SLFTD: A Subjective Logic Based Framework for Truth Discovery

Danchen Zhang[1(✉)], Vladimir I. Zadorozhny[1], and Vladimir A. Oleshchuk[2]

[1] School of Computing and Information, University of Pittsburgh, Pittsburgh, USA
{daz45,viz}@pitt.edu

[2] Department of Information and Communication Technology, University of Agder, Kristiansand, Norway
vladimir.oleshchuk@uia.no

Abstract. Finding truth from various conflicting candidate values provided by different data sources is called truth discovery, which is of vital importance in data integration. Several algorithms have been proposed in this area, which usually have similar procedure: iteratively inferring the truth and provider's reliability on providing truth until converge. Therefore, an accurate provider's reliability evaluation is essential. However, no work pays attention to "how reliable this provider continuously providing truth". Therefore, we introduce subjective logic, which can record both (1) the provider's reliability of generating truth, and (2) reliability of provider continuously doing so. Our proposed methods provides a better evaluation for data providers, and based on which, truth are discovered more accurately. Our framework can handle both categorical and numerical data, and can identify truth in either a generative or discriminative way. Experiments on two popular real world datasets, Book and Population, validates that our proposed subjective logic based framework can discover truth much more accurately than state-of-art methods.

Keywords: Data fusion · Truth discovery · Subjective logic

1 Introduction

Data conflict is a common problem in data management area. For example, for a given flight, different websites may report different departing time. Figuring out the (most likely) truth from conflicting values provided by different sources is an important and challenging task. Naive methods, such as voting, do not consider the data provider's reliability, and hence may fail in particular cases. Therefore, many methods [1–11] paying attention to accurately evaluate trustworthiness of data provider are proposed. With provider reliability considered, these methods then identify the truth usually by selecting the value with the maximum averaged provider reliability.

However, in past studies, provider's reliability is usually evaluated in a probabilistic logic, which uses an evidence based probability (ranging from 0 to 1) to

T. Welzer et al. (Eds.): ADBIS 2019, CCIS 1064, pp. 102–110, 2019.
https://doi.org/10.1007/978-3-030-30278-8_13

represent people's opinion. For example, after observing the flipping the coin for hundreds of times, people believe the probability of "head" is 0.5, and believe the probability of "tail" is 0.5, too. However, when sample size is too small, the probability is unreliable. In such a situation, Subjective Logic (SL), proposed by Jøsang [12], can provide more information for this situation. With SL, an opinion from a person p towards a statement s can be represented by a triple $\omega_s^p = \{t, d, u\}$, with $t, d, u \in [0, 1]^3$, and $t + d + u = 1$, where t means trust, d means distrust, and u means uncertainty. With a too small sample, we may use $\{0.3, 0.3, 0.4\}$ to describe our uncertainty and impression towards the coin than a simple 0.5. In terms of truth discovery, SL allows us to record our trust (of provider providing truth) and certainty (of provider continuously doing so) towards each provider. In turn, we can identify truth more accurately.

To summarize, our paper has following major contributions: (1) our study is the first to pay attention to "how reliable the provider is able to continuously provide truth"; (2) SL is first introduced to truth discovery area, and it can perfectly records above mentioned two kinds of reliability; (3) The experiments on two popular real world dataset show that, compared with state-of-art methods, our framework can improve the truth discovery performance by a large degree.

2 Related Works

In truth discovery area, the simplest mechanism is voting, which does not consider the provider's reliability. Many studies show that a good evaluation of provider reliability can improve the performance largely. In [1], Dong et al. proposed to use Accuracy, which is calculated as the probability of each value being correct, and average the confidence of facets provided by the source as the provider trustworthiness. After that, they proposed the concept of AccuracySimilarity, which further considers the similarity of two values. In [2], authors proposed POPAccuarcy, which differs from Accuracy by releasing the assumption that false value probability is uniformly distributed. Another popular method is the TruthFinder, proposed by Yin et al. [5], which differs from Accuracy by not normalizing the confidence score of each entity. In [8], Pasternack et al., proposed three methods: (1) AverageLog is a transformation of Hub-Authority algorithm, with source trustworthiness being the averaged confidence score of provided values multiplying the log of provided value count; (2) Investment, where the confidence score of the value grows exponentially with the accumulated providers' trustworthiness. (3) PooledInvestment, where the confidence score of the value grows linearly. In [4] authors proposed a semi-supervised reliability assessment method, SSTF. It is basically a PageRank method assuming that there is a set of entities having the true value, which will affect the result in the PageRank iteration. [3], proposes 2-Estimates, which is a transformation of Hub-Authority algorithm, whose provider trustworthiness is the average instead of the sum of the vote count. They further proposed 3-Estimates, which additionally considers the value's trustworthiness. Another group, involving CRH [9], CATD [10] and GTM [11], can generate the truth for data with numerical values, and they can also be adapted to categorical dataset with slight modification. They share a similar idea, trying to generate/select the true value of each entity to minimize the difference between "estimated true value" and the "observed input value". Additionally, CATD is designed to smoothly predict truth on the long tail data with chi-squared distribution.

