

A Effective Truth Discovery Algorithm with Multi-Source Sparse Data

Jiyuan Zhang^{1,2}, Shupeng Wang^{1,*}, Guangjun Wu^{1,*}, Lei Zhang^{1,*}

¹*Institute of Information Engineering, CAS, Beijing, 100093, China*

²*School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100031, China*

**Corresponding Authors {zhangjiyuan, wangshupeng, wuguangjun, zhanglei1}@iie.ac.cn*

Abstract. The problem to find out the truth from inconsistent information is defined as Truth Discovery. The essence of truth discovery is to estimate source quality. Therefore the measuring mechanism of data source will immensely affect the result and process of truth discovery. However the state-of-the-art algorithms don't consider how source quality is affected when null is provided by source. We propose to use the Silent Rate, True Rate and False Rate to measure source quality in this paper. In addition, we utilize Probability Graphical Model to model truth and source quality which is measured through null and real data. Our model makes full use of all claims and null to improve the accuracy of truth discovery. Compared with prevalent approaches, the effectiveness of our approach is verified on three real datasets and the recall has improved significantly.

Keywords: Truth discovery · Data fusion · Multi source data confliction

1 Introduction

With the development of information technology, the Internet has penetrated into all corner of human social life. The data on internet have accumulated sharply, and these data have been integrated into an information ocean. One of the important features of this information is diversity, so for any object, heterogeneous descriptions can be found on internet from multiple sources. The inconsistency or conflict of these diverse descriptions causes great confusion for us to identify true information from each other. Therefore, identifying the accurate and complete information from conflicting descriptions is the key factor for information integration. The problem is the Truth Discovery proposed in document [9].

In order solve the problem, this paper makes use of the Hub Authority method [4] [2] to solve the problem by the quality difference of source. And we will redesign the metrics of source quality, and measure the quality of sources with three indexes, such as silent rate, true rate and false rate, to improve the accuracy of truth discovery. The main work of this paper presents is as follows:

1. Redesign metrics for source quality. The quality of source is measured by three indexes, such as silent rate, true rate and false rate. The silent rate in the new metrics can make full use of the null data provided by the source, and can describe the source quality more comprehensively.
2. Fertilize the plate model of probabilistic graph to construct model. The relationship among the data source, the object and the truth is constructed, and a probabilistic graph model is established to deduce the truth. Combing the conditions and methods of influence propagation in probabilistic graphs, the truth probability of every claim is deduced.

2 Related work

The problem of truth discovery is defined for the first time in document [9], and a TruthFinder algorithm is proposed for solving this kind of problem. The method is similar to the iterative mechanism of Hub Authority, which synchronously infer the truth of object and the quality of the source. Inspired by it, a series of similar methods have been developed to study various factors that affect truth discovery [4] [2] [3]. The logical criterion of this method is: the higher the source quality is, the more likely it provides truth, at the same time, the more truth it provides, the higher the source quality is. These kind of algorithm is called heuristic method.

Based on the above processing logic, recent research has transformed truth discovery into a framework for optimization. Each source is assigned a weight according to its credibility, and the optimization objective is to minimize the distance between the target value and the truth. Document [6] which is extended in [5] uses the weight to represent the distribution of the credibility of each source, and uses a variety of loss functions to deal with heterogeneous data objects, so that the truth discovery of heterogeneous data is integrated into an objective function. Document [11] builds an optimization framework based on min-max entropy to estimate the truth of objects with noise. In a word, this kind of method updates the truth and the weight of the data source iteratively until the weight converges, and obtains the truth and the weight of all the data sources.

The other method is probability method. These methods solve the problem of truth discovery in document [1] [7] [8] [10]. The core idea is that multi-source data is considered as a mixed distribution, and the credibility of the source is integrated into the probabilistic model in the form of random variables. The probabilistic model is proposed by [10], and the quality of data sources is modeled by using two types of errors – false positive and false negative. Document [7] extends the truth discovery to the field of social group awareness. In document [8], the authors propose to use stochastic Gauss models to represent sources. So the mean represents the truth and variance represents the credibility of source. Theoretically, both solutions are proven to be contractive to an ϵ -ball around the maximum likelihood estimate.

