ontotext

LDBC Semantic Publishing Benchmark Evolution

March 2015

Outline

- **Motivation**
- The New Reference Datasets
- Changes in the generated metadata
- Ontologies and Required Rule-sets
- The Concrete Inference patterns
- Changes in the Basic Interactive Query Mix
- GraphDB Configuration Disclosure
- Future plans

Change is Needed

SPB needed to evolve in order to:

- Allow for retrieval of semantically relevant content
  - Based on rich metadata descriptions and diverse reference knowledge
- Demonstrate that triplestores offer simplified and more efficient querying

And this way to:

- Cover more advanced and pertinent usage of triplestores
- Present a bigger challenge to the engines

Directions for Improvement

- *Background and issues with SPB 1.0:*
  - *"Reference data" = master data in Dynamic Semantic Publishing*
    - *Essentially, taxonomies and entity datasets used to describe "creative works"*
  - *SPB 1.0 has very small set of reference data*
  - *Reference data is not interconnected – few relations between entities*
  - *All descriptions of content assets (articles, pictures, etc.) refer to the Reference Data, but not to one another*
  - *Star-shaped graph with small reference dataset in the middle*
    - *This way SPB is not using the full potential of graph databases*
- Bigger and better connected Reference Data
  - More reference data and more connections between the entities
- A better connected dataset
  - Make cross references between the different pieces of content

Directions for Improvement (2)

- Make use of inference
  - In SPB 1.0 it is really trivial: couple of subclasses and sub-properties
  - It would be relevant to have some transitive, symmetric and inverse properties
  - Such semantics can foster retrieval of relevant content
- More interesting and challenging query mixes
  - This can go in plenty of different directions, but the most obvious one is to evolve the queries so that they take benefit from changes above
  - On the bigger picture, there should be a way to make some of the queries nicer, i.e. simpler and cleaner
- Testing high-availability
  - FT acceptance tests for HA cluster are great starting point
  - Too ambitious for SPB 2.0

Outline

- Motivation
- **The New Reference Datasets**
- Changes in the generated metadata
- Ontologies and Required Rule-sets
- The Concrete Inference patterns
- Changes in the Basic Interactive Query Mix
- GraphDB Configuration Disclosure
- Future plans

Extended Reference Dataset

- 22M explicit statements total in SPB v.2.0
  - vs. 127k statements in SPB 1.0; BBC lists + tiny fractions of GeoNames
- Added reference data from DBPedia 2014
  - Extracts alike: **`DESCRIBE ?e WHERE { ?e a dbp-ont:Company }`**
  - Companies (85 000 entities)
  - Events (50 000 entities)
  - Persons (1M entities)
- Geonames data for Europe
  - All European countries, w/o RUS, UKR, BLR; 650 000 locations total
  - Some properties and locations types irrelevant to SPB excluded
- ***owl:sameAs*** links between DBpedia and Geonames
  - 500k owl:sameAs mappings

Interconnected Reference Dataset

- Substantial volume of connections between entities
- Geonames comes with hierarchical relationship, defining nesting of locations, ***gn:parentFeature***
- DBPedia inter-entity relationships between entities*:

	To Company	To Person	To Place / gn:Feature	To Event
Company	40,797	26,675	218,636	18
Person	89,506	1,324,425	3,380,145	145,892
Event	5,114	154,207	140,579	35,442

* numbers shown in table are approximate

RDF-Rank for the Reference Data

- GraphDB's RDFRank feature is used to calculate a rank for all the entities in the Reference Data
  - RDFRank calculates a measure of importance for all URIs in an RDF graph, based on Google's PageRank algorithm
- These RDFRanks are calculated using GraphDB after the entire Reference Data is loaded
  - This way all sorts of relationships in the graph are considered during calculations, including such that were logically inferred
- RDFRanks are inserted using predicate

  ```
  <http://www.ldbcouncil.org/spb#hasRDFRank>
  ```
- These ranks are available as a part of the Ref. Data
  - No need to get compute them again during loading

RDF-Rank Used for Popular Entities

- The Data Generator uses a set of "popular entities"
  - Those are referred in 30% of the content-to-entity relations/tags
  - This is one of the heuristics used by the data generator to produce more realistic data distributions
  - This is implemented in SPB 1.0, no change in SPB 2.0
- Popular entities are those with top 5% RDF Rank
  - This is the new thing in SPB v.2.0
  - Before that the popular entities were selected randomly
  - This way, we get a more realistic dataset where those entities which are more often used to tag content also have better connectivity in the Reference Data
- In the future, RDF-ranks can be used for other purposes also

