



Combining Truth Discovery and RDF Knowledge Bases to Their Mutual Advantage

Valentina Beretta¹(✉), Sébastien Harispe¹, Sylvie Ranwez¹,
and Isabelle Mougnot²

¹ LIGI2P, IMT Mines Ales, Univ Montpellier, Ales, France
valentina.beretta@mines-ales.fr

² UMR 228 Espace Dev UM, Maison de la Télédétection, Montpellier, France

Abstract. This study exploits knowledge expressed in RDF Knowledge Bases (KBs) to enhance Truth Discovery (TD) performances. TD aims to identify *facts* (true claims) when conflicting claims are made by several sources. Based on the assumption that true claims are provided by reliable sources and reliable sources provide true claims, TD models iteratively compute value confidence and source trustworthiness in order to determine which claims are true. We propose a model that exploits the knowledge extracted from an existing RDF KB in the form of rules. These rules are used to quantify the evidence given by the RDF KB to support a claim. This evidence is then integrated into the computation of the confidence value to improve its estimation. Enhancing TD models efficiently obtains a larger set of reliable facts that *vice versa* can populate RDF KBs. Empirical experiments on real-world datasets showed the potential of the proposed approach, which led to an improvement of up to 18% compared to the model we modified.

Keywords: Truth discovery · RDF KBs · Rule mining
Source trustworthiness · Value confidence

1 Introduction

Several popular initiatives, such as DBpedia [2], Yago [17] and Google Knowledge Vault [5], automatically populate Knowledge Bases (KBs) with Web data. The performance of this Knowledge Base Population (KBP) process is critical to ensuring the quality of the KB. In particular, it requires dealing with complex cases in which several conflicting data are extracted from different sources, e.g. different automatic extractors will provide different birth places for Pablo Picasso. Approaches based on voting or naive strategies that only consider the most frequently provided data value are *de facto* limited. Such approaches are unable to deal with spam-based attacks or duplicated errors, which are common on the Web. Dealing with this problem therefore requires distinguishing values

according to their sources. In this study, we propose an approach that serves KBP integrating potentially conflicting data provided by multiple sources; it relies on a general framework that can be used to address conflict resolution problems by exploiting prior knowledge defined in existing KBs.

Several techniques based on Knowledge Fusion have been proposed in order to automatically obtain reliable information. Most of them suppose that information veracity strictly depends on source reliability. Intuitively, the more reliable a source is, the more reliable the information it provides. In turn they also assume that source reliability depends on information veracity, i.e. reliable information is provided by reliable sources. Truth Discovery (TD) methods are unsupervised approaches based on these assumptions aimed at identifying the most reliable of a set of conflicting triples – for *functional* predicates, i.e. when there is a single true value for a property of a real-world entity. This study aims to enhance the TD framework using knowledge extracted from an existing RDF KB to obtain a larger set of correct *facts* that could be used to populate RDF KBs. More precisely, it makes the following contributions:

- A novel approach that can be used to enrich traditional TD models by incorporating additional information given by recurrent patterns extracted from a KB. A state-of-the-art rule mining system is used to extract rules that represent these patterns. A method is proposed for selecting the most useful rules to be used to evaluate veracity of triples. Moreover, since each rule contributes to TD performances according to its quality, a function that aggregates the existing rule quality metrics is also defined. High-quality rules will have a higher weight than low-quality rules;
- An extensive evaluation of the proposed approach; interestingly, it shows that the TD framework can benefit from information derived by rules. As a consequence, we point out how the creation of high quality RDF KBs may benefit from the use of highly reliable TD models. The datasets and source code proposed in this study are open-source and freely accessible online.¹

The paper is organized as follows. Section 2 presents an overview of the TD framework and how it can be applied in the RDF KB context. It also describes the state-of-the-art rule mining techniques that are used in our work to detect interesting recurrent patterns. Section 3 explains how additional information extracted from KBs is integrated into the TD framework. The proposed approach is evaluated and discussed in Sect. 4. Finally, Sect. 5 reports the main findings and discusses perspectives.

2 Related Work and Preliminaries

In this section we introduce the formal aspect of TD, its goal and the key elements required to achieve it. We then formally present rules and their quality metrics. We will then be able to use them to exploit identified recurrent patterns to increase confidence in certain triples.

¹ <https://github.com/lgi2p/TDwithRULES>.

In this study, we assume that sources provide their claims in the form of RDF triples $\langle \text{subject}, \text{predicate}, \text{object} \rangle \in I \times I \times (I \cup L)$ where I is the set of Internationalized Resource Identifiers (IRIs) and L the set of literals.

The following definition introduces all TD components (source, data items and values). Since the TD and Linked Data (LD) fields use different notations, this definition aims at clarifying the correspondence between terms belonging to each field.

