Knowledge-Based Systems 187 (2020) 104828

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Truth finding by reliability estimation on inconsistent entities for heterogeneous data sets[☆]

Hui Tian[a], Wenwen Sheng[b], Hong Shen[b,c,*], Can Wang[a]

[a] School of Information and Communication Technology, Griffith University, Australia
[b] School of Information Science and Technology, Sun Yat-Sen University, China
[c] School of Computer Science, University of Adelaide, Australia

ARTICLE INFO

Article history:
Received 30 March 2017
Received in revised form 27 June 2019
Accepted 29 June 2019
Available online 9 July 2019

Keywords:
Truth finding
Attribute reliability
Object reliability
Group reliability
Entity hardness
Probability graphical model

ABSTRACT

An important task in big data integration is to derive accurate data records from noisy and conflicting values collected from multiple sources. Most existing truth finding methods assume that the reliability is consistent on the whole data set, ignoring the fact that different attributes, objects and object groups may have different reliabilities even wrt the same source. These reliability differences are caused by the hardness differences in obtaining attribute values, non-uniform updates to objects and the differences in group privileges. This paper addresses the problem how to compute truths by effectively estimating the reliabilities of attributes, objects and object groups in a multi-source heterogeneous data environment. We first propose an optimization framework TFAR, its implementation and Lagrangian duality solution for Truth Finding by Attribute Reliability estimation. We then present a Bayesian probabilistic graphical model TFOR and an inference algorithm applying Collapsed Gibbs Sampling for Truth Finding by Object Reliability estimation. Finally we give an optimization framework TFGR and its implementation for Truth Finding by Group Reliability estimation. All these models lead to a more accurate estimation of the respective attribute, object and object group reliabilities, which in turn can achieve a better accuracy in inferring the truths. Experimental results on both real data and synthetic data show that our methods have better performance than the state-of-art truth discovery methods.

1. Introduction

With the rapid developments of big data and smart city, the need to integrate the true values on heterogeneous data observed from multiple sources together is becoming an urgent task because of the increasing unreliability in object data and observation sources. Reliability inconsistency exists widely in different levels and dimensions. First, apparently observations from different sources for an object may differ from each other due to the differences in data capture ability of the sources, resulting in a many-to-many relationship among Source-Value-Object as illustrated in Fig. 1. Moreover, reliabilities of different attributes of an object set wrt the same source may also be different because of the observation hardness differences of the attributes wrt the source (e.g. an RFID reader may have 0.99 reliability for barcode but only 0.1 for TID). Similarly in an orthogonal dimension, different objects (records) may also carry different reliabilities wrt the same source because of their differences in data entry and maintenance (e.g. the records updated frequently may have a higher reliability than those updated infrequently). Finally, we also observe that object reliability is consistent within a group if we divide objects into groups such that all objects in the same group have the same reliability. Examples of group reliability includes privilege groups for online services and user groups in social networks. These reliability inconsistencies will result in source data conflicts and increase the hardness for obtaining the truths for objects.

For truth finding from conflicting data, most existing methods [1–3] based on majority voting and mean computation for categorical and continuous data respectively took no consideration of source reliabilities and unrealistically treated all observations from all sources equally. Voting selects the majority claims among all the observations as the truth, while mean computation takes the mean of all observations as the truth.

When taking into account of source reliabilities, different truth discovery methods have been proposed [4–9], all aimed to utilize

[☆] No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.06.036.

[*] Corresponding author at: School of Computer Science, University of Adelaide, Australia.

E-mail addresses: hui.tian@griffith.edu.au (H. Tian), 771920866@qq.com (W. Sheng), hongsh01@gmail.com (H. Shen), can.wang@griffith.edu.au (C. Wang).

https://doi.org/10.1016/j.knosys.2019.06.036
0950-7051/© 2019 Elsevier B.V. All rights reserved.

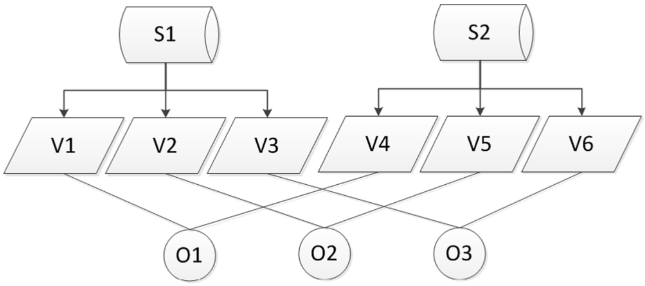

Fig. 1. The many-to-many relationship among sources, objects and values.

Table 1
Quiz answers of Susan, Mike and Leo.

