Truth Finding by Attribute Reliability Estimation for Heterogeneous Data

Wenwen Sheng and Hong Shen
School of Information Science and Technology, Sun Yat-Sen University, China
Email: 771920866 [AT] qq.com

Hong Shen
School of Computer Science, University of Adelaide, Australia

Abstract—**In the era of big data, data veracity is one of the most challenging problems. One important task in big data integration is to derive the most accurate records from noisy and conflicting data records collected from multiple sources. However, data sources may process a set of properties with inconsistent reliabilities, e.g., height and weight of a patient are more likely to be true than profession in medical records, departure and landing time of a flight are more likely to be true than weather in airline records. In a cloud computing environment, discrepancies among data describing the same object appear more common because of the increased degree of data replication and unknown trustiness of servers storing the data in a cloud. Besides, we observed that the difficulty to provide truth for different entity is quite different. In this paper, we propose an ARTF model to estimate attribute reliabilities with heterogeneous data types and update it with the entity hardness automatically. The property trustworthiness will be more precise in describing source reliability, which in turn will achieve a better precision in inferring the truth. We compare the performance of our method to the state-of-art truth discovery methods through a real world dataset and a synthetic dataset respectively, the experimental results show that our algorithm can process source conflicts much more accurately while reducing the convergence rate.**

Keywords-***truth finding, heterogeneous data types, entity hardness, attribute reliability estimation***

I. INTRODUCTION

The truth discovery problem can be formulated as follows: given a set of assertions claimed by multiple sources for a set of objects, find the truth claim for each object and compute the reliability of each source. Fig. 1 show the many-to-many relationship among sources, values and objects. In the era of big data, more and more sources automatically publish information with unknown credibility, which leads the veracity problem more challenging. For example, personal contact lists stored in multiple servers in the cloud environment undergo asynchronous updates, information extraction tools return one or more answers to an information extraction task and different tools might lead to different answers, many independent users with different trustworthiness or expertise assign tags to objects in social networks, multiple agencies report weather at the same time, multiple trading venues provide stock market transactions, and a number of companies provide real-time traffic conditions and so on. Observation errors over the data may occur due to the asynchronous update, sensor damage, personal skills, transmission faults and malicious modification by hackers, all leading to data conflict among data sources. Retrieval of inaccurate data will result in the inability to do the right schedule in real-life applications, such as personal communication, object recognition, weather forecast, stock trend predication, avoidance to crowded traffic. So finding the most trustworthy data from the conflicting sources is important.

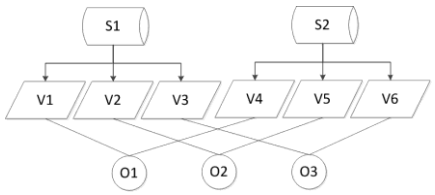

Figure 1: The many-to-many relationship among sources, values and objects.

Several ways have been proposed to resolve the source conflict problem [2], [17], [27]. The initial methods are majority voting for categorical data and mean or median computation for the continuous values. All these methods assume that the reliability of all the sources is equal and all the observations for the same object are equally treated. But in reality, some sources are more trustable than others, such as experts are more reliable than amateurs in crowdsourcing, students specialize in different subjects, data centers are partly paralyzed and the increased degree of data replication under the cloud computing environment and so on. Majority voting may result in inaccuracy aggregation results when there are more unreliable values than reliable ones. So it will help us infer the truth more accurate if we consider the source reliability which is not known a prior.

Motivated by the advantage of estimating the source reliability, many truth discovery methods have been proposed to process source reliability and truth deduction iteratively iteratively[1], [2], [8], [5], [6], [10], [11], [15], [16], [14], [20], [21], [22], [23], [24], [28], [31], [32], [33]. Most of these methods was designed for single data type, or calculate the source reliability including all properties. But in reality, there are usually multiple heterogeneous data types (categorical, continuous, text, graph or more complicated data types) in one source. In this case, many of the former methods will not be

applicable. Furthermore, the reliability of attributes may inconsistent in one source. For example, workers may have different reliability levels based on their expertise on different crowdsourcing tasks; students are good at different subjects other than every subject. So using attribute reliability rather than source credibility will help us getting more accurate truths. Besides, the hardness of every fact (the propensity of sources to be wrong on this fact) is different, just as the basic and reinforcement questions in one exam. Generally, if a student is good at the reinforcement questions, he should be more credible in this subject.

