Relational concept analysis for link key extraction

⋆ Jerôme Euzenat

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
Firstname.Lastname@inria.fr

Linked data aims at publishing data expressed in RDF (Resource Description Framework) at the scale of the worldwide web [4,8]. These datasets interoperate through links which identify individuals across heterogeneous datasets. Data interlinking, the problem of linking pairs of nodes in RDF graphs corresponding to the same resource, is an important task for linked open data.

Different approaches and methods have been proposed to address the problem of automatic data interlinking [6,10]. Most of them are based on numerical methods that measure a similarity between entities and consider that the closest the entities, the more likely they are the same [15,11]. A few other works take a logical approach to data interlinking and can leverage reasoning methods [13,1,9].

We introduced the notion of link keys as a way to identify such node pairs [5,2]. Link keys generalise keys in relational algebra in three ways: (1) they apply across two data sets instead of a single one, (2) they take into account multiple values for the same attribute, and (3) attribute values may be other objects. The latter makes link keys eventually dependent on each others.

Link keys specify the pairs of properties to compare for linking individuals belonging to different classes of the datasets. An example of a link key is:

\( \{\langle \text{auteur}, \text{creator} \rangle \} \{\langle \text{titre}, \text{title} \rangle \} \ linkkey \langle \text{Livre}, \text{Book} \rangle \)

stating that whenever an instance of the class Livre has the same values for property auteur as an instance of class Book has for property creator and they share at least one value for their property titre and title, then they denote the same entity.

Clearly, such a link key may depend on another one as, for instance, properties auteur and creator have values in the Écrivain and Writer classes respectively. Identifying their values will then resort to another link key:

\( \{\langle \text{prénom}, \text{firstname} \rangle \} \{\langle \text{nom}, \text{lastname} \rangle \} \ linkkey \langle \text{Écrivain}, \text{Writer} \rangle \)

This situation may be rendered even more intricate if Écrivain and Writer were instead identified from the values of their properties ouvrages and hasWritten referring to instances of Livre and Book. We would then face interdependent link keys.

⋆ Work developed in collaboration with Manuel Atencia, Jérôme David, Amedeo Napoli and Jérémy Vizzini
We have already proposed an algorithm for extracting some types of link keys [2]. This method may be decomposed in two distinct steps: (1) identifying link key candidates, followed by (2) selecting the best link key candidates according to quality measures. We have previously shown how to encode the functional link key extraction problem in relational databases into Formal Concept Analysis (FCA [7]) so that candidate link keys correspond to formal concepts [3].

In this talk we will show how to use Relational Concept Analysis (RCA, [12]) for dealing with cyclic dependencies across classes and hence to extract directly families of interdependent link keys from RDF data sets. This method generalises directly those presented for non dependent link keys [2] and link keys over the relational model [3].

We will consider the extensions of quality measures for families of interdependent link keys.

Finally, we will discuss linky, a prototype implementation of this framework [14] and the evaluation modalities and data sets for logical data interlinking.

Acknowledgements

This work has been partially funded by the ANR Elker project (ANR-17-CE23-0007-01).

References