Outline

- Motivation
- The New Reference Datasets
- **Changes in the generated metadata**
- Ontologies and Required Rule-sets
- The Concrete Inference patterns
- Changes in the Basic Interactive Query Mix
- GraphDB Configuration Disclosure
- Future plans

Changes in the Metadata

- Generated three times more relationships between Creative Works and Entities, than in SPB 1.0
  - More recent use cases in publishing, adopt rich metadata descriptions with more than 10 references to relevant entities and concepts
  - In SPB 1.0 there are on average a bit less than 3 entity references per creative work, based on distributions from an old BBC archive
  - While developing SPB 2.0 we didn't have access to substantial amount of rich semantic metadata from publisher's archive, to allow us derive content-to-concept frequency distributions
  - So, we decided to use the same distributions from BBC, but to triple the concept references for each of them
  - In SPB 2.0 there are on average about 8 content-to-concept references
  - This results in about 25 statements/creative work description
  - All other heuristics and specifics of the metadata generator were preserved (popular entities, CW sequences alike storylines, etc.)

Changes in the Metadata (2)

- Geonames URIs used for content-to-location references
  - As in SPB v.1.0, each creative work description refers (through ***cwork:Mention*** property) to exactly one location
    - These references are exploited in queries with geo-spatial constraints
  - While most of the geo-spatial information in the Reference Data comes from Geonames, for about 500 thousand locations there are also DBPedia URIs, that come from the DBPedia-to-Geonames ***owl:sameAs*** mappings
  - Intentionally, DBPedia URIs are not used when metadata about creative works is generated
    - In contrast, the substitution parameters for locations in the corresponding queries use only DBpedia locations
    - This way the corresponding queries would return no results if proper SameAs expansion doesn't exist

Changes in the Metadata - Stats

- As a result of changes in the size of the Reference Data and the metadata size, there are differences in the number of Creative Works included at the same scale factor in SPB v.1.0 and in SPB v.2.0

	SPB 1.0	SPB 2.0
Reference Data size (explicit statements)	**170k**	**22M**
Creative Work descr. size (explicit st./CW)	**19**	**25**
Metadata in 50M dataset (explicit statements)	50M	28M
Creative Works in 50M datasets (count)	2.6M	1.1M
Creative Work-to-Entity relationships in 50M	7M	9M
Metadata in 1B dataset (explicit statements)	1B	978M
Creative Works in 1B datasets (count)	53M	39M
Creative Work-to-Entity relationships in 1B	137M	313M

Outline

- Motivation
- The New Reference Datasets
- Changes in the generated metadata
- **Ontologies and Required Rule-sets**
- The Concrete Inference patterns
- Changes in the Basic Interactive Query Mix
- GraphDB Configuration Disclosure
- Future plans

The Rules Sets

- Complete inference in SPB 2.0 requires support for the following primitives:
  - RDFS: subPropertyOf, subClassOf
  - OWL: TransitiveProperty, SymmetricProperty, sameAs
- Any triplestore with OWL 2 RL reasoning support will be able to process SPB v2.0 correctly
  - In fact, a much reduced rule-set is sufficient for complete reasoning in as in OWL 2 RL there are host of primitives and rules that SPB 2.0 does not make use of. Such examples are all onProperty restrictions, class and property equivalence, property chains, etc.
  - The popular (but not W3C standardized) OWL Horst profile is sufficient for reasoning with SPB v 2.0
    - OWL Horst refers to the pD* entailment defined by Herman Ter Horst in: *Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary. J. Web Sem. 3(2-3): 79-115* (2005)

New and Updated Ontologies

- Modified versions of the BBC Ontologies
  - As in SPB v.1.0, BBC Core ontology defines relationships between Person – Organizations – Locations - Events
  - Those were mapped to corresponding DBPedia and Geonames classes and relationships
    - bbccore:Thing is defined as super class of dbp-ont:Company, dbp-ont:Event, foaf:Person, geonames-ontology:Feature (geographic feature)
  - dbp-ont:parentCompany is defined to be owl:TransitiveProperty
  - These extra definitions are added at the end of the BBC Core ontology
- Added a SPB-modified version of the Geonames ontology
  - Stripped out unnecessary pieces such onProperty restrictions that impose cardinality constraints