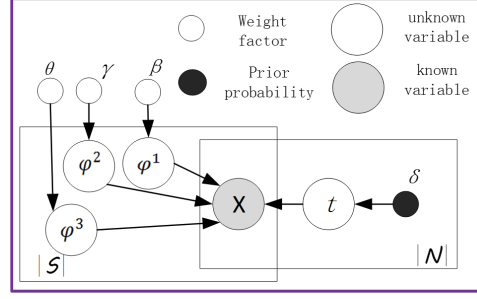
3 Implementation

In this section, we first introduce some terms and annotations. Suppose that S is a collection of sources, and j is one of the sources; N is the set of observation objects, and n is one of the objects. If a source j can only provide a claim for an observation object, then the n 's claim obtained from j is denoted as x_n^j . The n 's truth is denoted as t_n . The n 's claim set is $X_n = \{x_n^j\}_{j \in S_n}$ and N 's claim set is $X = X_1 \cup \dots \cup X_n$. In view of the fact that it is impossible to obtain the claims of object n from some sources, so the claims are only obtained from S_n i.e. a subset of S . We denote the collection of non-repeated claims of n as $V_n = \{v_{ni}\}, i = 1, \dots, N'$.

In order to simplify the solution model, two hypotheses are given for the source and the object.

Hypotheses1 : The data sources are independent of each other, and they provide claims independently.
Hypotheses2 : The objects are independent of each other. The alteration of one object's claim do not affect others'.

We construct a Bayesian network model as shown in Fig. 1 to calculate the probability of truth. In the network, each node represents a random variable, the dark node represents the known variable, and the light colored node represents the unknown variable. The solid black point δ is the prior probability of the truth, and the hollow points represent the weight of the source measurement indexes. β, γ, Θ is a group of super parameters. β represents the weight of the false rate when calculating the truth. The γ is the weight of the true rate, and the Θ is the weight of the silent rate, satisfying equation $\beta + \gamma + \Theta = 1$. The directed edges that connect nodes represent dependencies among variables. $|S|$ and $|N|$ in the corner of the box represent the number of sources and objects.


Fig. 1. Probability Graphical Model of FTS

3.1 Source Quality

The most important factor that affects the accuracy of truth discovery is the quality of sources, and the previous algorithm does not consider the null data how to affect source quality. In this paper, true rate, false rate to measure source quality.

FalseRate(FR): The rate of all claims is not truth, which is provided by source j , i.e. $FR = \frac{Fp}{Empt + Fp + Tp}$.

TrueRate(TR): The rate of all claims is truth, which is provided by source j , i.e. $TR = \frac{Tp(1-FR)}{Empt + Fp + Tp}$.

SilentRate(SR): The rate of all claims is null, which is provided by source j , i.e. $SR = \frac{Empt(1-FR)}{Empt + Fp + Tp}$.

$Empt$ is the number of null among all the claims that source j provides i.e. $Empt = \sum_{n=1}^N Signal(x_n^j = null)$. Fp is the expectation of false claims that the source j provides when the probability of truth is $p(t_n = v_{ni})$. Then, $Fp = p(t_n = v_{ni}) \times \sum_{n=1}^N Signal(x_n^j \neq v_{ni} \& \& x_n^j \neq null)$. Tp is the expectation of true claims that the source j provides when the probability of truth is $p(t_n = v_{ni})$. Then, $Tp = p(t_n = v_{ni}) \times \sum_{n=1}^N Signal(x_n^j == v_{ni})$. The $Signal(\cdot)$ is an indicator function in the formula:

$$Signal(t) = \begin{cases} 1, & t=true \\ 0, & t=false \end{cases}$$

The quality of source j is $\phi_j = (\phi_j^1, \phi_j^2, \phi_j^3)$ which is a tri-tuple. The value of them is $\phi_j^1 = FR$, $\phi_j^2 = TR$, $\phi_j^3 = SR$.

3.2 Truth Inference

According to the hypothesis 2, all $|N|$ objects are independent of each other. From the probability graph model of Fig. 1, the probability of the observed values of all observed objects is:

$$P(X|\phi_S) = \prod_{n=1}^{|N|} P(x_n^1, \dots, x_n^{|S|}|\phi_S) \quad (1)$$

$P(x_n^1, \dots, x_n^{|S|}|\phi_S)$ is a joint probability density function of Object n when $|S|$ sources provide claims.

