Data Veracity Estimation with Ensembling Truth Discovery Methods

Laure Berti-Équille
Qatar Computing Research Institute
Doha, Qatar
Email: lberti@qf.org.qa

***Abstract*—Estimation of data veracity is recognized as one of the grand challenges of big data. Typically, the goal of truth discovery is to determine the veracity of multi-source, conflicting data and return, as outputs, a veracity label and a confidence score for each data value, along with the trustworthiness score of each source claiming it. Although a plethora of methods has been proposed, it is unlikely a technique dominates all others across all data sets. Furthermore, the performance evaluation of the methods entirely depends on the availability of labeled *ground truth* data (*i.e.*, data whose veracity has been manually checked). In the context of Big Data, acquiring the complete ground truth data is out-of-reach. In this paper, we propose an ensembling method that mitigates the two problems of method selection and ground truth data sparsity. Our approach combines the results of a set of truth discovery methods and preliminary experiments suggest that it improves the quality performance over the single methods when samples of ground truth data are used.**

I. Introduction

In the Web, a massive amount of user-generated contents is available through various channels (e.g., tweets, blogs, forum, online encyclopedia, commercial websites, etc.). Conflicting information, rumors, erroneous and fake contents can be easily spread across multiple sources, making it hard to distinguish between what is true and what is not. With the advent of big data, data quality has become more important than ever. Typically, *volume*, *velocity* and *variety* are commonly used to characterize the salient features of big data. The estimation of the fourth "V" *–veracity–* is being more and more recognized as one of the key challenges of big data [19].

Since the early work of [23], the problem of data veracity estimation has been tackled by a large body of research work on *truth finding* (also referred as *fact-checking*) and has also been extended to Web-scale [14], [4], [22], [8]. In the last decade, many fact-finder algorithms have been proposed for fusion of structured data and textual content extraction for Knowledge Base Population [24]. The basic principle of fact-checking is to iteratively compute sources' reliability from the data and estimate the veracity of data from the reliability score of the sources claiming them. A single fact-checking algorithm typically performs very differently for different data sets, as demonstrated by the extensive study of [20]. Typically, quality performance of truth discovery methods is evaluated based on traditional metrics such as accuracy, F1-measure, MAE (Mean Average error), and MSE (Mean Square Error) that rely on the availability of *ground truth* data, *i.e.,* manually verified data, labeled as true or false, for the purpose of the method evaluation.

In this paper, we address two problems: (1) Method selection: Among so many methods for truth discovery, it is very difficult to determine which techniques are best suited for a given data set; (2) Ground truth data sample bias. In practice, only very small samples of ground truth data are available as manual labeling is too costly and time-consuming or simply out-of-reach for Web-scale applications (e.g., [4]). Interestingly, the problem of ground truth data sparsity is related to the problem of sample selection bias: a small set of ground truth data may be considered as a "biased" sample of the true distribution. However, the impact of this problem on the validity of the quality performance measures used in the evaluation of fact-checking methods has been largely understudied.

To address the first problem, ensemble methods, which combine various competing models, have demonstrated to be effective in many disciplines [1]. In this paper, we cast the problem of truth discovery

(a) Affiliations

		S_1	S_2	S_3	S_4	`Ground Truth`	`Conflicts`
d_1	*Bernstein*	MSR	-	AT&T	-	**MSR**	2
d_2	*Carey*	UCI	-	BEA	BEA	**UCI**	2
d_3	*Halevy*	Google	-	UWisc	MSR	**Google**	3
d_4	*Stonebraker*	MIT	UWisc	-	MIT	**MIT**	2
	`Coverage`	1	.25	.75	.75		

