Assignment 1: Schema Matching

Deadline: 20.05.2018, Presentation: 22.05.2018[1]

[1]The 800 Pound Gorilla in the Corner: Data Integration
Technische Universität Berlin

Abstract. *The goal of this assignment is to find the correspondences between columns in two datasets, which is called schema matching.*
Notice that:

- *The short questions usually require short answers too.*
- *One submission per each group is enough.*
- *You are free to use whatever tools/programming languages you intend to.*
- *Put your names and group number in the beginning of your final PDF report please.*
- *Submit a compressed file that contains (1) the PDF report and (2) your source code files please.*

1. Setup

1.1. Datasets

We have two relational datasets [1] with the following schemas:

$$Imdb(Id, Name, YearRange, ReleaseDate, Director, Creator, Cast, Duration, RatingValue, ContentRating, Genre, Url, Description),$$
$$rotten\_tomatoes(Id, Name, Year, Release\ Date, Director, Creator, Actors, Cast, Language, Country, Duration, RatingValue, RatingCount, ReviewCount, Genre, FilmingLocations, Description).$$

The datasets have been attached to this document.

1.2. Ground Truth

The set of actual correspondences is as follows:

$$G = \{\langle Imdb.Name, rt.Name\rangle, \langle Imdb.YearRange, rt.Year\rangle, \langle Imdb.ReleaseDate, rt.\text{"}Release\ Date\text{"}\rangle, \langle Imdb.Director, rt.Director\rangle, \langle Imdb.Creator, rt.Creator\rangle, \langle Imdb.Cast, rt.Cast\rangle, \langle Imdb.Duration, rt.Duration\rangle, \langle Imdb.RatingValue, rt.RatingValue\rangle, \langle Imdb.Genre, rt.Genre\rangle, \langle Imdb.Description, rt.Description\rangle\}.$$

Overall, we have 10 pairs of columns that are corresponded.

1.3. Evaluation Measures

In the following tasks, you should evaluate your implemented solutions in terms of precision and recall:

$$precision = \frac{Number\ of\ the\ discovered\ correspondences\ that\ are\ in\ G}{Number\ of\ all\ the\ discovered\ correspondences},$$
$$recall = \frac{Number\ of\ the\ discovered\ correspondences\ that\ are\ in\ G}{Number\ of\ all\ the\ actual\ correspondences = 10}.$$

2. Tasks

2.1. Label-Based Schema Matching

Here, we want to find the correspondences between the columns from the two datasets with the help of only *schema headers*.

1. Provide an algorithm. Specify the input, output, similarity function, and time complexity.
2. Implement the algorithm and report the results. Is there any parameter that affects the results?
3. What is the upsides and downsides of this method? When does it work and when not?

2.2. Instance-Based Schema Matching

Here, we want to find the correspondences between the columns from the two datasets with the help of only *data values*.

1. Provide an algorithm. Specify the input, output, similarity function, and time complexity.
2. Implement the algorithm and report the results. Is there any parameter that affects the results?
3. What is the upsides and downsides of this method? When does it work and when not?

References

[1] Sanjib Das, AnHai Doan, Paul Suganthan G. C., Chaitanya Gokhale, and Pradap Konda. The magellan data repository. `https://sites.google.com/site/anhaidgroup/projects/data`.