Object	Text Analysis	Digital	Logical	Material
Question 1	essay21	9	A	picture21
Question 2	essay22	12	B	picture22
Question 3	essay23	13	C	picture23

Object	Text Analysis	Digital	Logical	Material
Question 1	essay11	8	B	picture11
Question 2	essay12	12	B	picture12
Question 3	essay13	14	A	picture13

(a) Susan database
(b) Mike database

Object	Text Analysis	Digital	Logical	Material
Question 1	essay31	8	A	picture31
Question 2	essay32	12	C	picture32
Question 3	essay33	11	C	picture33

(c) Leo database

Table 1: Quiz answers of Susan, Mike and Leo

Object	Text Analysis	Digital	Logical	Material
Question 1	essay1	8	C	picture1
Question 2	essay2	14	B	picture2
Question 3	essay3	11	C	picture33

Table 2: Ground Truth of Quiz

The survey above motivates us to propose an Attribute Reliability Estimation and Truth Finding in Heterogeneous Data (ARTF) framework to infer the truths from multiple conflicting sources by calculating property reliabilities, and the hardness of each question. The framework also can take all data types into consideration. The property reliability will be more precise in describing source trustworthiness, which in turn will achieve a better precision in inferring the truth.

We formulate the problem as an optimization problem to minimize the overall weighted deviation between the identified truths and the input. We find out the truths and attribute reliability by solving the joint optimization problem. In our evaluation, we show the performance of our method outperforms the previously-proposed CATD framework and several other fact finders because these methods either applied consistent reliability on all properties, or didn't take the hardness of facts into consideration.

We summarize the main contributions of our work as follows.

- We propose a general optimization framework to model the conflict resolution problem on inconsistent property reliability in one source by taking property weights and fact hardness into consideration. The objective function calculates the overall deviation between observations and unknown truth while modeling the property weights.

- We propose an algorithm to solve the optimization problem by iteratively updating truths, property weights. We propose methods to calculate the deviation of different kinds of data type, including continuous data, discrete data, text, image and video.

- We preform experiments on both real-world and synthetic data sets, and the results show that our method performs better in resolving conflicts from multi-source and multi-property data, which demonstrates the interest of taking into account the reliability of attributes and the hardness of facts.

The rest of the paper is organized as follows: Related work is reviewed in Section 2, and then we will introduce our model and algorithms in Section 3. Evaluation results will be presented in Section 4. We will conclude the paper in Section 5.

II. RELATED WORK

Truth discovery also known as Veracity Problem [31], source trustworthiness estimation [11], information Corroboration [24], data fusion [7], conflicting data Integration [33], or knowledge fusion [7], has been extensively studied. [2] surveyed early approaches and [17], [27] compared recent approaches.

The Veracity problem about how to discover true facts from conflicting information was first formalized by Yin et al. [31] and they proposed a iterative method called TRUTHFINDER to jointly infer the truth values and source quality. [5] applied Bayesian analysis to detect dependence between data sources. [12] derived a model to evaluate trustworthiness based on attribute groups because they think attribute reliability is inconsistent in one source. Pasternack and Roth developed several web-link based algorithms and proposed a linear programming based algorithm [23]. They also introduced a generalized fact-finding framework to incorporate additional information into the truth finding process [22]. [9] studied how to select a subset of sources before integration to balance the quality of integrated data and integration cost. [25] solved the problem of source selection by considering dynamic data sources whose content changes over time. [14] proposed a transition model to capture sophisticated patterns of value transitions. [33] proposed a probabilistic graphical model to automatically infer true records and source quality without any supervision, this paper also was the first to merge multivalued attribute types. [30] proposed a Bayesian approach to tackle the multi-truth-finding problem. [24] discussed several correlation types between the

sources, such as copying, negative correlation and positive correlation. [7] studied the applicability and limitations of different fusion techniques on knowledge fusion problem. [16] proposed a CRH model to resolve conflicts among multiple sources with heterogeneous data types by modeling all properties in a unified model. They also proposed a CATD model to detect truths from conflicting data with long-tail phenomenon by considering the confidence interval of the estimation [15]. [29] proposed an approximate truth discovery approach via dividing sources and values into groups to deal with the large scale challenge.

There are also some works on streaming data. Jia and Wang proposed an incremental strategy adaptive for different update situations, boosting like ensemble classifier [13]. Zhao and Cheng proposed a probabilistic model that transforms the problem of truth discovery over data streams into a probabilistic inference problem [34]. Wang and Kaplan proposed a streaming fact-finding method that recursively updates previous estimates based on new data [28]. [18] proposed an incremental truth discovery framework to dynamically update object truths and source weights upon the arrival of new data.

The above works jointly inferred truth and source quality using different approaches on different aspects. Our approach is different in two aspects. First, we minimize the deviations of all sources by referring to attribute reliability and fact hardness. Second, we adapt inferring truth from various data types to generate accurate estimates of reliability. We will compare this approach with some other methods in our experiment.

III. METHODOLOGY

In this section, we present our ARTF model. The model iteratively updates property weights and truths from multi-source data. We resolve the truth finding problem as an optimization problem. The optimization solution updates the truths and attribute reliabilities by minimize the weighted deviation summation between the truth and observations. We present several hardness calculation methods and loss functions to complete the attribute weight assignment and truth computation procedure.

