# 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 for Informatics, Saarland Informatics Campus, Saarbrücken, Germany
{htran,dstepano,gadelrab,weikum}@mpi-inf.mpg.de
[b]Università degli Studi di Bari "Aldo Moro", Bari, Italy - francesca.lisi@uniba.it

**Abstract.** Recent advances in information extraction have led to the so-called knowledge graphs (KGs), i.e., huge collections of relational factual knowledge. Since KGs are automatically constructed, they are inherently incomplete, thus naturally treated under the Open World Assumption (OWA). Rule mining techniques have been exploited to support the crucial task of KG completion. However, these techniques can mine Horn rules, which are insufficiently expressive to capture exceptions, and might thus make incorrect predictions on missing links. Recently, a rule-based method for filling in this gap was proposed which, however, applies to a flattened representation of a KG with only unary facts. In this work we make the first steps towards extending this approach to KGs in their original relational form, and provide preliminary evaluation results on real-world KGs, which demonstrate the effectiveness of our method.

## 1 Introduction

**Motivation**. Recent advances in information extraction have led to the so-called *knowledge graphs* (KGs), *i.e.* huge collections of *triples* $\langle subject\ \ predicate\ \ object \rangle$ according to the RDF data model [17]. These triples encode facts about the world and can be straightforwardly represented by means of unary and binary first-order logic (FOL) 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 refer to $researcher(alice)$ and $isMarriedTo(bob, alice)$ respectively. Notable examples of KGs are NELL [4], DBpedia [1], YAGO [23] and Wikidata [9].

Since KGs are automatically constructed, they are inherently *incomplete*. Therefore, they are naturally treated under the Open World Assumption (OWA). The task of *completion* (also known as *link prediction*) is of crucial importance for the curation of KGs. To this aim, rule mining techniques (*e.g.*, [5,12]) have been exploited to automatically build rules able to make predictions on missing links. However, they mine Horn rules, which are insufficiently expressive to capture exceptions, and might thus deduce incorrect facts. For example, the following rule

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

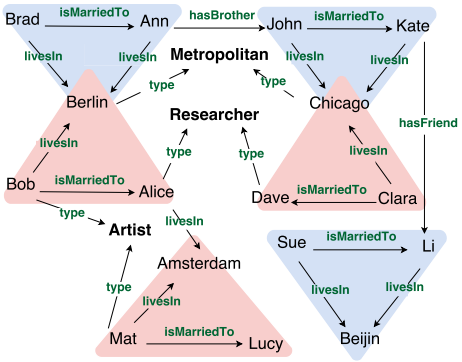


Fig. 1: Example of a Knowledge Graph

can be mined from the KG in Fig. 1 and used to produce the facts $livesIn(alice, berlin)$, $livesIn(dave, chicago)$ and $livesIn(lucy, amsterdam)$. Observe that the first two predicted facts might actually be wrong. Indeed, both $alice$ and $dave$ are researchers, and the rule $r1$ could be suspected to have $researcher$ as a potential exception.

**Challenges**. Exception handling has been traditionally faced in ILP by learning *non-monotonic logic programs*, i.e., programs with negations [15,26,25,6,18]. However, there are several important obstacles that prevent us from using the off-the-shelf non-monotonic ILP algorithms. First, the *target predicates* can not be easily identified, since we do not know which parts of the considered KG need to be completed. A standard way of addressing this issue would be just to learn rules for all the different predicate names occurring in the KG. Unfortunately, this is unfeasible in our case given the huge size of KGs. Second, the *negative examples* are not available, and they can not be easily obtained from, *e.g.*, domain experts due to - once again - the huge size of KGs. A natural solution to cope with this issue is to learn rules from positive examples only. Third, the definition of a *language bias* turns out to be cumbersome since the schema of the KG is usually not available.

To overcome the obstacles mentioned above, it turns out to be appropriate to treat the KG completion problem as an unsupervised relational learning task, and exploit algorithms for relational association rule mining such as [12]. In [11] these techniques are applied to first learn a set of Horn rules, which subsequently can be revised by adding negated atoms to their bodies in order to account for exceptions. However, the proposed approach applies only to a flattened representation of a KG containing just unary facts.

**Contributions**. In this work we extend the results from [11] to KGs in their original relational form. More specifically, we reformulate the KG completion problem as 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*, such that the revised ruleset is more accurate for link prediction than the original one. Essentially, we are interested in tackling a theory revision problem, in which, as possible revision operations, we are only allowed to add negated atoms to the antecedents of the rules.

Our approach combines standard relational association rule mining techniques with a FOIL-like supervised learning algorithm, which is used to detect exceptions. More specifically, we propose a method that proceeds in four steps as follows: First, for every Horn rule we determine the normal and abnormal substitutions, i.e., substitutions that satisfy (resp. do not satisfy) the considered rule. Second, we compute the so-called exception witness sets, i.e., sets of predicates that are potentially involved in explaining why abnormal substitutions fail to follow the rule (e.g., *researcher* in our example). Third, we construct candidate rule revisions by adding a single exception at a time. We devise quality measures for nonmonotonic rules to quantify their strength w.r.t the KG. We consider the crosstalk between the rules through the novel *partial materialization* technique instead of revising rules in isolation. Fourth, we rank rule revisions according to these measures to determine a ruleset that not only describes the data well but also shows a good predictive power by taking exceptions into account.

