Towards Nonmonotonic Relational Learning from Knowledge Graphs

Hai Dang Tran[a], Daria Stepanova[a], Mohamed H. Gad-Elrab [a], Francesca A. Lisi[b], Gerhard Weikum[a]

[a]Max Planck Institute of Informatics Saarbrücken, Germany
{htran,dstepano,gadelrab,weikum}@mpi-inf.mpg.de
[b]Università degli Studi di Bari "Aldo Moro", Bari, Italy - francesca.lisi@uniba.it

1 Introduction

Recent advances in information extraction have led to the so-called *knowledge graphs* (KGs), *i.e.* huge collections of *triples* in the form of $\langle subject\ predicate\ object \rangle$ according to the RDF data model [11]. These triples encode positive *relational* facts about the world, and they are naturally treated under the Open World Assumption (OWA).

In this work, we focus on KGs without blank nodes or schema. For simplicity, we represent the triples using unary and binary predicates. The unary predicates are the objects of the RDF $type$ predicate, while the binary ones correspond to all other RDF predicates, *e.g.*, $\langle alice\ type\ researcher \rangle$ and $\langle bob\ isMarriedTo\ alice \rangle$ from the KG in Fig. 1 correspond to the facts $researcher(alice)$ and $isMarriedTo(bob, alice)$. Notable examples of KGs are NELL [4], DBpedia [1], YAGO [13] and Wikidata [7].

As such KGs are automatically constructed, they are often erroneous and incomplete. To complete and curate KGs, data mining techniques (*e.g.*, [5,15,8]) have been exploited to infer Horn rules. These, however, are insufficiently expressive to capture exceptions, and might therefore make incorrect predictions on missing facts. For example, the application of the Horn rule

$$r1:\ livesIn(Z, Y) \leftarrow isMarriedTo(X, Z), livesIn(X, Y)$$

mined from the KG in Fig. 1 produces the following facts: $livesIn(alice,berlin)$, $livesIn(dave,chicago)$ and $livesIn(lucy, amsterdam)$. However, the first two facts might be actually false; indeed, both $alice$ and $dave$ are researchers, and the rule $r1$ could be suspected to have $researcher$ as a potential exception. In [14], a rule-based method for addressing this issue was proposed. It takes into account a cross-talk between the rules and aims at revising them to obtain a ruleset that describes the observed data well and predicts the unseen data consistently. However, it applies only to a flattened representation of a KG.

In this paper we extend the results from [14] to KGs in their original relational nature. Basically, we consider a *theory revision* problem, where, given a KG and a set of (previously learned) Horn rules, the task is to compute a set of *nonmonotonic rules* by adding default negated atoms into their bodies, such that the revised ruleset is more accurate than the original one. Note that in our case, and in contrast to the standard formulation of the problem (see, *e.g.*, [16]), the negative examples are not available due to the OWA. Thus, traditional methods cannot be applied.

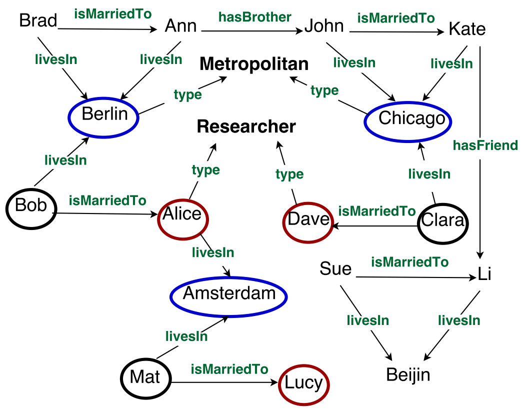

Fig. 1: Example of a Knowledge Graph

We implemented the developed extension and evaluated it on a real-world KG. Preliminary experiments demonstrate the effectiveness of our method and gains for the rule quality as well as the fact quality when performing KG completion.

2 A Theory Revision Framework

We start with introducing necessary notions and defining our problem formally.

The factual representation of a KG $\mathcal{G}$ is defined over the signature $\Sigma_{\mathcal{G}} = \langle \mathbf{C}, \mathbf{R}, \mathcal{C} \rangle$, where $\mathbf{C}$, $\mathbf{R}$ and $\mathcal{C}$ are sets of unary predicates, binary predicates and constants, resp. Following [6], we define the gap between the available graph $\mathcal{G}^a$ and the ideal graph $\mathcal{G}^i$.

