Mining Symbolic Rules To Explain Lung Cancer Treatments

Disha Purohit[1,2][0000−0002−1442−335X] and
Maria-Esther Vidal[1,2,3][0000−0003−1160−8727]

[1] Leibniz University, Hannover, Germany
[2] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{disha.purohit,maria.vidal}@tib.eu
[3] L3S Research Center, Hannover, Germany

Abstract. Knowledge Graphs (KGs) represent the convergence of data and knowledge as factual statements; they allow for the enrichment of decision-making semantically. Symbolic inductive learning enables uncovering relevant patterns, expressed, for example, as Horn clauses. Albeit powerful, existing symbolic inductive learning frameworks may mine many rules, being difficult for a user to extract actionable insights. This demo illustrates a pipeline to analyze mined logical rules toward discovering meaningful insights. The demo puts into perspective the role of semantic types in guiding the exploration of mined rules. Participants will observe strategies to traverse the mined logical statements and how the outcomes reveal patterns in the prescription of lung cancer treatments. A video is available online[4], a Jupyter notebook executes a live demos [5], and source-code is available in GitHub [6].

1 Introduction

Knowledge Graphs (KGs) are widely used to represent real-world data in the form of entities and relations. Several open KGs like DBpedia, and Wikidata, are already available to severe the Semantic Web community. KGs are frequently created from heterogeneous sources, which can vary significantly in terms of structure and granularity [3]. The open research challenge of mining Horn rules from facts and analyzing the mined logical rules to uncover meaningful insights has received numerous contributions from the Semantic Web. Exemplary rule mining approaches (e.g., AMIE[1,2,4], AnyBURL[5]) are devised to operate under OWA and mine logical rules. However, these approaches must still be designed to deal with KGs that include semantics and ignore the importance of analyzing the rules. We demonstrate DIGGER, a framework for analyzing mined logical rules. We will illustrate the impact of incorporating semantics and the relevant role of inductive learning in knowledge discovery.

[4] https://www.youtube.com/watch?v=CN4a3kUjfJ4&ab_channel=TIBSDMGroup
[5] https://mybinder.org/v2/gh/SDM-TIB/DIGGER-ESWC2023Demo/HEAD?labpath=Mining%20Symbolic%20Rules%20To%20Explain%20Lung%20Cancer%20Treatments.ipynb
[6] https://github.com/SDM-TIB/Mining_Symbolic_Rules_ESWC2023Demo

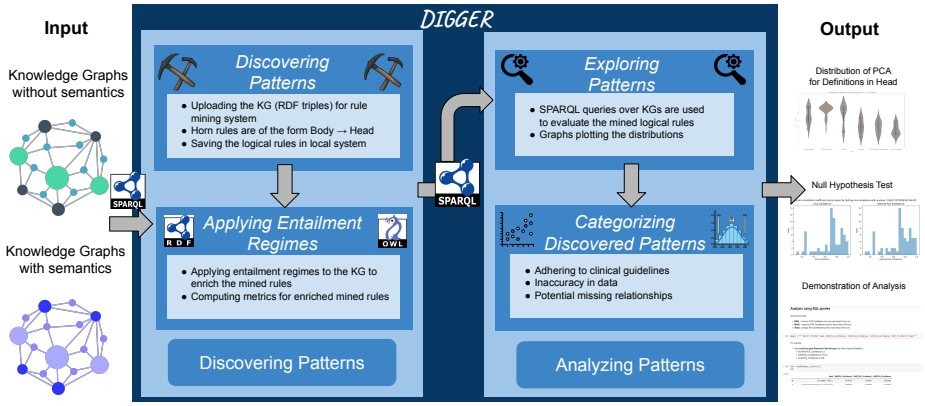

Fig. 1. Architecture. Input includes KGs, ontologies, and entailment regimes. SPARQL queries explore mined rules and perform analysis over KGs. DIGGER demonstrates clinical guidelines' validation, data errors, and missing relationships.

2 The DIGGER Architecture

We aim at providing DIGGER, a framework able to analyze mined logical rules from which true missing facts can be predicted. As a result, KG completion can be achieved with facts inferred from the mined rules and entailment regimes (e.g., RDFS or OWL). DIGGER currently relies on AMIE [4] to efficiently mine logical rules over KGs. Figure 1 depicts the DIGGER architecture; it receives as input KGs and outputs visualizations depicting the results of the analysis of the mined rules on top of the input KGs. DIGGER comprises two steps: a) *Discovering Patterns* and b) *Analyzing Patterns*. The former mines logical rules and then, expands them by applying the entailment regimes. Applying the W3C-recommended Web Ontology Language *(OWL)* and *RDFS* entailment regimes, helps to derive new insights from the KGs. Currently, DIGGER is considering entailment regimes but in the future incorporating *SHACL Integrity Constraints* and *Deductive Systems* to enhance the mined logical rules. Thus, it clearly illustrates that by incorporating the semantics of the KGs the mined logical rules are enriched. The logical rules are loaded in any relational database management system to expedite the process of analysis that can be performed over the mined logical rules. On the other hand, *Analyzing Patterns* explores the mined logical rules towards the discovery of unknown patterns. SPARQL queries are used to explore the mined logical rules. The Partial Completeness Assumption (PCA) assumes that heuristic-based negative edges are possible incomplete edges. Evaluating the logical rules based on the PCA Confidence metrics computed for the rules provides the possibility of identifying the potential prediction of edges over the KGs. Further, using the framework, identifying data errors, missing relationships, and lung cancer patients violating the clinical guidelines is clearly demonstrated which can help oncologists discover insights.

