Truth Finding from Multiple Data Sources by Source Confidence Estimation

Fan Zhang[1], Li Yu[2], Xiangrui Cai[2], Ying Zhang[1,2,*], Haiwei Zhang[1,2]
[1]College of Software, Nankai University
[2]College of Computer and Control Engineering, Nankai University
{zhangfan,yuli,caixiangrui,zhangying,zhanghaiwei}@dbis.nankai.edu.cn

***Abstract*—The volume of data on the Web has been growing at a dramatic pace in recent years and people rely more and more on the Web to fulfill their information needs. Numerous different descriptions of the properties towards the same objects can be obtained from a variety of data sources. This will inevitably lead to data incompleteness, data conflicts and out-of-date information problems. These issues make truth discovery among multiple data sources non-trivial. However, most of previous works consider only one single property, or deal with different properties separately by ignoring several characteristics of the properties, which will often cause unexpected deviations. In this paper, we propose a modified method to find the most trustable source and identify the true information. Our goal is to minimize the distance between the true information and the overall observed descriptions through considering the accuracy and the coverage of all the data sources at the same time. The experiments on the real dataset demonstrate the efficacy of our method.**

***Keywords*—data fusion;truth discovery; source selection**

I. Introduction

The amount of useful information available on the Web has been growing at a dramatic pace in recent years, especially in Big data era. Because of different kinds of conditions, such as disk damage, lacking of domain knowledge and outdated update, quality of sources is jagged. Some sources may be not so accurate, authoritative or persuasive. These data accordingly can be erroneous, incomplete and obsolete. It is critical to identify the most trustworthy answers from multiple sources with conflicting information. Generally, real-world objects are not described by single property but multiple ones. For example, we want to know climatic conditions of mountains, such as the Mountain Everest and the Kilimanjaro. Most of previous work just focuses on single property, like height or humidity or temperature. Actually, there can exist some kind of correlation between data properties. For instance, temperature and height can influence each other. Generally, temperature at a higher spot is relatively lower. What's more, the unique characteristics of data properties can affect truth finding. For categorical data, there can be only two answers, right or wrong. For continuous data, closeness is more accurate for evaluation. If the true height of the Mount Everest is 8844.43 meters, apparently, observed value 8848 meters is more accurate than 8800 meters. What's more, from the point of sources, we trust accurate sources more, that is to say, the description these sources make is less doubtful and more likely to be true. For example, we believe in *Google* more than gossip site about geographic research. Further, sources with wider coverage should get more confidence. Compare two sources with same accuracy (assuming 0.8), while source A has a wider coverage (assuming 0.7) than source B (assuming 0.1). Obviously, source A is more reliable because wider coverage demonstrates more trustable accuracy and narrow coverage indicates great contingency.

In this work we consider unique characteristics of data properties and different properties jointly, while most of previous work just focuses on a single attribute or separately treats attributes. Besides, We focus on accuracy and coverage of data sources since these two factors can greatly influence truth finding. We intend to find these descriptions made by sources with high accuracy and wide coverage. In general, we make three following contributions.

1) We propose a modified model to capture unique characteristics of different data types.
2) Our algorithm considers both accuracy and coverage.
3) We demonstrate the algorithm performance on real data sets. And the experiment shows our algorithm is more stable.

The remainder of this paper is organized as follows. In section II, we talk about related work. Our method will be introduced in the following section III. And then section IV shows the experiment. Finally, we conclude this paper in section V.

II. Related work

Truth finding can be seen a part of data fusion and data fusion can be seen a part of data integration. So far, a lot of work has being focused on resolving conflicts and truth finding. Influence between values and iterative computation has been considered in [1], and [2] discusses the role of source dependence in resolving conflicting data and mitigated a copier source's count. [3] talks about the relative accuracy of attributes using correlation between properties. More specifically, resolving numerical data type reconciliation is the main point of [4]. [5] introduces deep learning to effectively find and resolve uneasily detected inconsistencies. [6] brings in a data sharing system to tackle conflicting information. [7] tells

* Corresponding Author

978-1-4673-9372-0/15 $31.00 © 2015 IEEE
DOI 10.1109/WISA.2015.45

how "order" can affect the performance of integration. The most approximate work to ours is [8], which combines two different data types to find truth. Our work considers more factors and has better performance.

III. Methodology

In this section, we define the truth finding problem of different data attributes, propose our algorithm and analyze properties of the formulas. We give more weight to sources whose descriptions are closer to truth and whose coverage is wider. Our goal is to minimize the distance between source descriptions and truth.