(b) Truth Discovery Results

	Depen [3]	TruthFinder [23]	SimpleLCA [17]	3-Estimates [10]	LTM [26]
(c_1, S_1, `Bernstein:AffiliatedTo,MSR`)	false ✗	true ✓	false ✗	false ✗	false ✗
(c_2, S_3, `Bernstein:AffiliatedTo,AT&T`)	true ✗	false ✓	true ✗	true ✗	true ✗
(c_3, S_1, `Carey:AffiliatedTo,UCI`)	false ✗	true ✓	false ✗	false ✗	false ✗
(c_4, S_3, `Carey:AffiliatedTo,BEA`)	true ✗	false ✓	true ✗	true ✗	true ✗
(c_5, S_4, `Carey:AffiliatedTo,BEA`)	true ✗	false ✓	true ✗	true ✗	true ✗
(c_6, S_1, `Halevy:AffiliatedTo,Google`)	true ✓	true ✓	false ✗	false ✗	false ✗
(c_7, S_3, `Halevy:AffiliatedTo,UWisc`)	false ✓	false ✓	false ✓	false ✓	false ✓
(c_8, S_4, `Halevy:AffiliatedTo,MSR`)	false ✓	false ✓	true ✗	true ✗	true ✗
(c_9, S_1, `Stonebraker:AffiliatedTo,MIT`)	true ✓	true ✓	true ✓	true ✓	true ✓
(c_{10}, S_2, `Stonebraker:AffiliatedTo,UWisc`)	false ✓	false ✓	false ✓	false ✓	false ✓
(c_{11}, S_4, `Stonebraker:AffiliatedTo,MIT`)	true ✓	true ✓	true ✓	true ✓	true ✓
Algorithm Accuracy with 100% `Ground Truth`	6/11	11/11	4/11	4/11	4/11
Algorithm Accuracy with 50% `Ground Truth`	6/6	6/6	4/6	4/6	4/6

Table I: Illustrative Example

into a classification problem and propose an ensemble method to improve the accuracy of truth discovery results despite the selection bias of ground truth data samples. To the best of our knowledge, this work is the first attempt to apply an ensembling approach to truth discovery and also to mitigate the ground truth data sample selection bias.

The rest of the paper is structured as follows: Section II gives an illustrative example demonstrating the need for methods that can combine the results of multiple fact-checking methods. It also illustrates that the choice and the size of ground truth data samples are critical. Section III defines the notations and formalizes the problem of ensembling truth discovery methods. Section IV describes the ensembling method. Preliminary experiments are described in Section V. Section VI presents related work. Finally, this paper is concluded in Section VII.

II. Illustrative Example

We consider the truth discovery algorithms that take, as input data, a set of structured claims in the form of quadruplets `(claimID,sourceID,dataItemID,claimedValue)` and infer, as output result, a Boolean truth label for each claim. In addition, the truth discovery algorithms may also return the truthworthiness of each source, and the confidence of each value. Consider the four sources in Table I adapted from [6] providing eleven claims about the affiliation of four researchers. In this example, the ninth claim `(c9,S1,Stonebraker:AffiliatedTo,MIT)` in Table I(b) means that source S_1 claims that Stonebraker is affiliated to MIT (data item d_4). In five cases, the sources have no value for the researcher affiliation. Source coverage is reported in the last line of Table I(a) and represents the number of values claimed by a source over the total number of data items, e.g., the coverage is 1 (100%) for S_1 because it provides a value for each data item from d_1 to d_4. The correct true values are given in bold in `Ground Truth` column of Table I(a). Only S_1 actually provides a correct value for each data item, in conformance with the ground truth data. `Conflicts` column reports the number of distinct values per data item. To discovery the true values without ground truth data and prior knowledge, we could apply majority voting but this would lead to a random guess for data item d_1 (MSR or AT&T) and a wrong decision for data item d_2 (BEA).

In Table I(b), five fact-checking algorithms from the literature have been applied to this very simple data set: Depen [3], TruthFinder [23], SimpleLCA [17], 3-Estimates [10], and LTM [26]. For the sake of simplicity, we voluntarily limited the method outputs to the veracity labels. Table I(b) shows the Boolean veracity labels returned by the methods either in red with ✗ mark when the method result is incorrect with respect to the ground truth data or with ✓when the result is correct. A basic accuracy metric can be computed for each method based on the number of true positives using the available ground truth data set. As illustrated by this example, the truth discovery algorithms perform very differently: e.g., for 100% of the ground truth data, TruthFinder accuracy is 11/11 whereas accuracy is 6/11 for Depen and 4/11 for the other algorithms. Suppose now that the ground truth data is reduced by 50% to only labeled affiliations of Halevy and Stonebraker, respectively at Google and MIT. In the example, the accuracy measures will then be updated and increased for most of the methods (see the last line of Table I(b)): e.g., accuracy increases from 6/11 to 6/6 for Depen and from 4/11 to 4/6 for SimpleLCA, LTM and 3-Estimates. As illustrated by this example, we claim that the ground truth sample selection can tremendously bias the quality evaluation measures, in particular for big data sets where a small size of ground truth data sample cannot be representative enough to reflect accurately the true distribution of the original data set (see discussion in [20]). Such ground truth data samples are not statistically significant to be legitimately used for experimental evaluation and comparative studies.