Definition 1 (Incomplete data source). *An incomplete data source is a pair* $G = (\mathcal{G}^a, \mathcal{G}^i)$ *of two KGs, where* $\mathcal{G}^a \subseteq \mathcal{G}^i$ *and* $\Sigma_{\mathcal{G}^a} = \Sigma_{\mathcal{G}^i}$.

Our solution is based on the application of rules within the framework of nonmonotonic logic programming under answer set semantics (see [9] for details).

Definition 2 (Rule-based KG completion). *Let* $\mathcal{G}$ *be a KG over* $\Sigma_{\mathcal{G}} = \langle \mathbf{C}, \mathbf{R}, \mathcal{C} \rangle$ *and* $\mathcal{R}$ *be a set of rules mined from* $\mathcal{G}$*, i.e. rules over the signature* $\Sigma_{\mathcal{R}} = \langle \mathbf{C} \cup \mathbf{R}, \mathcal{C} \rangle$*. Then the* completion *of* $\mathcal{G}$ *is a graph* $\mathcal{G}_{\mathcal{R}}$ *constructed from any answer set of* $\mathcal{R} \cup \mathcal{G}$.

Note that $\mathcal{G}^i$ is the perfect completion of $\mathcal{G}^a$, which is supposed to contain all correct facts with entities and relations from $\Sigma_{\mathcal{G}^a}$ that hold in the current state of the world. Given a potentially incomplete graph $\mathcal{G}^a$ and a set $\mathcal{R}_H$ of Horn rules mined from $\mathcal{G}^a$, our goal is to add default negated atoms (exceptions) to the rules in $\mathcal{R}_H$ and obtain a revised ruleset $\mathcal{R}_{NM}$ such that the set difference between $\mathcal{G}^a_{\mathcal{R}_{NM}}$ and $\mathcal{G}^i$ is as small as possible. Normally, the ideal graph $\mathcal{G}^i$ is not available, and due to the OWA applying standard measures from ILP to evaluate the quality of a revised ruleset is impossible. For that reason instead we exploit measures from predictive association rule mining (see [2] for a survey) to approximately estimate the quality of a revised ruleset. We devise two quality functions q_{rm} and $q_{conflict}$, that take a ruleset $\mathcal{R}$ and a KG $\mathcal{G}$ as input and output a real value, reflecting the suitability of $\mathcal{R}$ for data prediction. The former generalizes standard predictive association rule measures rm to rulesets. More specifically,

$$q_{rm}(\mathcal{R},\mathcal{G}) = \frac{\sum_{r\in\mathcal{R}} rm(r,\mathcal{G})}{|\mathcal{R}|}. \quad (1)$$

Conversely, the latter estimates the number of conflicting predictions that the rules in a set generate. To measure $q_{conflict}$ for a given $\mathcal{R}$, we create an extended set of rules $\mathcal{R}^{aux}$, which contains each rule r in $\mathcal{R}$ together with its auxiliary version r^{aux}, constructed as follows: 1) transform r into a Horn rule by removing not from negated body atoms, and 2) replace the head predicate a of r with a newly introduced predicate not_a which intuitively contains instances which are *not* in a. Formally,

$$q_{conflict}(\mathcal{R},\mathcal{G}) = \sum_{p\in pred(\mathcal{R}^{aux})} \frac{|\{\boldsymbol{c} \mid p(\boldsymbol{c}), not\_p(\boldsymbol{c}) \in \mathcal{G}_{\mathcal{R}^{aux}}\}|}{|\{\boldsymbol{c} \mid not\_p(\boldsymbol{c}) \in \mathcal{G}_{\mathcal{R}^{aux}}\}|} \quad (2)$$

We are now ready to define our problem.