A. Basic Definitions

In this part, let us introduce the related concepts and define the problem to be solved. We assume that every object has only one correct value and many possible wrong values. We use an example on Susan database (Table 1(a)) to explain these concepts. The related terminologies and mathematical notations are shown in Table 3 and Table 4.

Given all the data sources, we aim to find the most trustworthy value for every entity, and infer the reliability degree of each property simultaneously. Note that a higher w_{kn} in Table 4 indicates the attribute n is more reliable than other attribute in source k and observations from this attribute are more likely to be accurate. This is under the observation that if a fact is provided by many trustworthy sources, it is more likely to be true. Furthermore, a source that provides mostly true facts will likely provide true facts for other objects. The source reliability and fact confidence are determined by each other and true facts are more consistent than false facts, so it is likely to find the true facts at the end.

As for the approaches that incorporate source reliability estimation, they either conduct on one type property or on all the properties together. The former will result in inaccurate reliability because of insufficient observations while the later cannot distinguish the quality of different properties in one source.

In the example shown in Table 1, if we only deal with concrete and continuous data types , there are two attributes (Text Analysis, Material Analysis) can't be processed. If we use the source reliability, the reliability degrees of Source 1 (Susan database) and Source 3 (Leo database) is approximate nevertheless Source 1 is more accurate in Text Analysis property and Source 3 is more accurate in Digital and Logical Analysis property. The answers to Question 3 in Digital Analysis are different from each other, which increases the hardness to get the truth. So the attribute who gets answer for harder questions should acquire a higher reliability relatively. In contrast, the attribute who gets answer from easier questions should acquire a lower reliability.

To characterize this phenomenon, the proposed framework calculated property reliability and every entry truth by iteratively minimizing the deviation summation of different type entries using attribute reliability which is regulated by fact hardness. In our framework, we take the collection of observations made by all the sources as the input. The outputs are property weight list and a truth table. The initial truths are generated by voting and mean methods. The hardness and deviation calculation methods will be stated later. The iteration procedure will stop if the successive truth table difference is under the threshold δ. We will discuss about δ later.

B. The ARTF Framework

We propose the following optimization framework ARTF that utilizes property weight to describe the reliability of source.

With property reliability updated periodically, the more reliable the attribute is, the closer the observations to the truths. Thus we should minimize the summation of weighted deviations from the truths to the multi-source observations, where the weights reflect the reliability degree of properties. Summing up, we propose the following optimization framework:

$$min_{S^{(*)},W} f\left(S^{(*)}, W\right) = \sum_{k=1}^{K}\sum_{n=1}^{N} w_{kn} d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right), \text{s.t.}\, \varepsilon(\mathrm{W}) = 1 \quad (1)$$

Concept	Explanation
object	A person or thing of interest. e.g., "Question 1".
property	An attribute to describe the object. e.g., "Text Analysis".
source	Describes the place where information about objects' properties can be collected. e.g., Susan database.
observation	The data describing a property of an object from a source. e.g., Text Analysis's Questioin 1 from Susan database is essay11"
entry	A property of an object. e.g., "Text Analysis's Question 1"
truth	Accurate information of an entry, which is unique. e.g., the real answer of Text Analysis's Question 1.

Table 3:Summary of terminologies

Notation	Description
K	Number of sources
N	Number of property
M	Number of objects
W	The trustworthiness list of all attributes in all sources $\{w_{11} \dots w_{1N}, w_{21} \dots w_{2N}, \dots \dots w_{K1} \dots w_{KN}\}$
$v_{nm}^{(k)}$	The observation of the n-th property for the m-th object made by the in k-th source
$v_{nm}^{(*)}$	The truth for the n-th property of the m-th object
d_n	The deviation function for the n-th property
w_{kn}	The trustworthiness of the n-th attribute in the k-th source
$S^{(k)}$	The collection of observations made on all the objects by the k-th source $\{v_{11}^k \dots v_{1M}^k, \dots \dots, v_{N1}^k \dots v_{NM}^k\}$
$S^{(*)}$	Set of truth for all objects on all properties $v_{11}^* \dots v_{1M}^*, \dots \dots, v_{N1}^* \dots v_{NM}^*$
δ	The threshold of successive truth table difference

Table 4: Summary of notations

Algorithm 1: Truth estimation Algorithm

Input: Observations made by K sources
$\{S^1, \dots, S^K\}$
Output: The true value for each object
$S^* = \{v_{nm}^*\}_{n=1,m=1}^{N,M}$
And property weights
$\{w_{11} \dots w_{1N}, w_{21} \dots w_{2N}, \dots \dots w_{K1} \dots w_{KN}\}$
1: Initialize the truths $S^{(*)}$; //using voting method
2: Calculate hardness of every entry using (2);
3: repeat
4: Update attributes weights according to (3) to Reflect attributes' reliability based on the Estimated truths and the hardness of claims;
5: for k=1 to K do
6: for n=1 to N do
7: Update the truth of the m-th object on The n-th property $v_{nm}^{(*)}$ according to (4) based on the current estimation of Attributes weights;
8: end for
9: end for
10: until Convergence criterion is satisfied;
//the successive truth table difference is under the
// threshold
11: return $S^{(*)}$ and W;

