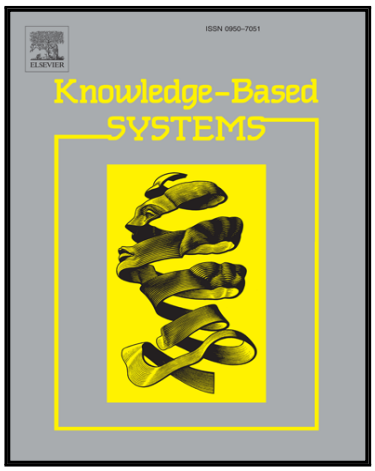
Truth selection for truth discovery models exploiting ordering relationship among values

Valentina Beretta, Sébastien Harispe, Sylvie Ranwez, Isabelle Mougenot



Truth Selection for Truth Discovery Models Exploiting Ordering Relationship Among Values

Valentina Beretta, Sébastien Harispe, Sylvie Ranwez, Isabelle Mougenot

PII: S0950-7051(18)30332-0
DOI: 10.1016/j.knosys.2018.06.023
Reference: KNOSYS 4396

To appear in: *Knowledge-Based Systems*

Received date: 19 December 2017
Revised date: 25 June 2018
Accepted date: 28 June 2018

Please cite this article as: Valentina Beretta, Sébastien Harispe, Sylvie Ranwez, Isabelle Mougenot, Truth Selection for Truth Discovery Models Exploiting Ordering Relationship Among Values, *Knowledge-Based Systems* (2018), doi: 10.1016/j.knosys.2018.06.023

Truth Selection for Truth Discovery Models Exploiting Ordering Relationship Among Values

Valentina Beretta[a,*], Sébastien Harispe[a], Sylvie Ranwez[a], Isabelle Mougenot[b]

[a] *LGI2P, IMT Mines Ales, Univ Montpellier, Ales, France.*
[b] *UMR 228 Espace Dev UM, Maison de la Télédétection, 500 rue JF Breton, 34093 Montpellier Cedex 5, France.*

Abstract

Data veracity is one of the main issues regarding Web data. Truth Discovery models can be used to assess it by estimating value confidence and source trustworthiness through analysis of claims on the same real-world entities provided by different sources. Many studies have been conducted in this domain. True values selected by most models have the highest confidence estimation. This naive strategy cannot be applied to identify true values when there is a partial order among values that is considered to enhance the final performance. Indeed, in this case, the resulting estimations monotonically increase with respect to the partial order of values. The highest confidence is always assigned to the most general value that is implicitly supported by all the others. Thus, using the highest confidence as criterion to select the true values is not appropriate because it will always return the most general values. To address this problem, we propose a post-processing procedure that, leveraging the partial order among values and their monotonic confidence estimations, is able to identify the expected true value. Experimental results on synthetic datasets show the effectiveness of our approach.

Keywords: Truth Identification, Truth Discovery, Conflicting values, Value Relationships, Ontology

*Corresponding author
Email addresses: valentina.beretta@mines-ales.fr (Valentina Beretta), sebastien.harispe@mines-ales.fr (Sébastien Harispe), sylvie.ranwez@mines-ales.fr (Sylvie Ranwez), isabelle.mougenot@umontpellier.fr (Isabelle Mougenot)

1. Introduction

Developing systems able to automatically evaluate the veracity of the avalanche of data produced by modern information society is of critical importance. Data veracity can be determined comparing information provided by multiple sources on the same subject [1–3]. Numerous scientific communities contribute to studying this complex issue, most notably with respect to (w.r.t.) data integration in information systems and databases.Among the several difficult tasks that data integration addresses and the different approaches that can be used to solve them [4–7], this paper focuses on automatic truth discovery for solving situations in which different sources provide potentially conflicting data about a specific property of an entity of interest, e.g. on the place of birth of a person.
Truth Discovery (TD) consider the accuracy associated with the data sources as an important factor to discriminate data veracity [6, 7].
The main aim of the TD models is to identify true information. They intend to automatically solve, in an unsupervised manner, conflicts that may occur among claims. They leverage both the redundancy of the data and the information that is possible to derive from sources (particularly their reliability). More precisely, the backbone of TD is based on the postulate that reliable sources provide true information and that, conversely, true information is given by reliable sources [3]. To identify reliable sources and true information, TD approaches estimate both source trustworthiness and value confidence; the true value is then considered to be the one with the highest confidence. Note that approaches that leverage information about data sources to check data veracity are currently the focus of a lot of attention in several domains such as social sensing[7] and question-answering [8].
Here we address the problem of selecting the truth for functional predicates when *a priori* knowledge in the form of a partial order of values (e.g. subsumption relationship in an ontology) is considered to improve value confidence and source trustworthiness estimations. A partial order highlights when different values are not conflicting, but they represent the same concepts with different levels of granularity. Indeed, conflict and granularity are two different aspects to consider when identifying the most reliable information. While conflict values produce inconsistency, different granularities only indicate imperfection in data [9]. In formal logic, a predicate p is consid-

ered functional if for any *subject* there is a unique value $v \in V$ for which $p(subject, v)$ is true[1] – birthplace is an example of a functional predicate. Note that this definition does not take subsumption relationships among values into account. This is in accordance to the fact that, for instance, everyone was born in a specific location, but it does not consider that this place can be described using different levels of precision, e.g. district or region. In this case, multiple values can be true given the same *subject*. Thus, considering partial order of values and closed world assumption (our knowledge of the world is complete), a predicate is functional if for any *subject* there is a unique value $v \in V$, for which $p(subject, v)$ is true, such that there is not another value $v' \in V$ subsumed by v for which $p(subject, v')$ is true. As far as we know, we are the first to propose to take ontologies (as *a priori* knowledge) into account [10]. In this situation, the traditional final value selection step in the majority of TD approaches cannot be applied. Indeed, in this case, since more abstract values will *de facto* be associated with a higher confidence value in accordance with the partial ordering of values modeling implications among them, the true value cannot be defined as that with the highest confidence. This is due to the hypotheses used by approaches that leverage information related to the structure that may exist among values. Briefly, sources that explicitly claim a value implicitly support all of its generalizations. Therefore, if a source claims that "Pablo Picasso was born in Malaga", it also implicitly supports the assertion that he was born in Spain, Europe, etc (considering that it is in agreement with the ontology). Thus the most general value is implicitly supported by all the others. Hence, its confidence will always be the highest. However, considering the most general value to be the only truth is not trustworthy (since it is a tautology), e.g. stating that Pablo Picasso was born in a Location is not meaningful. This paper proposes to overcome this problem by studying a solution able to identify more specific true answers (than the most general one) that may exist.

Our contribution consists of:

- proposing a post-processing approach able to identify the truth given the confidence estimations returned by any TD model that considers structured values;

[1] Elements of V are here considered to be independent.

- performing empirical experiments on synthetic datasets – this evaluation uses estimations returned by an adaptation of *Sums* able to take prior knowledge in the form of a partial order among values into account [10] – and comparing the proposed approach with existing ones evaluating identified true values.

The rest of the paper is structured as follows. Section 2 presents an overview of TD approaches taking advantage of potential relationships among sources or claims. This section ends with a discussion about the consequences of using a partial order among values as relationship information. In Section 3, notations are introduced and the problem is formalized. The solution strategy we propose is detailed in Section 4. The model is assessed via several experiments reported in Section 5 and discussed in Section 6. Section 7 summarizes the main findings of the study and the results that have been obtained; while the perspectives opened by our contribution are finally discussed.

2. Related work

Truth discovery aims to solve conflicts among data provided by several sources. The data treated in this domain consists of claims specifying the values that sources associate with certain data items (i.e. a data item represents a particular aspect of a real-world entity). Values can be numerical or categorical/strings. The main assumption of TD is that true information is provided by reliable sources and reliable sources provide true information ([3, 11]). This rationale can be modeled by defining the value confidence and source trustworthiness.

Many studies have been proposed in this field [3]. The baseline model consists of a voting strategy. For each data item it regards the value which is the most frequently claimed as truth. All sources are therefore implicitly considered similarly in this model. Otherwise TD estimates for each source a different trustworthiness level based on the claims it provides. Some models deal with numerical values [12], others with categorical claims [7, 13–15] and others with both [11, 16, 17]. While basic TD models limit their complexity to correctly estimate confidence and trustworthiness with different formulas, other approaches incorporate additional information to improve the overall performances. The latter group of approaches is the most relevant for our study and is detailed hereafter.

Among all of these methods, we focus our attention on those that use, as additional knowledge, correlations[2] that may exist among sources ([17–19]), data items ([7, 20, 21]) or values ([11, 22]).