Definition 3 (Quality-based Horn theory revision (QHTR)). *Given a set $\mathcal{R}_H$ of Horn rules, a KG $\mathcal{G}$, and the quality functions q_{rm} and $q_{conflict}$, the quality-based Horn theory revision problem is to find a set $\mathcal{R}_{NM}$ of rules obtained by adding default negated atoms to $Body(r)$ for some $r \in \mathcal{R}_H$, such that (i) $q_{rm}(\mathcal{R}_{NM},\mathcal{G})$ is maximal, and (ii) $q_{conflict}(\mathcal{R}_{NM},\mathcal{G})$ is minimal.*

Following the common practice, we consider only rules with linked variables [10]. Prior to tackling QHTR problem we introduce the notions of r-(ab)normal substitutions and Exception Witness Sets (EWSs) that are used in our revision framework.

Definition 4 (r-(ab)normal substitutions). *Let $\mathcal{G}$ be a KG, r a Horn rule mined from $\mathcal{G}$, and let $\mathcal{V}$ be a set of variables occurring in r. Then*

- $NS(r,\mathcal{G}) = \{\theta \mid H(r)\theta, B(r)\theta \in \mathcal{G}\}$ *is an r-normal set of substitutions;*
- $ABS(r,\mathcal{G}) = \{\theta' \mid B(r)\theta' \in \mathcal{G},\ H(r)\theta' \notin \mathcal{G}\}$ *is an r-abnormal set of substitutions,*

where $\theta, \theta' : \mathcal{V} \to \mathcal{C}$.

Example 1. For the KG from Fig. 1 and $r:\ livesIn(Y,Z) \leftarrow isMarriedTo(X,Y), livesIn(X,Z)$ we have

- $NS(r,\mathcal{G}) = \{\theta_1,\theta_2,\theta_3\}$, where $\theta_1 = \{X/Brad,\ Y/Ann,\ Z/Berlin\}$, $\theta_2 = \{X/John,\ Y/Kate,\ Z/Chicago\}, \theta_3 = \{X/Sue,\ Y/Li,\ Z/Beijin\}$
- $ABS(r,\mathcal{G}) = \{\theta_4,\theta_5,\theta_6\}$, where $\theta_4 = \{X/Bob,\ Y/Alice,\ Z/Berlin\}, \theta_5 = \{X/Clara,\ Y/Dave,\ Z/Chicago\}, \theta_6 = \{X/Mat,\ Y/Lucy,\ Z/Amsterdam\}$.

Intuitively, if the given data was complete, then the r-normal and r-abnormal substitutions would exactly correspond to substitutions for which the rule r holds (resp. does not hold) in $\mathcal{G}^i$. However, some r-abnormal substitutions might be classified as such due to the OWA. In order to distinguish the "wrongly" and "correctly" classified substitutions in the r-abnormal set, we construct *exception witness sets* (EWS).

Definition 5 (Exception Witness Set (EWS)). *Let $\mathcal{G}$ be a KG, let r be a rule mined from it, let $\mathcal{V}$ be a set of variables occurring in r and $\boldsymbol{X} \subseteq \mathcal{V}$. Exception witness set for r w.r.t. $\mathcal{G}$ and $\boldsymbol{X}$ is a maximal set of predicates $EWS(r, \mathcal{G}, \boldsymbol{X}) = \{e_1, \ldots, e_k\}$, s.t.*

- $e_i(\boldsymbol{X}\theta_j) \in \mathcal{G}$ *for some* $\theta_j \in ABS(r, \mathcal{G})$, $1 \leq i \leq k$ *and*
- $e_1(\boldsymbol{X}\theta'), \ldots, e_k(\boldsymbol{X}\theta') \notin \mathcal{G}$ *for all* $\theta' \in NS(r, \mathcal{G})$.

Example 2. For the KG in Fig. 1 and r from Ex. 1 we have that $EWS(r, \mathcal{G}, X) = \{researcher\}$. If $artist(bob)$, $artist(clara)$, $artist(mat)$ were in $\mathcal{G}$ then it would hold that $EWS(r, \mathcal{G}, Y) = \{artist\}$. Moreover, if $brad$ with ann and $john$ with $kate$ lived in cities different from $berlin$ and $chicago$, then $EWS(r, \mathcal{G}, Z) = \{metropolitan\}$.

In general when binary atoms are allowed in the rules, there might be potentially too many possible EWSs to construct and consider. For a rule with n distinct variables, n^2 candidate EWS sets can exist.