The contributions of our paper are:

- A theory revision framework, based on nonmonotonic relational rule learning, for capturing exceptions in rule-based approaches to KG completion.
- A methodology for computing exception candidates, measuring their quality, and ranking them taking into account the interaction among the rules.
- Experiments with the YAGO3 and IMDB KGs, which demonstrate the gains of our method for rule quality as well as fact quality when performing KG completion.

**Structure**. Sec. 2 introduces preliminaries on nonmonotonic logic programming and relational association rule mining. Sec. 3 describes our theory revision framework and the methodology. Sec. 4 reports on experimental results, while Sec. 5 and Sec. 6 discuss the related work and conclude the paper respectively.

## 2 Preliminaries

**Nonmonotonic Logic Programming**. We consider logic programs in their usual definition [22] under the answer set semantics. In short, a *(nonmonotonic) logic program* $P$ is a set of *rules* of the form

$$H \leftarrow B, not\ \ E \tag{1}$$

where $H$ is a standard first-order atom of the form $a(\boldsymbol{X})$ known as the rule head and denoted as $head(r)$, $B$ is a conjunction of positive atoms of the form $b_1(\boldsymbol{Y_1}), \ldots, b_k(\boldsymbol{Y_k})$ to which we refer as $body^+(r)$, and $not\ \ E$, with slight abuse of notation, denotes a conjunction of atoms $not\ \ b_{k+1}(\boldsymbol{Y_{k+1}}), \ldots, not\ \ b_n(\boldsymbol{Y_n})$. Here, $not$ is the so-called *negation as failure (NAF)* or *default negation*. The negated part of the body is denoted as $body^-(r)$. The rule $r$ is *positive* or *Horn* if $body^-(r) = \emptyset$. $\boldsymbol{X}, \boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}$ are tuples of either constants or variables whose length corresponds to the arity of the predicates $a, b_1, \ldots, b_n$ respectively. The signature of $P$ is given as $\Sigma_P = \langle \mathbf{P}, \mathcal{C} \rangle$, where $\mathbf{P}$ and $\mathcal{C}$ are resp. sets of predicates and constants occurring in $P$.

A logic program $P$ is *ground* if it consists of only ground rules, i.e. rules without variables. Ground instantiation $Gr(P)$ of a nonground program $P$ is obtained by substituting variables with constants in all possible ways. The *Herbrand universe* $HU(P)$ (resp. *Herbrand base* $HB(P)$) of $P$, is the set of all constants occurring in $P$, i.e.

$HU(P) = \mathcal{C}$ (resp. the set of all possible ground atoms that can be formed with predicates in $\mathbf{P}$ and constants in $\mathcal{C}$). We refer to any subset of $HB(P)$ as a *Herbrand interpretation*. By $MM(P)$ we denote the set-inclusion minimal Herbrand interpretation of a ground positive program $P$.

An interpretation $I$ of $P$ is an *answer set* (or *stable model*) of $P$ iff $I \in MM(P^I)$, where $P^I$ is the *Gelfond-Lifschitz (GL) reduct* [13] of $P$, obtained from $Gr(P)$ by removing (i) each rule $r$ such that $body^-(r) \cap I \neq \emptyset$, and (ii) all the negative atoms from the remaining rules. The set of answer sets of a program $P$ is denoted by $AS(P)$.

*Example 1.* Consider the program

$$P = \left\{ \begin{array}{l} (1)\ livesIn(brad, berlin);\ (2)\ isMarriedTo(brad, ann); \\ (3)\ livesIn(Y, Z) \leftarrow isMarriedTo(X, Y), livesIn(X, Z), not\ \ researcher(Y) \end{array} \right\}$$

The ground instantiation $Gr(P)$ of $P$ is obtained by substituting $X, Y, Z$ with $brad$, $ann$ and $berlin$ respectively. For $I = \{isMarriedTo(brad, ann),\ livesIn(ann, berlin),\ livesIn(brad, berlin)\}$, the GL-reduct $P^I$ of $P$ contains the rule $livesIn(ann, berlin) \leftarrow livesIn(brad, berlin), isMarriedTo(brad, ann)$ and the facts (1), (2). As $I$ is a minimal model of $P^I$, it holds that $I$ is an answer set of $P$. □

Following the common practice in ILP, we consider only safe rules (i.e., variables in the negated part must appear in some positive atoms) with linked variables [14].

**Relational association rule mining**. Association rule mining concerns the discovery of frequent patterns in a data set and the subsequent transformation of these patterns into rules. Association rules in the relational format have been subject of intensive research in ILP (see, e.g., [8] as the seminal work in this direction) and more recently in the KG community (see [12] as the most prominent work). In the following we adapt basic notions in relational association rule mining to our case of interest.

