Data Linking

Angelica Lo Duca
IIT-CNR
angelica.loduca@iit.cnr.it

Linked Open Data:
a paradigm for the
Semantic Web

Data Linking is the process of finding relationships or correspondences between resources of different datasets.

Data Linking

- Cannot be carried out manually at Web scale
- Automatic approaches
  - Ontology Matching
    - find schema matching
  - Instance Matching
    - find instances matching
    - use the **owl:sameAs** property to link resources

Ontology Matching

- Ontologies can be highly specialized
  - e.g. DBpedia has classes for *Educational Institutions, Bridges, Airports, etc.*
- But some can be rudimentary
  - e.g. in Geonames all instances only belong to a single class – ‘Feature’
- There might not exist exact equivalences between classes in two sources
  - Only subset relations possible

Ontology Matching Basic techniques

- name-based
    - string-based
    - language-based
- structure-based
- ...

Name-based techniques

- They can be applied to the name, the label or the comments of classes in order to find those which are similar
- Useful if conceptual schemas (or ontologies) use very similar strings to denote the same concepts
- Yield many false positives, if pairs of strings with low similarity are selected

Name-based techniques (cont.)

- string-based
  - compare strings with a metrics
  - the metrics maps two strings to a real number
- language-based
  - exploit linguistic transformations to compare strings

String-based techniques (cont.)

- Levenshtein distance
  - Measure the similarity between two strings by the minimum number of insertions, deletions, and substitutions of characters required to transform one string into the other
- Token-based distance
  - Treats strings as a bag of words (multisets of substrings)
  - May split strings into independent tokens
  - Calculate words frequency for each string
  - Compare bag of words with a metric (e.g. cosine-similarity)

Language-based techniques

- Intrinsic methods
  - reduce each term to a normal form to facilitate matching
  - use traditional natural language processing techniques
    - stopword elimination
    - tokenization: segment strings into sequences of tokens
    - lemmatization: reduce words to normal forms
      - suppress tense, gender and number

Language-based techniques (cont.)

- Extrinsic methods
  - Use dictionaries, lexicons and terminologies to help match terms from different schemas or ontologies
    - e.g. a terminology - a thesaurus which very often contains phrases rather than single words
    - deal with synonyms
    - word sense disambiguation

Ontology Matching Basic techniques

- name-based
  - string-based
  - language-based
- **structure-based**
- ...

Structure-based techniques

- Internal structure (constraint-based approaches)
  - based on the internal structure of classes
  - calculate the similarity between two classes based on
    - the set of their properties, including keys
    - the range of their properties (attributes and relations)
    - the cardinality of their properties
    - the transitivity or symmetry of their properties
- extensional techniques
  - When two ontologies share the same set of individuals, matching is highly facilitated

Internal Structure-based techniques (cont.)

- Relational Structure
  - similarity between two concepts
  - based on the relations between the concepts with other concepts
  - similar concepts should have similar related concepts
- Taxonomic Structure
  - Similarity between two concepts
  - Based on the graph of the subClassOf relation

Extensional Structure-based techniques

- Jaccard Similarity
    - The Jaccard index is defined as the size of the intersection of two sets divided by the size of the union of the two sets.
    - Given two sets A and B, let P(X) be the probability of a random instance to be in the set X
    - Note that the Jaccard Similarity reaches 1 when A = B and 0 when they are disjoint.

Instance Matching

- Some techniques used for Ontology matching can be used also for instance matching
- Existing tools can be used
    - Open Refine
    - Silk

Open Refine (also known as Google Refine)

- Powerful tool to work with messy data
  - cleaning
  - transforming
  - extending
- Download
  - http://openrefine.org
- Run
  - http://127.0.0.1:3333
- Plugin for Linked Data Linking
  - RDF Extension for Open Refine
  - http://refine.deri.ie

RDF extension for Open Refine

- Setup & Run
  - Make sure "extensions" folder exists in your Open Refine workspace
  - Download the extension
  - Extract the downloaded zip file to the "extensions" folder
  - Restart Google Refine

General Refine Expression Language (GREL)

- GREL is a language to manipulate data in Open Refine
- Documentation
  - https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language