3 Methodology

Our methodology for solving the QHTR problem comprises four steps, which we now discuss in details.

Step 1. After mining Horn rules using a state-of-the-art algorithm (e.g., [5] or [8]), we compute for each rule r the *r-normal* and *r-abnormal* substitutions.

Step 2 and 3. Then for every rule $r \in \mathcal{R}_H$ with the head atom $h(X, Y)$ we determine three EWS sets: $EWS(r, \mathcal{G}, X)$, $EWS(r, \mathcal{G}, Y)$ and $EWS(r, \mathcal{G}, \langle X, Y \rangle)$. From the obtained EWSs we create $|EWS(r, \mathcal{G}, X)| + |EWS(r, \mathcal{G}, Y)| + |EWS(r, \mathcal{G}, \langle X, Y \rangle)|$ potential revisions by adding every determined exception in the form of a negated atom to the rule r at hand.

Steps 4. After all candidate revisions are constructed we rank them and determine the resulting set $\mathcal{R}_{NM}$ by selecting for every rule the revision that is ranked the highest.

To find such globally best revised ruleset $\mathcal{R}_{NM}$ too many candidate combinations have to be checked, which is impractical due to the large size of the KG $\mathcal{G}$ and EWSs.

Thus, instead we incrementally build $\mathcal{R}_{NM}$ by considering every $r_i \in \mathcal{R}_H$ and choosing the locally best revision r_i^j for it. For that we exploit three ranking functions: a naive one and two more sophisticated ones, which invoke the novel concept of *partial materialization* (***PM***). Intuitively, the idea behind it is to rank candidate revisions not based on $\mathcal{G}$, but rather on its extension with predictions produced by other (selectively chosen) rules (grouped into a set $\mathcal{R}'$), thus ensuring a cross-talk between the rules. We now describe the ranking functions in more details.

- ***Naive (N)*** ranker is the most simple function, which prefers the revision r_i^j with the highest value of $rm(r_i^j, \mathcal{G})$ among all revisions of r_i.
- ***PM*** ranking function prefers r_i^j with the highest value of

$$score(r_i^j, \mathcal{G}) = \frac{rm(r_i^j, \mathcal{G}_{\mathcal{R}'}) + rm({r_i^j}^{aux}, \mathcal{G}_{\mathcal{R}'})}{2} \tag{3}$$

where $\mathcal{R}'$ is the set of rules r'_k, which are rules from $\mathcal{R}_H \backslash r_i$ with all candidate exceptions for r_k incorporated at once. Informally, $\mathcal{G}_{\mathcal{R}'}$ contains only facts that can be safely predicted by the rules from $\mathcal{R}_H \backslash r_i$, i.e., there is no evident reason (candidate exceptions) for avoiding predictions of these facts.

- ***OPM*** is similar to ***PM***, but the selected ruleset $\mathcal{R}'$ contains only those rules whose Horn version appears above the considered rule r_i in the ruleset $\mathcal{R}_H$, ordered (***O***) based on some chosen measure (e.g., the same as rm).

The algorithms for Step 1 and 2 are extended versions of those from [14]. We neglect further details here for space reasons.

4 Evaluation

Our revision approach is implemented in Java in a system prototype[1]. We conducted preliminary experiments on a multi-core Linux server with 40 cores and 400GB RAM and evaluated how our revision approach impacts (1) the quality of the rules and (2) the correctness of the predictions they produce. To estimate (1), we compare the average conviction of the Horn rules with their revisions, and measure $q_{conflict}$ for the revised ruleset. To evaluate (2), we count the number of true (resp. false) facts that were correctly preserved (resp. neglected) by the revised rules among all Horn rule predictions.

Dataset. For the approximation $\mathcal{G}^i_{appr}$ of the ideal graph we take a KG extracted from the IMDB [2] dataset with 111783 entities, 38 relations and 583380 facts [3]. We constructed the available graph $\mathcal{G}^a$ with 511482 facts by removing from $\mathcal{G}^i_{appr}$ 20% of the facts for each binary predicate. As a side constraint we ensure that in $\mathcal{G}^a$ every node is connected to at least one other node.

