An Ensemble Approach for Better Truth Discovery

Xiu Susie Fang[1(✉)], Quan Z. Sheng[1], and Xianzhi Wang[2]

[1] School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia
{xiu.fang,michael.sheng}@adelaide.edu.au
[2] School of Computer Science and Engineering, UNSW Australia, Sydney, NSW 2052, Australia
xianzhi.wang@unsw.edu.au

Abstract. Truth discovery is a hot research topic in the Big Data era, with the goal of identifying true values from the conflicting data provided by multiple sources on the same data items. Previously, many methods have been proposed to tackle this issue. However, none of the existing methods is a clear winner that consistently outperforms the others due to the varied characteristics of different methods. In addition, in some cases, an improved method may not even beat its original version as a result of the bias introduced by limited ground truths or different features of the applied datasets. To realize an approach that achieves better and robust overall performance, we propose to fully leverage the advantages of existing methods by extracting truth from the prediction results of these existing truth discovery methods. In particular, we first distinguish between the *single-truth* and *multi-truth* discovery problems and formally define the ensemble truth discovery problem. Then, we analyze the feasibility of the ensemble approach, and derive two models, i.e., *serial model* and *parallel model*, to implement the approach, and to further tackle the above two types of truth discovery problems. Extensive experiments over three large real-world datasets and various synthetic datasets demonstrate the effectiveness of our approach.

Keywords: Truth discovery · Big data · Multi-truths · Ensemble approach

1 Introduction

In the *Big Data* era, various sources may provide description of the same data items (i.e., properties of certain objects). Due to the existence of possible errors, out-of-date data, and missing records, the data collected from different sources may conflict. This makes it of paramount importance to discover the truth from these data to facilitate reliable knowledge discovery and decision making. To this end, tremendous research efforts have been paid to the truth discovery problem from both artificial intelligence and database communities under the

J. Li et al. (Eds.): ADMA 2016, LNAI 10086, pp. 298–311, 2016.
DOI: 10.1007/978-3-319-49586-6_20

topics of *information corroboration* [6], *information credibility* [13], *conflicting data integration* [5], *fact-checking* [7], *data fusion* [8], and *knowledge fusion* [4].

Despite the various truth discovery methods, such as those handling different data types (e.g., categorical and continuous data), and *source dependency* (e.g., copying relation among sources), those considering *source quality* (e.g., source accuracy/recall, specificity, sensitive, and freshness of data) and *object properties* (e.g., the difficulty of and relation between data objects), and those taking into account *value implications* (e.g., *complementary vote*[1]) and *truth properties* (e.g., *multiple truths* and "unknown" truths), no single method can fit or constantly outperform the others in all application scenarios [11] (our experiments on three real-world datasets and various synthetic datasets validate this conclusion). In addition, a recent investigation [10] shows that even an improved method does not always beat its original version.

Although an appropriate truth discovery method can be selected for each specific scenario [11,14], it is challenging to find a method that achieves generally good performance due to the technical limitations and biases of each specific method. As the ensemble approach has been proven to be effective for enhancing the robustness and overall performance of algorithms in many disciplines [2], in this paper, we study on the feasibility of ensembling existing methods for better truth discovery. Realizing such an ensemble truth discovery approach is a tricky task due to the complexity and diversity of existing truth discovery methods. In a nutshell, we make the following contributions in this paper:

- We distinguish between two types of truth discovery problems, i.e., the *single-truth* and *multi-truth* discovery problems, and formally define the ensemble truth discovery problem.
- We analyze the feasibility of the ensemble truth discovery approach, and propose two models, i.e., *serial* and *parallel model*, to implement the approach.
- We empirically evaluate our ensemble approach. Extensive experimental results show that our approach outperforms traditional methods on both real-world and synthetic datasets. In particular, the synthetic datasets with complete ground truths show the improved performance of the ensemble approaches without being biased by the sparsity of limited ground truths.

The rest of the paper is structured as follows. Section 2 reviews the related work. Section 3 defines the ensemble truth discovery problem. Section 4 analyzes the feasibility of the ensemble approach and presents two implementation models, namely the serial and parallel models. Finally, we report the experimental results in Sect. 5, and provide some concluding remarks in Sect. 6.