The first class is related to source interdependences. These models consider source relationships mainly by analyzing the pattern of similar claims with correlated accuracy estimations. They also usually assume that sources sharing common false values are more likely to be dependent than sources sharing common true values. Indeed, it is difficult to identify dependencies between sources stating different false values [23]. Most of these studies only analyze static correlations. To the best of our knowledge, time-course dependency relationship patterns has been only considered in [24]. In this case, dependency among sources is captured by studying the similarity between patterns of updates associated with sources [23]. Several methods take advantage of dependency in terms of the copying relationship [17–19, 22]. Other correlations among sources may also occur, such as the common errors made by different extractors that use the same extraction rules or the common values identified by the extractor that use different rules, and so on[17]. Moreover, the dependence relationship is often considered between source pairs, but dependencies may also occur at the group level [18].

The second correlation class is related to data items. The first body of works in this context proposed to deal with the social sensing problem. In crowd sensing, humans coupled with their smartphones become sensors that explicitly or implicitly provide observations about their physical environment. Then it becomes necessary to understand the validity of data sent by sensors. TD models applied to this domain take advantage of both physical [25] and temporal [26] correlations as well as causal relationships [21]. For physical correlations, they assume that co-located data items should have similar values. For instance, gas stations located in the same area should have similar gas prices. For temporal correlations, the assumption is that two temporally close observations cannot have very different values. This kind of correlation is especially useful when analyzed data has a long-tail characteristic, i.e. many data items observed by few sources and few data items observed by many sources. Indeed, in this case the estimations can easily deteriorate if the few sources that provide claims for a data item are also unreliable. Using

[2]We mean by correlation the interdependences between entities, the relationships that may influence them.

correlations, information associated with data items having a high number of observations provided by reliable sources can be propagated to data items having only a few claims associated with them. The findings of the two studies [25] and [26] permit to partition data items into small groups without considering any dependency among groups, but the complexity of their solutions is exponential w.r.t. the maximum group size. Alternative models have been proposed to overcome this limitation to be able to deal with a large number of dependencies, e.g. [20] [21]. The former classifies the problem as an optimization problem, and the latter, modelling the problem as a Bayesian network, leverages potential conditional independences among data items. Moreover, in this study they also take into account a second kind of correlation related to the data items: i.e. the category. In this case, a trustworthiness level may be attributed to each source w.r.t. the category a data item belongs to. The main limitation of this approach is that the Bayesian network has to be known or empirically learned from historical data by specific algorithms.

The third type of correlation regards the values. The basic idea is that two correlated values support each other. If one of them is considered true, then the other has a high probability to be true. In order to evaluate value correlations, previous studies such as [11], [22], and [16] use value similarity. For instance, they compute the edit distance of strings, similarity among sets, and difference among numerical values. Otherwise, in our previous study we took advantage of value correlations in the form of partial ordering that may exist among the provided values [10]. An example of partial order is shown in Fig. 1. Given this partial order, we adapted the *Sums* approach to incorporate this new information. The rationale is that if a source associates the value *Spain* to an aspect of a real-world entity, then it also implicitly supports all more general values, e.g. *Europe*. Thus we modified the formula to estimate the confidence of a value changing the set of sources used for the calculus. In *Sums*, confidence is computed by considering all sources that provide an analyzed value, while in the adapted model (*AdaptedSums*) sources that explicitly claim a more specific value than the one being analyzed are also taken into account as well. Indeed, these sources implicitly support all more general values than the one they provide. Using the adapted approach, at the end of the iterative procedure, the value confidence estimations monotonically decrease w.r.t. the partial order among values. In other words, for each value more specific than another, the confidence of the former is lower than or equal to the confidence of the latter. As a result, when considering

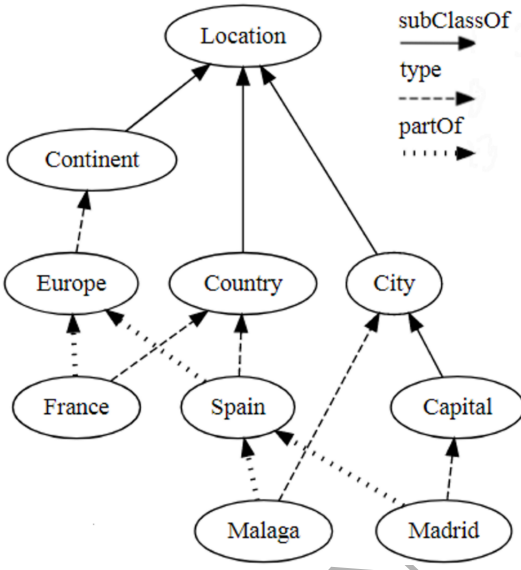


Figure 1: Example of a partial order that may exist among values.

the partial order of values, the highest confidence score will be always assigned to the most general value. In this context, using the usual strategy adopted by the existing models is not worthy. Indeed, selecting as truth the value having the highest confidence, they will always return as true value the most general one. In the rest of the paper, we describe a refined post-processing strategy able to select the true value leveraging these monotonic estimations.

3. Problem Formulation

Let's consider a set of data items D such that each $d \in D$ is composed of a pair $(subject, predicate)$ where the *subject* represents a real-world entity and the *predicate* represents its aspect of interest, e.g. $(picasso, bornIn)$. S denotes the set of sources, V the set of values, $V^s \subseteq V$ the set of values provided by $s \in S$ (for each data item for which s provided values), and $V_d \subseteq V$ the set of values associated with data item $d \in D$.
Formally, TD models first aim to identify the set of true values $V_d^* \subseteq V$ for each data item $d \in D$. In the case of a data item d characterized by a functional *predicate*, we have $|V_d^*| = 1$ if elements from V are disjoint, i.e. $\forall (v, v') \in V^2, \neg(v \implies v') \wedge \neg(v' \implies v)$. Note that a value v implies

($\Longrightarrow$) a value v' when v is subsumed by v'. To ease the formal introduction, and in accordance with the literature, we will as often as possible consider the special case of data items associated with the functional predicate.

Dealing with data items composed of functional predicates TD identifies the true values $v_d^* \in V$ for each data item d estimating value confidences $c : V \rightarrow [0, 1]$ – how an information is likely to be true – and source trustworthiness $t : S \rightarrow [0, 1]$ – how reliable is a source. This is done through an iterative procedure that alternatively estimates them. The execution of the model finishes when the stopping criteria is verified, e.g. convergence of estimations, maximum number of iterations, and so on. Hence, each value $v_d \in V$(w.r.t. a data item) is associated with a confidence level $c(v_d)$ and each source $s \in S$ with a trustworthiness level $t(s)$.

Existing approaches usually assume that for a specific data item d, elements of V_d are disjoint/independent and they therefore recognize the true value of d is that with the highest confidence score. This straightforward procedure cannot be applied using adapted models that consider ordering among values. Incorporating this information into the model, each value more general than a true value can only be considered as true as well[3]. Therefore, considering all values associated with a data item, the estimated confidence scores monotonically increase w.r.t. the partial ordering of values, i.e. $\forall v, v' \in V$: if $(v \implies v')$, then $c(v) \leq c(v')$. Consequently, the highest confidence score is always assigned to the most general value (that is implicitly supported by all provided claims). To solve this problem, we propose a post-processing procedure able to select the true value for each data item given the estimated confidence scores and the relationships that may exist among values.

We assume that the value dependencies are known *a priori* in the form of a partial order modelled by an ontology $O = (\preceq, V)$. Note that even if the domain knowledge is not available, partial order can be automatically constructed [27]. The partial order $\preceq$ can be represented by a Directed Acyclic Graph (DAG), $G_O = (V, E)$, where $V = \{v_0, v_1, \cdots, v_m\}$ is the set of values representing our vocabulary according to our knowledge of the world (all possible values that can be claimed by sources), and $E = \{(x, y) \in V \times V | x \preceq y\}$ is the set of edges specifying the partial ordering that exists between values. Specifically $x \preceq y$ when there is a directed path from x to y in the DAG; i.e. when y is reachable from x [28]. Note that a path from x to y is defined

[3]Assuming that the value ordering is consensual.

as a non-empty sequence of n different nodes $\langle v_0, v_1, \cdots, v_{n-1}\rangle$ with $x = v_0$, $y = v_{n-1}$ and for which $\forall i \in [0, n-2]$ $(v_i, v_{i+1}) \in E$. An important characteristic of the graph G_O is that it has to be transitively reduced. This is not a problem because by considering any DAG its transitive reduction can be obtained [29].
Here we introduce several functions that will be useful for manipulating the graphs (with $\mathcal{G}$ a set of graphs):

- *ancestors*: $\mathcal{G} \times V \rightarrow \mathcal{P}(V)$ such that $ancestors(G_O, x) = \{y | x \preceq y\}$.
- $children : \mathcal{G} \times V \rightarrow \mathcal{P}(V)$ designed as $children(G_O, x) = \{z | (z, x) \in E\}$
- $root : \mathcal{G} \rightarrow V$ such that as $root(G_O) = \{x | \forall y \in V, y \preceq x\}$

These properties enable us to easily explain our procedure to traverse the partial value ordering graph in the next section.
Further important information that can be derived from any ontology is the Information Contents (IC) of its concepts (e.g. [30]). This quantity, related to the concept specificity (see Section 3.3 in [31]) represents the degree of abstraction/concreteness of a concept w.r.t an ontology. One of main IC property is that the IC score monotonically decreases from the root to the leaves, i.e. if $x \preceq y$, then $IC(x) \geq IC(y)$ ($IC(root) = 0$). This score will help us to discriminate between different values w.r.t. their granularity.
All of the elements presented in this section will help define the approach used to select the true values that is described in Section 4.