Object	Digital analysis	Logical A	Material A
(a) Susan database			
Question 1	8	B	picture11
Question 2	12	B	picture12
Question 3	14	A	picture13
(b) Mike database			
Question 1	9	A	picture21
Question 2	12	B	picture22
Question 3	13	C	picture23
(c) Leo database			
Question 1	8	A	picture31
Question 2	12	C	picture32
Question 3	11	C	picture33

Table 2
Ground truth of Quiz.

Object	Digital analysis	Logical A	Material A
Question 1	8	C	picture1
Question 2	14	B	picture2
Question 3	11	A	picture3

some sort of specifications about the sources and applied the same basic heuristic idea: a claim is likely to be true if it is provided by trustworthy sources (especially if by many of them) and a source is trustworthy if most its claims are true. Based on this idea, most methods attempted to assign larger weight to reliable sources as they are more important when inferring the truths. These methods however applied the same source reliability to all attributes for each source and are hence unable to distinguish the quality of observations to different attributes from the same source.

We use an example in Table 1 Quiz answers to explain these concepts. *In the data sources shown in Table 1, if we only deal with concrete and continuous data types, the Material attribute cannot be processed. If we use the source reliability, the reliability degrees of Source 1 (Susan database) and Source 3 (Leo database) are approximate. Nevertheless, Source 1 is more accurate in Logical Analysis attribute and Source 3 is more accurate in Digital Analysis attribute. The answers to Question 3 in Digital Analysis are different from each other, which increases the hardness to get the truth. So the attributes that get answers for harder questions should have a higher reliability and for easier questions a lower reliability* wrt the ground truth in Table 2.

Existing methods ignored the fact that the same source's reliability may vary significantly among different attributes or objects (records). This motivates our work of this paper to investigate more effective methods for truth finding by reliability estimation on heterogeneous data. We first propose an optimization model TFAR, Truth Finding by Attribute Reliability estimation, to infer the truths by estimating the reliabilities of heterogeneous attributes, and the hardness of attribute observation. We obtain a solution for computing an optimal attribute weight (reliability) assignment that minimizes the total deviation between the truths and the observed values. Then we propose a Truth Finding by Object Reliability estimation model (TFOR) using a Bayesian probabilistic graphical model to infer the object reliabilities and truths. We formulate the derivation of the model's parameters as a Maximum Likelihood Estimation problem and apply Collapsed Gibbs Sampling to jointly infer the object reliabilities and truths. Finally we propose another optimization model TFGR for Truth Finding by Group Reliability Estimation to detect trustworthy claims from conflicting observations by estimating the (object) group reliability for the given group properties. We obtain its solution by minimizing the overall weighted deviation between inferred truths in the *i*th time (iteration of the deductive procedure) and the source observations to find the final truths. The above three models achieve a more accurate fine-grained source reliability estimation on attributes, objects and object groups respectively.

In our experimental evaluation, we show that our methods outperform the state-of-the-art truth-finding baselines that considered neither attribute reliability differences among all attributes nor object reliability differences among different objects for a source.

The main contributions of this paper are the proposed three mathematical models with their detailed implementation algorithms and solutions to solve the reliability conflict resolution problem for truth finding at attribute, object and object group three levels respectively, as summarized below:

- We propose a general optimization framework for truth finding on inconsistent attribute reliabilities by taking attribute weights and fact hardness into consideration.
- We propose a probabilistic graphical model for truth finding on inconsistent object reliabilities by incorporating quality measurement into object reliability.
- We propose a general optimization framework for truth finding on inconsistent object group reliabilities by iteratively updating group weights.
- We empirically show that our models outperform the existing methods for conflict resolution with three real-world datasets, which demonstrates the importance of taking into consideration reliability differences among attributes, objects and object groups for truth finding on heterogeneous data.

The reminder of this paper is organized as follows: In Section 2 we review the related work. Our proposed models and algorithms are introduced in Section 3, Sections 4 and 5. Section 6 presents the evaluation results. Section 7 concludes the paper.

2. Related work

The truth finding (conflict resolution) problem was first studied by Yin et al. [10] who proposed a TRUTHFINDER method to iteratively infer the truth values and source quality, and it has now been extensively studied. Existing work can be classified according to the specifications used to measure the source reliability.

Data source specification. The source selection problem identifies the subset of sources that maximizes the profit from integration. Rekatsinas et al defined a set of time-dependent metrics to characterize the quality of integrated data [11]. Dong et al. proposed an approaches of applying Bayesian analysis to decide dependence between sources [12] and select a subset of sources before integration to balance the quality of integrated data and integrated cost [13]. Li et al. studied the long-tail phenomenon