SPARQLytics: Multidimensional Analytics for RDF

Michael Rudolf

Database Technology Group, Technische Universität Dresden
March 8, 2017
Agenda

Motivation

RDF and SPARQL

Multidimensional Analytics for RDF
Motivation
Focus of Interest

**Focus moved from single entity (OLTP)**
- Bookkeeping
- Where is what?

**To aggregations over sets of entities of the same kind (OLAP)**
- Reporting
- What are the sales figures?

**To connections between entities**
- Who likes what and why?
- What do the friends of your customers buy?
Business Use Cases

Supply Chain Management

- Transportation & logistics: routing, tendering, tracking, auditing, payment

The Boeing Extended Global Supply Chain

783 million parts are procured in one year

- 737: 400 thousand parts
- 767: 3.1 million parts
- 787: 2.3 million parts
- 777: 3 million parts
- 747-8: 6 million parts

$28 Billion spend • 5,400 factories • 500,000 people

http://787updates.newairplane.com/787-Suppliers/World-Class-Supplier-Quality
Business Use Cases

Supply Chain Management
- Transportation & logistics: routing, tendering, tracking, auditing, payment

The Boeing Extended Global Supply Chain
- 783 million parts are procured in one year
- 737 767 787 777 747-8
- 400 3.1 2.3 3 6 million parts
- $28 Billion spend • 5,400 factories • 500,000 people

Track & Trace
- Pinpoint product recalls
- Mandated by law for certain industries (e.g. pharmaceuticals, food, waste)

EU Commission’s Rapid Alert System
<table>
<thead>
<tr>
<th></th>
<th>non-food (RAPEX)</th>
<th>food &amp; feed (RASFF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>2364</td>
<td>3137</td>
</tr>
<tr>
<td>2014</td>
<td>2435</td>
<td>3157</td>
</tr>
</tbody>
</table>

http://787updates.newairplane.com/787-Suppliers/World-Class-Supplier-Quality
RDF and SPARQL
Resource Description Framework (RDF) [WLC14]

- **Subjects** name an entity
- **Predicates** describe the relationship
- **Objects** can be literals or name

```

```product:1 amazon:capacity "64 GB" .
product:1 amazon:color "black" .
category:7 amazon:name "Tablets" .
user:8 amazon:country "FR" .
user:8 amazon:rates product:1 .
```

- no built-in schema
- can re-use vocabularies and ontologies
- suitable for inferencing facts
SPARQL Protocol and RDF Query Language [HS13]

- Built around pattern matching, produces pattern variable bindings
- Grouping and aggregation, CRUD operations
- No multidimensional concepts → complex and error-prone queries

PREFIX amazon: <http://www.amazon.com/#>
SELECT (AVG(?capacity) AS ?avgCap) (?name AS ?categoryName)
WHERE {
  ?category amazon:name  ?name .
  ?product amazon:capacity  ?capacity
}
GROUP BY ?categoryName
Multidimensional Analytics for RDF
Multidimensional Data Model [KR13]

(Base) Facts
- Describe events and measurements
- Mostly numeric and continuous

Dimensions
- Provide context for facts
- If numeric, then often discrete
- Can embody structure

Measures
- Are computed from grouped facts
- Are “arranged” in (hyper-)cubes
**Multidimensional Data Model [KR13]**

**BASE FACTS**
- Describe events and measurements
- Mostly numeric and continuous

**DIMENSIONS**
- Provide context for facts
- If numeric, then often discrete
- Can embody structure

**MEASURES**
- Are computed from grouped facts
- Are “arranged” in (hyper-)cubes
Multidimensional Data Model [KR13]

**Base Facts**
- Describe events and measurements
- Mostly numeric and continuous

**Dimensions**
- Provide context for facts
- If numeric, then often discrete
- Can embody structure

**Measures**
- Are computed from grouped facts
- Are “arranged” in (hyper-)cubes
From Intensional to Extensional Analytics

**Data Transformation**

- Intension fixed by domain expert or metadata
- Import data using ETL process
Data Transformation

- Intension fixed by domain expert or metadata
- Import data using ETL process

Query Generation

- Intension fixed by metadata
- Generate SPARQL queries from model
From Intensional to Extensional Analytics

**Data Transformation**
- Intension fixed by domain expert or metadata
- Import data using ETL process

**Query Generation**
- Intension fixed by metadata
- Generate SPARQL queries from model

**Extensional**
- Intension not fixed up-front
- Generate graph queries from user-specified intension
SPARQLytics for the Data Enthusiast

**SPARQLytics WorkFlow**

1. Create artifacts in repository
2. Start session re-using artifacts
3. Iteratively explore data, optionally create additional artifacts

Example 12
SPARQLytics for the Data Enthusiast

SPARQLytics Workflow

1. Create artifacts in repository

Example

USING REPOSITORY "myrepo";
SELECT FACTS {
  ?person rdf:type snvoc:Person ;
  snvoc:birthday ?birthday .
  FILTER (YEAR(NOW()) - YEAR(?birthday) >= 18)
};
DEFINE DIMENSION "Location" FROM ( ?person snvoc:isLocatedIn ?city .
  ?country snvoc:isPartOf ?continent
) WITH ( LEVEL "City" AS ?city,
  LEVEL "Country" AS ?country,
  LEVEL "Continent" AS ?continent
);
DEFINE MEASURE "Avg. No. Languages"
  AS COUNT(DISTINCT ?language) WHERE ( ?person snvoc:speaks ?language
  WITH "AVG";
CREATE CUBE "QB" FROM "Location", ... WITH "Avg. No. Languages", ...;
SPARQLytics for the Data Enthusiast

SPARQLytics Workflow

1. Create artifacts in repository
2. Start session re-using artifacts

Example

USING CUBE "QB" OVER <http://localhost:3030/ds/sparql>;
SLICE("Location", "Country", dbpedia:Italy);
COMPUTE ("Avg. No. Languages");
1. Create artifacts in repository
2. Start session re-using artifacts
3. Iteratively explore data, optionally create additional artifacts

**Example**