SL is a powerful decision making tool extending the probabilistic logic by including uncertainty and subjective belief ownership [12]. It is widely used in trust network analysis, conditional inference, information provider reliability assessment, trust management in sensor networks, etc. SL uses subjective opinions to express subjective beliefs about the truth of propositions with degrees of uncertainty.To the best of our knowledge, our work is the first one applying it to area of reliable truth discovery.

3 SL Based Framework for Truth Discovery (SLFTD)

Consider a dataset that contains a set of entities $E = \{e_1, e_2, ..., e_n\}$, and a set of data providers $P = \{p_1, p_2, ..., p_m\}$, the value of entity e_i provided by provider p_j is named as v_{ij}, constructing the value set V. Different providers may provide different values for same entity, and truth discovery aims to find the true value for each entity.

In the procedure of identifying the truth, our proposed framework involves three steps: (1)evaluate the provider's reliability score and entity's discrimination score in an iterative way, then (2) for each provider, SL based opinions are constructed based on the converged scores, and (3) the true value are inferred based on the fused opinions in either a discriminative manner or a generative manner. In the framework, two SL operations are utilized, and more detail can be referenced in [12].

- **Recommendation**. Assume two persons, A and B: A has an opinion towards B, and B has an opinion towards a statement s. Then according to B's recommendation, A can generate an opinion towards this statement s, described as $\omega_s^{AB} = \omega_B^A \otimes \omega_s^B = \{t_s^{AB}, d_s^{AB}, u_s^{AB}\}$.
- **Consensus**. If two persons A and B have opinions towards one statement s, then consensus operator $\oplus$ can be used to combine their opinions, described as $\omega_s^{A,B} = \omega_s^A \oplus \omega_s^B = \{t_s^{A,B}, d_s^{A,B}, u_s^{A,B}\}$.

3.1 Accurately Infer Providers' Reliability

Our first step is to iteratively evaluate the provider's reliability score and entity's discrimination score. In this study, the degree, to which the algorithm can infer the true value of the entity in an undisputed convincing manner, is defined as the **entity discrimination ability**. The entity whose majority candidate values are from reliable providers and are very similar/close to each other should be given a higher score. Thus discrimination score of entity E_i is defined as:

$$Disc(E_i) = \frac{\sum_{P_j, P_l \in P_{E_i}} Rel(P_j)Rel(P_l)Imp(V_{il} \rightarrow V_{ij})}{\sum_{P_j, P_l \in P_{E_i}} Rel(P_j)Rel(P_l)}, \tag{1}$$

where P_{E_i} is the set of providers that gives value on entity E_i; $Rel(P_j)$ is the reliability score of provider P_j, which will be described later. Please notice that $Imp(V_{il} \rightarrow V_{ij})$ reflects the implication from $V_{i.}$ to V_{ij}, introduced from [5]. It is a value (ranging from 0 to 1) reflecting to what degree V_{ij} is (partially) true if V_{il} is correct. At the first iteration, all provider have equal weights; and after each round, normalization is conducted with weights summing up to 1.

When measure the **providers' reliability**, more attention is paid to providers' performance on entities with higher discrimination score. Given such entities, if the values from this provider obtain lots of implications from other values of same entity, this provider reliability should be boosted; otherwise, should be lower down. Such impact from entities with low discrimination score should be relatively discounted. Thus reliability score of provider P_j is defined as:

$$Rel(P_j) = \frac{\sum_{E_i \in E_{P_j}} Disc(E_i)Imp(V_{i.} \rightarrow V_{ij})}{\sum_{E_i \in E_{P_j}} Disc(E_i)}, \tag{2}$$

where E_{P_j} is the set of entities to whom P_j gives value; and $V_{i.}$ consists of all candidate values of entity E_i. Also, normalization is conducted in the end of each iteration. This iterative procedure will continue until all scores converge.