A. Term Definition

We explain some important terminology at first.
Example 1, website *Accuweather* forecasts the high temperature of weather of Memphis at 8 in Jan 29, 2010 as 33 °F and weather condition is "cloudy".

A *source* is a database or a website (or something else which produces data) makes description about the world, real or virtual.
There exist many weather forecast organization, each of which is a source, such as *www.yahoo.com* and national meteorological center of CMA. In example 1, *Accuweather* is a source.

An *object* is a kind of entity, described by many properties.
Continued with example 1, weather of Memphis at 8 in Jan 29, 2010 is an object.

A *property* is an aspect of objects. Different properties come in different data types, such as categorical data, continuous data, graph data etc.
In example 1, high temperature and weather condition are properties. We can treat temperature as continuous data while condition as continuous.

A *claim* is a source's description about a property of an object.

A *fact* is the most representative value for claims about some property of some object.
Continued with Example 1: We think the high temperature of weather of Memphis at 9 in Jan 29, 2010 is actually 33 °F after running our algorithm.

B. Problem statement

Usually there are many sources and each source makes large amount of claims. However, these claims may be contradictory to each other.
Example 2, as shown in Table I, three websites (*climaton*, *accuweather*, *yahoo*) forecast weather of Fort Worth (F.W. for short), Memphis (Me. for short), Los Angeles (L.A for short) in 8 a.m. on Jan 29 2010. We select two typical properties: high temperature (H.T. for short, °F used as measure unit) and weather condition (Wea. Con. for short).

TABLE I: Weather forecast data

	accuweather		**climaton**		**yahoo**	
	H. T.	Wea. Con.	H.T.	Wea. Con.	H.T.	Wea. Con.
F.W.	46	rain	46	moderate rain	46	rain
Me.	33	cloudy	34	most cloudy	33	cloudy
L.V.	47	clear	46	clear sky	47	fair

We are aimed at finding what is the exact weather situation of these three cities. A naive way is to calculate average for temperature(continuous data type) and choose the most frequent value for weather condition (categorical data type). This method has an obvious drawback as these sources are not equally important. Put it in another way, their weight (contribution to corresponding value of claims) varies from one to another, other than equivalent. That is to say, if *climaton* had a bigger accuracy 0.8, while *accuweather* and *yahoo* had smaller accuracy 0.3. Then *climation* would be overwhelming than others. Therefore, deciding source weight is essential. What's more, coverage is another factor that affects our decision. Even though *accuweather* and *yahoo* has the same accuracy, their coverage is different (*accuweather* is more specific in weather forecast area and covers more cities than *yahoo*), so we give a little bigger weight to *accuweather*. In short, our main work is to decide accuracy and coverage of each source so that we can get truth about each claim.

C. Algorithm

The $X^{(k)}$ is the set of claims made on properties objects by the k-th source. It is denoted as a matrix whose im-th claim is $v_{im}^{(k)}$, which means m-th property of i-th object. $X^{(1)}$, $X^{(2)}$, ..., $X^{(k)}$ are the K source claim sets. In Table V, *accuweather*, *climaton*, *yahoo* are the three sources and the claims below them respectively. more specifically, $v_{32}^{(2)}$is "clear". Our goal is to get set source weights which satisfy the formulas below.

$$\min_{X^*,\mathcal{A}} \; f(X^*,\mathcal{W}) = \sum_{k=1}^{K} a_k \cdot c_k \sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)}) \qquad (1)$$
$$s.t. \quad \delta(\mathcal{A}) = 1, \; \mathcal{A} \in S$$

In the equations above, f is a loss function which we means the total weighted difference between observed value and ground truth. We hope to find the minimal one. a_k is the accuracy of the k-th source and shows how much probability the claims the source makes is true. c_k is the coverage of the k-th source. K, N, M are the amount of sources, objects, properties respectively (Here for easy expression, we assume all sources make all claims about all properties of all objects and missing or insufficient claims do not influence performance). $v_{im}^{(k)}$ represents the truth of m-th property of ith object at current iteration and $v_{im}^{(*)}$is the corresponding value from k-th source. d_m is a distance metric which measure the difference between $v_{im}^{(*)}$ and $v_{im}^{(k)}$. Moreover, d depends on data type of m-th property. Details will be discussed later. S is the set of source accuracy and $\mathcal{A}$ is vector space of accuracy. δ is a kind of normalization function which constrains weight value in case value range is too wide and we can map weight value between 0 and 1. We will talk more particular about this below.