Definition 1 (Truth Discovery). *Let $D \subseteq I \times I$ be a set of data items where each $d \in D$ is a pair $(\text{subj}, \text{pred})$ that refers to a functional property ($\text{pred} \in I$) of an entity ($\text{subj} \in I$). Let $V \subseteq I \cup L$ be a set of values that can be assigned to these data items and S be the set of sources. Each source $s \in S$ can associate a value $v \in V$ (corresponding to $\text{obj} \in I \cup L$) to a data item $d \in D$, hence providing a claim v_d that corresponds to the RDF triple $\langle \text{subj}, \text{pred}, \text{obj} \rangle$. **Truth Discovery** associates a value confidence to each claim and a trustworthiness score to each source. It then iteratively estimates these quantities to identify the true value v_d^* for each data item.*

Several TD approaches have been proposed, as detailed in recent surveys [4, 10]. The models differ from one another in the way they compute the value confidence of claims and the trustworthiness of sources. Some of them use no additional information, while others attempt to improve TD performances using external support such as extractor information (i.e. the confidence associated with extracted triples), the temporal dimension, hardness of facts, common sense reasoning or correlations. Models that take correlations into account can be divided according to the kinds of correlations they consider: source correlations, value correlations or data item correlations. To the best of our knowledge, no existing work takes advantage of data item correlations in the form of recurrent patterns to improve TD results. The idea is that the confidence of a certain claim can increase when recurrent patterns occur which are associated with the considered data item. This kind of correlation can be used to enhance existing TD models. In this study, a rule mining procedure is used to identify patterns in data. We specify the major aspects of the rule mining below.

2.1 Recurrent Pattern Detection from RDF KBs

Several techniques can be used to identify regularities in data. For instance, link mining models are often used for that purpose in knowledge base completion [13]. In this study we prefer to use rule mining techniques because they are easily interpretable [1]. Rules generalize patterns in order to identify useful suggestions that can be used to generate new data or correct existing data [6]. We therefore propose to exploit these suggestions in order to solve conflicts among triples provided by different sources. Given our problem setting, where rules are used to reinforce the confidence of a claim, we are particularly interested in Horn rules. Considering Datalog-style, a Horn rule $r : B_1 \wedge B_2 \wedge \dots \wedge B_n \rightarrow H$, i.e. $r : \widehat{B} \rightarrow H$, is an implication from a conjunction of atoms called the body to a single atom

called the head [12]. An atom is usually denoted $pred(subj, obj)$, where $subj$ and obj can be variables or constants. Considering that an instantiation of an atom is a substitution of its variables with IRIs, an atom a holds under an instantiation σ in a KB K if $\sigma(a) \in K$. Moreover, a body \widehat{B} holds under σ in K , if each atom in \widehat{B} holds [7]. Note that in our setting each instantiated atom $pred(subj, obj)$ can also be represented as an RDF triple $\langle subj, pred, obj \rangle$.

Rule extractors rely on the Closed World Assumption (CWA). This means that when a fact is not known (does not belong to the KB) it is considered to be false. This assumption is more often appropriate when KBs are complete. On the contrary, RDF KBs are based on the Open World Assumption (OWA). When dealing with incomplete information the OWA is preferable. If information is missing we need to distinguish between false and unknown information. A triple that does not appear in the KB is not systematically false. In this context, methods have recently been proposed that mine rules from RDF KBs such as DBpedia or Yago, taking the OWA into account [15]. An example of a rule mining system that considers the OWA is AMIE [8]. It is based on the Partial Completeness Assumption (PCA): if a KB contains some object values for a given pair (*subject*, *predicate*), it is assumed that all object values associated with it are known. This assumption can generate counter-examples, required for rule mining models, but do not appear in RDF KBs, which often contain only positive facts. Alternative assumptions and metrics have been proposed to extract rules under the OWA [9, 13, 18]. In this study, we use AMIE because it is a state-of-the-art system and its source code is freely available online.

2.2 Rule Quality Metrics

Any rule, independently of the system used to extract it, can be evaluated by several quality metrics; among them the most well-recognized measures are *support* and *confidence* [1, 11, 19]. Support represents the frequency of a rule in a KB, while confidence is the percentage of instantiations of a rule in the KB, compared to the instantiations of its body. Based on the formal definition given in [8], for the sake of coherence and clarity, we present how these metrics are computed below. In the rest of the paper we do not make a comparison of the different quality metrics because it is out of the scope of this study. The primary aim here is to evaluate the potential of integrating knowledge extracted from an RDF KB into a TD process. However, since we are aware that robust metrics could have an impact on TD results, we plan to study such a comparison in future studies.

Considering a Horn rule $r : \widehat{B} \rightarrow H$ where H is composed of a single atom $p(x, y)$, its support is defined by:

$$supp(\widehat{B} \rightarrow p(x, y)) := \#(x, y) : \exists z_1, \dots, z_n : \widehat{B} \wedge p(x, y) \quad (1)$$

where z_1, \dots, z_n are the variables contained in the atoms of the rule body \widehat{B} apart from x and y , and $\#(x, y)$ is the number of different pairs x and y .

Its confidence is computed using the following formula:

$$conf(\widehat{B} \rightarrow p(x, y)) := \frac{supp(\widehat{B} \rightarrow p(x, y))}{\#(x, y) : \exists z_1, \dots, z_n : \widehat{B}} \tag{2}$$

This formula was introduced to evaluate the quality of rules using the CWA. It is too restrictive when dealing with the OWA. For this reason Galarraga et al. defined a new confidence, called $conf_{PCA}$ [8]. It makes a distinction between false and unknown facts based on PCA. In this setting, if a predicate related to a particular subject, never appears in the KB, then it can neither be considered as true nor false. This new confidence based on PCA is evaluated as follows:

$$conf_{PCA}(\widehat{B} \rightarrow p(x, y)) := \frac{supp(\widehat{B} \rightarrow p(x, y))}{\sum_j supp(\widehat{B} \rightarrow p(x, j))} \tag{3}$$

where j 's are all instantiations of the object variable related to predicate p and having subject x . Using PCA, $conf_{PCA}$ normalizes the support by the set of true and false facts that does not include the unknown ones.