Through minimizing the function, we are going to obtain two sets of variables $S^{(*)}$ and W alternately. $S^{(*)}$ correspond to the set of truths and W represent property weights. Loss function d_n measures the deviation from the observation $v_{nm}^{(k)}$ to the truth $v_{nm}^{(*)}$. It output a high value if the deviation is high and low value otherwise. Weight w_{kn} reflects the trustworthiness of the n-th property in the k-th source. The higher of w_{kn}, the more trustable of the property. Naturally, the truths will rely on the property with higher weights to minimize the overall deviations. $\varepsilon(W)$ reveals the distributions of property weights. It constrains the weights into a certain range to rationalize the optimization problem.

We iteratively conduct three steps to get attribute weights and the truths through a joint procedure.

First, fact hardness calculation. We calculate the hardness of every observation in truth table by computing the dispersion degree of the corresponding observations in INPUT sources. We will discuss dispersion calculation method in Section 3.3.

$$\text{Hardness}\left(v_{nm} = \text{f}\left(\text{Dispersion}\left(v_{nm}^k\right)\right)\right) \quad (2)$$

Second, attribute weight update. We fix the truths value, and compute attribute weights based on the difference between the truths and the observations made by the attribute, and then adjust the weight according to the hardness of corresponding observations:

$$W \leftarrow \arg minf\left(S^{(*)}, Hardness(v_{nm})\right) \quad (3)$$

Third, truth update. We fix the weight w_{nm} of each attribute, and we update the truth for each entry to minimize the weighted difference between the truth and the attributes'

observations. By computing the truth for every entry, we can obtain the collection of truths $S^{(*)}$.

$$v_{nm}^{(*)} \leftarrow \arg minf(W, \{d_{kn}\}) \quad (4)$$

Implementation of this framework is given in Algorithm 1. We will elaborate the three steps using example functions in the following.

C. Hardness Calculation

Proposition 1. The entity hardness is presented by the dispersion level of the observations. The higher of dispersion level, the harder of the entity.

Example 2. *The answers' selection probabilities are same for one question. If the dispersion level is high, it indicates that the correct rate is low. If most of the students' answers are consistent, it is more likely to indicate that this question is quite easy. But we don't deny there are some exception cases that popular answers are wrong which is quite rarely.*

There are K sources in the INPUT altogether, so there are K observations for one entity at most. Now we present several hardness calculation methods for different data types.

As for categorical data, we add up the occurrence frequency of each term. If the maximum frequency is less than K/2, then the dispersion level is high, and this entity will be labeled as hard. Otherwise, the entity will be labeled as easy. As for continuous data, first we divide the values into several numerical intervals, and then we add up the occurrence frequency of each interval. As for text data, we draw keywords of each text, and add up the occurrence frequency of each keyword. As for image data, first we extract features, then build index, at last we search for the features and add up the occurrence frequency of each feature. If the maximum frequency is less than K/2, then the dispersion level is high, and the entity will be labeled as hard. Otherwise, the entity will be labeled as easy.

Example 3. *There are three sources in table 1, so K=3, K/2 =2. For Digital Analysis attribute, the first entity maximum frequency is 2 (value 8), not less than K/2, not hard. The second entity maximum frequency is 3 (value 12), above K/2, not hard. The third entity maximum frequency is 1, less than K/2, hard. So there is 1 hard label and 2 easy labels in Digital attribute. Similarly, there are 3 easy labels in Logical attribute, 3 hard labels in Text attribute, and 3 hard labels in Material attribute.*

D. Attribute Weight Assignment

First, attribute weight assignment. Since attribute weight assignment is similar to source weight assignment, we use the following regularization function to compute the property weight by constraining the summation of formula $\exp(-w_{kn})$:

$$\varepsilon(W) = \sum_{k=1}^{K}\sum_{n=1}^{N} \exp(-w_{kn}) \quad (5)$$

Theorem 4. *Suppose the truths are fixed, the optimization problem (1) with constraint (5) is convex, and the global optimal solution is given by*

$$w_{kn} = \log\left(\frac{\sum_{n=1}^{N}\sum_{m=1}^{M} d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right)}{\sum_{k'=1}^{K}\sum_{n'=1}^{N}\sum_{m'=1}^{M} d_n\left(v_{n'm'}^{(*)}, v_{n'm'}^{k'}\right)}\right) \quad (6)$$

Proof. Since the truths are fixed, (1) has only one set of variables W. We assume a variable $t_{kn} = \exp(-w_{kn})$ to prove the convexity of the optimization problem (1). Then (1) can be expressed as follows:

$$\min f(t_{kn}) = \sum_{k=1}^{K}\sum_{n=1}^{N} -\log(t_{kn}) . \sum_{n=1}^{N}\sum_{m=1}^{M} d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) \quad (7)$$

$$\text{s.t.} \sum_{k=1}^{K}\sum_{n=1}^{N} t_{kn} = 1 \quad (8)$$

The objective function of (7) is a linear combination of negative logarithm functions, and the constraint is linear in t_{kn}, so (7) is convex. Thus, the optimization problem (1) with constraint (8) is convex, and any local optimum is also global optimum [26].

