KDI : **Knowledge and Data Integration**

Karmalinker

Data linking tool

W6.L11.M4.T11.1.2

Contents

Contents

Karmalinker

It is a tool for *data linking* which is enhanced with language understanding capabilities, which allow us to align data to reference ontology.

This tool is an enhanced version of Karma developed by USCLA.

Workflow

There are three ways to carry on the process of data linking:

- manual linking of new data sources;
- semi-automated linking of new data sources;
- automated integration of known data sources.

Manual linking

Input

- Ontology
- Dataset
- Data scientist's knowledge

Output

- RDF
- EML
- Mapping file
- Training sets

This process is the most important one, in many case this is enough to perform a "one-shot" data integration.

Notes

1. some data types are requiring special attention: like *NLString*, Concept, Dates;
2. we have special function for extracting Concepts and Dates.
3. save often, history sometimes fails;
4. if the file is very big we can choose a small number of rows to perform this task, then use *"automated linking"* to perform the task on the entire dataset.

Semi-automated linking

Input

- Ontology
- Dataset
- Data scientist's knowledge
- Training sets

Output

- RDF
- EML
- Mapping file
- Training sets

In this process the data scientist asks KarmaLinker to infer the right mapping for the columns, based on the links that has been already done.

Notes

1. Not much different from the first modality;
2. is mostly a time-safer alternative to complete the mapping manually.
3. empirically this work well when you have mapped at least 4 similar datasets with your ontology.

Automated linking

Input

- Mapping file
- Dataset

Output

- RDF
- EML

This is important if you have to create a *data integration process*: this allows to automate extractions of data *in a known format*. In case those

data are very big is possible to use Apache Spark to run this task (will not be shown).

Demo

See the associated demo video.

KDI : Knowledge and Data Integration

W6.L11.M4.T11.1.2

Karmalinker

Data linking tool