In this paper, our goal is to address the two problems of method selection and ground truth sample selection bias to improve the quality performance of truth discovery. In the next section, we will describe our ensembling approach for estimating the veracity of data.

III. Problem Definition

We consider the following notations:

- c is a claim value from the set of claims, $c \in C$ and $P(c)$ is the probability distribution of the claim values;
- t is a truth label in $\{\texttt{true}, \texttt{false}\}$ and $P(t)$ is the prior truth probability;
- $\mathcal{R}$ is the set of possible pairs of claim and veracity label, $(c, t) \in \mathcal{R}$, output result of the truth discovery methods;
- $\mathcal{G}$ is the set of labeled ground truth data, $(g, t) \in \mathcal{G}$ with $g \in C$ and $Q(g, t)$ is the probability distribution of the claims in the sample ;
- $P(c, t)$ is the joint probability of having claim c and truth label t, and $P(t|c)$ is the conditional probability for c to have the truth label t.

Suppose k methods for truth discovery $\{M_1, M_2, \ldots, M_k\}$ are executed on the set of claims C. Each method outputs an estimated posterior probability $P(t|c, M_i)$ for each claim c in C.

For the sake of simplicity, we use M to denote any of the k methods M_i and use $\mathcal{M}$ to represent the collection of the k methods. Then any method M's expected mean squared error is the difference between its predicted probability and the true probability integrated over all claims:

$$\begin{aligned} Err^{M} &= \sum_{(c,t)\in\mathcal{R}} P(c,t)(P(t|c) - P(t|c,M))^2 \qquad (1) \\ &= E_{P(c,t)}\left[P(t|c)^2 - 2P(t|c)P(t|c,M) + P(t|c,M)^2\right]. \end{aligned}$$

Suppose each method M has probability $P(M)$ on the input data set and a probability $Q(M)$ on the ground truth data, then in the case of complete ground truth data: $P(M) = Q(M)$ and the expected error incurred by randomly choosing a truth discovery method to do veracity estimation is the above error Err^{M} integrated over all methods:

$$\begin{aligned} Err^{Random} &= \sum_{M\in\mathcal{M}} \sum_{(c,t)\in\mathcal{R}} P(c,t)(P(t|c) - P(t|c,M))^2 \qquad (2) \\ &= E_{P(M),P(c,t)}\left[P(t|c)^2 - 2P(t|c)P(t|c,M) + P(t|c,M)^2\right]. \end{aligned}$$

In the case of partial ground truth data, the error of method M_i can be evaluated based on the ground truth data sample $\mathcal{G}$ such as:

$$Err^{M_i/\mathcal{G}} = \sum_{(c,t_i)\in\mathcal{G}} Q(c,t_i)(Q(t_i|c) - Q(t_i|c,M_i))^2. \quad (3)$$

Assumption: Ground Truth Data Sample Selection Bias. Our assumption is each and every instance of the ground truth data sample is drawn from distribution $Q(c, t)$ where $Q(c, t)$ is a biased distribution, whereas $P(c, t)$ is the true unbiased distribution of all the claims manually verified and labeled in an exhaustive way, $Q(c, t) \neq P(c, t)$.

We can estimate the divergence of the two distributions using a normalized version of the Kullback-Leibler divergence denoted $D(. \parallel .)$ between P and Q in [0,1] such as:

$$D(P \parallel Q) = 1 - \exp\left(-\sum_{(c,t)} P(c,t) \log \frac{P(c,t)}{Q(c,t)}\right) \quad (4)$$

To illustrate this with our previous example, we can compute the divergence between the two distributions of true and false data in the two cases of complete versus partial ground truth data distributions such as:

$$\begin{aligned} & D(P_{GT100\%} \parallel Q_{GT50\%}) \\ &= 1 - \exp\left(- \frac{6}{11} \log \frac{6/11}{3/6} - \frac{5}{11} \log \frac{5/11}{3/6} \right) \\ &= 0.00179547 \end{aligned}$$

where there are 5 true and 6 false claims among 11 in the case of the complete ground truth data set versus 3 false and 3 true claims in the case of the 50% ground truth data set.