The V_n is the claim set of Object n whose truth, t_n , has $|V_n|$ possible values. Then:

$$\begin{cases} P(x_n^1, \dots, x_n^{|S|}|\phi_S) = \sum_{i=1}^{|V_n|} \delta_{ni} p(x_n^1, \dots, x_n^{|S|}|t_n = v_{ni}, \phi_S) \\ \sum_{i=1}^{|V_n|} \delta_{ni} = 1 \end{cases} \quad (2)$$

According to the hypothesis 1, if the sources are independent of each other to provide claims, and there is no mutual replication of claims, the probability of obtaining all the claims of object n from $|S|$ sources is

a joint probability function:

$$\begin{aligned}
 p(x_n^1, \dots, x_n^{|S|} | t_n = v_{ni}, \phi_S) &= \prod_{j=1}^{|S|} P(x_n^j | t_n = v_{ni}, \phi_j) \\
 &= \prod_{j=1}^{|S|} (\phi_j^1)^{\text{Signal}(x_n^j \neq v_{ni} \& \& x_n^j \neq \text{null})} \times (\phi_j^2)^{\text{Signal}(x_n^j = v_{ni})} \times (\phi_j^3)^{\text{Signal}(x_n^j = \text{null})}
 \end{aligned} \tag{3}$$

According to the definition of the problem, the truth probability of the object n is actually the truth probability under the condition of the current claim and the source quality, i.e.

$$p(t_n = v_{ni}) = p(t_n = v_{ni} | x_n^1, \dots, x_n^{|S|}, \phi_S) \tag{4}$$

According to Bayes formula,

$$p(t_n = v_{ni} | x_n^1, \dots, x_n^{|S|}, \phi_S) = \frac{p(x_n^1, \dots, x_n^{|S|} | t_n = v_{ni}, \phi_S) \times \delta_{ni}}{P(X | \phi_S)} \tag{5}$$

In formula (5), the denominator is exactly the same for all objects. Then, the truth probability $p(t_n = v_{ni})$ is proportional to the molecular term. After substituted formula (3), formula (5) changed into formula (6).

$$p(t_n = v_{ni}) = p(t_n = v_{ni} | x_n^1, \dots, x_n^{|S|}, \phi_S) \propto \delta_{ni} p(x_n^1, \dots, x_n^{|S|} | t_n = v_{ni}, \phi_S) \tag{6}$$

Formula (6) is a continued product. The likelihood of the truth probability is obtained.

$$r_{ni} = \log p(t_n = v_{ni}) \propto \ln \delta_{ni} +$$

$$\sum_{j=1}^{|S|} \{ \text{Signal}(x_n^j \neq v_{ni} \& \& x_n^j \neq \text{null}) \times \ln \phi_j^1 + \text{Signal}(x_n^j = v_{ni}) \times \ln \phi_j^2 + \text{Signal}(x_n^j = \text{null}) \times \ln \phi_j^3 \} \tag{7}$$

3.3 Iterative Computation

The model uses silent rate, truth rate and false rate to measure source quality. β, γ and θ is the weight to adjust impact on three index. Therefore, β, γ and θ is substituted into the formula (7). The recursive formula of the final truth probability can be obtained.

$$\begin{aligned}
 r_{ni}^{(k+1)} &= \ln \delta_{ni} + \sum_{j=1}^{|S|} \{ \beta * \text{Signal}(x_n^j \neq v_{ni} \& \& x_n^j \neq \text{null}) \times \ln \phi_j^{1(k)} + \\
 &\quad \gamma * \text{Signal}(x_n^j = v_{ni}) \times \ln \phi_j^{2(k)} + \theta * \text{Signal}(x_n^j = \text{null}) \times \ln \phi_j^{3(k)} \}
 \end{aligned} \tag{8}$$

Give an assumption: $f1 = \sum_{i=1}^{|X_n|} \sum_{n=1}^N r_{ni}^{(k)} \times \text{Signal}(x_n^j \neq v_{ni} \& \& x_n^j \neq \text{null})$, $f2 = \sum_{n=1}^N \text{Signal}(x_n^j = \text{null})$, $f3 = \sum_{i=1}^{|X_n|} \sum_{n=1}^N r_{ni}^{(k)} \times \text{Signal}(x_n^j = v_{ni})$

