

Clustering Knowledge Graphs Using Concept Lattices (Extended Abstract)

Fabiola Hodo, Sai Pranav and Barış Sertkaya

Frankfurt University of Applied Sciences


Keywords

Knowledge Graphs, clustering, Concept Lattices

We propose an approach for identifying clusters in a knowledge graph based on commonalities of entities and for computing logical descriptions of these clusters. Our approach is based on computing formal concepts from data, which is a standard data analysis technique in Formal Concept Analysis (FCA) [1]. In FCA, attributes are unary predicates which are either satisfied or not satisfied by objects. A formal concept is a pair (A, B) , where A is the set of objects satisfying all of the attributes in B and B is the set of common attributes of the objects in A . Ordered w.r.t. set inclusion, formal concepts form a lattice, called the concept lattice, which is a hierarchy of formal concepts. In [2, 3, 4, 5, 6, 7] FCA was employed in combination with DLs for computing the base of valid axioms of an application domain and for enriching existing DL knowledge bases. In [8] it was used to compute a data-driven schema from knowledge graphs. In the present work, we make use of similar ideas for a different purpose, namely for identifying clusters in knowledge graphs. Our motivation is to learn logical descriptions of entities that share common properties. As properties we use the notion of a model-based most specific concept (mmsc) introduced in [9, 3]. The resulting formal concepts, which are clusters of entities, are then described by \mathcal{EL}^\perp concept descriptions. Varying the role-depth allows us to compute clusters of variable granularities from the knowledge graph. Shallow role-depths result in coarse clusters, higher role-depths result in more fine-grained clusters. Our approach is comparable to Relational Concept Analysis (RCA) [10], which is an extension of FCA for dealing with relational data. As data structure it uses a set of formal contexts and a set of relations. The distinguishing feature of our work is that we use existing FCA data structures and algorithms of the shelf, and do not need to extend them for relational data. This has the advantage that we can use existing FCA libraries just out of the box.

1. Identifying Clusters in Knowledge Graphs


In order to depict our approach on a simple example consider the knowledge graph given in Figure 1. It contains the 5 entities a, b, c, d, e , and some facts on their class memberships and on their relationships to each other. For identifying formal concepts in a knowledge graph,

 DL 2023: 36th International Workshop on Description Logics, September 2–4, 2023, Rhodes, Greece

 fabiola.hodo@fb2.fra-uas.de (F. Hodo); sai.margasahayamvenkatesh@stud.fra-uas.de (S. Pranav); sertkaya@fb2.fra-uas.de (B. Sertkaya)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

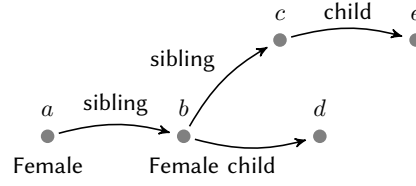


Figure 1: Toy knowledge graph with family relationships

the first step is to build a formal context whose objects are entities from the knowledge graph and whose attributes are properties of these entities. In principle, we can select class names as attributes. A formal concept obtained this way will have the common set of class names as its intent. More precisely, a formal concept (A, B) will generalize the entities in A just by conjuncting the classes that they belong to. As a result, the formal concepts, or the clusters we obtain will not be detailed enough since we do not take into account the relationships of the entities in A to other entities.

Alternatively, as the set of attributes we can select every combination of relation and class names until some certain role-depth. This would be the other extreme. It would result in a highly fine-grained clustering, but would have the drawback that the number of attributes is huge. For a set of class names N_C , a set of relation names N_R and for depth d , the number of attributes (or concept descriptions) we obtain this way is super-exponential in the sizes of N_C and N_R . The reason is that, the set of attributes for role depth d is generated using the power set of attributes at role depth $d - 1$. (For more details see [5] or [3].)

Our aim is clustering the entities in a knowledge graph based on their common properties, which requires use of attributes that generalize sets of entities, yet at the same time are as specific as possible. To this purpose, we use the notion of model-based most specific concept description introduced in [9, 3] as attributes of a formal context. For generalizing concept descriptions from individuals, least common subsumers (LCS) and most specific concepts (msc) have intensively been studied in the literature [11, 12, 13].