In the next section, we describe how these quality measures are combined into a single measure. Having a more robust metric is important because it is the quality of each rule that will determine its contribution to the computation of the overall evidence that supports a certain claim.

3 Incorporating Rules into the Truth Discovery Framework

This section presents how extracted rules are integrated into truth discovery models. To that end, we define the concepts of *eligible* and *approving* rules, which will be used to identify the most useful rules that need to be taken into account when evaluating the confidence of a claim. Then we describe how information associated with these rules is quantified to further introduce the new confidence estimation formulas used by our TD framework.

3.1 Eligible and Approving Rules

It may not be useful to consider the entire set of extracted rules (denoted R) in order to improve value confidence. For instance, some rules could have a body that is not related to a given data item. Therefore, given a claim $\langle d, v \rangle$, i.e. v_d , where $d = (subj, pred)$, only *eligible* rules are used as potential evidence to improve its confidence estimation. They are defined in the following way.

Definition 2 (Eligible Rule). *Given a KB K , a set of rules $R = \{r : \widehat{B} \rightarrow H\}$ extracted from K where $H = p(x, y)$ and a claim $\langle d, v \rangle$ where $d = (subj, pred)$, a rule $r \in R$ is an **eligible rule** when its body holds, i.e. all of its body atoms appear in K when all rule variables are instantiated w.r.t. the data item subject. Moreover, its head predicate has to correspond to the one in the claim under examination, i.e. $(\sigma(\widehat{B}) \in K) \wedge (H = pred(subj, y))$.*

In our context, the eligibility of a rule depends on the subject and the predicate that compose a data item d . Thus, all claims related to the same data item $d = (subj, pred)$ have the same set of eligible rules, denoted $R_d = \{r \in R \mid (\sigma(\widehat{B}) \in K) \wedge (H = pred(subj, y))\}$.

Once eligible rules for a claim v_d have been collected, the proposed approach checks how many of these rules endorse (approve) v_d , i.e. how many rules support v_d .

Definition 3 (Approving Rule). *Given a KB K , a set of eligible rules $R_d = \{r : \widehat{B} \rightarrow H\}$ where $H = pred(subj, y)$ and a claim $\langle d, v \rangle$ where $d = (subj, pred)$, a rule $r \in R_d$ is an **approving rule** when the value predicted by r corresponds to the claimed value v , i.e. $(\sigma(\widehat{B}) \in K) \wedge (H = pred(subj, v))$.*

The set of approving rules for v_d is represented by $R_d^v \subseteq R_d$ where d indicates that the rules are eligible for a certain data item d and v indicates that the rules predict/support value v . Formally, we obtain $R_d^v = \{r \in R_d \mid (\sigma(\widehat{B}) \in K) \wedge (H = pred(subj, v))\}$.

Example. Given a KB K , reported in Table 1, and the rules:

- $r_1 : speaks(x, z) \wedge officialLang(y, z) \rightarrow bornIn(x, y)$
- $r_2 : residentIn(x, w) \wedge cityOf(w, y) \rightarrow bornIn(x, y)$

Given the following claims about the birth location of some painters $\langle Picasso, bornIn, Spain \rangle$, $\langle Picasso, bornIn, Málaga \rangle$ and $\langle Monet, bornIn, France \rangle$, the set of eligible rules for data item $d_A = (Picasso, bornIn)$ is $R_{d_A} = \{r_1, r_2\}$. The predicate in the head corresponds to the predicate in the claim and when all occurrences of variable x are replaced by *Picasso* in r_1 's and r_2 's body, they are both verified. However, when $d_B = (Monet, bornIn)$ the set of eligible rules is $R_{d_B} = \{r_2\}$ because, even though the head and claim predicate are the same using both rules, if the x variable is substituted by *Monet* the body of r_1 is not verified.

The set of approving rules for the first, second and third claims are respectively $R_{d_A}^{Spain} = \{r_1\}$, $R_{d_A}^{Málaga} = \emptyset$ and $R_{d_B}^{France} = \{r_2\}$.

Before explaining how additional information related to approving and eligible rules is quantified and then incorporated into the TD framework, we describe a function used to integrate the two quality aspects we are interested in, for each rule. This enables better weighting of each rule's contribution during the evaluation of a claim.

Table 1. Illustrative set of triples.

<i>predicate</i>	<i>subject</i>	<i>object</i>	<i>predicate</i>	<i>subject</i>	<i>object</i>
officialLang	(Spain,	Spanish)	residentIn	(Picasso,	Paris)
speaks	(Picasso,	Spanish)	cityOf	(Paris,	France)
residentIn	(Monet,	Vétheuil)	cityOf	(Vétheuil,	France)

3.2 Combining Rule Quality Measures

Support and $conf_{PCA}$ represent different aspects of a rule, see Sect. 2.2. We propose an aggregate function to combine them into a single quality metric since, in our context, it is important to take both aspects into account. It may happen that two rules r_1 and r_2 have the same confidence, but different supports. For instance, if $conf_{PCA}(r_1) = conf_{PCA}(r_2) = 0.8$, $supp(r_1) = 5$ and $supp(r_2) = 500$, then r_2 deserves a higher level of *credibility* than r_1 since r_2 has been observed more often than r_1 .