We can get the formulas of $\phi_j^{1(k)}$, $\phi_j^{2(k)}$ and $\phi_j^{3(k)}$.

$$\phi_j^{1(k)} = \frac{f1}{f1 + f2 + f3}, \quad \phi_j^{2(k)} = \frac{f3}{f2 + f3} (1 - \phi_j^{1(k)}), \quad \phi_j^{3(k)} = \frac{f2}{f2 + f3} (1 - \phi_j^{1(k)}) \tag{9}$$

Algorithm 1: FTS truth discovery algorithm

input : X the whole claim set for all objects, the set of sources S , the set of objects N ; threshold of converges α
output: The truth of all object T ;The source quality of all sources ϕ_s

```

1 for  $o_n \in N$  /*initiate the prior probability of all claims*/
2     initiate the prior probability of  $o_n$ ,  $\delta_{ni}, i = 1, \dots, |V_n|$ 
3 end for
4 for  $s \in S$  /*initiate source quality*/
5      $\phi_j = (\phi_j^1, \phi_j^2, \phi_j^3), j = 1, \dots, |S|$ 
6 end for
7  $k \leftarrow 0$ 
8  $\phi'_S \leftarrow 0$ 
9  $c \leftarrow 1$ 
10 while  $|c| > \alpha$  /* convergence conditions is not satisfied*/
11      $k \leftarrow k + 1$ 
12     for  $o_n \in N$ 
13         calculate  $r_{ni}^{(k)}, i = 1, \dots, |V_n|$  from formula (9)
14     end for
15      $\phi'_S \leftarrow \phi_S$ 
16     for  $s \in S$ 
17         calculate source quality from formula (10)
18          $\phi_j = (\phi_j^{1(k)}, \phi_j^{2(k)}, \phi_j^{3(k)}), j = 1, \dots, |S|$ 
19     end for
20      $c \leftarrow \phi'_S - \phi_S$ 
21 end while
22 for  $o_n \in N$  /*calculate truth*/
23      $i \leftarrow \max_{i=1, \dots, |X_n|} r_{ni}^{(k)}$ 
24      $t_n = v_{ni}$ 
25 end for
26 return  $T, \phi_S$  /*Output truth and source quality*/

```

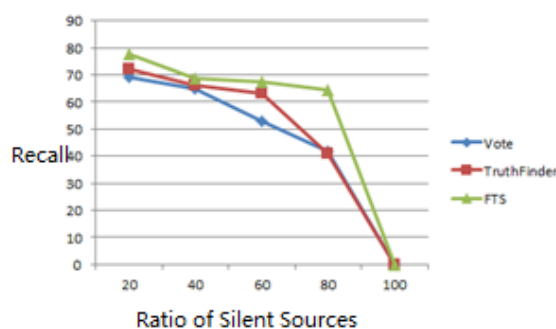
The whole procedure of the FTS is shown in algorithm 1. First, the prior probabilities of the claims of each object are initialized by uniform distribution and source quality ϕ is initialized by standard normal distribution. Then, according to the initial source quality, the whole iteration process is started. Finally, the (k+1)-th r_{ni} i.e. truth probability is calculated with k-th iteration of source quality i.e. ϕ_j^k . The new source quality can be produced with new truth probability. The iterative algorithm is repeated until the source quality updates less than the threshold and the algorithm converges. The truth and source quality is calculated synchronously.

4 Experiment

In this section the truth discovery algorithm will run on three real data sets to verify the performance of the designed FTS model. As a comparison, the FTS and two state-of-art algorithms are tested under the same conditions to acquire respective recall.

Table 1. Recall of all algorithms on different datasets.

Algorithm	Book-Author	Weather	Flight
Vote	76	70.59	54.32
TruthFinder	79	66.7	41.2
FTS	81	79.7	56.58

**Fig. 2.** Recall of different number of data source.

4.1 Experimental Setup

The configuration of all experiments in this paper: CPU is Intel(R) Core(TM) i7-6700 3.40GHz. Memory is 64G and OS is Windows 7. In this paper, three real datasets (<http://lunadong.com/>) such as *WeatherDataset*, *WeatherDataset*, and *FlightDataset* are used to verify the effect of the algorithm.