4. Proposed Approach

The entire truth-discovery procedure, from the input consisting of a set of claims to the output consisting of the true values and the degree of reliability associated with each source, is presented in Fig. 2. In this section, we propose a post-processing procedure that selects the true values given the estimations obtained by TD models that relax the assumption related to the disjointness of values.
It involves three steps: (i) selection of the best true value candidates; (ii) ranking of selected values; and (iii) filtering of ranked values w.r.t. defined desirable properties. For instance it may be useful to return a set of solutions that share an ordering relationship or, on the contrary, to return a value set composed only of "alternatives" that are not ordered. The choice related to

the appropriate features of the solution set depends mainly on the application scenario.

The first step of the process: (i) permits to retrieve the most specific possible true value(s) and all of its ancestors using available information, such as confidence scores and partial ordering of values. The second step: (ii) orders the selected values based on predefined criteria. The third step: (iii) is required to filter the top k results. For TD, the final aim k should be equal to 1, but in cases where there is uncertainty it may be useful to return a set of values, even if the predicate is functional. Moreover, answers that do not have defined desirable properties (see Section 4.3 for further details) are removed from the result list. Those three steps are detailed hereafter.

4.1. True value selection

The first part of the post-processing procedure concerns the selection of the promising candidate(s) as the most expected value(s) for each data item. We have defined a selection strategy that takes advantage of the partial order of values and step by step refines the granularity of the correct value associated with each data item. Now we will give an overview of the approach followed by Algorithm 1.

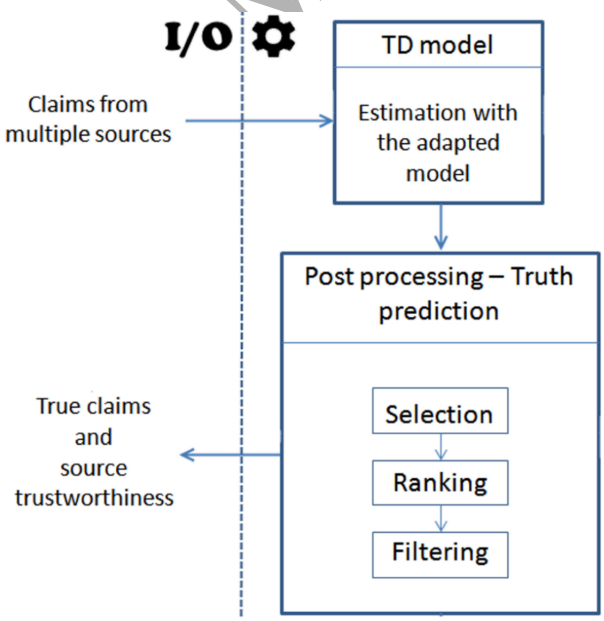


Figure 2: Diagram of the overall TD procedure.

Starting from the most general value (implicitly supported by all provided values and surely true), the process aims to detect the most specific expected value(s). A traversing procedure was thus applied on the graph that represents the partial order of values. It starts from the root, it selects the best alternatives among the children of the considered node and moves forward through the selected values. Our assumption is that values with the highest confidence locally should be the most likely to be true. Therefore the choice of the best alternative(s) is done by comparing the confidence scores associated with the children of the previously selected node.
In the case of functional predicate, the values can be partially ordered by their granularity, see Fig. 1. Therefore the selection procedure refines, at each step, the level of precision used to describe the single true value associated with a data item. The semantics of each selected node expresses the fact that the node subsumes the correct solution (i.e. the expected true value). The last selected nodes should correspond to the most specific answers that can be identified through the selection process.
The selection process has to handle two main undesirable situations that may occur: (1) selection of values with a confidence score too low to be considered as true, and (2) difficulty in discriminating the best alternative(s) among the children of a node since their confidence scores are not significantly different. As a solution, two thresholds have been defined: θ and δ.
The threshold $\theta \in (0, 1]$ enables us to specify the confidence lower bound required for a value to be part of the set of true values. Note that the value 0 is not included in the θ interval. Indeed, considering claims with confidence scores equal to 0 makes no sense because it would mean considering, as truth, values provided by totally unreliable sources (all with trustworthiness equal to 0). The confidence score that is compared to θ has to be previously normalized w.r.t. each data item, i.e. the confidence score associated with the most general value of each data item always has to be equal to 1. This normalization step is required to avoid the definition of an inconsistent threshold w.r.t. the different data items.
The threshold $\delta \in [0, 1]$ represents the minimum difference that has to exist among values with the highest confidence and all the others so that one prevails over the others. In particular, if the difference between the confidences of two values is less than or equal to δ, then it is hard to make a choice among them. This comparison is done among values that are children of the same father to select the best alternatives.
The definition of different parameter settings produces different behaviours

Table 1: Interesting settings for the selection procedure.

Config	θ	δ	**Selection procedure behaviour**
1	α	0	Naive greedy procedure that maximizes the confidence score at each step.
2	α	1	Greedy procedure that selects all values greater than α. Since δ is equal to 1, all values with confidence higher than α are selected.
3	α	β	At each iteration a value is collected only if the difference of its confidence and the highest confidence at the current step is lower or equal to β. All values in the returned set have confidence that is greater than α.

of the selection phase ending in the possibility of obtaining different kinds of solution sets. The main parameter settings are summarized in Table 1.
Configuration 1 ($\theta = \alpha, \delta = 0$) reproduces a naive greedy algorithm that, at each step, selects values with the highest confidence greater than α without performing any other control.
Configuration 2 ($\theta = \alpha, \delta = 1$) is able to return all claimed values with confidence higher than α. It may seem useless, but it is a selection configuration necessary to obtain a particular set of values at the end of the post-processing procedure. A set composed of "promising" alternatives that allow to report values that are, as much as possible, fine-grained and semantically different. In this way, we increase the probability of finding the correct value since we increase the number of different concepts that are considered. Therefore this strategy is useful to deal with cases in which there is a lot of uncertainty. The idea is to return all claimed values and their ancestors and then, using the ranking phase to position in the first places, the most promising alternatives with the properties we have just explained.
Configuration 3 ($\theta = \alpha, \delta = \beta$) is a generalization of the two previous configurations. It selects the set of values that are greater than a threshold θ and they differ, at each step, more than δ from the confidence of the other alternatives.

Algorithm 1 reports the pseudo-code of the selection procedure. The algorithm starts performing a transitive reduction of the graph representation of the partial ordering (line 2). We thus ensure that the choice of the best alternative is done among a set of children that do not share ordered relations. Moreover, this avoids useless comparison of a large number of confidence

Algorithm 1 True value set computation for any $d \in D$ considering a partial order of values represented as a DAG $G_O = (V, E)$, a threshold $\theta \in (0, 1]$, a threshold $\delta \in [0, 1]$, and a function $c : V \to [0, 1]$, i.e. confidence of each value

1: **procedure** SELECTIONTRUEVALUES($d, G_O, c, \theta, \delta$)
2: $G \leftarrow transitive\_reduction(G_O)$
3: $V^*_{visited} \leftarrow \{\ \}$
4: $queue \leftarrow list(root(G))$
5: **while** !($queue.isEmpty()$) **do**
6: $v \leftarrow queue.pop()$
7: $V^*_{visited} \leftarrow V^*_{visited} \cup \{v\}$
8: $V_{ch} \leftarrow children(G, v)$
9: $conf_{max} \leftarrow \max_{child \in V_{ch}} (c(child))$
10: $V_{ch^*} = \{v' \in V_{ch} : c(v') \geq \theta \wedge (conf_{max} - c(v')) \leq \delta\}$
11: $queue.addAll(V_{ch^*} \setminus V^*_{visited})$
12: **return** $\bigcup_{v \in V^*_{visited}} ancestors(G, v)$

scores. Then, at each iteration, the algorithm applies a greedy search by maximizing the confidence of the values (lines 5 – 9). It selects all values having confidence higher than or equal to θ whose scores are not significantly different from the highest confidence (line 10). Then, it adds them to the queue (line 11). Note that the the confidence scores were computed applying *AdaptedSums*. This model computes the value confidences summing up all trustworthiness of those sources providing the considered value or one of its possible specializations. The procedure stops when the last selected value has no more specific values to be visited. In order to be in accordance with our assumption and problem settings, all values that are more general than that selected will compose the set of true values – due to multiple inheritances some of those values may not have been visited by the greedy procedure (line 12). The fact that confidence score monotonically increases w.r.t. the partial order ensures that the scores related to ancestors of the visited values are higher than or equal to θ.