To address this issue, a function $score : R \rightarrow [0, 1]$ is defined. It is based on Empirical Bayes (EB) methods [16]. EB adjusts estimations resulting from a limited number of examples that may happen by chance. Estimations are modified in function of available examples and prior expectations. When many examples are available, estimation adjustments are small. On the contrary, when there are only few examples, the adjustments are greater. They are corrected w.r.t. the average value that is expected by *a priori* knowledge. Given a family of the prior distribution of available data, EB is able to directly estimate its hyper parameters from the data. Then, it updates the prior belief with new evidence. In other words, the estimation that can be computed from the new examples is modulated w.r.t. prior expectation. The new estimation corresponds to the expected value of a random variable following the updated distribution. In our case, a more robust $conf_{PCA}$, i.e. the proportion of positive examples among all examples considered, needs to be estimated. The prior expectation on our data can be modelled using a *Beta* distribution that is characterized by parameters α and β . Once the model has estimated them, it uses this distribution as prior to modulate each individual estimate. This estimation will be equal to the expected value of the updated distribution $Beta(\alpha + X, \beta + (N - X))$, where X is the number of new positive examples and N is the total number of new examples. The new expected value is $(\alpha + X)/(\alpha + \beta + N)$. This value is returned by the aggregation function. In summary, given the hyper parameters α_S and β_S , the value returned by $score$ for a rule $r : \hat{B} \rightarrow p(x, y)$ is computed as follows:

$$score(r) = \frac{\alpha_S + supp(r)}{\alpha_S + \beta_S + \sum_j supp(\hat{B} \rightarrow p(x, j))} \tag{4}$$

where $supp(r)$ is the support of r and $\sum_j supp(\hat{B} \rightarrow p(x, j))$ is the number of triples containing data item (x, p) . The returned score appears to be similar to $conf_{PCA}$, but it takes the cardinality of the examples into account.

Once this score has been estimated for each rule, the proposed approach sums up all this new information and integrates it into the value confidence estimation formula.

3.3 Assessing a Rule’s Viewpoint on Claim Confidence

All the evidence provided by rules for a claim v_d is summarized in a *boosting factor* that can be seen as the confidence that is assigned by these rules to

v_d . More precisely, it represents the proportion of eligible rules that confirm a given claim v_d . In other words it evaluates the percentage of approving rules out of the entire set of eligible rules, i.e. $|R_d^v|/|R_d|$. It is returned by a function $boost : D \times V \rightarrow [0, 1]$. As anticipated, the proposed model weights each rule differently w.r.t. its quality *score*. The higher the *score* of a rule, the stronger its impact should be on computing the *boosting factor*. Intuitively, given a claim v_d where $d = (subj, pred)$ and a set of rules R extracted from a KB K , the proposed model evaluates the *boosting factor* in the following way:

$$boost(d, v_d) \approx \frac{\sum_{r \in R_d^v} score(r)}{\sum_{r \in R_d} score(r)} \quad (5)$$

where R_d^v is the set of approving rules, R_d is the set of eligible rules and $score : R \rightarrow [0, 1]$ represents the quality score associated with a rule (as detailed in Sect. 3.2). Since the *boosting factor* consists in evaluating a proportion, EB is used also in this case to obtain a better estimation, less likely to be the result of chance. As explained in Sect. 3.2, when applying EB, initially the parameters α_b and β_b of a *Beta* distribution are estimated from the available data using methods of moments. Then this prior is updated based on evidence associated with a specific v_d . Thus, the *boosting factor*, corresponding to the expected value of the updated prior, is equal to:

$$boost(d, v_d) = \frac{\alpha_b + \sum_{r \in R_d^v} score(r)}{\alpha_b + \beta_b + \sum_{r \in R_d} score(r)} \quad (6)$$

where α_b and β_b are the hyper parameters of the Beta distribution representing the available examples. Since AMIE does not consider any *a priori* knowledge such as the partial order of values to extract rules, we decided to use it to further exploit rule information and compute a more refined boosting factor. More precisely, considering a partial order $\mathcal{V} = (V, \preceq)$, when a rule r explicitly predicts a value v , we assume that it implicitly supports all more general values v' such that $v \preceq v'$. In other words, the evidence provided as support by a rule to a value is propagated to all its generalizations. Therefore, in this case the boosting factor $boost_{PO}(d, v_d)$ indicates the percentage of approving rules out of all eligible rules, for both the value under examination and all of its more specific values. The subscript *PO* in the name of the boosting factor underlines the fact that the Partial Order among values is taken into account. The set R_d^v in Eq. 6 is replaced by the set $R_d^{v^+} = \{r \in R_d \mid \widehat{B} \wedge H = p(x, v'), v' \preceq v\}$.