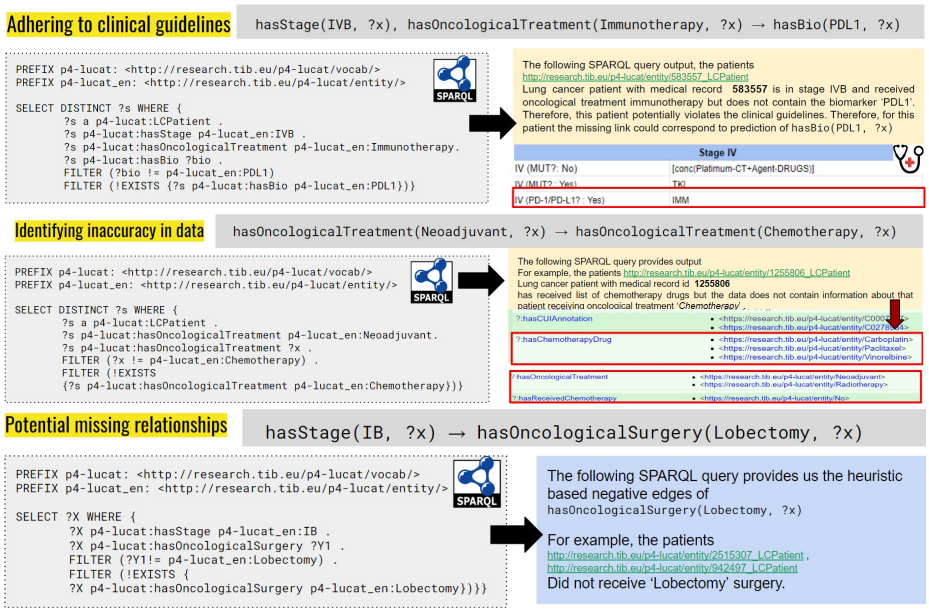

Fig. 2. Use cases: Illustration of use cases with example rules to demonstrate the usage of SPARQL queries over KGs to analyze the mined logical rules. The figure shows all three use cases exhibited in the video demonstration of *DIGGER*.

3 Demonstration of Use Case

The demonstration aims at illustrating the importance of considering the semantics of KGs in mining logical rules from ground facts. Also, use cases in Figure 2 that help in analyzing the mined logical rules are described in this section.

Adhering to clinical guidelines: Attendees will be able to observe the mined logical rules that are used to explain lung cancer treatments. Similarly, attendees will be able to discover the mined logical rules utilized in identifying patients that violate the clinical guidelines. Clinical guidelines are oncologists' treatment protocols for lung cancer. For instance, the logical rule `hasStage(IVB, X), hasOncologicalTreatment(Immunotherapy, X)` $\Rightarrow$ `hasBio(PDL1, X)` states if a patient is in cancer stage *IV* and receives an oncological treatment *Immunotherapy* then it is most likely that the patient is positive for biomarker *PDL1*. This rule complies with the clinical guidelines established by oncologists. The PCA Confidence score computed for the above-mentioned rule was 0.966 which states the KG is partially complete and has heuristic-based negative edges. Therefore, there are few patients that do not abide by the clinical guidelines. By running SPARQL queries over the KG, our framework can identify a patient who is not following the guidelines. As a result, we can conclude that the identified patient may be violating clinical guidelines, and the missing link for that patient may correspond to the prediction of `hasBio(PDL1, X)`; to complete KG.

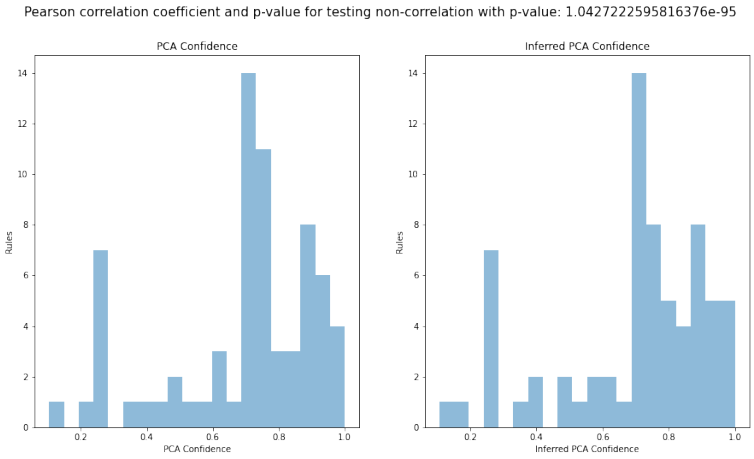

Fig. 3. Distribution of PCA Confidence: Analysis to demonstrate the usage of entailment regimes and inference over KGs. The experimental results of the probability of the correlation between *PCA Confidence* and *Inferred PCA Confidence*. It is represented by the p-value which states that the metrics are statistically significant.