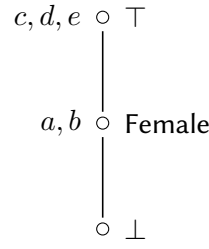
Definition 1 ([3, 5]). Let \mathcal{G} be a knowledge graph and $\Delta_{\mathcal{G}}$ be the set of entities in \mathcal{G} . We denote the most specific concept description for the entities in $X \subseteq \Delta_{\mathcal{G}}$ at depth d as $X^{\mathcal{G}_d}$. Based on this, we define a set of attributes for X at depth d as:

$$M_{\mathcal{G},X,d} = N_X \cup \{\perp, \top\} \cup \{\exists r. X^{\mathcal{G}_{d-1}} \mid X \subseteq \Delta_{\mathcal{G}} \text{ and } X \neq \emptyset\}$$

where N_X is the set of types entities in X . More precisely, it is defined as $N_X = \{C \mid C \in N_C \text{ and } \exists x \in X \text{ s.t. } x \text{ is an instance of } C\}$. For depth $d = 0$ we define $M_{\mathcal{G},X,d} = N_X \cup \{\perp, \top\}$. The formal context induced by \mathcal{G} and X at depth d is then $\mathbb{K} = (X, M_{\mathcal{G},X,d}, \mathcal{IR})$ where the incidence relation \mathcal{IR} is defined as $\mathcal{IR} = \{(x, m) \mid x \in X, m \in M_{\mathcal{G},X,d} \text{ and } x \text{ is an instance of } m\}$.

Example 1. Consider the toy knowledge graph \mathcal{G} in Figure 1 with $\Delta_{\mathcal{G}} = \{a, b, c, d, e\}$, $N_C = \{\text{Female}, \perp, \top\}$ and $N_R = \{\text{sibling}, \text{child}\}$. The formal context \mathbb{K}_0 obtained with $X = \Delta_{\mathcal{G}}$ and depth $d = 0$ has the attribute set $M_{\mathcal{G},\Delta_{\mathcal{G}},0} = \{\perp, \text{Female}, \top\}$.

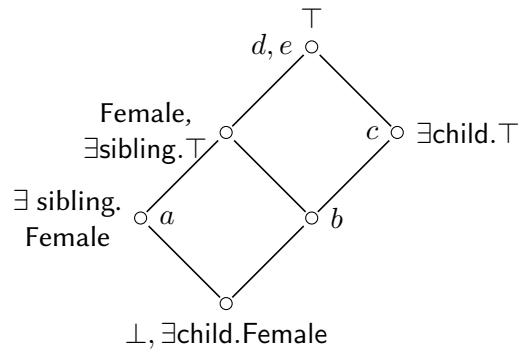
This formal context gives rise to three formal concepts: $(\emptyset, \{\perp\})$, $(\{a, b\}, \{\text{Female}\})$ and $(\{c, d, e\}, \{\top\})$. When ordered w.r.t. inclusion between their extents (first set in the pair), these formal concepts form the depicted hierarchy. For depth 0, we get a rather rough clustering of the knowledge graph. The only commonality of the individuals a and b is that they are both females.



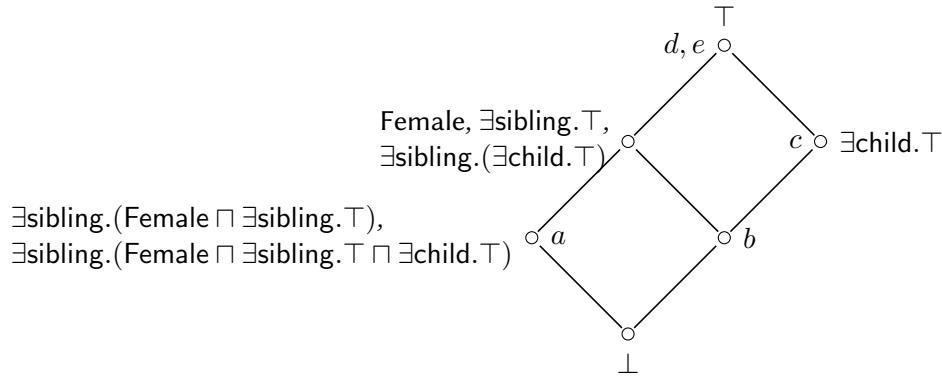
As next, we increase the role-depth d to 1 and using definition 1 compute the attribute set