Setup. The general workflow of the experiment is as follows. First, we mine Horn rules of the form $h(X,Y) \leftarrow p(X,Z), q(Z,Y)$ from $\mathcal{G}^a$ and rank them w.r.t. *conviction* [3].

$$conv(r) = \frac{1 - supp(r)}{1 - conf(r)}, \tag{4}$$

where $supp(r)$ and $conf(r)$ are resp. the support and the confidence of the rule r. We use this measure, since it appears to be well-suited for prediction as claimed in [2].

We then compute 3 EWSs: $EWS(r,\mathcal{G},X)$,$EWS(r,\mathcal{G},Y)$ and $EWS(r,\mathcal{G},\langle X,Y\rangle)$ over the variables appearing in the head of r. We take conviction as the rm function and rank exceptions e in every EWS using the ***OPM*** ranker (the most advanced method described in Sec. 3). For every rule r, we pick a single exception out of all exception candidates in the three EWSs for which the revised rule gets the highest score.

In Table 1, we report our experimental results for the top-k ($k = 5, 10..20$) revised rules $\mathcal{R}_{NM}$ and their Horn versions $\mathcal{R}_H$. The second and the third columns contain average conviction for $\mathcal{R}_H$ and $\mathcal{R}_{NM}$ resp., while the fourth column stores the value of

[1] https://github.com/htran010589/nonmonotonic-rule-mining
[2] http://imdb.com
[3] https://dl.dropboxusercontent.com/u/12978004/ILP2016_IMDB_data.zip

k	avg. conv.		confl.	number of predictions				
				all		in $\mathcal{G}^i_{appr}$		corr. negl.
	$\mathcal{R}_H$	$\mathcal{R}_{NM}$	$\mathcal{R}_{NM}$	$\mathcal{R}_H$	$\mathcal{R}_{NM}$	$\mathcal{R}_H$	$\mathcal{R}_{NM}$	$\mathcal{R}_{NM}$
5	4.08	6.16	0.28	345	331	161	156	0
10	2.91	4.21	0.08	2178	2118	456	450	27
15	2.5	3.42	0.09	3482	3348	629	622	86
20	2.29	3.0	0.13	5278	5046	848	835	157

Table 1: Rule revision results

$$r_1 : \mathit{writtenBy}(X, Z) \leftarrow \mathit{hasPredecessor}(X, Y), \mathit{writtenBy}(Y, Z), \textbf{not } \mathit{is\_American\_film}(X)$$
$$r_2 : \mathit{actedIn}(X, Z) \leftarrow \mathit{isMarriedTo}(X, Y), \mathit{directed}(Y, Z), \textbf{not } \mathit{is\_silent\_film\_actor}(X)$$
$$r_3 : \mathit{bornIn}(X, Z) \leftarrow \mathit{direct}(X, Y), \mathit{producedIn}(Y, Z), \textbf{not } \mathit{is\_American\_film\_director}(X)$$

Fig. 2: Examples of the revised rules

$q_{conflict}(\mathcal{R}_{NM}, \mathcal{G}^a)$ as defined in Eq. 2. The results show that the revision process consistently enhanced the average conviction. Moreover, the value of $q_{conflict}(\mathcal{R}_{NM}, \mathcal{G}^a)$ decreases almost smoothly with the increase of the number of the considered rules.

The last five columns present the prediction results of applying the rulesets $\mathcal{R}_H$ and $\mathcal{R}_{NM}$ on $\mathcal{G}^a$ using the DLV system [12] to generate the extended graphs $\mathcal{G}^a_{\mathcal{R}_H}$ and $\mathcal{G}^a_{\mathcal{R}_{NM}}$ respectively. Here, "**all**" refers to the number of all predictions produced by the rulesets, and "**in** $\mathcal{G}^i_{appr}$" stores the number of predictions in the approximation of the ideal graph. Among the rest of the predicted facts outside $\mathcal{G}^i_{appr}$, in "**corr. negl.**" we count the number of erroneous predictions made by $\mathcal{R}_H$, that were **correctly neglected** by $\mathcal{R}_{NM}$. Since these predictions are not present in the original KG, we had to assess them manually by consulting such Web resources as Wikipedia and IMDB.