3.4 Integrating Rules' Viewpoints into Confidence Computation

All the elements required to integrate information given by recurrent patterns into TD models have been defined. Since the *boosting factor* depends on the

claim, only the confidence formula has been updated. As proof of concept, in this study we modified *Sums* [14] whose estimation formulas are:

$$t^i(s) = \frac{1}{\max_{s' \in S'} \sum_{v'_d \in V^{s'}} c^{i-1}(v'_d)} \sum_{v_d \in V^s} c^{i-1}(v_d) \quad (7)$$

$$c^i(v_d) = \frac{1}{\max_{v'_d \in V} \sum_{s' \in S^{v'_d}} t^i(s')} \sum_{s \in S^{v_d}} t^i(s) \quad (8)$$

We modified Eq. 8 proposing *SumsRULES*. This new model integrates the additional information given by rules into the confidence formulas as follows:

$$c_{rules}^i(v_d) = \frac{1}{norm_{v_d}} [(1 - \gamma)c^i(v_d) + \gamma boost(d, v_d)] \quad (9)$$

where $\gamma \in [0, 1]$ is a weight that calibrates the influence assigned to sources and KB for estimating value confidences. For the sake of coherence, when using *boost_{PO}* we considered the partial order also for the computation of the confidence formula, as suggested in a previous study [3]. We refer to the model that uses confidence formula $c_{PO}^i(v_d)$, taking the partial order into account, as *Sums_{PO}*. It computes the confidence of v_d considering all the trustworthiness of sources that provide the value v for the data item d , i.e. the claim v_d under examination, or a more specific value than v . Indeed as highlighted above when claiming a value, we also consider that a source implicitly supports all its generalizations. Similarly, the model that integrates both the *boost_{PO}* and rules is indicated as *SumsRULES&PO* and is defined as follows:

$$c_{RULES\&PO}^i(v_d) = \frac{1}{norm_{v_d}} [(1 - \gamma)c_{PO}^i(v_d) + \gamma boost_{PO}(d, v_d)] \quad (10)$$

Note that, while *Sums* and *SumsRULES* return a true value for each data item selecting the value with the highest confidence, *SumsRULES&PO* and *Sums_{PO}* required a more refined and greedy procedure to select the most informative true value. Indeed, considering the partial order of values, the highest confidence is always assigned to the most general value (it is implicitly supported by all the others). Thus, since systematically returning the most general value each time is not worthwhile, the selection procedure leverages the partial order to identify the expected value. Starting from the root, at each step it selects the closest specialization of the value with the highest confidence. The procedure stops when there are no more specific values, or when the confidence of the selected values is lower than a given threshold θ defining the minimal confidence score required to be considered as a true value. For further details see [3].

4 Experiments and Results

In order to obtain an extended overview of the proposed approach, several experiments were carried out on synthetic and real-world datasets. First of all, experiments were conducted using synthetic datasets to determine the improvement

obtained by $Sums_{RULES}$ (Eq. 9) and $Sums_{RULES\&PO}$ w.r.t. their baseline, i.e. $Sums$ [14] (Eq. 8) and $Sums_{PO}$ (Eq. 10) considering different scenarios. Note that, in both cases, the baseline corresponds to set $\gamma = 0$ in the new confidence formula of the proposed models. A second set of experiments was conducted using a real-world dataset to test the proposed approach in a realistic scenario. A comparison with existing models is also presented.

The rules used in the experiments, as well as their support and $conf_{PCA}$ were extracted from DBpedia by AMIE. To ensure that the rules considered are abstractions of a sufficient number of facts, we selected those with the highest head coverage. We selected 62 rules for the predicate *birthPlace*. Examples of these rules are reported in Table 2.

Table 2. Examples of rules extracted by AMIE from DBpedia for *birthplace* predicate.

@prefix db: <http://dbpedia.org/resource/>.	
@prefix db-owl: <http://dbpedia.org/ontology/>.	
?a db-owl:deathPlace ?b	→ ?a db-owl:birthPlace ?b
?a db-owl:country ?b	→ ?a db-owl:birthPlace ?b
?a db-owl:deathPlace ?b ∧ ?b db-owl:language db:English_language	→ ?a db-owl:birthPlace ?b

4.1 Experiments on Synthetic Data

The synthetic datasets were used to evaluate the proposed model on various scenarios depending on the granularity of the true values provided. Experts usually provide specific true values. Non-expert users provide general values, which remain true. To evaluate the performance in these contexts, we measured the expected value rate/recall (returned values that correspond to expected ones), the true but more general value rate (returned values that are more general than the expected ones) and the erroneous value rate (values that are neither expected nor general) obtained by different model settings.

Generation. The main elements required to generate these datasets are: a ground truth, a partial order and a set of claims provided by several sources on different data items [3]. The ground truth was generated by selecting a subset of 10000 DBpedia instances having the *birthPlace* property, considering the related value as the true one. Also the partial order of values was constructed using the DBpedia ontology. Partial order relationships were added between all classes subsumed, i.e. `rdfs:subClassOf`, by `dbpedia-dbo:Place` class and between those classes and their instances. Moreover, the relationships were added to all instances for which the property `dbpedia-dbo:isPartOf` or `dbpedia-dbo:country` exists. Since `dbpedia-owl:Thing` is the most abstract concept in DBpedia, all the values belonging to the partial order graph were rooted to it. In order to obtain a partial order of values respecting the properties of a Directed Acyclic Graph,

all cycles induced by incorrectness on the part-of property were removed.² For the generation of the claims, 1000 sources and 10000 data items were considered. Table 3 reports all the features regarding the generation of the claim set. The main feature is related to the distribution used to select the granularity of the true values provided. Based on this feature, three types of dataset were generated: EXP, LOW_E and UNI figuring, respectively, the behaviors of experts, a mix of experts and non-experts, and non-expert users. Considering that Picasso was born in Málaga, for example, in the case of EXP datasets, the sources tend to provide true values such as Málaga, Andalusia, Spain, while in the case of UNI datasets they will also provide general values such as Europe or the Continent. For each scenario, 20 synthetic datasets were generated.

