WWW 2017 Tutorial:
Semantic Data Management in Practice

Part 8: Integrating

Olaf Hartig
Linköping University
olaf.hartig@liu.se
@olafhartig

Olivier Curé
University of Paris-Est Marne la Vallée
olivier.cure@u-pem.fr
@oliviercure

LINKÖPING UNIVERSITY

UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Goals

- Achieve global understanding of semantic integration:
  - What are the main problems?
  - Which approaches are in use?

- Understand the main features of some commercial and open source systems

Overview

- (semantic) integration
- Two approaches
- Systems
- Demo

Data integration

- is a core component of information technology
- enables the combination of data contained in multiple data sources
- has to deal with
  - **discovering** and **representing** mapping assertions between source schemata, e.g., database tables
  - **answering queries** using multiple data sources, e.g., using SQL

Semantic Integration ?

- Bringing together diverse, possibly heterogenous, sources of information and interrelating them by leveraging the **semantic** information that is embedded inside them

- Interrelation occur at the ontology/vocabulary level
  - Recall that ontologies aim for knowledge sharing

- For example, integrate data across DBpedia, Wikidata, or any other Linked Data sources

Why it is Important ?

- Combine data and knowledge from multiple sources to:
  - answer queries using multiple sources of data, e.g., in SPARQL
  - support interoperability between different systems and thus sharing knowledge
  - reason over multiple data sources using knowledge sources

Mapping Ontologies is a Hard Problem

- Too many large ontologies to consider manual mappings
- Semantic integration has to deal with different levels of mismatches:
  - Ontology: syntax, expressiveness
  - Linguistic: terms used in ontology
  - Modeling: conventions, granularity
  - Domain: coverage

Overview

- (semantic) integration
- Two approaches
- Systems
- Demo

Two Main Approaches

- Existence of a shared ontology which is extended to relate external ontologies via some mappings

- No shared ontology is available:
  - Heuristics-based or machine learning techniques are used to relate ontologies

Shared Ontology

- Several types of ontologies:
  - **Top-level** ontologies formalize general notions (e.g., processes, events, time, space, physical objects, etc.). An example is DOLCE which aims at capturing “ontological categories underlying naturel language and human common-sense”
  - **Domain** ontologies describe a specific domain in terms of concepts and properties
  - **Application** ontologies specify terms for a given application. They depend on a domain ontology.

Shared Ontology (2)

- Top-level ontologies are designed to support ontology matching
  - If two ontologies extend the same top-level ontology then it is easier to find correspondences between them. The top-level ontology serves as a bridge.

Heuristics and ML approaches

- Heuristic approaches are usually based one or a combination of structure, element or instance analysis. Examples are the PROMPT Suite (developed by the Protégé team)

- Machine learning approaches can combine different learners using a probabilistic model to discover correspondences between ontologies. Examples are GLUE, FCAMerge

Overview

- (semantic) integration
- Two approaches
- Systems
- Demo

Systems

- Commercial products
  - Semaflora systems
  - TopQuadrant’s TopBraid
  - Cambridge Semantics

- Free, open-source solutions
  - RDF Refine
  - Silk
  - Limes
  - Karma

Commercial products

- Ontoprise was one of the early software suite created with Semantic Web standards in view.
  - Some products, including semantic integration tools were acquired in 2012 by Semaflora systems

- Since the early 2000s, TopQuadrant proposes a Semantic ecosystem around the TopBraid suite. It is composed of an advanced ontology editor, TopBraid Live to integrate data

- Cambridge Semantics was founded in 2007. It proposes the Anzo suite which supports integration, cleaning, management of metadata

LIMES

- LInk discovery framework for MEtrix Spaces
- Academic project with software maintenance
- Composed of several discovery approaches
  - Link discovery for approximation of similarity between instances
  - Machine learning (supervised and unsupervised)
- Easily configurable through files or a Graphical User Interface

SILK

- A Linked data integration framework
- With commercial support by the Eccenca start-up
- Active since 2010, latest version is 2.7.1
- Main features
  - Generate links between related data items
  - Apply data transformations to structured data sources (e.g., generate RDF triples from csv files)
  - Link RDF triples to data sources on the Web (e.g., LOD)

KARMA

- An information integration tool
- Developed and maintained at University of South California (USC)
- Karma learns to recognize mapping of data to ontologies
- Provides a Graphical User Interface to interact with data sets and ontologies

Overview

- (semantic) integration
- Two approaches
- Systems
- Demo

Demo

- Using Karma
  - Website: http://usc-isi-i2.github.io/karma/
  - Download: https://github.com/usc-isi-i2/Web-Karma
  - Unzip web-karma-master.zip
  - Go to Web-Karma-master folder and run mvn clean install
  - To run karma, go to karma-web folder and run mvn -Djetty.port=8086 jetty:run
  - http://localhost:8086

Demo

- Scenario: integrate data obtained from a french organization (CSV file)
  - Create an ontology on the fly
  - Annotate data with this ontology
  - Annotate other data elements with the igeo ontology
  - Generate an RDF document

Wrap-up

- Mature tools in the Semantic data management ecosystem
- High quality open source systems are available
- Large companies are already present or entering the market

Wrap-up (2)

- Still many open issues to address
  - Efficient partitioning and RDF storage
  - Reasoning and high performance query answering
  - SPARQL query processing and analytics
  - Understanding, visualizing very large graphs
  - Data cleansing
  - ...