A *conjunctive query* $Q$ over $\mathcal{G}$ is of the form $Q(\boldsymbol{X}) : -p_1(\boldsymbol{X1}), \ldots, p_m(\boldsymbol{X_m})$. Its right-hand side (i.e., body) is a finite set of possibly negated atomic formulas over $\mathcal{G}$, while the left-hand side (i.e., head) is a tuple of variables occurring in the body. The *answer* of $Q$ on $\mathcal{G}$ is the set $Q(\mathcal{G}) := \{f(\boldsymbol{Y}) \mid \boldsymbol{Y}$ is the head of $Q$ and $f$ is a matching of $Q$ on $\mathcal{G}\}$. Following [8], the *(absolute) support* of a conjunctive query $Q$ in a KG $\mathcal{G}$ is the number of distinct tuples in the answer of $Q$ on $\mathcal{G}$. The support of the query

$$Q(X, Y, Z) : -isMarriedTo(X, Y), livesIn(X, Z) \tag{2}$$

over $\mathcal{G}$ in Fig. 1 asking for people, their spouses and living places is equal to 6.

An *association rule* is of the form $Q_1 => Q_2$, such that $Q_1$ and $Q_2$ are both conjunctive queries and the body of $Q_1$ considered as a set of atoms is included in the body of $Q_2$, i.e., $Q_1(\mathcal{G}') \subseteq Q_2(\mathcal{G}')$ for any possible KG $\mathcal{G}'$.

For example, from the above $Q(X, Y, Z)$ and

$$Q'(X, Y, Z) : -isMarriedTo(X, Y),\ livesIn(X, Z), livesIn(Y, Z) \tag{3}$$

we can construct the rule $Q => Q'$.

In this work we exploit association rules for reasoning purposes, and thus (with some abuse of notation) treat them as logical rules, i.e., for $Q_1 => Q_2$ we write $Q_2 \backslash Q_1 \leftarrow Q_1$, where $Q_2 \backslash Q_1$ refers to the set difference between $Q_2$ and $Q_1$ considered as sets. E.g., $Q => Q'$ from above corresponds to *r1* from Sec. 1.

We exploit the rule evaluation measure called *conviction* [3], as it is accepted to be appropriate for estimating the actual implication of the rule at hand, and is thus particularly attractive for our KG completion task. For $r: \ H \leftarrow B, not \ E$, with $H = h(X, Y)$ and $B, E$ involving variables from $\boldsymbol{Z} \supseteq X, Y$, the *conviction* is given by:

$$conv(r, \mathcal{G}) = \frac{1 - supp(h(X, Y), \mathcal{G})}{1 - conf(r, \mathcal{G})} \tag{4}$$

where $supp(h(X, Y), \mathcal{G})$ is the *relative support* of $h(X, Y)$ defined as follows:

$$supp(h(X,Y), \mathcal{G}) = \frac{\#(X,Y) : h(X,Y) \in \mathcal{G}}{(\#X : \exists Y \ h(X,Y) \in \mathcal{G}) * (\#Y : \exists X \ h(X,Y) \in \mathcal{G})} \tag{5}$$

and $conf$ is the confidence of $r$ given as

$$conf(r, \mathcal{G}) = \frac{\#(X, Y) : H \in \mathcal{G}, \exists \boldsymbol{Z} \ B \in \mathcal{G}, E \notin \mathcal{G}}{\#(X, Y) : \exists \boldsymbol{Z} \ B \in \mathcal{G}, E \notin \mathcal{G}} \tag{6}$$

*Example 2.* The conviction of the above rule *r1* is $conv(r1, \mathcal{G}) = \frac{1 - 0.3}{1 - 0.5} = 1.4$ □

# 3 A Theory Revision Framework for Rule-based KG Completion

## 3.1 Problem Statement

We start with defining the goal of this work formally. To this aim, let us introduce the factual representation of a KG $\mathcal{G}$ as the collection of facts 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 [7], we define the gap between the *available graph* $\mathcal{G}^a$ and the *ideal graph* $\mathcal{G}^i$, i.e., the graph containing all correct facts with constants and relations from $\Sigma_{\mathcal{G}^a}$ that hold in the current state of the world.

**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 goal is to learn a set $\mathcal{R}$ of nonmonotonic rules from the available graph, such that their application results in a good approximation of $\mathcal{G}^i$. Here, the application of $\mathcal{R}$ to a graph $\mathcal{G}$ refers to the computation of answer sets of $\mathcal{R} \cup \mathcal{G}$.

**Definition 2 (Rule-based KG completion).** *Let a factual representation of a KG $\mathcal{G}$ be given over the signature $\Sigma_{\mathcal{G}} = \langle \mathbf{C}, \mathbf{R}, \mathcal{C} \rangle$ 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 $\mathcal{G}$ w.r.t. $\mathcal{R}$ *is a graph $\mathcal{G}_{\mathcal{R}}$ constructed from any answer set $\mathcal{G}_{\mathcal{R}} \in AS(\mathcal{R} \cup \mathcal{G})$.*

Note that $\mathcal{G}^i$ is the perfect completion of $\mathcal{G}^a$, containing all correct facts over $\Sigma_{\mathcal{G}^a}$. 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. Intuitively, a good revision $\mathcal{R}_{NM}$ of $\mathcal{R}_H$ is the one that (i) neglects as many incorrect predictions made by $\mathcal{R}_H$ as possible, while still (ii) preserving as many correct predictions made by $\mathcal{R}_H$ as possible. Note that $\mathcal{G}^i$ is usually *not available*, thus we do

not know which predictions are actually correct and which are not. For this reason using standard ILP measures in our setting to evaluate the quality of a ruleset is impractical. To still make an estimate of the revision quality we exploit measures from association rule mining literature. According to our hypothesis, a good ruleset revision is the one for which the overall rule measure is the highest, while the added negated atoms are not over-fitting the data, i.e., the negated atoms are actual exceptions rather than noise.