Table 3. Features of synthetic datasets.

Feature	Description
Source coverage	Each source provides a number of claims that is exponentially distributed.
Source trustworthiness	The trustworthiness distribution is Gaussian with average 0.6 and standard deviation 0.4. This means that the sources are mostly reliable and only a few of them are always or never correct.
# of true claims per source	Each source provide a true value w.r.t. its trustworthiness level.
# of distinct true values per data item	$1..V_d^{true}$ where $V_d^{true} = \{v \in V : v_d^* \preceq v\}$
Granularity of the true value provided	Each source provides a true value having a granularity that approaches the granularity of the expected true value w.r.t. a high decay-rate exponential distribution (EXP), a low decay-rate exponential distribution (LOW_E) and a uniform distribution (UNI).
# of distinct false values per data item	$1..30$ values belonging to $V_d^{false} = V_d^{true} \setminus \{v v \preceq v_d^*\}$

Results. The results, summarized in Fig. 1, show that the proposed approach enables the definition of TD models that benefit from the use of *a priori* knowledge given by an external RDF KBs. Indeed, the number of correct *facts* identified by the proposed model usually increases w.r.t. the baseline. Intuitively, since

² We assumed that abstract concepts should have higher out-degree than less abstract ones. Thus, for each cycle, the edge whose target is the node with the highest out-degree was removed. Analysing the discarded edges, the heuristic works.

the number of correct *facts* increases, a new KB that is populated with the true claims identified by the improved TD will be of higher quality.

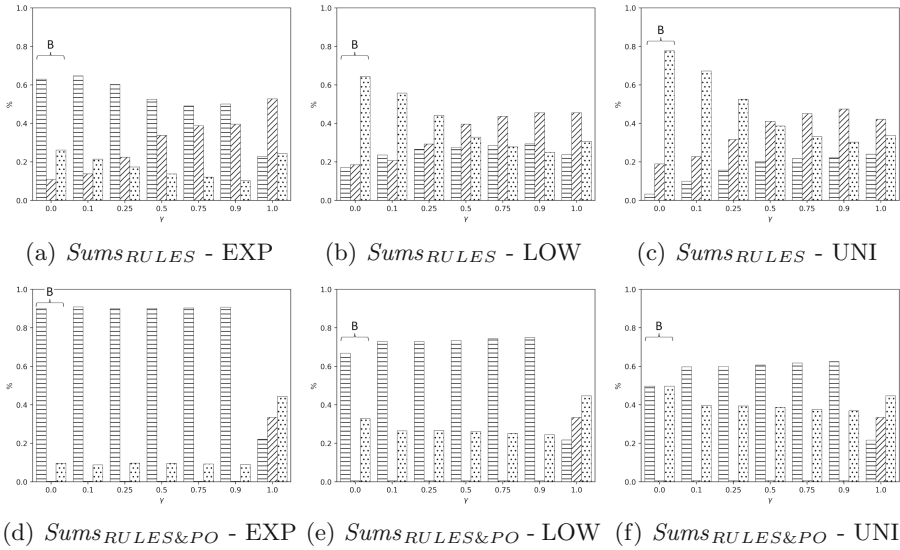


Fig. 1. Expected (horizontal line bars), true but more general (diagonal line bars) and erroneous values (dotted bars) obtained by Sum_{sRULES} and $Sum_{sRULES\&PO}$ on different datasets with several γ . The letter B indicates the baseline model results.

The improvement obtained by considering both Sum_{sRULES} and $Sum_{sRULES\&PO}$ was always greater for UNI datasets than for EXP or LOW_E ones. Since identifying true values in UNI settings was harder than in the other cases (the highest disagreement among sources on the true values was modeled by UNI), the baseline obtained the lowest recall. Using additional information tackles the high level of disagreement among sources and thus enables full exploitation of the higher scope for improvement that was available in the case of the UNI setting.

Considering Sum_{sRULES} the best recall was obtained with different γ values. For UNI datasets, the optimal configuration was when $\gamma = 1$. In such a case, it was considered that no information provided by sources was useful and that only rules should be used to solve conflicts among claims (when rules are available). This was true only for the extreme situation represented by UNI datasets where disagreement among sources was so high that the recall obtained by baseline model remained under 10%. Indeed, in the other cases it was advantageous to take both source trustworthiness and rule information into account. For EXP datasets, the optimal γ value was 0.1, while for LOW_E it was 0.9. Low γ values were preferred in EXP settings because in this case sources that provide true values are quite sure about the expected one, and it is thus less useful to consider

the rules' viewpoints. Moreover, this setting was the only situation where considering external knowledge was damaging in terms of recall. Nevertheless, the error rate obtained by *SumsRULES* when $0 < \gamma < 1$ was always lower than the error rate achieved when $\gamma = 0$. This is explained by the fact that the average Information Content³ (IC) of values inferred by rules extracted for the *birthPlace* predicate is around 0.53. This means that they often infer values that are general. Many returned values, selected with the highest value confidence criteria, were therefore more general than the expected one but not erroneous. In other words, the rules associated with the *birthPlace* predicate were more effective for discovering the country of birth than the expected location. However using rules were useful, as shown by the results the error rate decreased.