The termination of Algorithm 1 is ensured by line 6 and line 11. The complexity of the selection of the true value algorithm is related to the number of comparisons required to find the maximum value confidence traversing graph G_O. Therefore, the complexity of the algorithm is $O(E)$ which in turn

is $O(V^2)$. At each step, a number of comparisons equal to the number of node children is required. The worst case scenario is verified when the following conditions holds at the same time: (i) graph G_O has depth 2, (ii) its nodes are uniformly distributed between level 2 and 3, (iii) nodes at the same level have the same fathers and the same children and, moreover, (iv) they have equal or not significantly different confidence scores. The conditions related to the morphology of the DAG ensure that the number of comparisons is maximum, and the condition on the confidence score guarantees that the procedure traverses all nodes.
All of the configurations of the algorithm input parameters enable us to select a set of possible true values. Since the aim of TD is to find the most expected solution, a method able to choose it is required. The ranking phase described in the next section is devoted to this.

4.2. True value ranking

Given the true value set selected in the previous step, we have to define a ranking method in order to select the $k \in \mathbb{N}^+$ most expected values where k is a fixed number. In our investigations, k is experimentally set, at the most, at 5. The solution set of most expected true values is indicated as $V^* \subseteq \mathcal{P}(V^*_{candidates})$, where $V^*_{candidates}$ is the value set returned by the selection phase.
Now we propose to rank the values based on rather:

- their IC. This method is useful for situations in which specific answers are expected and when there is not much uncertainty on the data item under consideration. Note that in the following experiments IC is a measure computed according to the definition provided by Seco based on the analysis of the partial ordering topology [30]. In particular, it takes advantage of the number of descendants of a value:

$$IC_{Seco}(v) = 1 - \frac{log(|descendants(v)| + 1)}{log(|V|)} \tag{1}$$

  where $|V|$ is a non-empty set since an ontology is considered to have at least one concept, i.e. the root value.
  IC has been proposed because the user generally expects very precise answers. Often general true values for a data item are well known *a priori*, i.e. it is well known that a person is born in a place. If two or more true values have the same IC, then random selection can be

done or, alternatively, another criterion can be used to rank this value subset.

- their source average trustworthiness, denoted $WA_{trust}(v)$. The rationale is that if a lot of unreliable sources support a false value A (increasing its confidence score – *Sums* does not normalize based on the number of sources claiming a value, therefore its confidence estimation is biased), and there are only a few reliable sources that support a true value B, then sources providing B should have higher average trustworthiness scores. This measure is obtained by computing the average trustworthinesses associated with sources that explicitly or implicitly claim to have a particular value v and by weighting it by a normalization factor:

$$WA_{trust}(v) = (1 - \frac{1}{\eta + |S^{v_+}|}) * avg_{trust}(v) \quad (2)$$

where the average of source trustworthiness is represented by $avg_{trust}(v)$, S^{v_d+} is the set of sources that implicitly or explicitly provide the value v_d and η is a small number used to avoid that $WA_{trust}(v) = 0$ when v is provided by only one source. The first factor, i.e. normalization, was introduced in order to tune the average w.r.t. the number of sources providing the value. Indeed, inspired by the study presented in [32], the higher the number of sources providing a value, the higher our confidence in the computed average should be.
Moreover in this case, if two values have the same WA_{trust}, then another criterion can be used to rank them.

Once the values are ranked, the next and final step of the post-processing procedure can be performed.

4.3. Filtering of top-k true values

The filtering phase collects the top k values in the rank and returns them to end-users. Before performing selection of the top k values, all the ranked ones have to be controlled. This is necessary because truth discovery models can be applied to different scenarios: high or low uncertainty situations, high or low risky cases in which making an error is, respectively, very dangerous or not. For instance, if truth discovery models are used to populate a medical knowledge base containing, for each symptom, all possible correlated diseases,

then the end-users want to be really careful in accepting a value as true. Therefore, based on the possible application contexts, different properties that the solution set V^* has to respect can be defined. In this way various true value sets with different characteristics can be identified:

- the solution set V^*_{ord} contains only values sharing partial ordering relationships; formally $\forall(x, y) \in V^*_{ord} \times V^*_{ord}, x \preceq y \vee y \preceq x$. The procedure to create a set containing values that respect this property is as follows: it iteratively selects and removes the first element in the ranked list. Then it adds this value to the solution set only if it is an ancestor or descendant of all elements that are already present in it.

- the solution set V^*_{disj} is composed of values that do not share any partial ordering relationships; formally $\forall(x, y) \in V^*_{disj} \times V^*_{disj}, \neg(x \preceq y) \wedge \neg(y \preceq x) \wedge \nexists\, w, z \in V^*_{candidates} w \preceq x \vee z \preceq y$. This means that all values in the solution set are the most specific among those returned by the selection phase. Indeed, only values that do not have descendants in the returned true value set belong to the solution. The procedure used returns a set of alternatives that are as much as possible very specific and different. In other words, this set of values consists of elements that do not have any of their exclusive descendants in the sorted list. For example, if the values returned by the previous step are *Europe*, *Continent*, *Country*, *City*, *Location*, then, in accordance with the partial order in Fig. 1, the V^*_{disj} is composed only of *Europe*, *Country* and *City*.

The first kind of solution can be desirable when there is not much uncertainty (end-users expect to easily find the true answers) or the end-users do not want to deal with potentially different values in a domain where they are not experts. The second property can be adopted when there is a lot of uncertainty and especially when the application context could result in making errors without dangerous consequences. Indeed, when there is uncertainty, to postpone the selection of true values to the end-users, avoiding to automatically select only a specific value and its ancestor, may be useful. In order to support the end-users final choice, returning a set of values composed of the most promising alternatives is important.
Obtaining V^*_{ord} is suitable when $\delta = 0$. Indeed, taking the value with the highest confidence at each step, the process ends with the selection of only one specific true value (and its ancestors). Considering this set of returned values, the first property is often verified without filtering any value out. In

any case, very general values often are not returned since only the top-k values are selected after the verification of the property. Otherwise, the second property, i.e. V^*_{disj}, is not useful considering $\delta = 0$. Only the single most specific value contained in the set of returned values is selected when this property must hold. Indeed, all of the others share partial ordering relationships. This corresponds to consider that $\delta = 0$, $k = 1$ and a solution set V^*_{ord}. Obtaining V^*_{disj} is preferable when $\delta = 1$. Indeed, in those cases all values having confidence higher than θ are returned by the selection phase, but for the final aim of truth discovery (finding the truth) it is suitable to only keep the set of "promising" alternatives that correspond to a set of values that are different and specific as much as possible.

5. Experiments

In this section we describe all experiments performed on synthetic datasets to confirm the validity of our approach. First, we focus on the synthetic datasets: how we generated them and what are the parameter settings we tested. Then, we present the evaluation methodology we used to evaluate the approach and to compare it with existing models.

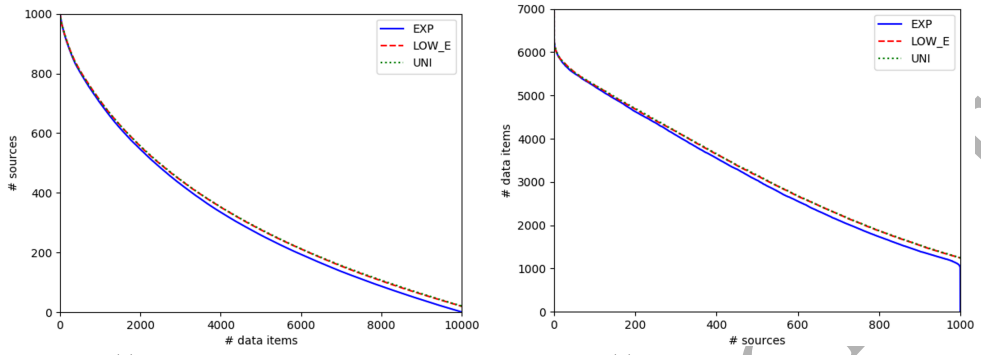
5.1. Datasets

In order to assess the behaviour of our approach w.r.t. the ontology used to derive the relationship among values, we integrate preliminary experiments carried out using the synthetic *birthPlace* datasets (see [10] for further details related to their generation) with additional ones performed using different partial ordering structures.

Each synthetic dataset contains a set of claims concerning a specific predicate, a set of sources and the subset of claims provided by each source. We

Table 2: Features of the different partial order structures.