$$M_{G, \Delta_G, 1} = \{\perp, \text{Female}, \exists \text{ sibling. Female}, \exists \text{ sibling. } \top, \exists \text{ child. Female}, \exists \text{ child. } \top, \top\}$$

which gives us the formal context \mathbb{K}_1 . The new formal context gives rise to 6 formal concepts seen in the concept lattice on the right. It is visible in the lattice, that the entities a and b now belong to the formal concept marked with the attributes Female and $\exists \text{ sibling. } \top$. (For reading the attributes satisfied by an object, start at the node with this object and follow the upwards lines in the lattice.)



For depth 1, the concept hierarchy is more detailed compared to the hierarchy we obtain with depth 0. Finally we build the formal context for the same knowledge graph for role-depth $d = 2$, which gives us the formal context \mathbb{K}_2 with 13 attributes. It gives rise to a new concept lattice, which as the previous one, has 6 formal concepts. This time the concepts are even more fine-grained. Consider again the concept with the extent $\{a, b\}$. This time it has the intent $\{\text{Female}, \exists \text{ sibling. } \top, \exists \text{ sibling. } (\exists \text{ child. } \top)\}$, which describes the set of females that have a sibling, which has a child. Intuitively, this concept corresponds to the notion of an aunt in natural language.



2. Experimenting with WikiData

For evaluating our approach we implemented it in Python and tested it on a small dataset extracted from WikiData. Our implementation¹ uses the Python library `rdflib`² for parsing and navigating a knowledge graph written in RDF. Additionally it uses the `concepts`³ library for computing the formal concepts of a given context. The tests are run on a laptop with an Intel i7 processor with 8 cores running at 2.8Ghz and with 16GB of RAM.

The data extracted from WikiData⁴ contains 1216 triples about the 27 member states of the European Union. The properties occurring in the dataset are `rdf:type`, P1344 (participant in), P463 (member of) and P47 (shares border with). Some of the 191 classes occurring in the data set are Q1065 (United Nations), Q8908 (Council of Europe), Q8268 (Eurozone), Q7817 (World Health Organization) Q8919 (European Atomic Energy Community) and Q7809 (UNESCO).

We ran experiments for 3,4,5,10,15 and 20 of the 27 EU-member countries each with role-depths 0, 1 and 2. Table 1 displays the number of attributes and number of clusters generated in these settings and also the execution time in seconds. For the object set of size 20 and role-depth 2, the implementation ran more than 30 minutes without producing a result. Profiling of the code revealed that the reason for this is the unoptimized implementation of the recursive function for computing the mmsc of a given set of entities. It also revealed that in all of the experiments, the major time consuming part was computing the attributes of the formal context, which is based on the computation of the mmsc.

3. Conclusion and Future Work

We have presented an approach for identifying clusters in a knowledge graph. Unlike other approaches for analyzing knowledge graphs, which are based on statistical and machine learning models, our approach uses a method based on DLs and FCA. It can be used to learn logical descriptions of classes from a knowledge graph and to build a hierarchy of these classes. An experimental evaluation with data sets extracted from WikiData showed that our implementation suffers from performance issues even for small data sets.

As future work, we are going to work on improving both the theory and the implementation of our approach. On the theory part, we are going to investigate how we can improve the computation of the attributes. One idea here could be to learn approximate descriptions of the clusters instead of exact descriptions. In the implementation part, we are going to work on optimizations so that we can compute clusters for larger fragments of WikiData with higher role-depths.