The limitation related to rules that support general values was in part overcome by considering *SumsRULES&PO*, which also takes the partial order of values into account. In this case rules can improve the selection of the correct value during the first steps of the selection procedure. They were able to handle and dominate the false general values supported by many sources. The selection process was then continued with the fine-grained values evaluated based only on source trustworthiness information since no evidence provided by rules was available. For *SumsRULES&PO* tested on EXP datasets, low γ values were preferred, while on LOW_E and UNI datasets high γ values led to the best performance.

The best overall recall was obtained by *SumsRULES&PO*, which considers both kinds of *a priori* knowledge: extracted rules and partial order of values.

4.2 Experiments on Real-World Data

These experiments were conducted to test the proposed model in a realistic scenario. Since the results of experiments on synthetic data showed that the most interesting results were obtained by considering both extracted rules and the partial order of values, we compared the results obtained in this case with those obtained by existing TD methods⁴ [20]. The evaluation protocol consisted in counting the number of values returned by a model that are equal to the expected values. In this setting, the number of general values returned were not analyzed since the main aim of TD models is to return the expected value, not its generalizations.

Generation. We collected a set of claims related to the predicate `dbo:birthPlace`, i.e. people's birth location. As data item subject, we randomly selected a subset of 480 DBpedia instances of type `dbo:Person` having the property `birthPlace` and having at least one eligible rule. For each data item we collected a set of webpages (up to 50) containing at least one occurrence of the

³ Information Content indicates the degree of abstraction/concreteness of a concept w.r.t. an ontology. It monotonically increases from the most abstract concept (its $IC = 0$) to the most concrete ones discriminating the granularity of different values.

⁴ For these models we used the implementation available at <http://www.github.com/daqcri/DAFNA-EA>.

subject’s full name and the words “was born”, i.e. the natural language expression that is usually used to introduce the birth location of a person. Given a webpage and its data item, we defined two procedures for extracting the provided claim. Procedure A selects, as claimed value, the location (identified by DBpedia-spotlight API) that co-occurs in the same sentence and is nearest to the word “born”. Procedure B adds a constraint to procedure A: a value can be selected only if it appears after the first occurrence of the subject’s full name in the text. Two different datasets were created based on procedures A and B, respectively DataA and DataB. For building our ground truth, we assumed that the values defined in DBpedia as birth location for each data item were the true ones. Since in the collected claims, values that were more specific than the expected one (contained in the ground truth) were provided, we manually checked if these specifications were true. For 20 instances that we manually checked, 10 were found to be true specifications. Note that as partial order we considered the same one as for the experiments on synthetic data. The procedures, source code and datasets obtained are available online at <https://github.com/lgi2p/TDwithRULES>.

Results. We can observe that for both datasets DataA and DataB we improved the performance by 18% and 14% respectively compared to the baseline, i.e. *Sums* – the approach we decided to modify. Table 4 shows the results obtained by the best configuration of parameters where both extracted rules and partial order were considered.

When comparing the proposed approach to existing TD models, it did not outperform the others, see Table 5. Note that our study focused on modifying *Sums* which is considered to be one of the most well studied models, but not necessarily the most effective one. After investigating the errors, we found out that it was mainly due to a limitation of *Sums*: it rewards sources having high coverage and, meanwhile, penalizes those with low coverage. Indeed *Sums* computes the trustworthiness of a source by summing up all the confidence of the claims it provides. Thus the higher the number of claims a source provides, the higher the trustworthiness of the source. The problem is that *Sums* does not distinguish between sources always providing true values, but having different coverage. While [Wikipedia.org](https://en.wikipedia.org) is correctly considered as a high reliable source, an actor’s fan club website is incorrectly considered as unreliable. Even if the information it provides is correct, because it covers only one data item its trustworthiness will be lower than the one of [Wikipedia.org](https://en.wikipedia.org) (source having a high coverage). In real-world datasets very few sources have high coverage, and most of them have low coverage – power law phenomenon. In this scenario the sources having high coverage dominate the specialized ones. Therefore, no extraction errors from high coverage sources are allowed. Indeed if an incorrect value is extracted from [Wikipedia.org](https://en.wikipedia.org) (for instance when the sentence refers to another person), this will be incorrectly considered as the true one. Since this cannot be guaranteed (the extraction procedures we defined are voluntarily naive), we propose a post-processing procedure that alleviates this problem. Before selecting the true value, it sets equal to 0 all the confidence of those values that are

Table 4. Recall obtained using *Sums* and its modifications on DataA and DataB.

Model	DataA	DataB
<i>Sums</i>	0.448	0.473
<i>Sums</i> _{PO} ($\gamma = 0.0, \theta = 0.05$)	0.517	0.566
<i>Sums</i> _{RULES&PO} ($\gamma = 0.3, \theta = 0.0$)	0.527	0.548
<i>Sums</i> _{RULES&PO} ($\gamma = 0.3, \theta = 0.05$)	0.565	0.590
<i>Sums</i> _{RULES&PO} +post-proc. ($\gamma = 0.3, \theta = 0.1$)	0.631	0.614

Table 5. Recall obtained using existing models on DataA and DataB.