Features	*CC*	*MF*	*BP*	*genre*	*birthPlace*
Values	3984	10243	28822	1838	682658
$\preceq$ depth (max depth)	12	15	16	8	14
Average depth	5,223	5,610	6,906	3,93	5,424
Average # of children	1,451	1,196	1,898	1,041	1,535
Max # of children	466	291	451	824	160194
Leaves	3016	8192	14797	1563	663373



(a) # of sources per data item (b) # of dataitems per source

Figure 3: Statistics of sources-data items for the *CC* datasets.

generate 5 different datasets considering the predicates *birthPlace* and *genre* from DBpedia [33], and the predicates *Cellular Component*(*CC*), *Molecular Function* (*MF*) and *Biological Process* (*BP*) from Gene Ontology [34]. All the datasets are randomly generated based on a ground truth (containing a set of true claims, for each predicate), a partial order among the values contained in the ground truth (an example is shown in Fig. 1) and a set of factitious sources. Note that Table 2 reports the features associated to the different partial ordering structures that we use. Given these elements the generation process can start.

First, a trustworthiness level is associated with each source. We assume that the majority of sources are sufficiently reliable and only a few of them are always or never correct. A Gaussian distribution with an average and standard deviation equal to, respectively, 0.6 and 0.4 was used to model the described behaviour.

Second, we reproduce the long-tail phenomena [35] for which many sources provide values for a few data items and a few sources provide values for many data items. This is modelled using a simple exponential distribution. It associates, with each source, the number of data items on which it has to provide a value. The statistic that confirm that this behaviour is respected by the datasets that were generated are reported in Fig. 3. In Fig. 3a we observe that approximatively 80% of data items are claimed by less than 500 sources. Fig. 3b shows that most of sources have provided at least 1000 data items.

Third, each source claims a true or false value for a specific data item w.r.t.

its trustworthiness. In case of true claims, the value is selected among the inclusive ancestors of the value contained in the data item. In the case of false claims, it is selected from the set of values that are neither inclusive ancestors nor descendants of the true one denoted v_d^*. In both cases the values are selected w.r.t. a similarity measure between the values and v_d^*.
For the selection of the true values, three different strategies were adopted: EXP, LOW_E, UNI (Fig. 4). EXP simulates cases in which sources are quite sure about the true values, so they tend to claim values similar to the expected one (contained in the ground truth) when they have to provide a true value. UNI reproduces a world where there is a lot of uncertainty, then the sources tend to indiscriminately select the value from the entire set of possible true values. LOW_E is a trade-off between the previous two types. Sources uniformly select the value from the set of possibilities, but there is a slightly higher probability of choosing values similar to the expected one. For instance, Fig. 4 reports, on the x axis, the values of Fig. 1 ordered according to their similarity measures w.r.t. the true value *Malaga*. Considering that v_d^* is *Malaga* and the EXP law, sources will more often provide values such as *Malaga*, *Spain* and *City* than values like *Continent* or *Location*. Otherwise, considering the UNI distribution, the probability of claiming these values will be the same.
For the selection of the false values, only a single strategy was considered. A source that has to provide a false claim tends to provide a false value that is

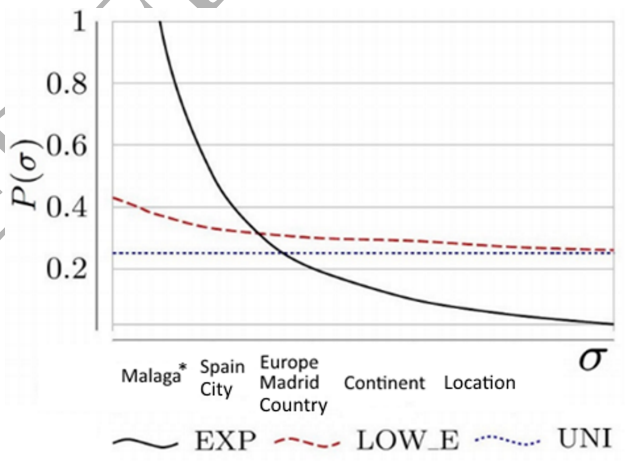


Figure 4: Distributions used to select true values.

Table 3: Set of experiments performed for each predicate.

Conf. Name	θ	δ	**Rank**		**Filter**
			1^{st}	2^{nd}	
TSbC$_{trust}$	0,..., 0.5	0	WA_{trust}	IC_{Seco}	V^*_{ord}
TSbC$_{IC}$	0,..., 0.5	0	IC_{Seco}	WA_{trust}	V^*_{ord}
TSaC$_{trust}$	0,..., 0.5	1	WA_{trust}	IC_{Seco}	V^*_{disj}
TSaC$_{IC}$	0,..., 0.5	1	IC_{Seco}	WA_{trust}	V^*_{disj}

similar to the expected true value. For instance, if the true value is *Malaga*, then a source provides the value *Portugal* with an higher probability than the value *Brazil*. Moreover, sources tend to claim the same false values. Therefore, the probability of a value to be selected as false one increases w.r.t. the number of sources that previously claimed it.
For each predicate, 20 datasets were produced w.r.t. the different laws that can be used to select the true values provided by sources. Further details on the generation of the ground truth and how the partial order of values has been derived are provided with the datasets at www.github.com/lgi2p/TDSelection.

5.2. Experiment settings

In order to provide robust results, considering each predicate, we generated 60 synthetic datasets (20 for each different distribution used to select the true values). Several experiments were conducted on them. Table 3 reports all of the experimental settings in which the datasets were tested. The name associated with each configuration indicates the delta setting. When $\delta = 0$ the approach is called TSbC (Truth Selection of the Best Child). Indeed, the selection algorithm chooses, at each step, the value with the highest confidence. In other words, it selects the best node among the children of the considered one. Otherwise, when $\delta = 1$ the approach is called TSaC (Truth Selection of all Children). Indeed, using this configuration the algorithm selects, at each step, all the children of the considered nodes. Moreover, the subscript specifies the first ranking criteria used, i.e. TSbC$_{IC}$ means that IC is used for the ranking phase as first criteria to order the values. For all the experiments, different threshold θ values were used: 0, 0.1, 0.2, 0.3, 0.4, 0.5. Note that when δ is equal to 0, we test only the property of the solution set indicating that its values share ordering relationships. Indeed, the selection procedure in this case chooses, at each step, only values having the highest confidence and therefore only a single most specific value and its ancestors

can be returned. No alternatives to the most specific value can be selected. When δ is equal to 1, we test only the property indicating that the values in the solution set do not share a partial order. The procedure may select more than one branch. In this situation, if we force the returned true values to share an ordering relationship, we oblige the algorithm to select only one path. Thus, the main advantage of this configuration, i.e. to propose a set of alternatives, is wasted.

For the confidence and trustworthiness estimations, we initialize the confidence value at 0.5 in order to start the iterative procedure, i.e. *AdaptedSums*. The stopping criteria used for the iterative procedure is the same as that used in the original paper of *Sums*[36]: the procedure was stopped after 20 iterations. The algorithms were implemented in Python 3.4. The experiments were performed on a PC with an Intel Core 2 Duo processor (2.93GHz×8GB). To give an idea, using the codebase developed for these experiments[4], memory consumption varies from 1.6 to 4.3 GB depending on the number of values composing the partial order. Using TSbC and TSaC, running times were, respectively, around 0.24 and 1.7 milliseconds per data item. Note that running times may increase significantly when partial orders have specific topological properties. In particular, optimizations have to be studied when dealing with partial orders having values with numerous children (hubs). For instance, the partial order used for the *birthPlace* predicate contains the value "Settlement" with 160 thousands children; running times using this partial order were 0.02 and 0.8 seconds per data item for TSbC and TSaC respectively. The source code and datasets associated with this study are open-sourced and published on the Web at the following link www.github.com/lgi2p/TDSelection. Note that for *Sums* and *AdaptedSums*, we used the code developed by [10]. Otherwise, for the experiments related to the other existing models, we used the DAFNA-EA[5] implementation [37]. This API provides the source code for the main existing models.

5.3. Evaluation

The evaluation of the model we proposed to select the true values was carried out using both traditional and hierarchical performance measures of classification problems.

[4]Note that this codebase has been developed for experimental purpose and was not optimized to lower memory consumption and running time.

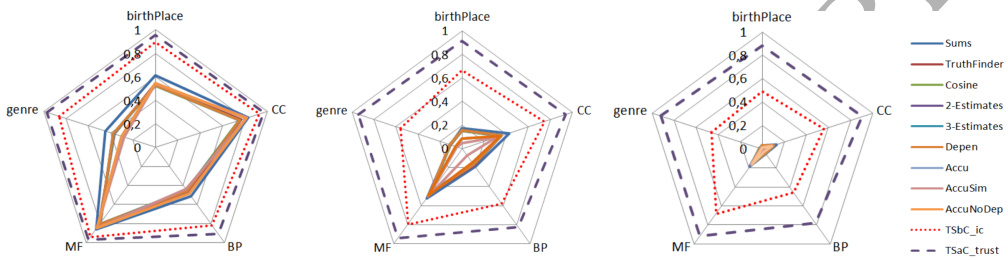
[5]www.github.com/daqcri/DAFNA-EA

Among traditional metrics, precision and recall were mainly used to compare our approach with the existing models that do not consider the partial order. Our positive class consists of all pairs $(d \in D, v_d^* \in V_d)$ where v_d^* is the value contained in the ground truth for the data item d, and the negative class is composed of all pairs $(d \in D, v_d \in V_d - v_d^*)$. Therefore, the precision is the proportion of pairs (d, v_d^*) returned by the approach among all the pairs it returns. The recall is the proportion of pairs (d, v_d^*) returned by the approach among all pairs contained in the ground truth.

