Faithful Rule Extraction for Differentiable Rule Learning Models (Extended Abstract)

Xiaxia Wang[1,*], **David J. Tena Cucala**[2], **Bernardo Cuenca Grau**[1], **Ian Horrocks**[1]
[1]Department of Computer Science, University of Oxford, UK
[2]Department of Computer Science, Royal Holloway, University of London, UK
xiaxia.wang@cs.ox.ac.uk

Abstract

There is growing interest in extracting rules from machine learning models to improve the interpretability of their predictions, yet many existing methods lack formal guarantees regarding the precise relationship between the extracted rules and the original model's behavior. We investigate rule extraction for DRUM, a rule learning model for knowledge graph completion. After identifying that its original rule extraction approach may yield unsound or incomplete rules, we propose a novel algorithm that ensures both soundness and completeness of extracted rules. To improve practicality, we also introduce dataset-specific rule extraction algorithms and simplified model variants that support faithful rule extraction with a trade-off on expressive power.

1 Introduction

Machine learning models have achieved notable success on a wide range of tasks over knowledge graphs (KGs). However, the lack of interpretability has become a limitation for their application in safety-critical or legally regulated domains. To address this, prior works have explored rule extraction techniques that aim to explain model predictions using symbolic rules. Nonetheless, most of these methods lack formal guarantees for the extracted rules to the model's behavior.

Recent works have proposed notions of soundness and completeness for rules extracted for models [Tena Cucala et al., 2022]. A set of rules is sound if all facts derived by them are also predicted by the model. Conversely, it is complete if all facts predicted by the model can also be derived by the rules. A set of rules that is both sound and complete is termed faithful to the model. Our work [Wang et al., 2024] focuses on DRUM [Sadeghian et al., 2019], an approach with superior performance in KG completion. First, we analyze the DRUM model and reveal that its original rule extraction approach may suffer from unsoundness and/or incompleteness. To address this, we propose a new algorithm for extracting faithful rules expressed in an extension of Datalog that incorporates inequalities and disjunctions in the rule body. To improve practical efficiency of rule extraction, we also introduce additional strategies for reducing computational complexity, including dataset-specific rule extraction and simplified model variants.

2 Faithful Rule Extraction for DRUM

The DRUM model is designed to learn chain-like Datalog rules that represent paths in a knowledge graph, which can be expressed in the form: $H(x, y) \leftarrow B_1(x, z_1) \wedge B_2(z_1, z_2) \wedge \cdots \wedge B_\ell(z_{\ell-1}, y)$. Our analysis begins by characterizing the behavior of a DRUM model. We demonstrate that the model's decision to predict a fact can be understood as *counting the distinct matches of chain-like conjunctions* of atoms within the input KG, where variables are mapped to specific constants. Based on that, we show that rules extracted by the original approach from DRUM may be unsound or incomplete. Specifically, while soundness could be ensured by choosing appropriate thresholds, completeness is fundamentally unattainable with standard Datalog.

To bridge this expressivity gap and enable faithful rule extraction, we characterize the model's behavior using an extension of Datalog with inequality atoms and disjunctions in the rule body, which supports explicit counting semantics by enforcing different variables mapped to distinct constants. We introduce 'multipath conjunction' as the basic building block in our extended rule bodies. A multipath conjunction

[*]Corresponding author.

IRIS-AI 2025: *International Research and Industry Symposium on Artificial Intelligence*, Montefiascone, Italy.

consists of multiple copies of a chain-like conjunction of binary atoms, with variables renamed and inequalities added to ensure distinct matches. Subsequently, we propose an algorithm that describes the procedure for extracting a faithful set of rules from a DRUM model. However, despite its theoretical contribution, the computational complexity of the algorithm makes it challenging to use in practice.

3 Practical Rule Extraction

To address the challenge of scalability, we propose two practical strategies to facilitate faithful rule extraction for real-world applications.

Dataset-Specific Rule Extraction. In contrast to the above algorithm being dataset-independent, we propose an alternative algorithm that focuses on *a fixed dataset*. Specifically, it extracts a small subset of sound rules that exactly replicate the predictions made by a DRUM model on the given dataset. This approach is particularly useful in scenarios where the underlying KG does not change frequently, offering a trade-off between generality and efficiency in realistic settings.

Model Simplification. We introduce two simplified variants of the DRUM model with reduced expressivity. SMDRUM removes the model's counting ability by altering the aggregation mechanism of the model from standard matrix product to max-product. Beyond that, MMDRUM further simplifies the model by replacing sum aggregations for paths with max operations, which enable efficient faithful rule extraction under standard Datalog semantics. These simplifications allow for faithful rule extraction within practical time constraints, albeit with some trade-off for expressive power.

4 Evaluation and Future Works

We compared the performance of DRUM, SMDRUM, and MMDRUM on multiple inductive KG completion benchmarks, including FB15k-237, NELL-995, WN18RR, and Family. Our results show that:

- DRUM generally outperforms its simplified variants; however, MMDRUM and SMDRUM still achieve competitive performance on certain datasets, indicating their practical usability.
- The algorithms for dataset-specific rule extraction from DRUM and faithful rule extraction from MMDRUM both finished within a few minutes across all KGs. This empirically validates the feasibility of our proposed methods for real-world applications.
- The original rule extraction approach for DRUM covered less than 7% of the model's predictions, underscoring their insufficiency for explaining model behavior and highlighting the necessity of faithful rule extraction.