1) Distance Metric: Distance metric depends on data type and how to choose a suitable measurement matters. Similarity is just another aspect of distance and some similarity metrics can also be applied.
For categorical data, we use 0-1 function

$$d_m(v_{im}^{(*)}, v_{im}^{(k)}) = \begin{cases} 1, & if\ v_{im}^{(*)} \neq v_{im}^{(k)} \\ 0, & otherwise \end{cases} \tag{2}$$

For continuous data, we choose normalized absolute deviation.

$$d_m(v_{im}^{(*)}, v_{im}^{(k)}) = \frac{|v_{im}^{(*)} - v_{im}^{(k)}|}{std(v_{im}^{(1)}, \cdots, v_{im}^{(k)})} \tag{3}$$

For multi-value data, Jaccard distance is applied.

$$d_m(v_{im}^{(*)}, v_{im}^{(k)}) = \frac{|v_{im}^{(*)} \cup v_{im}^{(k)}| - |v_{im}^{(*)} \cap v_{im}^{(k)}|}{|v_{im}^{(*)} \cup v_{im}^{(k)}|} \tag{4}$$

For text data, we can adopt Levenshtein distance. For graph data, we can use graph edit distance. In our paper, we mainly talk about the first two.

2) normalization function: normalization function is used to mitigate the effect of outliers and we can map corresponding parameters to a suitable range.

$$\delta(\mathcal{A}) = \sum_{k=1}^{K} exp(-a_k) \tag{5}$$

The reason we choose equation (5) is this function accords with (1). That is to say, the bigger a_k is, the less distance to truth the k-th source can have, otherwise (1) can not reach minimal (detailed explanation can be found in [8]). There are also some other options, such as L^p-norm regularization.

3) Algorithm Design: Our algorithm consists of two steps. The first step is to update source accuracy while truth is fixed. This step is to obtain source accuracy and tell good sources from bad ones. The less further the claims some source makes from truth, the better the source is. The second step is to update truth while weights (accuracy and coverage) of sources are constant. This step is to find the most representative value about the same claim from different sources. Obviously, sources with bigger accuracy and wider coverage should have greater confidence.

Step 1: accuracy update, namely, to solve equation (5)

$$\begin{aligned} &\mathcal{A} \leftarrow \underset{\mathcal{A}}{\operatorname{argmin}} f(X^{(*)}, \mathcal{A}) \\ &s.t. \quad \delta(\mathcal{A}) = 1,\ \mathcal{A} \in S \end{aligned} \tag{6}$$

From [8], we can attain optimal solution:

$$a_k = -log(\frac{c_k \sum_{i=1}^{N} \sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)})}{(\sum_{j=1}^{K} c_j \sum_{i=1}^{N} \sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(j)})} \tag{7}$$

TABLE II: Statistics of weather forecast data

	numbers
Claims	16038
Facts	2100
Ground Truths	1739

Step 2: truth update. In this step, both accuracy and coverage are constant. The key point here is to distinguish different data types and choose suitable distance metric. For categorical data ,$v_{im}^{(k)}$ will be

$$v_{im}^{*} \leftarrow \underset{v}{\operatorname{argmax}} \sum_{k=1}^{K} a_k \cdot c_k \cdot 1(v, v_{im}^{(k)}) \tag{8}$$

The $v_{im}^{(*)}$ actually is the most frequent one weighted by weight and coverage of corresponding sources.

For continuous data , $v_{im}^{(k)}$ will be weighted median and the choosed value v^j can keep the balance of observed value distribution.

$$\sum_{k:v^k<v^j} a_k{\cdot}c_m < \frac{1}{2}\sum_{k=1}^{K} a_k{\cdot}c_k \quad \& \sum_{k:v^k>v^j} a_k{\cdot}c_k < \frac{1}{2}\sum_{k=1}^{K} a_k{\cdot}c_k \tag{9}$$

IV. Experiments

In this section, we firstly introduce the data sets, the experimental settings and the evaluation indicators. Then we show performance of our method.

A. Data set

Weather forecast data. Daily weather situation consist of many property description as shown in Example 1. There are 9 sources in total and we choose high temperature, low temperature and weather condition to do our experiment. We treat the first two as continuous data, the last as categorical data. The details of the dataset can be found in [8]. Statistics are shown in Table II.

Ground truth. We abtain ground truth from real-world weather report about the corresponding day. We implemented our algorithm with matlab, and performed experiments on an Intel Core i7-2600 3.40GHz with 8GB RAM. To evaluate our method, we compute error rate for categorical data and MNAD(Mean Normalized Absolute Distance) for continuous data.