3.2 Construct SL Opinions

Then SL opinions of each provider is computed based on converged scores. Two kinds of provider reliability should be considered: (1) reliability of generating the true value; (2) reliability of continuously doing so. Provider's reliability in last subsection describes the first kind of reliability, and we propose a new concept, **certainty**, to describe the second one. Certainty of provider P_j is defined as:

$$Certainty(P_j) = \frac{\sum_{E_i \in E_{P_j}} Disc(E_i)}{|E_{P_j}|}. \tag{3}$$

With SL, we proposed to record the provider's reliability of generating the true value in **trust**, and record the reliability of continuously doing so with **uncertainty**. In this way, the algorithm's opinion towards the P_j is defined as $\omega_{P_j}^{Algo} = \{t_{P_j}^{Algo}, d_{P_j}^{Algo}, u_{P_j}^{Algo}\}$:

$$t_{P_j}^{Algo} = (1 - u_{P_j}^{Algo})Rel(P_j) \tag{4}$$

$$d_{P_j}^{Algo} = 1 - t_{P_j}^{Algo} - u_{P_j}^{Algo} \tag{5}$$

$$u_{P_j}^{Algo} = \gamma(1 - Certainty(P_j)) + \alpha. \tag{6}$$

where $Algo$ is short for "algorithm". α describe people's fundamental uncertainty, since even given by enough evidence, people can still be skeptical. γ is a parameter to limit the certainty to a certain range. Both parameters range from 0 to 1. In this way, provider's reliability can be accurately described.

3.3 Infer True Value in Generative Manner

This manner only fits numerical data, i.e., $V_{ij} \in R$. For each entity, we define a statement "true value of entity is the largest candidate value", and generate *Algo*'s opinion towards them. Higher trust means truth is close to the max candidate value; otherwise, truth is close to the min candidate value. First, on each entity E_i, we normalize all the candidate values in the following manner:

$$V_{ij}^{'} = \frac{V_{ij} - min(V_{i.})}{max(V_{i.}) - min(V_{i.})}, \tag{7}$$

so that $V_{ij}^{'} \in [0, 1]$. Then, the original statement is mapped to "true value of entity in the normalized space is 1". Thereby, given provider E_i, the provider P_j's opinion towards the statement can be defined as:

$$\omega_{truth(E_i)=1}^{P_j} = \{(1-\beta)V_{ij}^{'}, 1-(1-\beta)V_{ij}^{'} - \beta, \beta\}, \tag{8}$$

where β also describes people's fundamental uncertainty, similar to α. Second, the provider can recommend his opinion of the entity's truth to *Algo*. Thus, *Algo*'s opinion towards truth of E_i by P_j's recommendation is defined as:

$$\omega_{truth(E_i)=1}^{Algo,P_j} = \omega_{P_j}^{Algo} \otimes \omega_{truth(E_i)=1}^{P_j}. \tag{9}$$

Entity E_i has a set of candidate values from several providers $\{P_j, ..., P_k\}$, and *Algo* should have a summarized opinion based on all recommendations with Consensus operation. The algorithm's final opinion towards truth of E_i is defined as:

$$\omega_{truth(E_i)=1}^{Algo,P_j,...,P_k} = \omega_{truth(E_i)=1}^{Algo,P_j} \oplus ... \oplus \omega_{truth(E_i)=1}^{Algo,P_k}. \tag{10}$$

In the fused opinion, the trust reflects the true value of E_i in the normalized space, and final step is to map it to the original numerical space by:

$$V_{ij}^{true} = t_{truth(E_i)=1}^{Algo,P_j,...,P_k}(max(V_{i.}) - min(V_{i.})) + min(V_{i.}). \tag{11}$$

3.4 Infer True Value in Discriminative Manner

In this model, we generate a SL opinion for each candidate value, and then for each entity, select the value with highest trust as the truth. Given a provider P_j, the algorithm's opinion towards a value V_{ij} is defined as:

$$\omega_{V_{ij}}^{Algo,P_j} = \{t_{P_j}^{Algo}, d_{P_j}^{Algo}, u_{P_j}^{Algo}\}. \tag{12}$$