Identifying inaccuracy in data: The attendees will be able to discover the usage of mined logical rules in detecting errors in the data. Errors in clinical data are common and can result in incorrect outcomes. Mining logical rules over KGs aids in the detection of data errors. For example, the mined logical rule `hasOncologicalTreatment(Neoadjuvant, X)` ⇒ `hasOncologicalTreatment(Chemotherapy, X)` with computed PCA Confidence score of 0.971; one patient is identified as receiving *Neoadjuvant* oncological treatment but not receiving oncological treatment *Chemotherapy*. Further investigation revealed a data error in which chemotherapy drugs were administered to that patient but chemotherapy treatment was not recorded in the data. This type of error potentially leads to false conclusions about patients' treatments and is, thus scrutinized by *DIGGER* to help the oncologists make better decisions.
Potential missing relationships: Attendees will be able to explore the mined logical rules in identifying the potential incomplete edges of KGs using the *PCA*. The definition of negative edges in a labeled-edge graph $G = (V, E, L)$ is not precise under the OWA. The goal of the PCA is to make accurate predictions that can potentially complete the large incomplete KGs. To illustrate potential missing links in the KG a logical rule `hasStage(IB, X)` ⇒ `hasOncologicalSurgery(Lobectomy, X)` states that patients who are in cancer stage *IB* receive oncological surgery as *Lobectomy*. The PCA Confidence score computed for this rule was 0.976. Using a SPARQL query, DIGGER is able to identify two patients who were recorded to be in stage *IB* and did not receive *Lobectomy*. As a result, we can conclude that these two patients in the KG have missing links to onco-

logical surgery and can be regarded as potential predictions. *DIGGER* is domain agnostic but in order to evaluate the accuracy of the use cases used in this work clinical guidelines are considered and humans are in a loop. Another study aims at reporting the impact of injecting entailment regimes on the KGs. In contrast to naive approaches, *DIGGER* takes `rdfs:subPropertyOf` into account for the experiments in the current example. This yields higher metrics values and demonstrates potential true predictions. For example, higher Inferred PCA Confidence of a rule quantifies the KG's partial completion by identifying more productive rules. The mined logical rules with all the metrics are computed on the KGs first without considering the entailment regimes to obtain *PCA Confidence*. Further, the entailment regimes are injected into the same KG before mining the logical rules. A null hypothesis test (i.e., p-value) shown in Figure 3 is used to observe the difference in metrics value to compare the results.

4 Conclusion and Future Work

We demonstrate our framework that allows semantically identifying missing information in terms of relationships in the KG or data. Additionally, attendees will be able to observe how we use logical rules to discover potential errors, relationships, and protocol violations in the healthcare domain. To justify the demonstration, the oncologists from the P4-LUCAT project confirmed our experimental findings. By incorporating the analysis discussed in this paper, we aim to design a scalable rule mining system that takes into account all of the semantics of the KGs to discover more meaningful insights. More importantly, evidence of the work presented and the analysis methodology will be provided.

Acknowledgements This work has been supported by the project TrustKG - Transforming Data in Trustable Insights with grant P99/2020 and the EraMed project P4-LUCAT (GA No. 53000015).

References

1. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.: Fast Rule Mining in Ontological Knowledge Bases with AMIE+. The VLDB Journal (2015), https://hal-imt.archives-ouvertes.fr/hal-01699866
2. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.: Amie: Association rule mining under incomplete evidence in ontological knowledge bases. In: WWW 2013 (2013). https://doi.org/10.1145/2488388.2488425
3. Hogan, A., et al.: Knowledge Graphs (2021). https://doi.org/10.2200/S01125ED1V01Y202109DSK022
4. Lajus, J., Galárraga, L., Suchanek, F.: Fast and exact rule mining with amie 3. In: Harth, A., Kirrane, S., Ngonga Ngomo, A.C., Paulheim, H., Rula, A., Gentile, A.L., Haase, P., Cochez, M. (eds.) The Semantic Web (2020)
5. Meilicke, C., Chekol, M.W., Ruffinelli, D., Stuckenschmidt, H.: Anytime bottom-up rule learning for knowledge graph completion. In: IJCAI-19 (2019). https://doi.org/10.24963/ijcai.2019/435