2 Related Work

Truth discovery has been actively studied by the data integration community in the last few years. Early methods for tackling this issue consist of taking

[1] If a source claims value(s) for a certain object, it implicitly votes against other candidate values of this object.

the mean, median for continuous data, and majority voting for categorical data. These methods commonly neglect sources' quality differences, treat every source equally, and are therefore inaccurate in cases where the majority of sources provide false values. Based on this consideration, various methods incorporate source quality by applying a general principle: a source is more trustworthy if it provides more truths; meanwhile, a value has a bigger possibility of being selected as truth if it is claimed by more high-quality sources. The existing truth discovery methods generally fall into three groups.

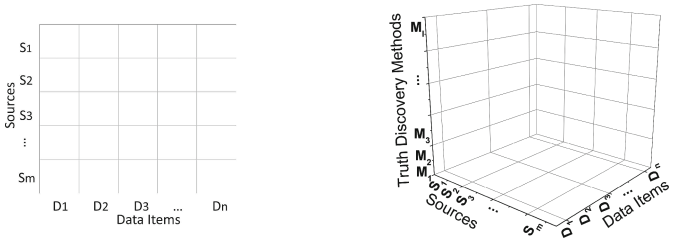
The *iterative* methods predict truths and estimate source reliability iteratively until certain convergence conditions are met. Typical work in this category includes: *TruthFinder* [17], which applies a Bayesian analysis to conduct the iterative processes. *AccuSim* proposed by Dong et al. [5,10] incorporates the implication of value similarity. They further extend AccuSim by additionally considering the copying relations among sources and introduce *AccuCopy*. *AverageLog*, *Investment*, and *PooledInvestment* are developed by Pasternack et al. [12] in order to prevent sources that make more claims from obtaining higher quality weights. *Cosine* and *2-Estimates* are proposed by Galland et al. [6] to adopt complementary vote, they further introduce an improved method 3-Estimates by incorporating "hardness of fact". *SSTF* [16] is a semi-supervised method, which refers to a small set of labeled truths as an additional input data. To relax the single-truth assumption, Wang et al. [15] introduce a Bayesian framework based method *MBM* for multi-truth discovery, in which they also incorporate a finer-grained copy detection technique. The second group is about *optimization based* methods. Both *CRH* [8] and *CATD* [9] model the truth discovery problem as an optimization problem. They differ in that the former is specially designed for handling heterogeneous data while the latter for the long-tail data. The third group is about *probabilistic graphical model based* methods. Methods in this category typically model truths as latent variables. For example, Zhao et al. design a *Gaussian Truth Model* (GTM) [20] for continuous data. *Latent Truth model* (LTM) [19] models source reliability using two metrics, i.e., specificity and sensitive, for multi-truth discovery. *Latent Credibility Assessment* (LCA) [13] additionally considers more factors such as the probability of guessing to facilitate more accurate truth discovery.

A recent survey [10] tests the performance of several methods on two real-world datasets, which shows that no single method always outperforms the others, and nearly half of the mistakes in the best truth discovery results can be avoided if the trustworthiness of sources is known in apriori. More surveys and experimental studies in [14] and [11] show the potential of improving the usability and repeatability of existing truth discovery methods via an ensemble approach. To the best of our knowledge, [1] is the only work that applies an ensemble approach in truth discovery. It proposes two ensemble methods, i.e., *Uniform Weight Ensemble* (UWE) and *Adjusted Weight Ensemble* (AWE), and proves that the ensemble approach can generally mitigate the biases introduced by sparse ground truth and outperform the traditional methods. Our work is the first to formally define the ensemble truth discovery problem and to provide in-depth comparisons of different ensemble methods over both single-truth and multi-truth scenarios.