If a value is provided by several providers $\{P_j, ..., P_k\}$, consensus operation is used to fuse opinions together. Thus we have algorithm's final opinion towards a value V_{ij}:

$$\omega_{V_{ij}}^{Algo,P_j,...,P_k} = \omega_{V_{ij}}^{Algo,P_j} \oplus ... \oplus \omega_{V_{ij}}^{Algo,P_k}. \tag{13}$$

4 Experiments

In this section, we evaluate our proposed framework on two popular real word datasets, **Book** and **Population**, one being categorical another being numerical. In the experiment, we name our proposed method SLFTD generating the true value in a generative way, as **SLFTD-Gen**; and name SLFTD selecting the true value from existing candidates in a discriminative manner, as **SLFTD-Dis**. Naive baselines include **Voting**, **Median**, **Average**. State-of-art methods include **TruthFinder** [5], **Accuracy** [6], **AccuracySim** [6], **Sums**, **Investment**, **PooledInvestment**, **Average.Log** [8], **CRH** [9], **CATD** [10] and **GTM** [11].

4.1 Finding True Book Author List

Dataset: Book. It is a popular categorical dataset in truth discovery area. Its data describes that for each book, online bookstores post author list in their web pages, but some data is wrong. It contains the information on ISBN, book name, authors, online bookstore name for 1265 books. Totally, there are 894 bookstores and they generate 26,494 author lists. In this study, we use two testing data: (1) the gold testing dataset in the original dataset consisting of 100 books, (2) a new silver testing dataset composed of 161 book, containing the first 100 books and other 61 books. The 61 books are selected because different methods appearing in our experiments gives different true data. Thus it is more challenging than the first one. The true author list of both testing data are manually assigned by people reading the cover page of the book. All the codes and preprocessed datasets in this paper are posted online[1].

Settings. The implication appeared in Eq. 1, is defined as $Imp(V_{il} \rightarrow V_{ij}) = \frac{\#|V_{il} \bigcap V_{ij}|}{\#|V_{ij}|}$, where $\#|V_{il}|$ is the amount of elements in V_{il}; the implication appeared in Eq. 2 is defined as $Imp(V_{i.} \rightarrow V_{ij})$ is defined as $\frac{\sum_{P_l,P_j \in P_{E_i}} Imp(V_{il} \rightarrow V_{ij})}{\sum_{P_l,P_j \in P_{E_i}} 1}$. Following past studies, the parameters of all methods are set with optimal performance on the testing data. In TruthFinder, $\{\gamma = 0.1, \rho = 0.7\}$. In AccuracySim, $\{\lambda = 0.9\}$. In SLFTD-Dis, $\{\gamma = 0.2, \alpha = 0.2\}$.

[1] https://github.com/daz45.

Table 1. Precision of all methods on dataset BOOK. Best results are in bold.

Method	Golden Testing	Silver Testing
SLFTD-Dis	**94%**	**77.6%**
TruthFinder	93%	**77.6%**
AccuracySim	91%	68.9%
Accuracy	89%	68.9%
PooledInvestment	87%	72.75%
Average.log	82%	62.1%
Voting	80%	62.1%
Investment	79%	63.4%
Sums	74%	55.3%

Results. From Table 1, we can see that our proposed method SLFTD-Dis has the best performance on both testing data. TruthFinder provides a nearly same good result. Also, methods (SLFTD-Dis, TruthFinder, and AccuracySim) that use value similarity, shows a better performance than those do not. Please note that our preprocessed dataset is cleaner than the data used in prior works [6], as the voting results is 82%, while past studies showed only 71%.

4.2 Finding True Population of the City

Dataset: Population. It is a sample of Wikipedia edit history of city population, proposed in [8], and algorithms need to identify the true population for each city of each year. This data is picked to test the system's performance on numerical data. It is preprocessed in the same way as that in [11], except σ_0 is set to be 0.91 instead of 0.9. The final data consists of 4,183 tuples on 1,172 city-year from 1,926 providers, and methods are evaluated on 277 city-year.