The hierarchical evaluation measures (HEM) were used to analyse the behaviour obtained by different parameter settings of our approach. Indeed, hierarchical metrics distinguish the severity of different errors taking the hierarchy of classes into account. Reasonably if *Malaga* is the true value, then an approach that returns *Portugal* should be less penalized than another that returns *Brazil*. Indeed *Portugal* is in the same continent than *Malaga*, i.e. Europe, while *Brasil* is in a different continent, i.e. America. A detailed study related to hierarchical measures was presented in [38]. They distinguish the main dimensions that characterize hierarchical classification problems and suggest, for each possible combination, which are the best evaluation metrics to use. They recommend $F_{LCA}, P_{LCA}, R_{LCA}$ and $MGIA$ when dealing with single-label problem and DAG hierarchy. This situation corresponds to our initial problem settings: for each data item there is a single expected true value and our partial order among values is represented using a DAG. F_{LCA}, P_{LCA} and R_{LCA} are set-based measures. They use hierarchical relations to augment the sets of returned and true values and to compute precision and recall. Since adding ancestors over-penalize errors that occur to nodes with many of them, $F_{LCA}, P_{LCA}, R_{LCA}$ use the notion of the Lowest Common Ancestor to limit this undesirable effect. $MGIA$ is a pair-based metric that uses graph distance measures to compare returned and true values. Its limitation is that it does not change with depth. For further details related to the computation of these measure please refer to [38]. Now, we briefly describe the main characteristics of these hierarchical measures through an illustrative example. This enable the reader to better understand the result discussion in the next section. Considering the DAG in Fig. 1 and *Malaga* as the true value, the HEMs related to several returned values are reported in Table 4. As shown, if the returned value is more general than the expected one, then P_{LCA} is not affected, while R_{LCA} decreases when increasing the distance from the expected value. Otherwise, if the returned value is an error (neither the expected value nor more general one),

Table 4: Example of HEM considering the DAG in Fig. 1 and *Malaga* as the true value.

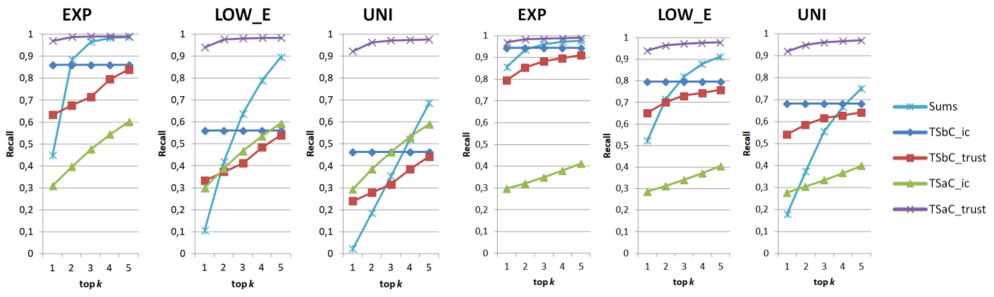
Returned value	P_{LCA}	R_{LCA}	F_{LCA}	$MGIA$
Malaga	1	1	1	1
Spain	1	0.5	0.7	0.9
Country	1	0.3	0.5	0.8
Madrid	0.5	0.5	0.5	0.8
France	0.5	0.3	0.4	0.7



(a) EXP (b) LOW_E (c) UNI

Figure 5: Recall obtained by applying our approaches TSbC_{IC} (dotted line) and TSaC_{trust} (dashed line), both with $k = 1$ and $\theta = 0$, and the models provided by DAFNA API (solid lines) on the synthetic datasets.

then P_{LCA} and R_{LCA} decrease w.r.t. the position of the returned value in the partial order. $MGIA$ indicates the distance among the returned value and the expected one without considering if one value is more general or specific than the other.



(a) genre (b) Molecular Function (MF)

Figure 6: Recall obtained by applying our approach and the proposed models (with $\theta = 0$) on the synthetic datasets w.r.t. the dataset type and number of returned values.

6. Results

All of the experimental settings presented in Section 5 were tested. Here, the results are presented and discussed. Note that a robust analysis were conducted given the artificial nature of the synthetic datasets.
Results show that our approach enables successfully addressing the problem of selecting true values. Recall that our study considers a setting where value confidence estimations w.r.t. the partial order of values monotonically increases. The most effective configuration settings of our selection procedure were TSaC_{trust} and TSbC_{IC} as shown in Fig. 6. These settings coupled with the *AdaptedSums* model were able to outperform, in terms of recall, existing truth discovery methods on the different datasets and predicates that were used for the experiments, see Fig. 5. Note that in these experiments we compared our post-processing strategies considering $k = 1$ with the other models. Indeed, the general aim of TD is to return a single answer for each data item.
In the rest of this section we detail the comparison of the proposed approach with existing truth discovery models and we study different configuration settings of the post-processing procedure analysing its behaviour considering different k, δ and θ values.
Both TSaC_{trust} and TSbC_{IC} obtained good performance, but TSaC_{trust} was the most robust approach independently of the predicate and dataset type, as shown in Fig. 5. It resulted to be only slightly influenced by source disagreement increase (UNI dataset case). Indeed, TSaC_{trust} aimed to analyse and compare the trustworthiness of sources providing the most specific values that do not share partial order relationships. This was done selecting and returning all provided values higher than θ, i.e. $\delta = 1$. Then ranking the values according to the weighted average trustworthiness of sources claiming them. Finally, filtering the first k values that did not share ordering relationships. Following this post-processing procedure, TSaC_{trust} performance was not affected when the number of sources providing true general values increased (UNI dataset). Precisely, analysing the recall obtained by the different models from EXP to UNI dataset types, we observed that, when increasing sources that provided general true values, TSaC_{trust} had a recall drop equal to 0.073 against a recall drop around 0.528 obtained by existing truth discovery models. Indeed, the average recall, over the different predicates, obtained by TSaC_{trust} was 0.954, 0.912 and 0.881 respectively for EXP, LOW_E and UNI dataset types. The average recall achieved by

existing truth discovery models was 0.595, 0.243 and 0.067 respectively for EXP, LOW_E and UNI dataset types.
On the contrary TSbC_{IC} performance was more influenced by source disagreement increase than TSaC_{trust} performance. It is the post-processing strategy that employed the greedy algorithm to select the true value, i.e. at each step the selection phase chooses the values with the highest confidence. Then it ordered them w.r.t. their IC. Finally, it kept only values that shared a partial order. Therefore, it used as selection criterion, at each step, the value confidence. When sources provided more general true values the information associated to these claims were propagated to less values. Thus the confidence estimations were less informative in the last steps of the procedure. Anyway, also TSbC_{IC} outperformed existing methods obtaining recall levels that were equal to 0.889, 0.670 and 0.531 for EXP, LOW_E and UNI dataset respectively (thus with a recall drop of 0.358).
Observing Fig. 5 we analysed for which predicates our approaches, TSaC_{trust} and TSbC_{IC}, obtained slightly lower performances. Even in these cases our models stills outperformed existing ones.
Considering TSaC_{trust}, the worst recall performance were achieved for *birthPlace* and *BP* predicate. Analysing the features shown in Table 2 related to the different predicate partial order, it is clear that this configuration setting was influenced by the average number of children in the partial order. Indeed *birthPlace* and *BP* were the two predicates with the highest children average number. Moreover, the ranking of predicates w.r.t. their recall corresponded to the predicate ranking w.r.t. the children average number in decreasing order.
Otherwise, when considering TSbC_{IC} approach the worst performance in terms of recall were obtained considering *genre* and *BP* predicate. We found out that TSbC_{IC} performance depended both on the children average number and the average depth of expected solutions w.r.t. the maximum depth. Indeed, at each step of TSbC_{IC} the probability of error is related to the number of alternatives among which the procedure can select a value. Moreover, it also related to the percentage of the partial order that the selection procedure has to traverse in order to reach the expected solutions w.r.t. the maximum depth. The probability of error increased when the part of the graph to traverse augmented. For instance *genre* predicate had the lowest children average number, but it obtained performance lower than *MF*, *CC* and *birthPlace* predicate. This because its expected values had a depth that required to traverse a bigger part of the partial order than in the other cases.