Then we use the Lagrange multipliers to solve (7). The Lagrangian of (7) is as follows:

$$L\left(\{t_{kn}\}_{k=1,n=1}^{K,N}, \gamma\right) = \sum_{k=1}^{K}\sum_{n=1}^{N} -\log(t_{kn}) . \sum_{n=1}^{N}\sum_{m=1}^{M} d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) + \gamma\left(\sum_{k=1}^{K}\sum_{n=1}^{N} t_{kn} - 1\right) \quad (9)$$

where γ is a Lagrange multiplier. Let the partial derivative of Lagrangian with respect to t_{kn} be 0, and we can get:

$$\sum_{n=1}^{N}\sum_{m=1}^{M} d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) = \gamma t_{kn} \quad (10)$$

From the constraint that $\sum_{n=1}^{N}\sum_{m=1}^{M} t_{kn} = 1$, we can derive that

$$\gamma = \sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{m=1}^{M} d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) \quad (11)$$

Plugging (11) and w_{kn} =- log (t_{kn}) into (10), we obtain (6).

Since we have calculated the deviation of all the entry in front, then we can compute the attribute weights directly using (8).

This weight calculation formula indicates that an attribute with observations which are closer to the truths will have a greater weight. Therefore, (7) is a reasonable constraint function by leading to meaningful attribute weight assignment formula.

Second, weight regulation. As we stated above, we should adjust the attribute reliability according to the fact hardness label obtained by Section 3.3 to acquire a more accurate truth table. If we get α hard labels and β easy labels in n-attribute, the reliability of the corresponding attributes can be adjusted as:

$$w_{kn} = w_{kn} * \left(\frac{M+\alpha}{M+\beta}\right) \quad (12)$$

Equation (12) shows that the attribute who get answer for harder questions should acquire a higher reliability relatively. In contrast, the sources who get answer for easier questions should acquire a lower reliability.

Example 5. *We calculated the deviations using (13)~(18), calculated the attribute weights through (6). The attribute weights are {(0.45, 0.83, 0.84, 0.52); (0.65, 0.73, 0.84, 0.68); (0.75, 0.69, 0.87, 0.74)} respectively. Given the hardness labels in Example 3 {(3,0); (2,1); (0,3); (3,0) / (α , β)}, with the (12), we can get the regulated weights are {(0.9, 1.04, 0.42, 1.04); (1.3, 0.91, 0.42, 1.36); (1.5, 0.86, 0.43, 1.48)}. Material analysis is the most reliable property in the Leo database as it provides few errors and answers more hard questions.*

E. Truth Computation

The attribute weights are fixed, and the truth computation (4) is depended on the data type and loss function. We will introduce several truth computation methods for categorical data, continuous data, text data, image data and video data.

The most commonly used loss function for categorical data is 0-1 loss in which an error is incurred if the observation is different from the truth. Formally, if the n-th property is categorical, the deviation from the truth $v_{nm}^{(*)}$ is defined as:

$$d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) = \begin{cases} 1 & v_{nm}^{(*)} \neq v_{nm}^{(k)}, \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

We plug (13) into the objective function in (1), and then we can obtain the following formula:

$$v_{nm}^{(*)} = \arg min_v \sum_{k=1}^{K} \sum_{n=1}^{N} w_{kn} . d\left(v, v_{nm}^{(k)}\right) \quad (14)$$

This formula indicate that based on 0-1 loss function, to minimize the objective function, the truth should be the value that receives the highest weighted votes among all possible values.

Similarly, the loss function for continuous data is (15), indicating that we can use weighted mean method to calculate the truth. The truth could be the weighted mean summation of all the observations.

$$d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) = \left|v_{nm}^{(*)} - v_{nm}^{(k)}\right| \quad (15)$$

As for text data, the loss function is (16), indicating that we can use weighted *cosine similarity* method [3] to calculate the truth.

$$d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) = \frac{v_{nm}^{(*)} * v_{nm}^{(k)}}{\left|v_{nm}^{(*)}\right| * \left|v_{nm}^{(k)}\right|} \quad (16)$$

As for image data, the loss function is (17), indicating that we can use weighted *SITF (Scale Invariant Feature Transform)* [19] method to calculate the truth.

$$d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) = \text{SITF}\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) \quad (17)$$

As for video data, the loss function is (18), indicating that we can use weighted *PSNR (Peak signal-to-noise ratio)* [4] method to calculate the truth.

$$d_n\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) = \text{PSNR}\left(v_{nm}^{(*)}, v_{nm}^{(k)}\right) \quad (18)$$

This computation follows the principle that an observation stated by reliable sources will be more likely to be regarded as the truth. If the difference between the successive truth table is below the threshold δ twice, then the iteration procedure ends. We assume δ is one tenth of the difference.

