering

Similarity Computation

Clustering
Clustering

- Similarity: 17 SPARQL features + Levenshtein distance
- Brute Force = $O(n^2) = 2.6$ Billion Comparisons for DBpedia
- LIMES Framework: 3% of brute force runtime
Clustering

- Time Efficient + Soft Clustering (a query can belong to several clusters)
- Borderflow Algorithm generated 622 multi-node clusters for DBpedia
Query List Generation

- Query Log
- Preprocessing
- Clustering
- Clean Queries
- Query Clusters
- Feature Selection
- Query List
Query Features and Variability

SPARQL Feature Selection

Query Variability
Query Features and Variability

- Number of triple patterns: \(|GP|\)
- Pattern constructors: \texttt{UNION, OPTIONAL}
- Solution modifiers: \texttt{DISTINCT}
- Filter conditions: \texttt{LANG, REGEX, STR, …}
Query Features and Variability

- **Original Query:**

```
SELECT ?var0 WHERE {
  ?var1 foaf:homepage ?var0.
  ?var1 rdf:type <http://rdfs.org/void#Dataset>.
}
```
Base Query:

```sparql
SELECT ?var0 WHERE {
  ?var1 foaf:homepage ?var0.
  ?var1 rdf:type %var%% .
}
```
Auxiliary Query:

```sql
SELECT DISTINCT ?var WHERE {
    ?var1 foaf:homepage ?var0.
    ?var1 rdf:type ?var.
} LIMIT 1000
```
Approach

Dataset

Query-Log

Triple Store

Benchmark Results
Dataset Generation

- Generation of different dataset sizes to judge scalability
- Should resemble the original RDF data
- Applicable to other datasets
Dataset Generation

\[ \text{Diagram: } C_0 \rightarrow \{I_{0,0}, I_{0,1}, I_{0,2}, \ldots, I_{0,n_0}\} \rightarrow C_1 \rightarrow \{I_{1,0}, I_{1,1}, I_{1,2}, \ldots, I_{1,n_1}\} \rightarrow \ldots \rightarrow \{I_{k,0}, I_{k,1}, I_{k,2}, \ldots, I_{k,n_k}\} \]
1- Select all classes
Dataset Generation

1- Select all classes
2- Select seed instances
1- Select all classes
2- Select seed instances
3- Follow links
Dataset Generation

1- Select all classes
2- Select seed instances
3- Follow links
1- Select all classes  
2- Select seed instances  
3- Follow links
Dataset Generation

1- Select all classes
2- Select seed instances
3- Follow links
Dataset Generation

1- Select all classes
2- Select seed instances
3- Follow links
4- Continue until required percentage reached
Experimental Setup

- Tested Triple Stores:
  - Sesame
  - Jena TDB
  - Virtuoso
  - Big Owlim

- Dataset sizes: 10%, 50%, 100%

- Similar configurations for all triples stores (main memory)

- Benchmark execution:
  - System Restart
  - Warm-Up Phase (20 Minutes)
  - Hot-Run Phase (60 Minutes)
Results: Queries per Second

- Similar Analysis for other Dataset Sizes
- Allows to Analyse Weak and Strong areas (relative to features and their combination) → details in ISWC’11 and AAAI’12 papers and [http://aksw.org/Projects/DBPSB](http://aksw.org/Projects/DBPSB)
Results: Query Mixes per Hour

- Virtuoso fastest and most scalable triple store
- Large Performance Differences unlike those observed in „relational database style“ RDF benchmarks
Conclusions & Future Work

- First triple store benchmark based on real queries and real data
- Allows to test variety of SPARQL features
- Applicable to arbitrary datasets
- Future work:
  - SPARQL Update (In Progress)
  - Evaluate Reasoning Performance (SPARQL 1.1 Entailment)
  - Improve Dataset Generation (Duan et al., „Apples and oranges: a comparison of RDF benchmarks and real RDF datasets”, SIGMOD 2011)
  - Application on other Datasets
  - Increase number of tested triple stores
Future of RDF Benchmarking

- Linked Data Benchmarking Council (LDBC)
- Vendor cooperation to establish accepted RDF/Graph database benchmarks:
  - LDBC Foundation of graph and RDF DB vendors
  - Good initial set of benchmarks and results
- DBpedia Benchmark will be part of LDBC
Thank you

Jens Lehmann
http://jens-lehmann.org

Group website: http://aksw.org
Project website: http://aksw.org/Projects/DBPSB