3 Problem Formulation

For the input of truth discovery, suppose M data sources (e.g., "*Wikipedia*"), $\mathbf{S} = \{S_1, S_2, \ldots, S_m\}$, provide values on N data items (e.g., "*the cast of Harry Potter*"), $\mathbf{D} = \{D_1, D_2, \ldots, D_n\}$. This input data can be visualized as an $M \times N$ data matrix (Fig. 1(a)). Each cell represents a *claim* that describes the value(s) claimed by a source on a data item (e.g., a claim "*July 9, 1956*" for the data item "*the birthday of Tom Hanks*" provided by source "*Wikipedia*"). The values in the cells of the same columns may conflict due to the different reliability of sources. The objective of the truth discovery problem is to predict the truth(s) for each data item (corresponding to a column), given the noisy data matrix, while estimating the reliability of each source (corresponding to a row). Since the numbers of true values may vary among data items in practice, e.g., "*the birthday of Tom Hanks*" contains only one date, but "*the cast of Harry Potter*" includes a team of actors, the truth discovery problem can be classified into two categories: (i) if we make the single-truth assumption by treating the values in each cell (claim) of the matrix as a joint single value, we have the *single-truth discovery problem*; and (ii) if we relax the assumption by treating each distinct value individually, meaning either each cell or the truths may involve several values, we have the *multi-truth discovery problem*. LTM [19] and MBM [15] are the only two methods that are applicable for multi-truth discovery, while all the rest belongs to single-truth discovery methods.

The input of the ensemble truth discovery problem can be formulated as adding a third dimension to the aforementioned data matrix, resulting in a cube (see Fig. 1(b)). The third dimension represents different truth discovery methods, which is denoted as $\mathbf{M} = \{M_1, M_2, \ldots, M_l\}$. Each cell of the cube contains values and their corresponding labels (true or false) provided by the corresponding method. For the single-truth discovery methods, they provide the same label to the value(s) in the same cell, while the multi-truth discovery methods label the value(s) individually. As the methods may have differed performance given a specific application scenario, their results may be conflicting and of varied quality. We formally define the ensemble truth discovery problem as follows:



(a) Truth Discovery (b) Ensemble Approach

Fig. 1. Input dimension comparison of the original and ensemble truth discovery

Ensemble Truth Discovery Problem. Given a 3-dimensional matrix (or cube), **L** truth discovery methods provide boolean labels on values claimed by **M** sources on **N** data items, the objective is to predict the truth of the **N** data items, while estimating the quality of different methods and sources.

4 Ensemble Approaches

4.1 Feasibility Analysis

Berti-Equille implements four approaches including *Simple Bayesian Ensemble* (SBE) [3], *Majority Voting* (MVE), *Uniform Weight* (UWE) and *Adjust Weight* (AWE) ensembles for combining twelve single-truth discovery methods. These approaches are straightforward, which simply unify the outputs of existing methods to the format of a triple {data item, true value(s), veracity score} and combine them directly. Although they are applicable for most of the existing methods, they neglect the useful intermediate results, such as source reliability obtained by the truth discovery methods, thus resulting in limited performance. Moreover, as one of the twelve combined methods, LTM is a special method which incorporates the enriched meaning of source reliability and can tackle multi-truth discovery problem. Naively combining LTM with other single-truth discovery methods and neglecting the two categories of truth discovery problems may further deteriorate the effectiveness of ensemble approaches. In this section, we analyze the feasibility of the ensemble approach and present the possible ways of ensembling the existing methods as follows.

Parallel Model. Although the output formats of existing truth discovery methods vary from one another, they can be transformed into the same format. Therefore, a possible way to ensemble the existing methods is to combine their outputs in a different manner, i.e., *parallel model* (to be detailed in Sect. 4.2).

Serial Model. As aforementioned, the existing methods realize truth discovery following the same general principle. Despite their different ways of implementations, they are generally mutually convertible in their ways of implementations. In particular, both the parameter inference in probabilistic graphical model based methods and the coordinate descent in optimization based methods require updating rules iteratively, which show their potential to be converted into iterative methods; meanwhile, some iterative methods can be formulated as parameter inference tasks or optimization problems. Thus, we can consider using one method's output as another method's input for initializing on the priors, forming the *serial model* (Sect. 4.3).

For either of the above models, we introduce two methods for the two categories of truth discovery problems, i.e., *single-truth discovery ensemble* (S-ensemble) and *multi-truth discovery ensemble* (M-ensemble).