Outline

- Motivation
- The New Reference Datasets
- Changes in the generated metadata
- Ontologies and Required Rule-sets
- **The Concrete Inference patterns**
- Changes in the Basic Interactive Query Mix
- GraphDB Configuration Disclosure
- Future plans

Concrete Inference Patterns

- Transitive closure of location-nesting relationships
  - ***geo-ont:parentFeature*** property is used in GeoNames to link each location to larger locations that it is a part of
  - ***geo-ont:parentFeature*** is ***owl:TransitiveProperty*** in the GN ontology
  - This way if Munich has ***gn:parentFeature*** Bavaria, that on its turn has ***gn:parentFeature*** Germany, than reasoning should make sure that Germany also appears as ***gn:parentFeature of Munich***
- Transitive closure of ***dbp-ont:ParentCompany***
  - Inference unveils indirect company control relationships
- ***owl:sameAs*** relations between Geonames features and DBPedia URIs for the same locations:
  - The standard semantics of ***owl:sameAs*** requires that each statement asserted using the Geonames URIs should be inferable/retrievable also with the mapped DBPedia URIs

Concrete Inference Patterns (2)

- The inference patterns from SPB v.1.0 are still there:
  - ***cwork:tag*** statements inferred from each of its sub-properties ***cwork:about*** and ***cwork:mentions***
  - ***< ?cw rdf:type cwork:CreativeWork >*** statements inferred for instances of each of its sub-classes
  - These two are the only reasoning patterns that apply to the generated content metadata; all the others apply only to the Reference Data

Inference Statistics and Ratios

- If brute-force materialization is used, the Reference Data expands from 22M statements to about 100M
  - If ***owl:sameAs*** expansion is not considered, materialization on top of the Reference Data adds about 7M statements – most of those coming from the transitive closure of ***geo-ont:parentFeature***
  - Several triplestores (e.g. ORACLE and GraphDB) that use forward-chaining have specific mechanism, which allow them to handle ***owl:sameAs*** reasoning in a sort of hybrid manner, without expanding their indices. After materialization, such engines will have to deal with 29M statements of Reference Data, instead of 100M
- Brute-force materialization of the generated metadata describing creative works would double it
  - In SPB v.1.0 the expansion factor was 1.6, but now there are more entity references and also ***owl:sameAs*** equivalents of the location URIs

Outline

- Motivation
- The New Reference Datasets
- Changes in the generated metadata
- Ontologies and Required Rule-sets
- The Concrete Inference patterns
- **Changes in the Basic Interactive Query Mix**
- GraphDB Configuration Disclosure
- Future plans

Modified Queries in Basic Interactive

- SPB 2.0 changes only the Basic Interactive Query-Mix
  - The other query mixes are mostly unchanged
- In most of the queries, ***cwork:about*** and ***cwork:mention*** properties that link creative work to an entity, have been replaced with their supper ***cwork:tag***
  - This applies also to the Advanced Interactive and the Analytical use cases
  - For most of the queries, both types of relations are equally relevant
  - Using the super-property make the query simpler; as compared to other approaches to make a query that uses both (e.g. UNION)

New Queries

- Basic Interactive query-mix now adds 2 new queries exploring the interconnectedness in reference data
- **Q10:** Retrieve CWs that mention locations in the same province (A.ADM1) as the specified one
  - There is additional constraint on time interval (5 days)
- **Q11:** Retrieve the most recent CWs that are tagged with entities, related to a specific popular entity
  - Relations can be inbound and outbound; explicit or inferred

Q10: News from the region

```
SELECT ?cw ?title ?dateModified {
  <http://dbpedia.org/resource/Sofia> geo-ont:parentFeature ?province .
  ?province geo-ont:featureCode geo-ont:A.ADM1 .

  {
    ?location geo-ont:parentFeature ?province .
  } UNION {
    BIND(?province as ?location) .
  }

  ?cw a cwork:CreativeWork ;
      cwork:tag ?location ;
      cwork:title ?title ;
      cwork:dateModified ?dateModified .

   FILTER(?dateModified >=
"2011-05-14T00:00:00.000"^^<http://www.w3.org/2001/XMLSchema#dateTime>
&& ?dateModified <
"2011-05-19T23:59:59.999"^^<http://www.w3.org/2001/XMLSchema#dateTime>)
  }
LIMIT 100
```