To this end, 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. In particular, $q_{rm}$ generalizes any rule measure $rm$ to rulesets as follows

$$q_{rm}(\mathcal{R}, \mathcal{G}) = \frac{\sum_{r \in \mathcal{R}} rm(r, \mathcal{G})}{|\mathcal{R}|}. \tag{7}$$

Conversely, $q_{conflict}$ estimates the number of conflicting predictions that the rules in $\mathcal{R}$ generate. To measure $q_{conflict}$ for a given $\mathcal{R}$, we create an extended set of rules $\mathcal{R}^{aux}$, which contains each nonmonotonic 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 $h$ of $r$ with a newly introduced predicate $not\_h$ which intuitively contains instances which are *not* in $h$. Formally,

$$q_{conflict}(\mathcal{R}, \mathcal{G}) = \sum_{p \in pred(\mathcal{R})} \frac{|\{\boldsymbol{c} \mid p(\boldsymbol{c}), not\_p(\boldsymbol{c}) \in \mathcal{G}_{\mathcal{R}^{aux}}\}|}{|\{\boldsymbol{c} \mid not\_p(\boldsymbol{c}) \in \mathcal{G}_{\mathcal{R}^{aux}}\}|}, \tag{8}$$

where $pred(\mathcal{R})$ is the set of predicates appearing in $\mathcal{R}$, and $\boldsymbol{c} \subseteq \mathcal{C}$ with $1 \leq |\boldsymbol{c}| \leq 2$. Note that $q_{conflict}$ is designed to distinguish real exceptions from noise, by considering the cross talk between the rules in a set, as illustrated in the following example.

*Example 3.* The predicate $researcher$ is a good exception for $r1$ w.r.t. $\mathcal{G}$ (Fig. 1) with $bornIn(dave, chicago)$ added, i.e. it explains why for 2 out of 3 substitutions marked with red triangles the rule $r1$ is not satisfied. However, this exception becomes less prominent, whenever $r2: livesIn(X, Y) \leftarrow bornIn(X, Y), not\ emigrant(X)$ is applied to $\mathcal{G}$. Indeed, after $livesIn(dave, chicago)$ is predicted, the substitution $X/clara$, $Y/dave, Z/chicago$ starts satisfying $r1$, but $researcher$ still holds for $dave$, which weakens the predicate $researcher$ as an exception for $r1$. □

We now define our theory revision problem based on the above quality functions.

**Definition 3 (Quality-based Horn theory revision (QHTR)).** *Given a set $\mathcal{R}_H$ of Horn rules over the signature $\Sigma$, 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 over $\Sigma$ 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.*

Prior to tackling the 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*

- $NS(r, \mathcal{G}) = \{\theta \mid head(r)\theta, body(r)\theta \subseteq \mathcal{G}\}$ *is an* $r$*-normal set of substitutions;*
- $ABS(r, \mathcal{G})$=$\{\theta' \mid body(r)\theta' \subseteq \mathcal{G}\ , head(r)\theta' \notin \mathcal{G}\}$ *is an* $r$*-abnormal one, where* $\theta, \theta' : \mathcal{V} \to \mathcal{C}$.

*Example 4.* For $\mathcal{G}$ from Fig. 1 and $r1$ we have $NS(r1, \mathcal{G}) = \{\theta_1, \theta_2, \theta_3\}$, where $\theta_1 = \{X/Brad, Y/Ann, Z/Berlin\}$; similarly, the most right and bottom blue triangles in Fig. 1 refer to $\theta_2$ and $\theta_3$ resp., while the red ones represent $ABS(r1, \mathcal{G})$.

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 5.* For $\mathcal{G}$ in Fig. 1 and $r1$ we have that $EWS(r, \mathcal{G}, Y) = \{researcher\}$. Furthermore, $EWS(r, \mathcal{G}, X) = \{artist\}$. If $brad$ with $ann$ and $john$ with $kate$ lived in cities different from $berlin$ and $chicago$ resp., then $EWS(r, \mathcal{G}, Z) = \{metropolitan\}$.

In general when binary atoms are allowed in the rules, there might be potentially too many possible $EWS$s to construct. For a rule with $n$ distinct variables, $n^2$ candidate $EWS$s might exist. Furthermore, combinations of exception candidates could be an explanation for some missing links, so the search space of solutions to QHTR problem is large. In this work, however, we restrict ourselves only to a single predicate as a final exception, and leave the extensions to arbitrary combinations for future research.

### 3.2 Methodology

Due to the large number of exception candidates to consider, determining the globally best solution to the QHTR problem is not feasible in practice especially given the huge size of KGs. Therefore, we aim at finding an approximately good solution. Intuitively, our approach is to revise rules one by one finding the locally best revision, while considering the predictive impact of other rules in a set. Our methodology for solving the QHTR problem comprises four steps, which we now discuss in details.