To better understand the best parametrization for the post-processing procedure several experiments were conducted w.r.t. the different settings reported in Table 3.

First of all, we compared the different post-processing strategies we proposed, evaluating the recall at different levels of k. The results are reported in Fig. 6. Note that we show the results for the predicates *genre* and *MF*, but a similar behaviour was obtained with all the others.

We observe that the best results were obtained by the TSaC_{trust} for any k value. It took advantage of the fact that it returned a set of alternatives as different as possible from each other and, at the same time, as specific as possible. Usually TSbC_{IC} also outperformed the baseline model (*Sums*), but for higher values of k it was worse than *Sums*. This is because we forced the result of TSbC_{IC} to share ordered relationships, while in the case of *Sums*, k values with the highest confidence were returned (no additional filter was applied on these values). Note that the recall of TSbC_{IC} did not improve when increasing the value of k. This means that a situation in which a returned value is more specific than the expected one never occurs. This is in accordance with the policy we adopted to generate the synthetic datasets. Given the expected value, we cannot say anything about its descendants. Each of them may be a true specification of the expected truth or not. Consequently, we removed all of the descendants from the set of possible true and false values. In other words, no sources provide a claim that contains one of the descendants of the expected value associated with the considered data item. Otherwise, in all the other configurations, increasing the number of values returned (k) enhanced the recall.

The TSaC_{IC} and TSbC_{trust} configurations were for the majority of cases worse than those of the baseline approaches. TSaC_{IC} consists of the selection strategy with $\delta = 1$, i.e. all provided values having confidence higher than θ are selected, and the use of IC as first ranking criterion. It obtained low performance because IC_{Seco} was not a good discriminator among values that did not share ordering relationships. Indeed it is based on the number of descendant values and it may happen in situations in which x is the expected value and y has the same father as x. If x has descendants, while y has none, y will be preferred by the ranking based on the IC_{Seco} even if it is not a true value. Thus, the WA_{trust} ranking is more suitable in these cases.

Otherwise TSbC_{trust} is a post-processing strategy with $\delta = 0$, i.e. at each step of the selection process only one value is selected, with the use of source average as ranking criterion. Obtaining low recall for this model means that

Table 5: HEM obtained for the different predicates w.r.t. the model and the threshold θ considered.

		Model											
		TSbC_{IC}						TSaC_{TRUST}					
Predicate	**HEM**	$\boldsymbol{\theta}$						$\boldsymbol{\theta}$					
		0	0.1	0.2	0.3	0.4	0.5	0	0.1	0.2	0.3	0.4	0.5
CC	F_{LCA}	**0.836**	0.826	0.770	0.694	0.617	0.561	**0.958**	0.890	0.783	0.690	0.613	0.560
	P_{LCA}	0.824	0.874	0.943	0.986	**0.991**	0.988	0.959	0.954	0.967	0.985	**0.989**	**0.989**
	R_{LCA}	**0.862**	0.812	0.693	0.568	0.469	0.407	**0.959**	0.861	0.701	0.563	0.465	0.406
	MGIA	0.879	**0.910**	0.907	0.890	0.855	0.818	**0.963**	0.945	0.917	0.887	0.851	0.816
MF	F_{LCA}	**0.878**	0.865	0.800	0.697	0.637	0.572	**0.962**	0.914	0.807	0.695	0.636	0.572
	P_{LCA}	0.870	0.907	0.960	0.990	**0.994**	**0.994**	0.964	0.965	0.971	0.989	**0.994**	**0.994**
	R_{LCA}	**0.898**	0.850	0.729	0.568	0.492	0.414	**0.963**	0.893	0.734	0.567	0.491	0.414
	MGIA	0.909	**0.937**	0.926	0.892	0.862	0.824	**0.966**	0.958	0.928	0.890	0.860	0.824
BP	F_{LCA}	**0.745**	0.689	0.620	0.540	0.484	0.438	**0.881**	0.725	0.607	0.527	0.477	0.436
	P_{LCA}	0.732	0.859	0.957	**0.979**	0.976	0.968	0.886	0.935	0.963	**0.976**	0.974	0.967
	R_{LCA}	**0.783**	0.624	0.494	0.391	0.335	0.293	**0.882**	0.642	0.481	0.379	0.329	0.291
	MGIA	0.792	**0.853**	0.836	0.774	0.707	0.641	**0.881**	0.865	0.815	0.754	0.696	0.635
birthPlace	F_{LCA}	**0.791**	0.773	0.709	0.640	0.587	0.532	**0.946**	0.855	0.713	0.627	0.576	0.530
	P_{LCA}	0.788	0.841	0.936	0.988	**0.993**	0.990	0.948	0.941	0.953	0.988	**0.991**	0.989
	R_{LCA}	**0.800**	0.744	0.601	0.483	0.424	0.372	**0.946**	0.813	0.602	0.469	0.414	0.369
	MGIA	0.909	**0.912**	0.897	0.877	0.845	0.807	**0.968**	0.948	0.900	0.869	0.838	0.805
genre	F_{LCA}	**0.784**	0.775	0.708	0.657	0.617	0.571	**0.963**	0.930	0.729	0.657	0.617	0.571
	P_{LCA}	0.781	0.791	0.855	0.979	0.995	**0.997**	0.966	0.952	0.878	0.980	0.994	**0.997**
	R_{LCA}	**0.793**	0.774	0.641	0.505	0.454	0.409	**0.962**	0.920	0.660	0.505	0.454	0.409
	MGIA	0.903	**0.904**	0.889	0.887	0.867	0.833	**0.974**	0.967	0.897	0.887	0.867	0.833

WA_{trust} was not a good discriminator to rank the values sharing partial order relationships returned by the selection phase.
Moreover, Fig. 6 shows that when disagreement among sources providing true values increased these two latter approaches (TSaC_{IC} and TSbC_{trust}) could be useful anyway. The recall they obtained for $k = 1$ was higher than the recall of *Sums* model. Therefore in case of high level of disagreement also a not optimal procedure can be advantageous.
As expected, in all the cases, the precision always decrease when increasing k. Moreover, comparing the different settings of the proposed approach, we observed that the ranking based on their precision performances was the same that the one obtained w.r.t. their recall. Therefore, we omit these repetitive results.
Our further analysis focused on models TSaC_{trust} and TSbC_{IC} since they were the models among the proposed ones that achieved the best performances. We examined the impact of different threshold values, setting $k = 1$, w.r.t. the hierarchical evaluation metrics: $F_{LCA}, P_{LCA}, R_{LCA}$ and $MGIA$. The results are reported in Table 5. Considering TSbC_{IC}, we noticed that, when slightly increasing θ, $MGIA$ increased in the majority of the cases. This occurred because there are expected values (supported by few reliable sources) with a confidence lower than false ones (supported by many unreliable sources), even though the former have a higher WA_{trust} than the latter. Thus, using TSbC_{IC} and $\theta = 0$, these values were selected as true values. Increasing θ allows the procedure to avoid a part of these errors. Indeed, eliminating the values with confidence score very low enables the procedure to return, with high probability, the father of the expected value. Anyway, further increasing the threshold caused a loss of $MGIA$ because the returned values result to be very general. This does not happen with TSaC_{trust} since this kind of errors are already overcome considering WA_{trust} as first ranking criterion.
Moreover, we observed that, in the majority of cases, when increasing θ the R_{LCA} always decreased, while the P_{LCA} always increased. Precisely, the highest R_{LCA} for both TSaC_{trust} and TSbC_{IC} was obtained with $\theta = 0$. The highest P_{LCA} was obtained for both approaches with different θ values depending on the predicate as shown in Table 5.
Summarising, the most effective configuration settings were TSaC_{trust} and TSbC_{IC}. They were both able to obtain better performance than existing truth discovery models. We noted that increasing the number of values returned for each data item allow increasing the performance. Nevertheless

this can be applied only in the case where a group of experts can select the true values among the ones proposed by the proposed approach for each data item. Otherwise, we have to force the parametrization $k = 1$. Regarding the threshold θ, a high θ value is recommended when the application scenario does not permit to assume many risks. In this case it is important to have a high precision. In other words, obtaining a general true value than the specific false one is preferred. Therefore, the different parameter settings of the proposed post-processing procedure allow dealing with different application scenarios taking their requirements into account.

7. Conclusion

In this paper, we have presented a post-processing procedure able to select true values after estimation of the value confidences using the *AdaptedSums* approach we proposed in our previous work. This general procedure can be used with any TD approach when partial order of values is taken into account as *a priori* knowledge. The post-processing process involves three main steps. The first one consists of the selection procedure. It aims to identify the set of possible true values using relationships among them and includes two parameters (δ and θ). Based on their tuning, different behaviours of the selection process can be obtained. The second step ranks the returned values of the selection phase. Finally, the third step permits to filter the top k values and ensure desirable properties (values that share or not relationships). The results confirmed our preliminary finding: using partial ordering of values helps to improve both source trustworthiness estimation, as already demonstrated by our preliminary study [10], and the true value identification. More precisely, the best results are obtained with the configuration of the algorithm that selects a set of alternatives, not sharing ordering relationships, and ranks them through the average trustworthiness of sources claiming those values. The results showed a similar behaviour on the datasets obtained by the two different ontologies (DBpedia and Gene Ontology).