Our future work will explore several promising directions.

- First, we plan to generalize rule learning and faithful rule extraction to handle higher-arity predicates, thereby broadening the applicability of faithful rule extraction to diverse data structures like tabular data and knowledge hypergraphs.
- Second, we aim to extend our faithfulness analysis to other rule learning approaches, such as ∂ILP.
- Third, we are interested in analyzing the extracted rules, such as their actual support rates within the input KG, as these insights reflect the model's behaviors and can inform subsequent improvements to the model.
- Furthermore, we intend to expand the applications of rule learning and faithful rule extraction beyond KG completion to more diverse tasks, such as entity resolution, and for downstream scenarios that require rules as input.

References

Maxime Buron, François Goasdoué, Ioana Manolescu, and Marie-Laure Mugnier. Reformulation-based query answering for rdf graphs with rdfs ontologies. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, pages 19–35. Springer, 2019. https://doi.org/10.1007/978-3-030-21348-0_2.

Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Tractable reasoning and efficient query answering in description logics: The dl-lite family. *Journal of Automated reasoning*, 39:385–429, 2007. https://doi.org/10.1007/s10817-007-9078-x.

Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Mariano Rodriguez-Muro, Riccardo Rosati, Marco Ruzzi, and Domenico Fabio Savo. The mastro system for ontology-based data access. *Semantic web*, 2(1):43–53, 2011. https://doi.org/10.3233/SW-2011-0029.

Diego Calvanese, Benjamin Cogrel, Sarah Komla-Ebri, Roman Kontchakov, Davide Lanti, Martin Rezk, Mariano Rodriguez-Muro, and Guohui Xiao. Ontop: Answering sparql queries over relational databases. *Semantic Web*, 8(3):471–487, 2016. https://doi.org/10.3233/SW-160217.

Melisachew Wudage Chekol, Jérôme Euzenat, Pierre Genevès, and Nabil Layaïda. Sparql query containment under rdfs entailment regime. In *International Joint Conference on Automated Reasoning*, pages 134–148. Springer, 2012. https://doi.org/10.1007/978-3-642-31365-3_13.

Gianluca Cima, Maurizio Lenzerini, and Antonella Poggi. Answering conjunctive queries with inequalities in *DL-Lite*$_{\mathscr{R}}$. pages 2782–2789, 2020. https://doi.org/10.1609/aaai.v34i03.5666.

Gianluca Cima, Marco Console, Roberto Maria Delfino, Maurizio Lenzerini, and Antonella Poggi. Answering conjunctive queries with safe negation and inequalities over rdfs knowledge bases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(14):14824–14831, Apr. 2025. https://doi.org/10.1609/aaai.v39i14.33625.

Jos De Bruijn and Stijn Heymans. Logical foundations of (e) rdf (s): Complexity and reasoning. In *International Semantic Web Conference*, pages 86–99. Springer, 2007. https://doi.org/10.1007/978-3-540-76298-0_7.

Roberto Maria Delfino, Maurizio Lenzerini, and Antonella Poggi. RDFS Knowledge Graphs Through the Lens of Logic: Semantics and Query Answering. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025)*, pages 1511–1518. IOS Press, October 2025. ISBN 9781643686318. https://doi.org/10.3233/faia250974.

Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about knowledge*. MIT press, 2004. https://doi.org/10.5555/995831.

Enrico Franconi, Claudio Gutierrez, Alessandro Mosca, Giuseppe Pirrò, and Riccardo Rosati. The logic of extensional RDFS. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha F. Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, volume 8218 of *Lecture Notes in Computer Science*, pages 101–116. Springer, 2013. https://doi.org/10.1007/978-3-642-41335-3_7.

François Goasdoué, Ioana Manolescu, and Alexandra Roatiş. Efficient query answering against dynamic rdf databases. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 299–310, 2013. https://doi.org/10.1145/2452376.2452412.

Sergio Greco, Cristian Molinaro, and Francesca Spezzano. *Incomplete Data and Data Dependencies in Relational Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2012. ISBN 978-3-031-00765-1. https://doi.org/10.2200/S00435ED1V01Y201207DTM029.

Víctor Gutiérrez-Basulto, Yazmin Angélica Ibáñez-García, Roman Kontchakov, and Egor V. Kostylev. Queries with negation and inequalities over lightweight ontologies. 35:184–202, 2015. https://doi.org/10.1016/j.websem.2015.06.002.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021. https://doi.org/10.1145/3447772.

Markus Krötzsch. Ontologies for knowledge graphs? In Alessandro Artale, Birte Glimm, and Roman Kontchakov, editors, *Proceedings of the 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017*, volume 1879 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017. URL https://ceur-ws.org/Vol-1879/invited2.pdf.

Maurizio Lenzerini, Lorenzo Lepore, and Antonella Poggi. Metamodeling and metaquerying in OWL 2 QL. *Artif. Intell.*, 292:103432, 2021. https://doi.org/10.1016/j.artint.2020.103432.

Sergio Munoz, Jorge Pérez, and Claudio Gutierrez. Simple and efficient minimal rdfs. *Journal of web semantics*, 7(3):220–234, 2009. https://doi.org/10.1016/j.websem.2009.07.003.