Problem definition. Our goal is to select, for each data item, the pair of labeled claim $(c,t)^*$ provided by the method with minimal expected error such as:

$$(c,t)^* = \arg\min_{(c,t)\in\mathcal{R}}(Err^M)$$

where (c,t) is the claim c labeled by method M.

Our goal is to decide the probability of method M being optimal, *i.e.,* with a minimal mean squared error and such that $P(M^*) = 1$ where M^* is the most accurate truth discovery method.

When we do not have access to the complete ground truth dataset, we can adopt a conservative approach since we do not know the true distributions of the methods $P(M)$ and set the KL-divergence equal to a constant noted δ $(0 \leq \delta \leq 1)$. In the next section, we will aggregate the results of the single methods and show that these ensembles can reduce the expected error.

IV. Aggregating the single methods' results

We define two variants of ensembling:

- the first variant uses simple averaging to combine the probability outputs of the different methods with uniform weights (UW) and we use the function $f^{UW}(c) = \frac{1}{k}\sum_{i=1}^{k} P(t|c, M_i)$;
- the second variant adjusts the method's weights (AW) from the ground truth data sample thus the function we use is $f^{AW}(c) = \sum_{i=1}^{k} w_i P(t|c, M_i)$ with $\sum_{i=1}^{k} w_i = 1$.

A. Uniform weights for combining the methods

In this first approach of ensembling, we use the following model: $f^{UW} = E_{P(M)}[P(t|c, M)]$. Then the expected error of this ensemble is the error integrated over the universe of claims:

$$\begin{aligned} Err^{UW} &= \sum_{(c,t)\in\mathcal{R}} P(c,t)(P(t|c) - f^{UW})^2 \qquad (5) \\ &= E_{P(c,t)}\left[P(t|c)^2 - 2P(t|c)f^{UW} + (f^{UW})^2\right]. \end{aligned}$$

When we replace f^{UW} by its expression $E_{P(M)}[P(t|c, M)]$ and compare Err^{UW} with Err^M in Eq. (1). We have $Err^{UW} \leq Err^M$. This result has been already demonstrated in [9] and [25]: the probability averaging of multiple methods is superior to any single method chosen at random with respect to reduction in expected errors. Therefore, $(E_{P(M)}[P(t|c, M)])^2 \leq E_{P(M)}\left[P(t|c, M)^2\right]$.

B. Adjusted weights

Different from the other ensemble approaches, the weight-adjusted ensemble approach assigns a weight to each single method which reflects its predictive accuracy on the labeled claims and the final estimation outputs are combined through weighted averaging. Weights are assigned to the k methods $\{w_1, w_2, \ldots, w_k\}$, each of which is from [0,1] and satisfies the constraint $\sum_{i=1}^{k} w_i = 1$. Ideally, the weight of method M_i $(1 \leq i \leq k)$ ought approximate its true probability $P(M_i)$ as well as possible. As previously defined, the KL-divergence should be minimal. We use the following model: $f^{AW}(c) = \sum_{i=1}^{k} w_i P(t|c, M_i)$ such as:

$$Err^{AW} = \sum_{(c,t)\in\mathcal{R}} \sum_{i=1}^{k} P(c,t)(P(t|c) - w_i P(t|c, M_i))^2. \quad (6)$$

When we don't have any prior knowledge about the optimality of a method, and no preference for some method over others, the constraint $\sum_{i=1}^{k} P(M_i) = 1$ should be satisfied and we can used samples of ground truth to initialize the methods' weights using $Q(M)$ over the sample.

Suppose there are three methods, M_1, M_2, and M_3 with unknown true distributions $P(M_1)$, $P(M_2)$ and $P(M_3)$ uniformly distributed within [0,1]. We have $Q_1(M_1) = 0.4$, $Q_1(M_2) = 0.5$, $Q_1(M_3) = 0.1$ for one ground truth data sample g_1, and $Q_2(M_1) = 0.2$ and $Q_2(M_2) = 0.3$ and $Q_2(M_3) = 0.5$ for another sample g_2. It is unclear that which method would be preferred depending on the ground truth sample.

