Resolving Entities in the Web

Kostas Stefanidis
kstef@ics.forth.gr

Vassilis Christophides, Vasilis Efthymiou

ICS-FORTH
Research data is spread across the Web

- Data about research projects, organisations, researchers or research outputs, such as publications or patents

Where: institutional and personal Web pages,

(semi-)open databases and information systems

Rich metadata, containing entities, e.g., persons, projects, organizations

- Linking such entities builds a foundation for the Semantic Web
  - Applications: biology, geology, digital libraries, digital humanities, cultural heritage
Resolving & Interlinking Entities
The Web of Data (Entities)
Web Knowledge Bases (KBs)

Host machine-readable entity descriptions
  – as instances, using concepts and relationships
  – expressed in RDF, overcoming the schema rigidity of traditional DBs

• Domain-specific
  – e.g., dedicated to music (e.g., BBCMusic), movies (e.g., LinkedMDB), scientific publications (e.g., DBLP)
• Cross-domain
  – containing encyclopaedic knowledge for a variety of entity types
    • e.g., DBpedia, YAGO2, Freebase.
Web Knowledge Bases (KBs)

Several questions naturally arise regarding the characteristics of the entity descriptions hosted by a KB on the Web.

• How many entities are described and by how many facts (aka triples)?

• What entity types (aka classes) are covered, and what property vocabularies are employed to describe them?

• Which semantic relationships (e.g., equivalence, relatedness) stand between the entities described within or across KBs?
Characteristics of Entity Descriptions in the Web of Data

- Big volume (1014 KBs, 600M entities in Knowledge Graph only)
- High heterogeneity
  - semantic (different naming policies, several semantic types)
  - structural (loose structuring, even within the same domain)
- Varying quality
  - inherent incompleteness (many missing links)
  - inconsistencies among KBs (contradicting values)
  - erroneous values (typos, extraction tools errors, linking errors)
- Strong redundancy
  - among KBs (inter-duplicates, due to common data sources)
  - within the same KB (intra-duplicates, due to wrong integration/bad curation)
In overall:
The scale, diversity and graph structuring of entity descriptions in the Web of data challenge the way two descriptions can be effectively compared to efficiently decide whether they are referring to the same real-world entity

• Entity resolution at the Web scale requires efficient techniques that go beyond deduplication algorithms in data warehouses

• Ontology and instance matching algorithms seek correspondences between the different ontology classes and properties, to match the instance descriptions, which are using them
  – In the Web of data, the same instance can be described using several vocabularies, refuting the assumption that ontologies play the role of schemas, imposing structural restrictions on the descriptions
Entity Resolution

The problem of identifying descriptions of the same real-world entity.
The Value of Entity Resolution: Entity-Centric Search

Figure 1.4: Searching for the entity "Stanley Kubrick" in the Web of data. The user can, moreover, select entities related to Kubrick and explore information about the movie *A Clockwork Orange*, actors of this movie, like Malcolm McDowell, or other movies of McDowell, like *The Artist*. The search results can be presented to the user either by propagating all the entity descriptions that are relevant, or after an additional fusion step has been applied, as implied in Figure 1.4 (bottom).
**Entity Resolution Workflow**

**Entity matching**: Based on which notion of similarity? Cannot use domain-specific knowledge at the cross-domain Web scale
**Entity Resolution Workflow**

**Tradeoff:** more comparisons we skip $\rightarrow$ more matches we miss

**Main Concern:** no missed matches (FNs)

**Blocking:** A technique to reduce the number of compared pairs
- Split entity descriptions into blocks
- Perform comparisons only between descriptions of the same block
  - TP: a suggested comparison between matching descriptions
  - FP: a suggested comparison between non-matching descriptions
  - FN: a pair of matching descriptions, not placed in any common block
Relational Blocking Is Not Enough

• Blocking has been extensively studied for tabular data
  – Approaches assume that the input data follow a schema which is available and known
• Thus, traditional approaches are built on the notion of a blocking key:
  – Criteria on a fixed set of attributes, based on which the descriptions are placed into blocks
  – The value of each description for this key, i.e., their blocking key value, determines the specific block(s) into which the description will be placed

The high heterogeneity of entity descriptions in the Web of data makes the use of blocking keys inapplicable!
• We do not even have pre-defined sets of attributes
Improving Recall

could entail more than $|E|^2$ comparisons!
Entity Resolution in the Web of Data

[WWW 2014 (tutorial)]


Joint work with Vassilis Christophides and Vasilis Efthymiou
Thank You!