Settings. The implication appeared in Eq. 1, is defined as: $Imp(V_{il} \rightarrow V_{ij}) = 1 - \frac{|V_{ij} - V_{il}|}{max(V_{i.}) - min(V_{i.})}$; the implication appeared in Eq. 2 is defined as $Imp(V_{i.} \rightarrow V_{ij}) = 1 - \frac{|V_{ij} - avg(V_{i.})|}{max(V_{i.}) - min(V_{i.})}$. Following past studies, the parameters of all methods are set based on optimal performance on the testing data. In TruthFinder, $\{\gamma = 0.3, \rho = 0.01\}$. In terms of GTM, we have two set of parameters, $(\alpha = 10, \beta = 10, \mu_0 = 0, \sigma_0^2 = 1)$ suggested by [11], and $(\alpha = 1, \beta = 1, \mu_0 = 0, \sigma_0^2 = 1)$ suggested by our experiment. For CATD, significance level α is set to be 0.03. For SLFTD-Gen, $(\alpha = 0, \beta = 0.01, \gamma = 0.001)$; for SLFTD-Dis, $(\alpha = 0.01, \gamma = 0.001)$.

Results. Experiment results are shown in Table 2. We can see that SLFTD-Dis gives best performance on three metrics, then TruthFinder provides second best on MAE and Error Rate, while SLFTD-Gen gives second best on RMSE.

Comparing two groups, Error Rate shows that discriminative methods generally makes less errors than generative models. Naive methods, especially Average, gives a much worse performance. Also, SLFTD-Dis, SLFTD-Gen, TruthFinder, CATD can find the true value with a smaller error.

Table 2. Experiment results on dataset Population. First group are discriminative models; second group are generative models. Best results are in bold; second best is labeled with *. Error Rate is based on mismatch 10%.

Methods	MAE	RMSE	Error Rate
SLFTD-Dis	**1489.57**	**5819.00**	**14.44**%
TruthFinder	1744.05*	8942.86	16.97*%
Voting	2511.04	11328.71	23.10%
Investment	2614.21	11378.42	25.99%
CRH - weighted median	3030.23	12696.96	25.99%
Median	2475.05	9759.71	33.57%
CATD	1796.67	8765.81	21.30%
SLFTD-Gen	2132.35	7070.54*	55.60%
GTM - parameters by us	2424.10	8659.36	57.04%
GTM - parameters in [11]	2710.30	9290.32	58.12%
Average	3336.49	9799.66	58.48%
CRH - weighted average	3805.10	11898.04	58.48%

5 Conclusion

In this study, we proposed a SL based framework for the truth discovery, which can predict truth either in a discriminative way or a generative manner. SL is introduced to more accurately describe the provider's reliability. Experiments on two real world datasets validates the effectiveness of our proposed methods.

References

1. Dong, X.L., Berti-Equille, L., Srivastava, D.: Integrating conflicting data: the role of source dependence. In: Proceedings of the VLDB Endowment 2.1, pp. 550–561 (2009)
2. Dong, X.L., Saha, B., Srivastava, D.: Less is more: selecting sources wisely for integration. In: Proceedings of the VLDB Endowment, vol. 6. no. 2. VLDB Endowment (2012)
3. Galland, A., et al.: Corroborating information from disagreeing views. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. ACM (2010)
4. Yin, X., Tan, W.: Semi-supervised truth discovery. In: Proceedings of the 20th International Conference on World Wide Web. ACM (2011)

5. Yin, X., Han, J., Philip, S.Y.: Truth discovery with multiple conflicting information providers on the web. IEEE Trans. Knowl. Data Eng. **20**(6), 796–808 (2008)
6. Dong, X.L., Srivastava, D.: Big data integration. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE), p. 2. IEEE (2013)
7. Pelechrinis, K., et al.: Automatic evaluation of information provider reliability and expertise. World Wide Web **18**(1), 33–72 (2015)
8. Pasternack, J., Roth, D.: Knowing what to believe (when you already know something). In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics (2010)
9. Li, Q., et al.: Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. ACM (2014)
10. Li, Q., et al.: A confidence-aware approach for truth discovery on long-tail data. Proc. VLDB Endow. **8**(4), 425–436 (2014)
11. Zhao, B., Han, J.: A probabilistic model for estimating real-valued truth from conflicting sources. In: Proceedings of QDB (2012)
12. Jøsang, A.: Subjective Logic. Springer, Heidelberg (2016)