IV. EXPERIMENTS

We apply our ARTF algorithm on a real-world dataset and a synthetic dataset. The experimental results show that the proposed method is efficient and outperforms state-of-the-art conflict resolution methods.

A. Experiment Setup

a) Performance Measures

The problem background is that we have multi-sets of observations with heterogeneous data types and the ground truths for each object on each property. Ground truths are only used for evaluation. All methods are implemented in an unsupervised form. In this experiment, we consider categorical, continuous and text data types. We use Error Rate, Distance and Cos as the performance measures of a method for these data types respectively. Error Rate is computed as the proportion of the method's output inconsistent with the ground truth for categorical data. Distance is computed as the mean of normalized absolute distance from the method's output to the ground truths. Cos is computed as the cosine similarity between the output of the method to the ground truths, then reverse the result. For all measures, the lower the value, the better performance of the method.

b) Baseline Methods

In our ARTF model, we use weighted voting, weighted median, and weighted cosine similarity as their deviation function. We compute attribute weights using (6). We use the following methods to compare with our approach. As CRH and CATD are the latest research results and outperform most approaches, we mainly compare our method with them.

- Voting: This method takes the value provided by a large percentage of sources as the truth without source reliability. This method can only be applied on categorical data.
- Mean: This method takes the mean of all observations as the truth without source reliability. This method can only be applied on continuous data.
- CRH: This method iteratively calculates the source weight and truth by minimizing the weighted deviation between the truths and observations. It can be applied on heterogeneous data types.
- CATD: This method detects truths from conflicting data with long-tail phenomenon by considering the source reliability and confidence interval of the estimation. It can be applied on numerical data type.

We re-implement all the baselines in MATLAB R2013a under a common implementation to test their performance as accurately as possible. We ran experiments on windows PC with Intel Core i7 processor.

B. Experiment Results

a) Real-world Data Sets

German Credit Data Set. This dataset was provided by Professor Dr. Hans Hofmann from Institut f"ur Statistik und "Okonometrie Universit" at Hamburg. It contains 1000 instances and 20 properties (8 categorical properties, 7 numerical properties, 5 text type properties) from 10 sources. The ground truths are also provided.

We summarize the performance of all the methods in table 5. We evaluate the performance separately on categorical, continuous and text data types, using Error Rate, Distance and Cos respectively. We can observe that the proposed ARTF approach achieves better performance comparing with all the baselines. This is because the baseline methods either failed to take entity hardness into consideration or can't deal with heterogeneous data types with imprecise source reliability. By the comparison we can see that ARTF can model source reliability more accurately by inferring attribute reliability and adjust the reliability by entity hardness. This also justifies our assumption that property reliability is more accurate than sources reliability.

Method	Error Rate	Distance	Cos
ARTF	**0.2651**	**2.547**	**0.3481**
CRH	0.3418	2.839	0.3642
CATD	0.3941	NA	NA
Voting	0.3519	NA	NA
Mean	NA	3.412	NA

Table 5:Performance Comparison on German Credit Data Set

Method	Error Rate	Distance	Cos
ARTF	**0.3816**	**4.4812**	**0.1587**
CRH	0.4641	4.5149	0.2375
CATD	0.4869	NA	NA
Voting	0.4921	NA	NA
Mean	NA	5.0322	NA

Table 6: Performance Comparison on Simulated Data Set

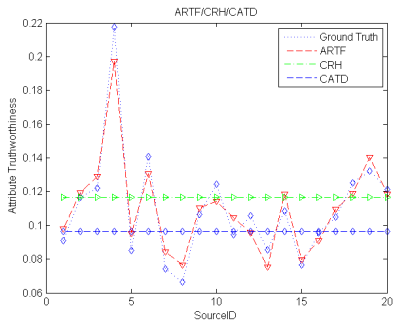

Figure 2: Comparison of Attribute Reliability Degrees with Ground Truth

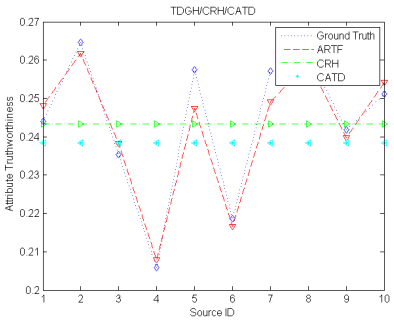

Figure 3: Comparison of Source Trustworthiness between Methods Applied on Diabetes Data