Q11: News on about related entities

```
SELECT DISTINCT ?cw ?title ?description ?dateModified ?primaryContent
{
 {
  <http://dbpedia.org/resource/Teresa_Fedor> ?p ?e .
 }
 UNION
 {
  ?e ?p <http://dbpedia.org/resource/Teresa_Fedor> .
 }

 ?e a core:Thing .

 ?cw cwork:tag ?e ;
  cwork:title ?title ;
  cwork:description ?description ;
  cwork:dateModified ?dateModified ;
  bbc:primaryContentOf ?primaryContent .
}
ORDER BY DESC(?dateModified)
LIMIT 100
```

Trivial Validation Opportunities

- The validation mechanisms from SPB v.1.0 remain unchanged
- In addition to this, the changes to the benchmark are designed such a way that in case that specific inference patterns are not supported by the engine, some of the queries will return zero results
  - If ***owl:sameAs*** inference is not supported Q10 will return 0 results
  - If transitive properties are not supported, Q10 will return 0 results
    - Q11 will return smaller number of results (higher ratio of zeros also)
  - If ***rdfs:subProperty*** is not supported, Q2-Q10 will return 0 results
  - If ***rdfs:subClass*** is not supported, Q1, Q2, Q6, Q8 will return no results

Outline

- Motivation
- The New Reference Datasets
- Changes in the generated metadata
- Ontologies and Required Rule-sets
- The Concrete Inference patterns
- The New Queries
- **GraphDB Configuration Disclosure**
- Future plans

GraphDB Configuration

1. Reasoning performed through forward-chaining with a custom ruleset
   - This is the rdfs-optimized ruleset of GraphDB with added rules to support transitive, inverse and symmetric properties
2. ***owl:sameAs*** optimization of GraphDB is beneficial
   - This is the standard behavior of GraphDB; but it helps a lot
   - Query-time tuning is used to disable expansion of results with respect to ***owl:sameAs*** equivalent URIs
3. Geo-spatial index in Q6 of the basic interactive mix
4. Lucene Connector for full-text search in Q8

Note: *Points #2-#4 above help query time, but slow down updates in GraphDB. As materialization does also*

GraphDB – First Results

- Experimental run using GraphDB-SE 6.1
- Using LDBC Scale Factor 1 - (22M reference data + 30M generated data)
- Hardware: 32G RAM, Intel i7-4770 CPU @ 3.40GHz, single SSD Drive Samsung 845 DC
- Benchmark configuration: 8 reading / 2 writing agents, 1 min warmup, 40 min benchmark
- Updates/sec : **13**, Selects /sec : **44**

Outline

- Motivation
- The New Reference Datasets
- Changes in the generated metadata
- Ontologies and Required Rule-sets
- The Concrete Inference patterns
- Changes in the Basic Interactive Query Mix
- GraphDB Configuration Disclosure
- **Future plans**

Future Plans

- Relationships between pieces of content
  - At present Creative Works are related only to Entities, not to other CWs
  - These could be StoryLines or Collections of MM assets, e.g. an article with few images related to it
- Better modelling of content-to-entity cardinalities
  - Based on data from FT
- Enriched query sets and realistic query frequencies
  - Based on query logs from FT
- Loading/updating big dataset in live instances
- Testing high-availability
  - FT acceptance tests for High Availability cluster are great starting point

Summary of SPB v.2.0 Changes

- **Much bigger reference dataset: from 170k to 22M**
  - 7M statement from GeoNames about Europe
  - 14M statements from DBpedia for Companies, Persons, Events
- **Interconnected reference data: more than 5M links**
  - owl:sameAs links between DBPedia and Geonames
- **Much more comprehensive usage of inference**
  - While still the simplest possible flavor of OWL is used
    - rdfs:subClassOf, rdfs:subPorpertyOf, owl:TransitiveProperty, owl:sameAs
  - Transitive closure over company control and geographic nesting
  - Simpler queries through usage of super-property
- **Retrieval of relevant content through links in the reference data**