**Step 1.** We start with a KG $\mathcal{G}$ and compute frequent conjunctive queries, which are then cast into Horn rules $\mathcal{R}_H$ based on some association rule measure $rm$. For that any state-of-the-art relational association rule learning algorithm can be used. We then compute for each rule $r \in \mathcal{R}_H$ the $r$*-normal* and $r$*-abnormal* substitutions.

**Step 2 and 3.** Then, for every $r \in \mathcal{R}_H$ with $h(X, Y)$ in the head, we determine $EWS(r, \mathcal{G}, X)$, $EWS(r, \mathcal{G}, Y)$ and $EWS(r, \mathcal{G}, \langle X, Y \rangle)$. The algorithm for computing $EWS$*s* is an extended version of the one reported in [11]. Here, we first construct $E^+ = \{not\_h(c, d), \text{s.t.}\, \theta = \{X/c, Y/d, \ldots\}$ is in $ABS(r, \mathcal{G})\}$ and $E^- = \{not\_h(e, f), \text{s.t.}$

$\theta' = \{X/e, Y/f, \dots\}$ is in $NS(r, \mathcal{G})\}$. A classical ILP procedure $learn(E^+, E^-, \mathcal{G})$ (*e.g.*, based on [28]) is then invoked, which searches for hypothesis with $not\_h(X, Y)$ in the head and a single body atom of the form $p(X)$, $p'(Y)$ or $p''(X, Y)$, where $p, p', p''$ are predicates from $\Sigma_{\mathcal{G}}$. The target hypothesis should not cover any examples in $E^-$, while covering at least some examples in $E^+$. From the bodies of the obtained hypothesis the predicates for $EWS$ sets are extracted.

Then, for every $r \in \mathcal{R}_H$ we create potential revisions by adding to $r$ a single negated atom from $EWS$ sets at a time. Overall for each rule this way we obtain $|EWS(r, \mathcal{G}, X)| + |EWS(r, \mathcal{G}, Y)| + |EWS(r, \mathcal{G}, \langle X, Y\rangle)|$ candidate revisions.

**Steps 4.** After all potential 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 both $\mathcal{G}$ and $EWS$s. 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.

The **Naive (N)** ranker is the simplest function, which prefers the revision $r_i^j$ with the highest value of $rm(r_i^j, \mathcal{G})$ among all revisions of $r_i$. This selection function produces a globally best revision with respect to (i) in Def. 3. However, it completely ignores (ii), and thus might return rules with overly noisy exceptions.

The **PM** ranker prefers $r_i^j$ with the highest value of

$$score(r_i^j, \mathcal{G}) = \frac{rm(r_i^j, \mathcal{G}_{\mathcal{R}'}) + rm(r_i^{j\,aux}, \mathcal{G}_{\mathcal{R}'})}{2} \tag{9}$$

where $\mathcal{R}'$ is the set of rules $r'_l \in \mathcal{R}_H \backslash r_i$ with candidate exceptions from all $EWS$s for $r_l$ 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 not making these predictions, and thus we can rely on them when revising $r_i$.

The **OPM** ranker 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 rule measure, which is not necessarily the same as $rm$.

## 4 Evaluation

Our revision approach aims at (1) enhancing the quality of a given ruleset w.r.t. conviction, and consequently (2) improving the accuracy of its predictions. Ideally, the set difference between $\mathcal{G}_{\mathcal{R}_{NM}}$ and $\mathcal{G}^i$ should be minimized (see Fig. 2 for illustration).

**Dataset**. An automatic evaluation of the prediction quality requires an ideal graph $\mathcal{G}^i$ which is known to be complete as a ground truth. However, obtaining a real life complete KG is not possible. Therefore, we used the existing KG as an approximation of $\mathcal{G}^i$ ($\mathcal{G}^i_{appr}$), and constructed the available graph $\mathcal{G}^a$ by removing from $\mathcal{G}^i_{appr}$ 20% of

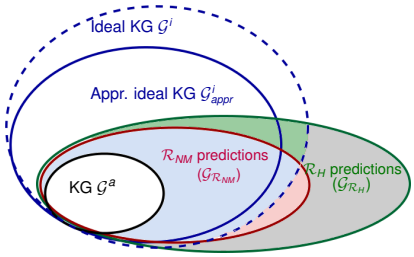


Fig. 2: Relations between the ideal, approximated and available slices of a KG.

| ***topk*** | **YAGO** | | | | **IMDB** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\mathcal{R}_H$ | $\mathcal{R}_N$ | $\mathcal{R}_{PM}$ | $\mathcal{R}_{OPM}$ | $\mathcal{R}_H$ | $\mathcal{R}_N$ | $\mathcal{R}_{PM}$ | $\mathcal{R}_{OPM}$ |
| 5 | 1.3784 | 1.3821 | 1.3821 | 1.3821 | 2.2670 | 2.3014 | 2.3008 | 2.3014 |
| 30 | 1.1207 | 1.1253 | 1.1236 | 1.1237 | 1.5453 | 1.5644 | 1.5543 | 1.5640 |
| 50 | 1.0884 | 1.0923 | 1.0909 | 1.0913 | 1.3571 | 1.3749 | 1.3666 | 1.3746 |
| 60 | 1.0797 | 1.0837 | 1.0823 | 1.0829 | 1.3063 | 1.3221 | 1.3143 | 1.3219 |
| 70 | 1.0714 | 1.0755 | 1.0736 | 1.0744 | 1.2675 | 1.2817 | 1.2746 | 1.2814 |
| 80 | 1.0685 | 1.0731 | 1.0710 | 1.0720 | 1.2368 | 1.2499 | 1.2431 | 1.2497 |
| 100 | 1.0618 | 1.0668 | 1.0648 | 1.0659 | 1.3074 | 1.4100 | 1.3987 | 1.4098 |