Existing model	DataA	DataB
Voting	0.640	0.625
TruthFinder	0.646	0.622
2-Estimates	0.631	0.635
3-Estimates	0.008	0.612
Cosine	0.636	0.635
AccuCopy	0.638	0.640
Accu	0.638	0.660
Depen	0.431	0.494
AccuSim	0.413	0.448
SimpleLCA	0.631	0.660
GuessLCA	0.644	0.646

provided by only a single source. We assume that it is highly improbable that the same extraction error occurs, i.e. the erroneous value should therefore be provided only once. This solution, indicated as *Sums*_{RULES&PO} + post-proc., obtained performances comparable with existing models for DataA and DataB. While it enables to avoid some of the extraction errors (occurring more with the most naive procedure A), it is still not capable of assigning lower trustworthiness levels to specialized sources.

Given these observations, in real-world settings it is very important to consider the power law phenomenon. The results show that *Sums* is not efficient in this kind of situation. Nevertheless, using additional information (partial order and extracted rules) improved the results w.r.t. the baseline approach, and this is promising for the principles introduced in this study. As shown in Table 4, the improvement due to taking this information into account was 18% for DataA and of 14% for DataB. Moreover, through this study we also show that correctness and the granularity of values in DBpedia can be improved using TD models. Claims on data items can easily be collected on the Web. When more specific values than the one contained in DBpedia are found, they can be verified using TD model.

5 Conclusion

Solving information conflicts in an automated fashion is critical for the development of large RDF KBs populated by heterogeneous information extraction systems. In this study, we suggest using TD models as unsupervised techniques to populate RDF KBs. In order to create high quality KBs and exploit current ones, we propose improving an existing TD model (*Sums*) using knowledge

extracted from an external RDF KB in the form of rules. Several experiments that show the validity of the proposed model were conducted. The performances of the proposed model show higher recall than baseline methods (up to 18% of improvement). The datasets, source code and procedures are all available online. We plan to apply the rationale of the proposed model to other TD models in order to outperform them all. In addition, we envisage extending the evaluation methodology in order to consolidate our results by considering other predicates and non-functional ones such as those used in ISWC Semantic Web Challenge 2017. Currently, we do not consider as negative evidence the fact that a rule predicts a different value than the one contained in a claim. In the future, we envisage studying how to incorporate this information, as well as explicit axioms, subjectivity information and contextual dependencies (such as diachronicity).

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD 1993, vol. 22, pp. 207–216. ACM (1993)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
3. Beretta, V., Harispe, S., Ranwez, S., Mougnot, I.: How can ontologies give you clue for truth-discovery? An exploratory study. In: WIMS 2016, p. 15. ACM (2016). <https://doi.org/10.1145/2912845.2912848>
4. Berti-Équille, L., Borge-Holthoefer, J.: Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics. Synthesis Lectures on Data Management. Morgan & Claypool Publishers (2015). <https://doi.org/10.2200/S00676ED1V01Y201509DTM042>
5. Dong, X., et al.: Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: KDD 2014, pp. 601–610. ACM (2014). <https://doi.org/10.1145/2623330.2623623>
6. Galárraga, L.: Interactive rule mining in knowledge bases. In: Actes des 31e Conférence sur la Gestion de Données (BDA 2015), Île de Porquerolles (2015)
7. Galárraga, L., Suchanek, F.M.: Towards a numeric rule mining language. In: Proceedings of Automated Knowledge Base Construction workshop (2014)
8. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.M.: Fast rule mining in ontological knowledge bases with AMIE+. VLDB J. **24**(6), 707–730 (2015). <https://doi.org/10.1007/s00778-015-0394-1>
9. Lehmann, J., Völker, J. (eds.): Perspectives On Ontology Learning, vol. 18. IOS Press (2014). <https://doi.org/10.3233/978-1-61499-379-7-i>
10. Li, Y., et al.: A survey on truth discovery. SIGKDD Explor. Newsl. **17**(2), 1–16 (2016). <https://doi.org/10.1145/2897350.2897352>
11. Maimon, O., Rokach, L.: Data Mining and Knowledge Discovery Handbook, vol. 2. Springer, Heidelberg (2005). <https://doi.org/10.1007/b107408>
12. Nebot, V., Berlanga, R.: Finding association rules in semantic web data. Knowl.-Based Syst. **25**(1), 51–62 (2012)
13. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A review of relational machine learning for knowledge graphs. Proc. IEEE **104**(1), 11–33 (2016)

14. Pasternack, J., Roth, D.: Knowing what to believe (when you already know something). In: COLING 2010, pp. 877–885. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
15. Quboa, Q.K., Saraee, M.: A state-of-the-art survey on semantic web mining. *Intell. Inf. Manag.* **5**(01), 1–10 (2013)
16. Robbins, H.: An empirical Bayes approach to statistics. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability Volume 1: Contributions to the Theory of Statistics*, pp. 157–163. University of California Press, Berkeley, California (1956)
17. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge. In: WWW 2007, pp. 697–706. ACM (2007). <https://doi.org/10.1145/1242572.1242667>
18. Pellissier Tanon, T., Stepanova, D., Razniewski, S., Mirza, P., Weikum, G.: Completeness-aware rule learning from knowledge graphs. In: d’Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10587, pp. 507–525. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68288-4_30
19. Ventura, S., Luna, J.M.: Quality measures in pattern mining. In: Ventura, S., Luna, J.M. (eds.) *Pattern Mining with Evolutionary Algorithms*, pp. 27–44. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-33858-3_2
20. Waguilh, D.A., Berti-Equille, L.: Truth discovery algorithms: an experimental evaluation. CoRR abs/1409.6428 (2014)