ABSTRACT

- Most of the data published on the web is unstructured or does not follow a standard.
- It makes the data harder to be retrieved and interchanged between different data sources.
- Linked Open Data (LOD) technologies are applied in a scenario that deals with a large amount of computer network measurement data.
- As a result, we generated more structured data, hence easier to be retrieved, analyzed, and more interoperable.
- The challenges of processing large amount of data to transform it into a standard format (RDF); link it to other data sources; and analyze and visualize the transformed data are discussed.
- An ontology that aims to minimize the number of triples is proposed and a discussion on how ontologies may impact query performance is presented.
- We emphasize the advantages of having the data in RDF format and show use cases on the scenario of the project.

SYSTEM MODELING AND RESULTS

Domain Analysis

Understanding PingER Project’s domain

- It involves data about network performance
  - Permanent
  - 60 monitoring nodes
  - 600 nodes monitored
  - 6000 nodes of mobiles (applied monitored)
  - 16 countries, several cities within each country
  - 16 network metrics (e.g., TCP throughput, packet loss, average RTT)
  - Hourly data, since 1998.
- Data can be applied to many different situations such as economical, geographical, and seasonal events.

Problem and Strategies

- Hard to query the CSV files to retrieve specific data, comparing traditional CSVs.
- Hard to produce informative graphs, reports, and identify data.
- Data not interoperable with other data sources.

Data stored in multiple flat CSV files

Simple Domain Model

How a ping measurement can be modelled

Ontology Engineering

- An Ontology is needed to model the domain following W3C recommendations.
- Resulting existing ontologies support the idea of standardization and interoperability with the LOD community.

- Semantic expressivity
- Completeness in relation to the domain
- Impact on query performance
- Ontologies being reused

Geonames Ontology

Geonames [9]

- RDF Time Ontology [4] (MOMENT) [9]

PingER LOD Ontology

ETL Process for General Data

- Parallel and Distributed approach to copy multiple CSV files
- Transforms 6 into Triples and Load into the RDF DBMS
- While the data is being transformed into triples, it is also being linked to external data sources in the LOD Cloud
- Each process is independent, hence can be simultaneously executed in different machines.

ETL Process for Measurement Data

- Each ETL process for measurement data is responsible for a single network metric
- A single time aggregation
- An execution takes at least an hour, packet loss, etc.
- 10 processors that can run in distributed machines
- Each process is further parallelized

Publication & Applications

Multiples Network Metrics

Network Metrics vs. University Metrics

- It takes advantage of well-structured data with a schema, in a very expressive format (JSON-LD).

- It supports the execution of SPARQL queries to capture precisely what is being searched.

- The possible combination of parameters is able to be retrieved.

- The advantage of PingER data with D瘩edia [8] data about universities (information about number of students, enrolment, etc.)

- Using this graph, one could visually verify that well-funded universities have better network connectivity.

REFERENCES


ACKNOWLEDGEMENT

- This work was supported in part by the Department of Energy Contract DE-FG02-07ER25701.
- This work was supported in part by the Rio de Janeiro State Science Foundation (FAPERJ).