Table 1: The average conviction for the *top-k* Horn rules and their revisions.

the facts for each binary predicate. As a side constraint, we ensure that every node in $\mathcal{G}^a$ is connected to at least one other node. We constructed two datasets for evaluating our approach: (i) YAGO3 [23], as a general purpose KG, with more than 1.8M entities, 38 relations, and 20.7M facts, and (ii) a domain-specific KG extracted from the IMDB [1] dataset with 112K entities, 38 relations, and 583K facts[2].

**Setup**. We have implemented our approach in a system prototype [3], and conducted experiments on a multi-core Linux server with 40 cores and 400GB RAM. We start with mining Horn rules of the form $h(X, Z) \leftarrow p(X, Y), q(Y, Z)$ from $\mathcal{G}^a$ and ranking them w.r.t. their *absolute support*. Then, we revise the rules as described in Sec. 3.2, taking *conviction* as the $rm$ measure. For every rule we rank the constructed revisions and pick the one with the highest score as the final result. This process is repeated for the proposed ranking methods, i.e., *Naive*, *Partial Materialization*, and *Ordered Partial Materialization* resulting in the rulesets $\mathcal{R}_N$, $\mathcal{R}_{PM}$, and $\mathcal{R}_{OPM}$ respectively.

**Ruleset quality**. In Tab. 1, we report the *average conviction* for the top-$k$ ($k$=5,...100) Horn rules $\mathcal{R}_H$ and their revisions for YAGO and IMDB. The results show that the revision process consistently enhances the avg. ruleset conviction. Moreover, while the conviction per ruleset naturally decreases with addition of lower quality rules, improvement ratios are increasing with the best enhancement (7.6%) for IMDB top-100 rules.

**Prediction quality**. To evaluate the quality of ruleset predictions, we sampled a set of 5 Horn rules $\mathcal{R}_H$ from the top-50 Horn rules both for IMDB and YAGO and compared

[1] `http://imdb.com`
[2] `http://people.mpi-inf.mpg.de/~gadelrab/downloads/ILP2016`
[3] `https://github.com/htran010589/nonmonotonic-rule-mining`

| predicate | predictions | | | | outside $\mathcal{G}^i_{appr}$ | | | | corr. removed,% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{R}_H$ | $\mathcal{R}_N$ | $\mathcal{R}_{PM}$ | $\mathcal{R}_{OPM}$ | $\mathcal{R}_H$ | $\mathcal{R}_N$ | $\mathcal{R}_{PM}$ | $\mathcal{R}_{OPM}$ | $\mathcal{R}_N$ | $\mathcal{R}_{PM}$ | $\mathcal{R}_{OPM}$ |
| $I$:$actedIn$ | 1231 | 1214 | 1230 | 1214 | 1148 | 1131 | 1147 | 1131 | 90 | 100 | 90 |
| $I$:$genre$ | 629 | 609 | 618 | 609 | 493 | 477 | 482 | 477 | 50 | 20 | 50 |
| $I$:$hasLang$ | 173 | 102 | 125 | 102 | 163 | 92 | 115 | 92 | 60 | 100 | 60 |
| $I$:$prodIn$ | 2489 | 2256 | 2327 | 2327 | 2488 | 2255 | 2326 | 2326 | 10 | 10 | 30 |
| | | | | | | | | | 52.50 | 45.16 | **57.75** |
| $Y$:$direct$ | 41079 | 39174 | 39174 | 39174 | 41021 | 39116 | 39116 | 39116 | 100 | 100 | 100 |
| $Y$:$grFrom$ | 3519 | 3456 | 3456 | 3456 | 3363 | 3300 | 3300 | 3300 | 100 | 100 | 70 |
| $Y$:$citizOf$ | 3407 | 2883 | 2883 | 2883 | 3360 | 2836 | 2836 | 2836 | 50 | 50 | 70 |
| $Y$:$bornIn$ | 110283 | 108317 | 109846 | 108317 | 109572 | 107607 | 109137 | 107607 | 90 | 90 | 100 |
| | | | | | | | | | 85 | 85 | 85 |

Table 2: Predictions of sampled rules and their revisions for IMDB (*I*) and YAGO (*Y*).

them against their revisions w.r.t. the predictive power. For that, we run DLV [20] with these rulesets and the facts in $\mathcal{G}^a$ and obtained resp. $\mathcal{G}_{\mathcal{R}_H}$, $\mathcal{G}_{\mathcal{R}_N}$, $\mathcal{G}_{\mathcal{R}_{PM}}$ and $\mathcal{G}_{\mathcal{R}_{OPM}}$. Tab. 2 reports for each head predicate appearing in the sampled rules the number of newly predicted facts, i.e. those not in $\mathcal{G}^a$ (second column) and the portion of predictions among them that are outside $\mathcal{G}^i_{appr}$ (third column).