Control Parameter	Value	Description
`Number of sources (S)`	1,000 to 10,000	The number of sources providing claims: from 1,000 to 10,000
`Number of data items (D)`	1,000 to 10,000	The number of data items, *i.e.*, pairs of (object,attribute) with claimed values

`Source Coverage (Cov)`	`U25; U75`
	`E`
		Uniform: The number of values provided by the sources is uniformly distributed on 25% and 75% of the data items.
		Exponential: The number of values provided by the sources is exponentially distributed across the data items.

`Ground Truth Distribution per Source (GT)`	`R`
	`U25; U75`
	`FP`
	`FO`
	`80P`
	`80O`
	`E`
		Random: The number of true positive claims per source is random.
		Uniform: Each source provides the same number of true positive claims.
		Fully Pessimistic: 80% of the sources provide always false claims and 20% of the sources provide always true positive claims.
		Fully Optimistic: 80% of the sources provide always true positive claims and 20% of the sources provide always false claims.
		80-Pessimistic: 80% of the sources provide 20% true positive claims. 20% of the sources provide 80% true positive claims.
		80-Optimistic: 80% of the sources provide 80% true positive claims. 20% of the sources provide 20% true positive claims.
		Exponential: The number of true positive values provided by the sources is exponentially distributed.

`Distinct Value Distribution per Data item (Conf)`	`U` `E`	**Uniform:** All data items have the same number of distinct values claimed by the set of sources. **Exponential:** Each data item has a number of distinct values that is exponentially distributed.
`Number of Distinct Values`	2...20	The number of distinct values per data item.

Table II: Parameters for Synthetic Data Sets Generation

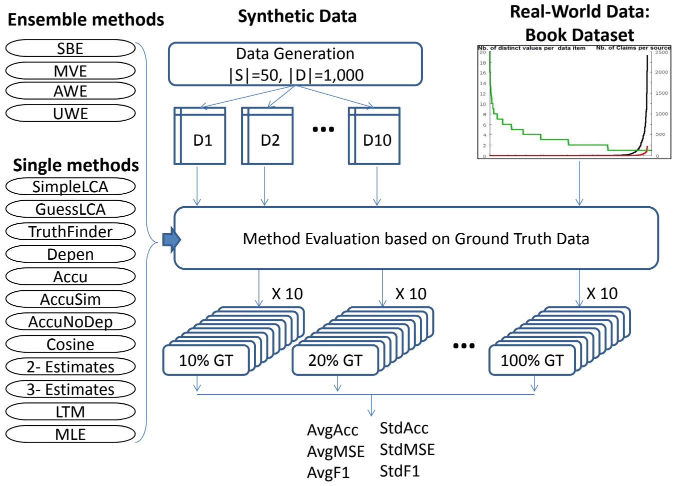

Figure 1: Experimental Pipeline

V. Experiments

We conduct preliminary experiments using both synthetic and real-world data sets to demonstrate the effectiveness of the uniform weight and adjusted weight ensemble methods. This study validates the following claim: the ensemble-based method reduces expected errors compared with single methods, thus is more accurate; and it is more effective than the other tested ensemble approaches.

We implemented twelve algorithms from the literature on truth discovery and four ensembling methods in Java 7 to test as accurately as possible their relative quality performance. We used the API available at: http://daqcri.github.io/dafna/ and data set generator provided by [20]. We ran experiments on 2 PCs with Intel Core i7-2600 processor (3.40GHz×8, 32GB).

A. *Evaluation metrics*

For each method, we evaluate the accuracy, the mean squared error (MSE), defined as the averaged difference between estimated probability and true posterior probability $P(t|c)$, and the F1 measure. In the case where we do not know the true probability, we set the KL-divergence and use various value of δ ($\delta = 0.001; 0.01; 0.1$) . We are comparing four ensemble approaches including Simple Bayesian Ensemble (SBE) from [2], Majority Voting (MVE), Uniform Weight (UWE) and Adjusted Weight (AWE) ensembles with twelve single methods for truth discovery, namely: TruthFinder [23], four variants of Depen [3], two variants of LCA [17], 3-Estimates, 2-Estimates, Cosine [10], LTM [26], and MLE [21]. For all the truth discovery algorithms, we use as default parameter setting the optimal one proposed by the algorithm's authors.

We illustrated the experimental pipeline in Figure 1.