¹<https://github.com/sertkaya/knowledge-graph-concept-learner>

²<https://rdflib.readthedocs.io>

³<https://concepts.readthedocs.io>

⁴<https://www.wikidata.org/>

Number of objects	Depth	Number of		Execution time (s)
		attributes	clusters	
3	0	10	6	0.00044
	1	25	8	0.0023
	2	31	8	0.019
4	0	12	7	0.00045
	1	30	10	0.0036
	2	39	11	0.039
5	0	13	8	0.00057
	1	34	15	0.0054
	2	49	17	0.085
10	0	14	12	0.00078
	1	47	39	0.186
	2	95	55	4.558
15	0	15	16	0.00112
	1	60	59	7.594
	2	141	144	166.55
20	0	16	18	0.00149
	1	67	67	290.02
	2	-	-	-

Table 1
Evaluation results on a WikiData data set

References

- [1] B. Ganter, R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer-Verlag, Berlin, Germany, 1999.
- [2] S. Rudolph, Relational exploration: Combining Description Logics and Formal Concept Analysis for knowledge specification, Ph.D. dissertation, Fakultät Mathematik und Naturwissenschaften, TU Dresden, Germany, 2006.
- [3] F. Distel, Learning description logic knowledge bases from data using methods from formal concept analysis, Ph.D. dissertation, Dresden University of Technology, Germany, 2011. URL: <https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa-70199>.
- [4] F. Dau, B. Sertkaya, Formal concept analysis for qualitative data analysis over triple stores, in: O. D. Troyer, C. B. Medeiros, R. Billen, P. Hallot, A. Simitsis, H. V. Mingroot (Eds.), Advances in Conceptual Modeling. Recent Developments and New Directions - ER 2011 Workshops FP-UML, MoRE-BI, Onto-CoM, SeCoGIS, Variability@ER, WISM, Brussels, Belgium, October 31 - November 3, 2011. Proceedings, volume 6999 of *Lecture Notes in Computer Science*, Springer, 2011, pp. 45–54.
- [5] D. Borchmann, F. Distel, F. Kriegel, Axiomatisation of general concept inclusions from finite interpretations, *Journal of Applied Non-Classical Logics* 26 (2016) 1–46.
- [6] F. Kriegel, Constructing and Extending Description Logic Ontologies using Methods of Formal Concept Analysis, Ph.D. thesis, Technische Universität Dresden, Dresden, Germany, 2019.
- [7] R. Guimarães, A. Ozaki, C. Persia, B. Sertkaya, Mining \mathcal{EL}_\perp bases with adaptable role

- depth, *J. Artif. Intell. Res.* 76 (2023) 883–924. URL: <https://doi.org/10.1613/jair.1.13777>. doi:10.1613/jair.1.13777.
- [8] L. González, A. Hogan, Modelling dynamics in semantic web knowledge graphs with formal concept analysis, in: P. Champin, F. Gandon, M. Lalmas, P. G. Ipeirotis (Eds.), *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, ACM, 2018, pp. 1175–1184. URL: <https://doi.org/10.1145/3178876.3186016>. doi:10.1145/3178876.3186016.
- [9] F. Baader, F. Distel, Exploring finite models in the description logic ELgfp, in: S. Ferré, S. Rudolph (Eds.), *Proceedings of the 7th International Conference on Formal Concept Analysis, (ICFCA 2009)*, volume 5548 of *Lecture Notes in Artificial Intelligence*, Springer-Verlag, 2009, pp. 146–161.
- [10] M. R. Hacene, M. Huchard, A. Napoli, P. Valtchev, Relational concept analysis: mining concept lattices from multi-relational data, *Ann. Math. Artif. Intell.* 67 (2013) 81–108. URL: <https://doi.org/10.1007/s10472-012-9329-3>. doi:10.1007/s10472-012-9329-3.
- [11] F. Baader, R. Küsters, R. Molitor, Computing least common subsumers in description logics with existential restrictions, in: *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, 1999, pp. 96–101.
- [12] A. Ecke, A. Turhan, Role-depth bounded least common subsumers for EL+ and ELI, in: Y. Kazakov, D. Lembo, F. Wolter (Eds.), *Proceedings of the 2012 International Workshop on Description Logics, DL-2012, Rome, Italy, June 7-10, 2012*, volume 846 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2012. URL: https://ceur-ws.org/Vol-846/paper_58.pdf.
- [13] J. C. Jung, C. Lutz, F. Wolter, Least general generalizations in description logic: Verification and existence, in: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, AAAI Press, 2020, pp. 2854–2861. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/5675>.