Relational concept analysis for link key extraction[⋆]

Jérôme 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:

$$\{\langle \mathsf{auteur}, \mathsf{creator}\rangle\}\{\langle \mathsf{titre}, \mathsf{title}\rangle\}\ linkkey\ \langle \mathsf{Livre}, \mathsf{Book}\rangle$$

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:

$$\{\langle \mathsf{prénom}, \mathsf{firstname}\rangle\}\{\langle \mathsf{nom}, \mathsf{lastname}\rangle\}\ linkkey\ \langle \mathsf{Écrivain}, \mathsf{Writer}\rangle$$

This situation may be rendered even more intricate if Écrivain and Writer were instead identified from the values of their properties ouvrages and hasWritten refering 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 methods 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

1. Al-Bakri, M., Atencia, M., Lalande, S., Rousset, M.C.: Inferring same-as facts from linked data: an iterative import-by-query approach. In: Proc. 29th AAAI Conference on Artificial Intelligence, Austin (TX US). pp. 9–15. AAAI Press (2015)
2. Atencia, M., David, J., Euzenat, J.: Data interlinking through robust linkkey extraction. In: Proc. 21st European Conference on Artificial Intelligence (ECAI). pp. 15–20. IOS Press (2014)
3. Atencia, M., David, J., Euzenat, J.: What can FCA do for database linkkey extraction? In: Proc. 3rd ECAI workshop on What can FCA do for Artificial Intelligence? (FCA4AI), Praha (CZ). pp. 85–92 (2014)
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked data — the story so far. International Journal of Semantic Web Information Systems 5(3), 1–22 (2009)
5. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (DE), 2nd edn. (2013)
6. Ferrara, A., Nikolov, A., Scharffe, F.: Data linking for the semantic web. International Journal of Semantic Web and Information Systems 7(3), 46–76 (2011)
7. Ganter, B., Wille, R.: Formal Concept Analysis. Springer, Berlin (1999)
8. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool (2011)
9. Hogan, A., Zimmermann, A., Umbrich, J., Polleres, A., Decker, S.: Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. Journal of Web Semantics 10, 76–110 (2012)
10. Nentwig, M., Hartung, M., Ngonga Ngomo, A.C., Rahm, E.: A survey of current link discovery frameworks. Semantic Web 8(3), 419–436 (2017)

11. Ngonga Ngomo, A.C., Auer, S.: LIMES: A time-efficient approach for large-scale link discovery on the web of data. In: Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI), Barcelona (ES). pp. 2312–2317. Barcelona (ES) (2011)
12. Rouane-Hacene, M., Huchard, M., Napoli, A., Valtchev, P.: Relational Concept Analysis: mining concept lattices from multi-relational data. Annals of Mathematics and Artificial Intelligence 67(1), 81–108 (2013)
13. Saïs, F., Pernelle, N., Rousset, M.C.: L2R: A logical method for reference reconciliation. In: Proc. 22nd National Conference on Artificial Intelligence (AAAI), Vancouver (CA). pp. 329–334. AAAI Press (2007)
14. Vizzini, J.: Data interlinking with relational concept analysis. Mémoire de master, Université Grenoble Alpes (2017)
15. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk – A link discovery framework for the web of data. In: Proc. WWW Workshop on Linked Data on the Web, LDOW, Madrid (SP). CEUR Workshop Proceedings, vol. 538. CEUR-WS.org (2009)