Method	AvgAcc	AvgMSE	AvgF1	StdAcc	StdMSE	StdF1
UWE	**0.7568**	**0.2432**	0.6712	0.2436	**0.0035**	**0.1033**
AWE	0.7468	0.2532	**0.7441**	0.2764	0.0465	0.1043
SBC	0.6978	0.3022	0.6852	0.3370	0.0038	0.1075
MVE	0.6757	0.3243	0.6457	**0.2336**	0.0572	0.1576
TruthFinder	0.6070	0.3930	0.6941	0.3299	0.1487	0.1825
SimpleLCA	0.7133	0.2867	0.6713	0.3673	0.0150	0.1916
GuessLCA	0.6533	0.3467	0.6110	0.3671	0.0250	0.1416
Depen	0.6992	0.3008	0.6657	0.2564	0.0381	0.1347
AccuNoDep	0.6333	0.3667	0.6510	0.3071	0.0150	0.1616
AccuSim	0.6406	0.3594	**0.7301**	0.3250	0.0235	0.2892
Accu	0.7233	0.2767	0.6400	0.3671	**0.0050**	0.1316
3-Estimates	**0.7392**	**0.2608**	0.7012	0.3063	0.0167	0.1348
2-Estimates	0.6792	0.3208	0.6794	**0.2436**	0.0471	**0.1081**
Cosine	0.7292	0.2708	0.7012	0.2964	0.0281	0.1148
MLE	0.7068	0.2932	0.6771	0.4157	0.0320	0.1676
LTM	0.7089	0.2911	0.6176	0.4426	0.0513	0.1549

Table III: Experimental Results on Synthetic Data

Method	AvgAcc	AvgMSE	AvgF1	StdAcc	StdMSE	StdF1
UWE	**0.7868**	**0.2132**	0.7212	0.2936	0.0435	0.1335
AWE	0.7768	0.2232	0.7541	0.2964	**0.0335**	**0.1235**
SBC	0.7678	0.2322	**0.7552**	0.3570	0.0438	0.1448
MVE	0.7657	0.2343	0.7157	**0.2636**	0.0428	0.1776
TruthFinder	**0.7792**	**0.2208**	0.7241	0.3399	0.1687	0.2625
SimpleLCA	0.7333	0.2667	0.6813	0.3773	0.0450	0.2016
GuessLCA	0.7333	0.2667	0.6810	0.3771	0.0450	0.2016
Depen	0.7692	0.2308	0.7257	0.3264	**0.0419**	0.1447
AccuNoDep	0.7333	0.2667	0.6810	0.3771	0.0450	0.2016
AccuSim	0.7106	0.2894	**0.7500**	0.3450	0.0835	0.2992
Accu	0.7333	0.2667	0.6900	0.3771	0.0450	0.2016
3-Estimates	0.7070	0.2930	0.7212	0.3263	0.0433	0.1548
2-Estimates	0.7782	0.2218	0.7094	**0.3236**	0.0429	**0.1381**
Cosine	0.7692	0.2308	0.7212	0.3264	0.0519	0.1448
MLE	0.7568	0.2432	0.7271	0.4257	0.0420	0.1776
LTM	0.7689	0.2311	0.7076	0.4526	0.0713	0.1649

Table IV: Experimental Results on Real-world Data

We evaluate a method M_h on a dataset D_i with ground truth sample G_i to obtain its accuracy p_{ih} and MSE e_{ih}. We report its average accuracy (AvgAcc), average MSE (AvgMSE), average F1 (AvgF1). Furthermore, we keep track of the standard deviations of these measures over 10 runs and variants of ground truth data sample size. A good algorithm will have higher accuracy and F1, a lower MSE and lower standard deviations.

B. Data sets

A first set of experiments has been conducted over synthetic data sets to evaluate the performance of our ensembling methods compared with the base algorithms. A second set of experiments has been conducted over one real-world data set.

First, we generated synthetic data to evaluate the algorithms under a wide range of truth discovery scenarios. Table II summarizes the parameters we used to control the characteristics of the synthetic data set generation.

