ABSTRA: Toward Generic Abstractions for Data of Any Model

Nelly Barret, Ioana Manolescu, Prajna Upadhyay
Institut Polytechnique de Paris & Inria
https://team.inria.fr/cedar/projects/abstra/

CONTEXT

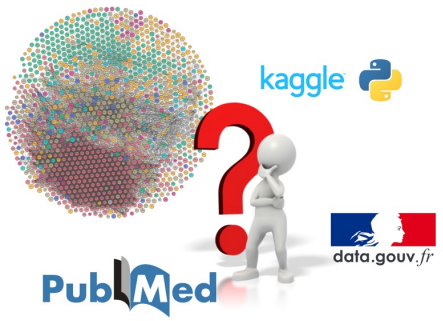
- Open data initiative has led to a set of big heterogeneous datasets
- Heterogeneous datasets are difficult to integrate and understand/exploit

How to help a human user grasp the content of a dataset?

1. Analyse and exploit the structure of the data
2. Compute a form of semantics on the content of the data

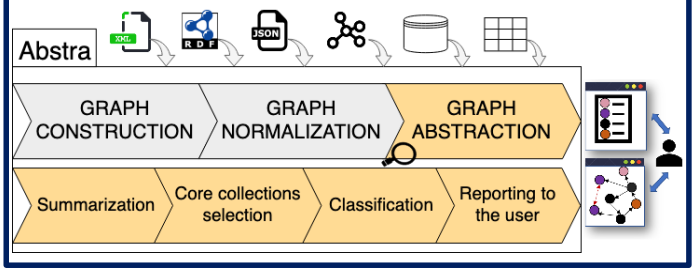
Existing works

- Summarizing semi-structured data (statistical, structured, logical approaches)
  → Based only on structure, not on content
- Schema inference
  → Specific to one data model



Goasdoué F, Guzewicz P, Manolescu I. RDF graph summarization for first-sight structure discovery. VLDBJ 2020.
Baazizi MA, Colazzo D, Ghelli G. *et al.* Parametric schema inference for massive JSON datasets. VLDB 2019.

OVERVIEW OF THE ABSTRA APPROACH



ASSUMPTIONS

We are given the following resources:

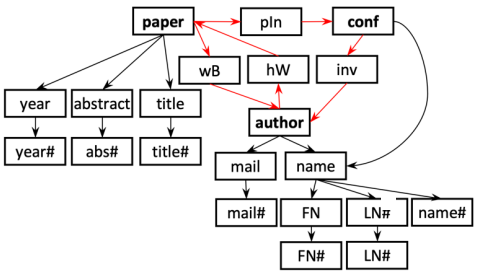
- A directed graph G containing:
  - Data labelled nodes and unlabelled edges
  - Extracted entities nodes
- A set of semantic categories: *Person, Event, Place, …*
  - For each category, a set of properties (e.g. *address*, *age*, …)

UNDERSTANDING WHAT A DATASET IS ABOUT (IN AN AUTOMATIC WAY)

Intuition: any dataset contains **entities/records/objects** grouped in **collections** and collections may be linked with **relationships**

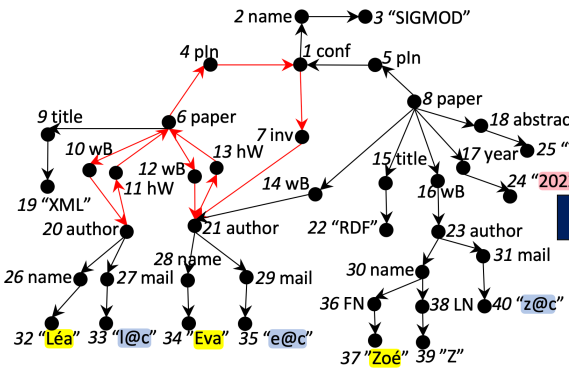
Method:

1. Summarize G and produce the *collection graph*
2. Determine the *core collections*:
   1. Assign a *weight* to each collection (*dw*-based or PageRank-based)
   2. Greedily select the most-weighted collection and define its *boundary*
   3. Update the collection graph to reflect the selection and return to step 1
3. Determine *relationships* between core collections
4. *Classify* each core collection among the semantic categories using its properties
5. Create the *textual description* based on the core collections and their relationships

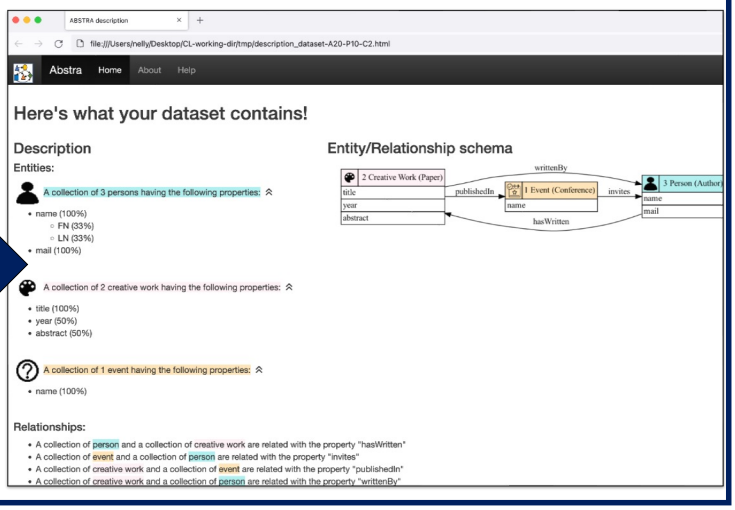


Example of a collection graph

A SCENARIO IN ABSTRA



ABSTRA



The graph produced out of a dataset describing papers that were written by some authors and then published in some conferences

FUTURE WORK

- Improve the methods or devise new ones to assign scores to collections
- Enrich the dataset to improve the understanding, e.g. using knowledge bases or web information
- Make the classification process bottom-up to take advantage of the data structure