As prospects, we envisage to incorporate our framework by adapting it to another existing model. Indeed we would like to show the flexibility of our approach and further enhance the results. Moreover, we intend to explore other kinds of additional information such as correlations among data items and values. For instance, usually the birth location of people is correlated with the language they speak. Therefore, if we know that a person speaks

Italian, we can increase the confidence in those claims that contain Italy as value for the bornIn predicate. More precisely, we plan to design a model that integrates, into existing approaches, information extracted from external knowledge bases in the form of rules. The idea is to add, in the confidence formula, a boosting factor indicating the confidence level of each claim according to the external knowledge base.

References

[1] J. Bleiholder, F. Naumann, Data fusion, ACM Computing Surveys (CSUR) 41 (1) (2009) 1.

[2] C. Li, V. S. Sheng, L. Jiang, H. Li, Noise filtering to improve data and model quality for crowdsourcing, Knowledge-Based Systems 107 (Supplement C) (2016) 96 – 103. doi:https://doi.org/10.1016/j.knosys.2016.06.003.
URL http://www.sciencedirect.com/science/article/pii/S0950705116301666

[3] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, J. Han, A survey on truth discovery, SIGKDD Explorations 17 (2) (2015) 1–16.

[4] T. Knap, J. Michelfeit, M. Necaskỳ, Linked open data aggregation: Conflict resolution and aggregate quality, Computer Software and Applications Conference Workshops (COMPSACW), 2012 IEEE 36th Annual (2012) 106–111.

[5] A. Guzman-Arenas, A.-D. Cuevas, A. Jimenez, The centroid or consensus of a set of objects with qualitative attributes, Expert Systems with Applications 38 (5) (2011) 4908 – 4919. doi:https://doi.org/10.1016/j.eswa.2010.09.169.
URL http://www.sciencedirect.com/science/article/pii/S0957417410011267

[6] L. Berti-Equille, J. Borge-Holthoefer, Veracity of data: From truth discovery computation algorithms to models of misinformation dynamics, Synthesis Lectures on Data Management 7 (3) (2015) 1–155.

[7] D. Wang, T. Abdelzaher, L. Kaplan, Social sensing: building reliable systems on unreliable data, Morgan Kaufmann, 2015.

[8] G. Zhou, J. Zhao, T. He, W. Wu, An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities, Knowledge-Based Systems 66 (2014) 136 – 145. doi:https://doi.org/10.1016/j.knosys.2014.04.032.
URL http://www.sciencedirect.com/science/article/pii/S0950705114001543

[9] B. Khaleghi, A. Khamis, F. O. Karray, S. N. Razavi, Multisensor data fusion: A review of the state-of-the-art, Information Fusion 14 (1) (2013) 28 – 44. doi:https://doi.org/10.1016/j.inffus.2011.08.001.
URL http://www.sciencedirect.com/science/article/pii/S1566253511000558

[10] V. Beretta, S. Harispe, S. Ranwez, I. Mougenot, How can ontologies give you clue for truth-discovery? an exploratory study, in: Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, WIMS 2016, Nîmes, France, June 13-15, 2016.

[11] X. Yin, J. Han, S. Y. Philip, Truth discovery with multiple conflicting information providers on the web, IEEE Transactions on Knowledge and Data Engineering 20 (6) (2008) 796–808.

[12] R. W. Ouyang, L. M. Kaplan, A. Toniolo, M. Srivastava, T. J. Norman, Aggregating crowdsourced quantitative claims: Additive and multiplicative models, IEEE Transactions on Knowledge and Data Engineering 28 (7) (2016) 1621–1634.

[13] R. W. Ouyang, M. Srivastava, A. Toniolo, T. J. Norman, Truth discovery in crowdsourced detection of spatial events, IEEE Transactions on Knowledge and Data Engineering 28 (4) (2016) 1047–1060.

[14] J. Pasternack, D. Roth, Latent credibility analysis, in: Proceedings of the 22nd international conference on World Wide Web, ACM, 2013, pp. 1009–1020.

[15] B. Zhao, B. I. Rubinstein, J. Gemmell, J. Han, A bayesian approach to discovering truth from conflicting sources for data integration, Proceedings of the VLDB Endowment 5 (6) (2012) 550–561.

[16] X. Yin, W. Tan, Semi-supervised truth discovery, Proceedings of the 20th international conference on World wide web (2011) 217–226.

[17] R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, D. Srivastava, Fusing data with correlations, Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD '14 (2014) 433–444arXiv:1503.00306, doi:10.1145/2588555.2593674.

[18] G.-J. Qi, C. C. Aggarwal, J. Han, T. Huang, Mining collective intelligence in diverse groups, Proceedings of the 22nd International Conference on World Wide Web, WWW '13 (2013) 1041–1052.

[19] X. Wang, Q. Z. Sheng, X. S. Fang, L. Yao, X. Xu, X. Li, An integrated bayesian approach for effective multi-truth discovery, Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15 (2015) 493–502.

[20] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, Y. Cheng, Truth discovery on crowd sensing of correlated entities, Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, SenSys '15 (2015) 169–182.

[21] S. Wang, L. Su, S. Li, S. Hu, T. Amin, H. Wang, S. Yao, L. Kaplan, T. Abdelzaher, Scalable social sensing of interdependent phenomena, Proceedings of the 14th International Conference on Information Processing in Sensor Networks, IPSN '15 (2015) 202–213.

[22] X. L. Dong, L. Berti-Equille, D. Srivastava, Integrating conflicting data: the role of source dependence, Proceedings of the VLDB Endowment 2 (1) (2009) 550–561.

[23] L. Berti-Equille, A. D. Sarma, Xin, Dong, A. Marian, D. Srivastava, Sailing the Information Ocean with Awareness of Currents: Discovery and Application of Source Dependence, CIDRarXiv:0909.1776.

[24] X. L. Dong, L. Berti-Equille, D. Srivastava, Truth Discovery and Copying Detection in a Dynamic World, Vldb 2 (1) (2009) 562–573.

[25] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, H. Liu, Exploitation of physical constraints for reliable social sensing, Proceedings of the 2013 IEEE 34th Real-Time Systems Symposium, RTSS '13 (2013) 212–223.

[26] S. Wang, D. Wang, L. Su, L. Kaplan, T. F. Abdelzaher, Towards cyber-physical systems in social spaces: The data reliability challenge, Real-Time Systems Symposium (RTSS), 2014 IEEE (2014) 74–85.

[27] A. Bronselaer, M. Szymczak, S. Zadrony, G. D. Tr, Dynamical order construction in data fusion, Information Fusion 27 (Supplement C) (2016) 1 – 18. doi:https://doi.org/10.1016/j.inffus.2015.05.001. URL http://www.sciencedirect.com/science/article/pii/S1566253515000391

[28] D. C. Kozen, The Design and Analysis of Algorithms, Springer-Verlag New York, Inc., 1992.

[29] A. V. Aho, M. R. Garey, J. D. Ullman, The transitive reduction of a directed graph, SIAM J. Comput. 1 (2) (1972) 131–137.

[30] N. Seco, T. Veale, J. Hayes, An intrinsic information content metric for semantic similarity in wordnet, Proceedings of the 16th European conference on artificial intelligence (2004) 1089–1090.

[31] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, Semantic similarity from natural language and ontology analysis, Synthesis Lectures on Human Language Technologies 8 (1) (2015) 1–254.

[32] P.-A. Jean, S. Harispe, S. Ranwez, P. Bellot, J. Montmain, Uncertainty detection in natural language: a probabilistic model, in: Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, WIMS 2016, Nîmes, France, June 13-15, 2016.

[33] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference, ISWC'07/ASWC'07, Springer-Verlag, Berlin, Heidelberg, 2007.

[34] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al., Gene ontology: tool for the unification of biology, Nature genetics 25 (1) (2000) 25–29.

[35] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, J. Han, A confidence-aware approach for truth discovery on long-tail data, Proceedings of the VLDB Endowment 8 (4) (2014) 425–436.

[36] J. Pasternack, D. Roth, Knowing what to believe (when you already know something), Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10 (2010) 877–885.

[37] D. A. Waguih, L. Berti-Equille, Truth discovery algorithms: An experimental evaluation, CoRR abs/1409.6428.

[38] A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, I. Androutsopoulos, Evaluation measures for hierarchical classification: a unified view and novel approaches, Data Mining and Knowledge Discovery 29 (3) (2015) 820–865.