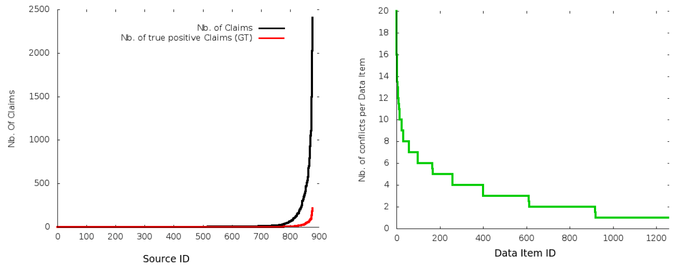

Figure 2: Book Data set

In particular, we control the percentage and distribution of data items for which a source claims a value (`Cov`) and the number and distribution model of distinct values per data item (`Conf`). We generated 10 data sets ($D1, \ldots, D10$), each with $|S| = 50$ sources, $|D| = 1,000$ data items, `Cov=E`, and `Conf=E` for a number of conflicts varying from 2 to 20 conflicts per data item as described in Table II. We selected exponential distribution of conflicts and exponential source coverage as these are the distributions the closest to real-world data distributions. For example, in the case of exponential conflict distribution, few data items may have lots of conflicting values whereas most of data items have few conflicts. In the case of exponential source coverage, very few sources cover most of the data items whereas the majority of the sources covers few data items.

Second, for the experiments on real-world data, we used the Book data set from [5] originally proposed by [23] which consists of 33,235 claims on the author names of 1,263 books by 877 book seller sources. In Figure 2, the distribution of conflicts is represented as a green line; it is exponential as well as the source coverage represented as a black line. The red line represents the ground truth data sample coverage.

C. Ground truth data samples

For synthetic data, we controlled the percentage and distribution model of true positive values per source (`GT` line in Table II). In this way, we generate the exhaustive ground truth data from which samples are randomly selected and used a posteriori for computing the quality metrics of the algorithms. Finally, we varied the size of each ground truth sample with a random selection of the claims to get 10 quality measures for each algorithm over the 10 data sets from 10% to 100% of the exhaustive ground truth data set size, as illustrated in the pipeline of Figure 1.

For the real-world data, the ground truth data sample consists of 100 randomly sampled books for which the book covers were manually verified by the authors of [5] actually representing 100/1263 = 7.91% of the complete ground truth. Similarly, we varied the size of the ground truth data samples from 10% to 100% of the original ground truth data set 10 times to compute the averages and standard deviations of the quality metrics.

D. Results

The results are presented in Table III for synthetic data and in Table IV for the real-world data respectively. They demonstrate that the the ensemble approaches based on uniform and adjusted weights have the best performances on average with higher accuracy and lower standard deviations compared with the best single methods (both in bold).

For both synthetic and real-world data, UWE proves to be the most accurate and stable method with the highest accuracy (0.7568 and 0.7868) and lowest MSE (0.2432 and 0.2132) respectively for synthetic and real-world datasets. The best single method is TruthFinder for the real-world dataset and 3-Estimates for synthetic data sets. However, when we consider the recall and the F1 quality metric, AWE has the highest F1 in both datasets over AccuSim which consistently outperforms other single methods in terms of F1. We can conclude that ensembling methods are more robust in average.

We observed that when facing the problem of small ground truth data sets available for validation, ensemble methods are the best choice to minimize the number of misclassification errors. There are no uniformly best single methods. The large variabilities of single methods result in their relatively higher expected errors in both cases of data sets and on average, ensemble approaches seem to be the most capable of estimating data veracity.

VI. Related Work

Estimating the veracity of data has recently become a very active research field. Since the work of [23], various models have been proposed for truth discovery to incorporate various aspects beyond source trustworthiness and claim belief such as: the dependence between sources [5], [3], the correlation of claims [18], the temporal dimension of evolving truth [7]. Recent contributions relaxing prior modeling assumptions have been proposed to deal with truth existence [28] and approximate truth discovery [22], [13]. Other new developments are related to truth evolution [15], incremental truth discovery [12], truth discovery from data streams [27], and highly dynamic applications of truth discovery in social media and crowd sourcing [11], [16].

However, the study of [20] showed that most of prior work lacks an analysis of computational complexity and usually suffer from scalability issues. Most importantly, it is currently unclear which techniques are the best suited as they are highly data-dependent and their evaluation depends on available samples of ground truth data. To the best of our knowledge, our work is the first one to apply an ensembling method to truth discovery and also the first to address the problem of ground truth data sample selection bias.

VII. Conclusion

Estimating data veracity is recognized as one of the grand challenges of big data. The main issues are related to the selection and evaluation of truth discovery methods when relatively small and biased samples of ground truth data are available for the purpose of validation. Due to the sample selection bias of such ground truth data, quality performance evaluation of truth discovery methods can be invalidated and the experimental results may not be not statistically significant to compare the competing approaches and legitimate the results.