B. Experiment

1) coverage of sources: Table III shows the coverage of each source about continuous and categorical data respectively. vspace-10pt Combining coverage and accuracy, we are less doubtful that the 4th source is the most reliable one, just as presented in Figure.1.

TABLE III: Coverage distribution about different data types of all sources

	continuous data	**categorical data**
Source1	0.8314	0.8314
Source2	0.8314	0.8314
Source3	0.8314	0.8314
Source4	0.8571	0.8571
Source5	0.8571	0.8571
Source6	0.8571	0.8571
Source7	0.8571	0.8571
Source8	0.8571	0.8571
Source9	0.8571	0.8571

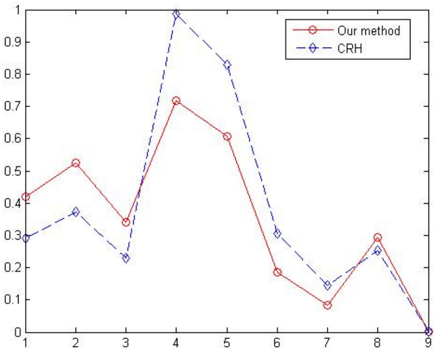

Fig. 1: Confidence degree of each source

2) Truth finding: Compare conflict resolving strategies.

- Voting: For categorical data, choose the most frequently occurred value as truth.
- Median: Calculates the median of all observations on each property of each object as the final output.
- CRH: Take two steps to update weight and truth iteratively by considering different data types together.
- our method: We introduce coverage to update truth.

We use distance to show method performance. In Table IV, error rate means how much percentage of claims is different from ground truth for categorical data and MNAD indicates how much further attained true is from ground truth. For continuous data, MNAD drops from 4.6947 to 4.6724 of 1159 continuous data facts .

TABLE IV: Error rate for each data type

	Error rate	MNAD
voting	0.4845	NA
median	NA	4.9878
CRH	0.3759	4.6947
our method	0.4190	4.6724

C. Analysis

Voting is just a naive way which considers all sources are equivalently important and ignores the property of each data type. CRH considers different data types jointly and regards their characteristic as key factors. Compared to considering accuracy alone of CRH, taking both accuracy and coverage into consideration can help find more reliable truth. The reason is coverage can mitigate contingency and distinguish sources more sharply.

V. Conclusion and Future work

It is challenging and significant to obtain accurate information from massive sources whose data qualities are different. How to tell good source from bad ones and choosing what to believe is a tricky problem. Most of existing work focuses on single attribute of objects or consider attributes separately. Actually, there can be some sort of correlation these attributes and these attributes can affect each other, positively or negatively. Given this, we adopt a model which considers different data types jointly and combines accuracy with coverage.

In future work, we want to take other factors into consideration and these factors including value similarity (how close of two values about same claim), time (nothing is absolutely right and truth is changing with time) and property correlation. Meanwhile, we want to implement the parallelization of the algorithm and run it on Big Data platform such as Spark to improve efficiency.

ACKNOWLEDGEMENT

This work is partially supported by Tianjin Municipal Science and Technology Commission under Grant No. 13ZCZDGX02200, 13ZCZDGX01098, and 13JCQNJC00100. This work is also partially supported by the Ph.D. Programs Foundation of Ministry of Education of China under Grand No. 20120031120038.

References

[1] X. Yin, J. Han, and P. Yu, “Truth discovery with multiple conflicting information providers on the web,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 6, pp. 796–808, June 2008.

[2] X. L. Dong, L. Berti-Equille, and D. Srivastava, “Integrating conflicting data: the role of source dependence,” *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 550–561, 2009.

[3] Y. Cao, W. Fan, and W. Yu, “Determining the relative accuracy of attributes,” in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 2013, pp. 565–576.

[4] Z. Jiang, “A decision-theoretic framework for numerical attribute value reconciliation,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 7, pp. 1153–1169, 2012.

[5] L. Ge, J. Gao, X. Li, and A. Zhang, “Multi-source deep learning for information trustworthiness estimation,” in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 766–774.

[6] N. E. Taylor and Z. G. Ives, “Reconciling while tolerating disagreement in collaborative data sharing,” in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. ACM, 2006, pp. 13–24.

[7] M. Salloum, L. Dong, D. Srivastava, and V. J. Tsotras, “Online ordering of overlapping data sources,” *Proceedings of the VLDB Endowment*, vol. 7, no. 3, 2013.

[8] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, “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, pp. 1187–1198.