Overall, we were able to predict 5278 facts using the top-20 Horn rules, from which 16% can be verified from $\mathcal{G}^i_{appr}$. Since the Horn versions of the top-5 rules are already of a high quality, their revisions did not prevent erroneous predictions. With the increase of the number of the considered rules, the percentage of the false facts correctly neglected by our revised rules increases reaching 67% for the top-20 rules. In the future work we plan to semi-atomatically derive the truthfulness of the unknown facts outside $\mathcal{G}^i_{appr}$ predicted both by $\mathcal{R}_H$ and $\mathcal{R}_{NM}$ exploiting properties like (a)symmetry or transitivity of the relational predicates as well as making use of information extraction techniques.

For the 134 Horn rules mined from $\mathcal{G}^a$ all EWSs with an average of 1649 exception candidates for each rule were computed within 10 sec. The overall best revision for 134 Horn rules was determined in 30 sec., and the predictions using the revised top-k rules on $\mathcal{G}^a$ were found whithin 1.5 min. via the DLV system.

Fig. 2 shows examples for our revised rules, e.g., r_1 states that the writers of movie plots stay the same throughout the sequel with the exception of American movies. The Horn version of r_2 reflects that the movie directors normally have their spouses on the cast. Our revision approach excluded the old silent movies, for which this practice was not common.

Acknowledgements. We would like to thank Jacopo Urbani for the constructive feedback that helped to improve this work.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: A nucleus for a web of open data. In: Proc. of ISWC. pp. 722–735 (2007)
2. Azevedo, P.J., Jorge, A.M.: Comparing Rule Measures for Predictive Association Rules. In: Proceedings of ECML. pp. 510–517 (2007)
3. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: SIGMOD 1997, Proc. ACM SIGMOD International Conference on Management of Data. pp. 255–264 (1997)
4. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E.R.H., Mitchell, T.M.: Toward an Architecture for Never-Ending Language Learning. In: Proc. of AAAI (2010)
5. Chen, Y., Goldberg, S., Wang, D.Z., Johri, S.S.: Ontological Pathfinding: Mining First-Order Knowledge from Large Knowledge Bases. In: in Proc. of SIGMOD 2016. p. to appear (2016)
6. Darari, F., Nutt, W., Pirrò, G., Razniewski, S.: Completeness Statements about RDF Data Sources and Their Use for Query Answering. In: Proceedings of ISWC. pp. 66–83 (2013)
7. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandecic, D.: Introducing Wikidata to the Linked Data Web. In: Proceedings of ISWC. pp. 50–65 (2014)
8. Galrraga, L., Teflioudi, C., Hose, K., Suchanek, F.M.: Fast Rule Mining in Ontological Knowledge Bases with AMIE+. In: VLDB Journal (2015)
9. Gelfond, M., Lifschitz, V.: The stable model semantics for logic programming. In: Proc. of 5th Int. Conf. and Symp. on Logic Programming, ICLP 1988. pp. 1070–1080 (1988)
10. Helft, N.: Induction as nonmonotonic inference. In: Proc. of the 1st Int. Conf. on Principles of KR (KR'89). pp. 149–156 (1989)
11. Lassila, O., Swick, R.R.: Resource description framework (RDF) model and syntax specification (February 1999)
12. Leone, N., Pfeifer, G., Faber, W., Eiter, T., Gottlob, G., Perri, S., Scarcello, F.: The dlv system for knowledge representation and reasoning. ACM Transactions on Computational Logic (TOCL) 7(3), 499–562 (2006)
13. Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: A knowledge base from multilingual wikipedias. In: Proceedings of CIDR (2015)
14. Mohamed Gad-Elrab, Daria Stepanova, J.U.G.W.: Exception-enriched rule learning from knowledge graphs. In: Proc. of the International Semantic Web Conference. TR available at `http://people.mpi-inf.mpg.de/~gadelrab/ExRules_TR.pdf` (2016)
15. Wang, Z., Li, J.: RDF2Rules: Learning rules from RDF knowledge bases by mining frequent predicate cycles. CoRR abs/1512.07734 (2015)
16. Wrobel, S.: Exception rules in association rule mining. In: Advances in Inductive Logic Programming. pp. 14–33 (1996)