In this paper, we consider the two problems of method selection and ground truth data sample bias and we propose an ensembling method for truth discovery. The experiments provide a preliminary validation of the approach and demonstrates that it can reduce expected errors and give the best quality performance on average for estimating the veracity of data with small ground truth datasets. For future work we will design a suite of experiments (both on synthetic and real-world data) to compare with other ensembling approaches on a larger variety and scale of truth discovery scenarios.

We hope our contribution will serve as a starting point, generating future in-depth research for ensembling of truth discovery methods which could be beneficial for estimating the veracity of big data.

References

[1] T. G. Dietterich. Ensemble methods in machine learning. In *Proc. of the First Workshop on Multiple Classifier Systems*, pages 1–15, 2000.

[2] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Mach. Learn.*, 29(2-3):103–130, Nov. 1997.

[3] X. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. SOLOMON: Seeking the Truth Via Copying Detection. *Proc. of the VLDB Endowment*, 3(2):1617–1620, 2010.

[4] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 601–610, 2014.

[5] X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global Detection of Complex Copying Relationships Between Sources. *Proc. of the VLDB Endowment*, 3(1-2):1358–1369, 2010.

[6] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. *Proc. of the VLDB Endowment*, 2(1):550–561, 2009.

[7] X. L. Dong, L. Berti-Equille, and D. Srivastava. Truth Discovery and Copying Detection in a Dynamic World. *Proc. of the VLDB Endowment*, 2(1):562–573, 2009.

[8] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang. From Data Fusion to Knowledge Fusion. In *Proc. of the VLDB Endowment*, 2014.

[9] W. Fan and I. Davidson. Reverse testing: An efficient framework to select amongst classifiers under sample selection bias. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 147–156, 2006.

[10] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating Information from Disagreeing Views. In *WSDM*, pages 131–140, 2010.

[11] J. Gao, Q. Li, B. Zhao, W. Fan, and J. Han. Truth discovery and crowdsourcing aggregation: A unified perspective. *Proc. of the VLDB Endowment*, 8(12):2048–2059, 2015.

[12] L. Jia, H. Wang, J. Li, and H. Gao. Incremental truth discovery for information from multiple data sources. In *Proc. of the Web-Age Information Management (WAIM) 2013 International Workshops*, pages 56–66, 2013.

[13] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. A confidence-aware approach for truth discovery on long-tail data. *Proc. of the VLDB Endowment*, 8(4):425–436, 2014.

[14] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Scaling up copy detection. *CoRR*, abs/1503.00309, 2015.

[15] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. On the discovery of evolving truth. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 675–684, 2015.

[16] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 745–754, 2015.

[17] J. Pasternack and D. Roth. Latent Credibility Analysis. In *WWW*, pages 1009–1020, 2013.

[18] R. Pochampally, A. D. Sarma, X. L. Dong, A. Meliou, and D. Srivastava. Fusing Data with Correlations. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, 2014.

[19] B. Saha and D. Srivastava. Data quality: The other face of big data. In *Proc. of IEEE 30th International Conference on Data Engineering*, pages 1294–1297, 2014.

[20] D. A. Waguih and L. Berti-Equille. Truth Discovery Algorithms: An Experimental Evaluation, QCRI Technical Report, May , 2014.

[21] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proc. of the 11th International Conference on Information Processing in Sensor Networks (IPSN)*, pages 233–244, 2012.

[22] X. Wang, Q. Z. Sheng, X. S. Fang, X. Li, X. Xu, and L. Yao. Approximate truth discovery via problem scale reduction. In *Proc. of the 24th ACM Conference on Information and Knowledge Management (CIKM)*, October 2015.

[23] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1048–1052, 2007.

[24] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. R. Voss, and M. Magdon-Ismail. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *Proc. of the 25th International Conference on Computational Linguistics (COLING)*, pages 1567–1578, 2014.

[25] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proc. of the Twenty-first International Conference on Machine Learning (ICML)*, 2004.

[26] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. *Proc. of the VLDB Endowment*, 5(6):550–561, 2012.

[27] Z. Zhao, J. Cheng, and W. Ng. Truth discovery in data streams: A single-pass probabilistic approach. In *Proc. of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM)*, pages 1589–1598, 2014.

[28] S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. Modeling truth existence in truth discovery. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1543–1552, 2015.