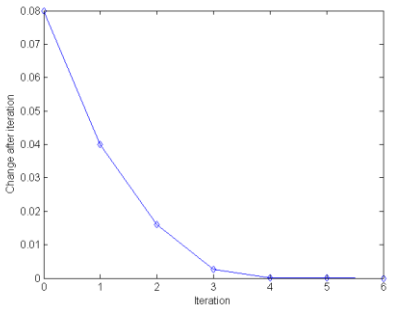

Figure 4: Convergence Rate

Now we show the source reliability degrees estimated for the 9 sources by all methods. First, we compute the true reliabilities by comparing how close the observations made by sources with the grounds truths. We show the source reliability degrees in Fig. 2.

We can observe from Fig. 2 (we normalize all the scores into the range [0, 1] and their total sum is 1 to make them comparable) that the property reliability degree estimated by ARTF is a bit more consistent with the true reliabilities. The difference is not obvious because the three properties are not particularly inconsistent. Yet by effectively characterizing different property reliabilities, our method can still distinguish reliable property from unreliable ones, and derive the truth based on credible attributes.

b) Noisy Multi-source Simulations

We conduct experiments on simulated data set to test the performance of our method by varying property reliabilities. The simulated dataset Diabetes and ReutersCon is obtained from Weka-3-7 (data mining software) datasets. It's also the ground truth of our experiment. This dataset has eight continuous properties, one categorical property and one text type property, 768 objects, and 6912 observations. We generate a dataset consisting of 10 multiple conflicting sources by injecting different kinds of noise into different properties of ground truth. We take the variation dataset as the input to our approach and baseline methods. We change the data randomly to generate the input data source. A parameter α is used to control the reliability degree of each property (a lower α indicates the property is altered in a lower chance, we use α

=0.1 to 0.5). In this way, we simulated a dataset with property reliability in various degrees in all data sources.

Table 5 shows all the results on this dataset. It can be observed that ARTF outperforms all the other methods in truth estimation by inferring accurately property reliability degrees. Baseline approaches cannot estimate source reliability accurately because they didn't take the property reliability inconsistent and entity hardness into consideration.

From the results we have the following observations: First, the plots show that the ARTF framework outperforms existing conflict resolution techniques, which ignore the unique reliability degree of each property in one source. When property reliabilities are approximate, ARTF will achieve similar performance as CRH and CATD. However, when the property reliability degree is inconsistent in one source, ARTF will have a much better performance. Second, it is more efficient to detect truths when we have more reliable properties in one source.

c) Efficiency

We now show the convergence rate using Diabetes dataset. Fig. 4 shows the change of the difference of the truth table variation. We can see that the variation decreases fast at the first four iterations and then reaches a stable stage, showing that the proposed method converges quickly in practice. As we propose the data in one pass, so that our approach has linear complexity in the number of observations.

V. CONCLUSION

In this paper, we proposed a method called ARTF to solve the truth finding problem with heterogeneous property data types. We observed that the property reliability is very likely to be inconsistent in a source. ARTF will describe the source trustworthiness more precisely by calculating property trustworthiness and regulating it with entity hardness. Experiments on one real world dataset and one simulated dataset showed that our method have better performance than the state-of-the-art truth finding methods when property reliability is not consistent with each other. We also introduced several deviation and hardness calculation methods.

There are still interesting challenges in this problem. Our method is based on the intuition that properties are independent with each other. However this assumption may not always apply (e.g., a person's title may have relationship with his age, a person ages 18 is more likely to be a student than a professor). We will extend our method to take into consideration of the relationship among properties in the future.

ACKNOWLEDGMENT

This work is supported by Research Initiative Grant of Sun Yat-Sen University under Project 985, National Science Foundation of China under its General Projects funding #61170232, Australian Research Council Discovery Project DP150104871. The corresponding author is Hong Shen.

REFERENCES

[1] Bahadir Ismail Aydin, Yavuz Selim Yilmaz, Yaliang Li, Qi Li, Jing Gao, and Murat Demirbas. Crowdsourcing for multiple-choice question answering. In Twenty-Sixth IAAI Conference, 2014.

[2] Jens Bleiholder, Karsten Draba, and Felix Naumann. Fusem: exploring different semantics of data fusion. In Proceedings of the 33rd international conference on Very large data bases, pages 1350–1353. VLDB Endowment, 2007.

[3] Courtney Corley and Rada Mihalcea. Measuring the semantic similarity of texts. Unt Scholarly Works, pages 13–18, 2005.

[4] Johannes F De Boer, Barry Cense, B Hyle Park, Mark C Pierce, Guillermo J Tearney, and Brett E Bouma. Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography. Optics letters, 28(21):2067–2069, 2003.

[5] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: the role of source dependence. Proceedings of the VLDB Endowment, 2(1):550–561, 2009.