```sparql
USING CUBE "QB" OVER <http://localhost:3030/ds/sparql>;
SLICE("Location", "Country", dbpedia:Italy);
COMPUTE ("Avg. No. Languages");
RESET FILTER("Location", "Country");
ROLLUP("Location", 1);
COMPUTE ("Avg. No. Languages");
...```
Summary

**Big Graph Data**
- Not just social networks, also business scenarios
- Not enough data scientists, enable data enthusiasts

**RDF and SPARQL**
- Linked Open Data a rich source of information
- SPARQL does not expose multidimensional concepts

**SPARQLytics**
- Re-use core SPARQL elements for defining multidimensional model
- Generate complex SPARQL queries from analytical session
- Stateful approach integrates well with data enthusiasts workflow
Additional Material & References
Charu C. Aggarwal and Haixun Wang.  
A Survey of Clustering Algorithms for Graph Data.  
In Charu C. Aggarwal and Haixun Wang, editors, Managing and Mining Graph Data, volume 40 of Advances in Database Systems, chapter 9, pages 275–301. Springer US, 2010.  

A framework and a language for on-line analytical processing on graphs.  

Peter Boncz.  
LDBC: Benchmarks for Graph and RDF Data Management.  

Fabio Crestani.  
Application of spreading activation techniques in information retrieval.  

Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, and Philip S. Yu.  
Graph OLAP: Towards Online Analytical Processing on Graphs.  

Hartmut Ehrig, Gregor Engels, Hans-Jörg Kreowski, and Grzegorz Rozenberg, editors.  
References II


Satu Elisa Schaeffer.  
Graph clustering.  

Yuanyuan Tian and Jignesh M. Patel.  
TALE: A Tool for Approximate Large Graph Matching.  

David Wood, Markus Lanthaler, and Richard Cyganiak.  
RDF 1.1 concepts and abstract syntax.  
W3C recommendation, W3C, February 2014.  
http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han.  
Graph Cube: On Warehousing and OLAP Multidimensional Networks.  

Ning Zhang, Yuanyuan Tian, and Jignesh M. Patel.  
Discovery-Driven Graph Summarization.  