First, observe that naturally relatively few predictions can be found in $\mathcal{G}^i_{appr}$ ($\approx$9% for IMDB and $\approx$2% for YAGO). This is expected as the latter graph is highly incomplete. Second, it is important to note that $\mathcal{R}_H$ and the revised rulesets produced roughly the same number of correct predictions within $\mathcal{G}^i_{appr}$. E.g., for YAGO we have $\mathcal{G}_{\mathcal{R}_H} \backslash \mathcal{G}_{\mathcal{R}_{PM}} \cap \mathcal{G}^i_{appr} = \emptyset$, meaning that the green area within the approximation of the ideal graph in Fig. 2 is empty, which shows that incorporated exceptions did not spoil the positive rules with respect to correct predictions in $\mathcal{G}^i_{appr}$.

To make the comparison between $\mathcal{R}_H$ and the revised rulesets fair, we need to ensure that $\mathcal{R}_H$ on its own is not completely inaccurate. Indeed, if $\mathcal{R}_H$ makes only false predictions, then adding even irrelevant exceptions will reduce the number of incorrect instances, thus, improving the ruleset predictive quality. The number of $\mathcal{R}_H$ predictions outside $\mathcal{G}^i_{appr}$ is large, and we do not know the ground truth for these predictions. Therefore, we had to verify these facts manually using web resources. Obviously such verification for all of the predictions is not feasible. Hence, we restricted ourselves to a uniform random sample of 20 predicted facts per head predicate in $\mathcal{R}_H$. Among the IMDB samples, the precision of 70% has been achieved, while for YAGO we have obtained precision of 30%. This shows that the rules in $\mathcal{R}_H$ are not completely erroneous.

To assess the impact of the revision methods, we also had to select a uniform sample due to the large size of the differences between $\mathcal{G}_{\mathcal{R}_H}$ and the graphs obtained by applying revised rulesets. More specifically, we have randomly sampled 10 predictions per head predicate from $\mathcal{G}_{\mathcal{R}_H} \backslash \mathcal{G}_{\mathcal{R}_N}$, $\mathcal{G}_{\mathcal{R}_H} \backslash \mathcal{G}_{\mathcal{R}_{PM}}$ and $\mathcal{G}_{\mathcal{R}_H} \backslash \mathcal{G}_{\mathcal{R}_{OPM}}$ resp. The 4th column in Tab. 2 reports the percentage of erroneous predictions among the sampled facts in the difference for each revision method (referred to as correctly removed), i.e., gray area in Fig. 2. For IMDB $\mathcal{R}_{OPM}$ achieved the best improvement. For YAGO, all of the revision methods performed equally well. Moreover, the effect of YAGO revisions is more visible, since $\mathcal{R}_H$ for YAGO is of a lower quality than for IMDB as reported earlier.

$r_1 : writtenBy(X, Z) \leftarrow hasPredecessor(X, Y), writtenBy(Y, Z),$ **not** $american\_film(X)$
$r_2 : actedIn(X, Z) \leftarrow isMarriedTo(X, Y), directed(Y, Z),$ **not** $silent\_film\_actor(X)$
$r_3 : isPoliticianOf(X, Z) \leftarrow hasChild(X, Y), isPoliticianOf(Y, Z),$**not** $vicepresidentOfMexico(X)$

Fig. 3: Examples of the revised rules

**Running times**. Our main goal was to evaluate the predictive quality of computed rules rather then the running times of the implemented algorithms. Therefore, the latter are only briefly reported. For the *top-100* Horn YAGO and IMDB rules mined from $\mathcal{G}^a$, $EWS$s with an average of 1.6K and 10.9K exception candidates per rule were computed within 7 and 68 seconds resp. As regards IMDB, the revisions $\mathcal{R}_N, \mathcal{R}_{PM}$, and $\mathcal{R}_{OPM}$ were determined in 9, 62, and 24 seconds resp., while for YAGO, they required 45, 177, and 112 seconds. Besides, the predictions of each of the rulesets on $\mathcal{G}^a$ were found via DLV, on average, within 8 seconds for IMDB and 310 seconds for YAGO.

**Example rules**. Fig. 3 shows examples of our revised rules, e.g., $r_1$ extracted from IMDB states that movie plot writers stay the same throughout the sequel unless a movie is American, and $r_3$ learned from YAGO says that ancestors of politicians are also politicians in the same country with the exception of Mexican vice-presidents.

## 5 Related Work

Approaches for link prediction are divided into statistics-based (see [24] for overview), and logic-based (e.g., [12]), which are the closest to our work. The latter basically extend and adapt previous work in ILP on relational association rule mining. However, algorithms such as [12] mine only Horn rules, rather than nonmonotonic as we do.

In the association rule mining community, some works studied (interesting) exception rules (e.g. [27]), i.e., rules with low support and high confidence. Our work differs as we do not necessarily look for rare rules, but care about their predictive power.