[6] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Truth discovery and copying detection in a dynamic world. Proceedings of the VLDB Endowment, 2(1):562–573, 2009.

[7] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wiko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. From data fusion to knowledge fusion. Proceedings of the VLDB Endowment, 7(10), 2014.

[8] Xin Luna Dong and Felix Naumann. Data fusion: resolving data conflict -s for integration. Proceedings of the VLDB Endowment, 2(2):1654–1655,2009.

[9] Xin Luna Dong, Barna Saha, and Divesh Srivastava. Less is more: Selecting sources wisely for integration. In Proceedings of the 39th international conference on Very Large Data Bases, pages 37− 48. VLDB Endowment, 2012.

[10] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating information from disagreeing views. In Proceedings of the third ACM international conference on Web search and data mining , pages 131–140. ACM, 2010.

[11] Liang Ge, Jing Gao, Xiaoyi Li, and Aidong Zhang. Multi-source deep learning for information trustworthiness estimation. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 766–774. ACM, 2013.

[12] Manish Gupta, Yizhou Sun, and Jiawei Han. Trust analysis with clustering. Www Ser Www ', pages 53–54, 2011.

[13] Li Jia, Hongzhi Wang, Jianzhong Li, and Hong Gao. Incremental truth discovery for information from multiple data sources. In Web-Age Information Management, pages 56–66. Springer, 2013.

[14] Furong Li, Mong Li Lee, Wynne Hsu, and Wang-Chiew Tan. Linking temporal records for profiling entities. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pag es 593–605. ACM, 2015.

[15] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. Proceedings of the VLDB Endowment, 8(4):425–436, 2014.

[16] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. 2014.

[17] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the deep web: is the problem solved? In Proceedings of the 39th international conference on Very Large Data Bases, pages 97–108. VLDB Endowment, 2012.

[18] Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. On the discovery of evolving truth. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 675–684. ACM, 2015.

[19] Tony Lindeberg. Scale invariant feature transform. Scholarpedia, 7(5):10491, 2012.

[20] Adway Mitra and Srujana Merugu. Reconciliation of categorical

opinions from multiple sources. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pages 1561–1564. ACM, 2013.

[21] Aditya Pal, Vibhor Rastogi, Ashwin Machanavajjhala, and Philip Bohannon. Information integration over time in unreliable and uncertain environments. In Proceedings of the 21st international conference on World Wide Web, pages 789–798. ACM, 2012.

[22] Jeff Pasternack and Dan Roth. Knowing what to believe (when you already know something). In Proceedings of the 23rd International Conference on Computational Linguistics, pages 877–885. Association for Computational Linguistics, 2010.

[23] Jeff Pasternack and Dan Roth. Making better informed trust decisions with generalized fact-finding. In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three, pages 2324–2329. AAAI Press, 2011.

[24] Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh Srivastava. Fusing data with correlations. Sigmod, 2014.

[25] Theodoros Rekatsinas, Xin Luna Dong, and Divesh Srivastava. Characterizing and selecting fresh data sources. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pages 919–930. ACM, 2014.

[26] S.Boyd and L.Vandenberghe. Convex optimization. Cambridge University Press, 2004.

[27] Dalia Attia Waguih and Laure Berti-Equille. Truth discovery algorithms: An experimental evaluation. arXiv preprint arXiv:1409.6428, 2014.

[28] Dong Wang, Tarek Abdelzaher, Lance Kaplan, and Charu C Aggarwal. Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In Distributed Computing Systems (ICDCS),2013 IEEE 33rd International Conference on, pages 530–539. IEEE,2013.

[29] Xianzhi Wang, Quan Z Sheng, Xiu Susie Fang, Xue Li, Xiaofei Xu, and Lina Yao. Approximate truth discovery via problem scale reduction. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pages 503–512. ACM, 2015.

[30] Xianzhi Wang, Quan Z Sheng, Xiu Susie Fang, Lina Yao, Xiaofei Xu, and Xue Li. An integrated bayesian approach for effective multi-truth discovery. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pages 493–502. ACM, 2015.

[31] Xiaoxin Yin, Jiawei Han, and Philip S Yu. Truth discovery with multiple conflicting information providers on the web. Knowledge and Data conflicting information providers on the web. Knowledge and Data Engineering, IEEE Transactions on, 20(6):796–808, 2008.

[32] Xiaoxin Yin and Wenzhao Tan. Semi-supervised truth discovery. In Proceedings of the 20th international conference on World wide web, pages 217–226. ACM, 2011.

[33] Bo Zhao, Benjamin IP Rubinstein, Jim Gemmell, and Jiawei Han. A bayesian approach to discovering truth from conflicting sources for data integration. Proceedings of the VLDB Endowment, 5(6):550–561, 2012.

[34] Zhou Zhao, James Cheng, and Wilfred Ng. Truth discovery in data streams: A single-pass probabilistic approach. 2014.