4.2 Experimental Result

In order to verify the accuracy of the algorithm, the truth discovery accuracy test was carried out on three real datasets, and the test results were listed in table 1.

In the experiment, the set of ground truth is split into two parts randomly, one is used for verification, and the other is used for testing. β, γ and $\theta = 1 - \beta - \gamma$ is used to adjust the weight of silent rate, true rate and false rate. During the processes of verification, β, γ is taken from the range(0,1) and the step is 0.02 and the optimal combination of , is obtained. With the optimal β, γ, θ , recall is calculated on testing set.

From the experimental data in Table 1, we can see that our FTS method is better than the classical truth discovery algorithm on recall. It means silent rate has positive impact on truth discovery. The recall on weather and flight datasets is significantly decreased. The main reason for this phenomenon is that the number of sources in these two datasets is small, which affects the effect of truth discovery.

In order to test the influence of silent data source on the recall, the Book-Author data set are processed, and randomly selected 20%, 40%, 60%, 80% of the sources to keep silent i.e. claims is null. The experimental results are shown in Fig. 2. With the increase of silent source, the recalls of all algorithm decline steadily. If 100% of the sources remain silent, no algorithm can predict the truth, so the recall is 0. Comparing with the three algorithms, the Vote algorithm drops the fastest, which shows that the accuracy of the algorithm heavily depends on the number of sources that provide claims. The TrueFinder algorithm is also affected by

the number of silent sources, and its performance on recall is lower than FTS. With comprehensive comparison, FTS algorithm is affected little by the number of silent source. The reason is that the effect of silent rate has been considered in FTS model.

5 Conclusion

First of all, the metrics of source quality has been redesigned. In FTS model, the reliability of data source quality is measured by true rate, false rate and silent rate. Thus the situation that source provides null is comprehensively comprised. Secondly, a probabilistic graph model is established to construct the relationship among the source, the object and the truth. Then the truth is deduced. In this model, the relationship among the source metrics, the truth and the claims of the source is presented in the form of figures. Using the conditions and methods of influence propagation in probabilistic graphs, the probabilities of every claim as truth are deduced. The experimental results on three datasets show that the new algorithm significantly improves the recall of truth discovery compared with the traditional classical algorithm.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No.61601458) and National Key Research and Development Program of China (No.2016YFB0801004,2016YFB0801305).

References

1. Blanco, L., Crescenzi, V., Merialdo, P., Papotti, P.: Probabilistic models to reconcile complex data from inaccurate data sources. In: International Conference on Advanced Information Systems Engineering. pp. 83–97 (2010)
2. Dian, Y., Hongzhao, H., Taylor, C., Ji, H., Chi, W., Shi, Z., Jiawei, H., Clare, V., Malik, M.I.: The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In: Proceedings of 2014 International Conference on Computational Linguistics. pp. 1567–1578 (2014)
3. Furong, L., MongLi, L., Wynne, H.: Entity profiling with varying source reliabilities. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1146–1155 (2014)
4. Galland, A., Abiteboul, S., Marian, A., Senellart, P.: Corroborating information from disagreeing views pp. 131–140 (2010)
5. Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., Fan, W., Han, J.: A confidence-aware approach for truth discovery on long-tail data. *very large data bases* **8**(4), 425–436 (2014)
6. Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., Han, J.: Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation pp. 1187–1198 (2014)
7. Wang, D., Kaplan, L.M., Le, H.K., Abdelzaher, T.F.: On truth discovery in social sensing: a maximum likelihood estimation approach pp. 233–244 (2012)
8. Xiao, H., Gao, J., Wang, Z., Wang, S., Su, L., Liu, H.: A truth discovery approach with theoretical guarantee. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1925–1934 (2016)
9. Yin, X., Han, J., Yu, P.S.: Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering* **20**(6), 796–808 (2008)
10. Zhao, B., Rubinstein, B.I.P., Gemmell, J., Han, J.: A bayesian approach to discovering truth from conflicting sources for data integration. *very large data bases* **5**(6), 550–561 (2012)
11. Zhou, D., Basu, S., Mao, Y., Platt, J.: Learning from the wisdom of crowds by minimax entropy pp. 2195–2203 (2012)