In the context of inductive and abductive logic [10], learning nonmonotonic rules from complete datasets was considered in several works ([26,25,6,18]) These methods rely on CWA and focus on describing a dataset at hand exploiting negative examples, which are explicitly given unlike in our setting. Learning nonmonotonic rules in presence of incompleteness was studied in hybrid settings in [16] and [21] respectively. There a background theory or a hypothesis can be represented as a combination of a DL ontology and Horn or nonmonotonic rules. While the focus of these works is on the complex interaction between reasoning components, we are more concerned with techniques for deriving rules with high predictive quality from huge KGs.

## 6 Conclusions and Future Work

We have presented an approach for mining relational nonmonotonic rules from KGs under OWA by casting this problem into a theory revision task and exploiting association rule mining methods to cope with the huge size of KGs. The approach extends our previous work [11], where this problem was studied for KGs with only unary predicates.

Further extensions to more complex combinations of exceptions as well as more general types of rules (e.g., with existentials in the head) are a natural future direction. Moreover enhancing our framework by partial completeness assumptions for certain (combinations of) predicates/constants is another orthogonal but interesting research stream. On the practical side, we plan to develop advanced evaluation strategies, which is very challenging due to the absence of the ideal graph and the large KG size.

**Acknowledgements**. We thank anonymous reviewers for their insightful suggestions and Jacopo Urbani for his helpful comments on an earlier version of this paper.

## 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: ISWC. pp. 722–735 (2007)
2. Azevedo, P.J., Jorge, A.M.: Comparing Rule Measures for Predictive Association Rules. In: 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, 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: AAAI (2010)
5. Chen, Y., Goldberg, S., Wang, D.Z., Johri, S.S.: Ontological Pathfinding: Mining First-Order Knowledge from Large Knowledge Bases. In: SIGMOD (2016)
6. Corapi, D., Russo, A., Lupu, E.: Inductive logic programming as abductive search. In: ICLP. pp. 54–63 (2010)
7. Darari, F., Nutt, W., Pirrò, G., Razniewski, S.: Completeness Statements about RDF Data Sources and Their Use for Query Answering. In: ISWC. pp. 66–83 (2013)
8. Dehaspe, L., De Raedt, L.: Mining Association Rules in Multiple Relations, In: ILP, pp. 125–132 (1997)
9. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandecic, D.: Introducing Wikidata to the Linked Data Web. In: ISWC. pp. 50–65 (2014)
10. Flach, P.A., Kakas, A.: Abduction and Induction: essays on their relation and integration, vol. 18. Applied Logic Series (2000)
11. Gad-Elrab, M. H., Stepanova, D. Urbani, J., Weikum, G.: Exception-enriched rule learning from knowledge graphs. In: Proc. of the ISWC, pp. 234–251 (2016)
12. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.M.: Fast Rule Mining in Ontological Knowledge Bases with AMIE+. In: VLDB J. (2015)
13. Gelfond, M., Lifschitz, V.: The stable model semantics for logic programming. In: ICLP, pp. 1070–1080 (1988)
14. Helft, N.: Induction as nonmonotonic inference. In: KR, pp. 149–156 (1989)
15. Inoue, K. and Kudoh, Y.:Learning Extended Logic Programs, In: IJCAI, pp. 176–181 (1997)
16. Józefowska, J., Lawrynowicz, A., Lukaszewski, T.: The role of semantics in mining frequent patterns from knowledge bases in description logics with rules. TPLP 10(3), 251–289 (2010)
17. Lassila, O., Swick, R.R.: Resource description framework (RDF) model and syntax specification (1999)
18. Law, M., Russo, A., Broda, K.: The ILASP system for learning answer set programs (2015)
19. Lehmann, J., Auer, S., Bühmann, L., Tramp, S.: Class expression learning for ontology engineering. J. Web Sem. 9(1), 71–81 (2011)
20. Leone, N., Pfeifer, G., Faber, W., Eiter, T., Gottlob, G., Perri, S., Scarcello, F.: The dlv system for knowledge representation and reasoning. ACM TOCL, 7(3), 499–562 (2006)
21. Lisi, F.A.: Inductive Logic Programming in Databases: From Datalog to DL+log. TPLP 10(3), 331–359 (2010)
22. Lloyd, J.W.: Foundations of Logic Programming, 2nd Edition. Springer (1987)
23. Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: A knowledge base from multilingual wikipedias. In: Proc. of CIDR (2015)
24. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A review of relational machine learning for knowledge graphs. Proc. of the IEEE 104(1), 11–33 (2016)
25. Ray, O.: Nonmonotonic abductive inductive learning. J. Appl. Log. 3(7), 329–340 (2008)
26. Sakama, C.: Induction from answer sets in nonmonotonic logic programs. ACM Trans. Comput. Log. 6(2), 203–231 (2005)
27. Taniar, D., Rahayu, W., Lee, V., Daly, O.: Exception rules in association rule mining. Appl. Math. and Comp. 205(2), 735–750 (2008)
28. Quinlan, J. R.: Learning logical definitions from relations. Machine Learn., 5:239-266 (1990)
29. Wrobel, S.: First Order Theory Refinement, In ILP, pp. 14–33 (1996)