4.2 Parallel Model

The parallel model unifies the format of and combines their outputs to ensemble existing methods. The ensemble truth discovery problem differs from the

traditional truth discovery problem in that it takes 3-dimensional rather than 2-dimensional matrix data as inputs. To realize the parallel ensemble model, we first reduce the dimension of the ensemble problem by regarding each distinctive $(Source, Method)$ pair as a virtual data source. Therefore, a value associated with a large number of $(Source, Method)$ pairs indicates that it is either supported by many sources or predicted as truth by various truth discovery methods. As each method only provides Boolean values to the values provided by sources, we can further remove the values labeled as false to reduce the solution space. After such reduction, the ensemble problem is converted into a traditional truth discovery problem and can be handled using existing methods.

Parallel S-ensemble. This approach first runs all the existing methods and formulates their outputs into a 3-dimensional matrix. Then, it trims the matrix by applying the above-mentioned reduction operations. Finally, it applies one of the existing truth discovery methods on this trimmed matrix to deliver the final results. We call these parallel S-ensemble methods "*PS-Method*" (e.g., PS-Accu). Specially, though there is no copying relation among the original methods, there might be complex latent relations among the sources. In such cases, the source dependence-aware methods, e.g., AccuCopy, are applicable for implementing the ensemble. This is another difference between our work and UWE/AWE, as they simply ensemble the outputs of the methods, and consider the methods to be combined as virtual sources without considering data sources. Thus, they neglect the copying relations among sources.

Parallel M-ensemble. This approach first revises the existing methods under the single-truth assumption so that they can be applied to the multi-truth discovery scenario[2]. In particular, it treats the values in each cell of the matrix individually, and run the original methods to output source reliability. Then, it counts the number of values provided by each source on each data item, and calculates the truth probability of each number as follows:

$$P^*_{D_i}(n) = \sqrt[|S_{D_i}|]{\prod_{n_s=n, s\in S_{D_i}} A(s) \prod_{n_s\neq n, s\in S_{D_i}} (1 - A(s))} \tag{1}$$

where $P^*_{D_i}(n)$ is the unnormalized probability[3] of truth number n of data item D_i, S_{D_i} is the set of sources which provides values on D_i, n_s is the number provided by source s, and $A(s)$ is the reliability of s. For each data item, it chooses the number with the biggest probability as the number of true values (denoted as N) and output the top-N values instead of choosing the value with the biggest confidence score as the outputs. It revises, if necessary, and runs all the truth discovery methods, formulates and trims their outputs as a 3-dimensional matrix. Finally, both the existing multi-truth discovery methods (LTM or MBM) and the revised single-truth discovery methods can be applied to this matrix to address the ensemble problem. We call these parallel M-ensemble methods "*PM-Method*" (e.g., PM-Accu).

[2] Hereafter we call the revised methods the modified single-truth discovery methods.
[3] Such values are then normalized to represent probabilities.

4.3 Serial Model

As an alternative, we can sequentially combine the existing methods, i.e., using one method's outputs as another method's a priori inputs to implement the ensemble approach leading to the serial ensemble model. Here, we simply omit the consideration of the impact of different orders of the single-truth discovery methods on the performance of the ensemble approach, but leave further research on this issue to our future work.

Most existing methods initialize source reliability by assigning uniform weights among the sources. There are some potential disadvantages of the uniform initialization: firstly, with uniform initialization of source reliability, the performance of methods may rely on the majority. This strategy works well for the case that the majority of sources are good. However, the real scenarios usually are not the case, as sources may copy from each other or provide out-of-date information. Moreover, when we apply truth discovery on challenging tasks, such as information extraction and knowledge graph construction, most of the sources are unreliable. For example [18] describes that in their task that "62 % of the true responses are produced only by one or two of the 18 systems (sources)"; secondly, for the scenario where tie cases (i.e., each source claims a unique value on a data item) exists, the results of the methods using uniform initialization are generally unrepeatable. This is because, for the tie cases, the methods would perform voting or averaging like operations and choose a random value as the truth at the beginning of the iteration, leading to randomized source reliability estimation. In contrast, "knowing the precise trustworthiness of sources can fix nearly half of the mistakes in the best fusion results" [10]. Both the above observations motivate us to ensemble existing methods based on a serial ensemble model, which utilizes the source reliability predicted by one method as the prior for initializing another method.

Serial S-ensemble. The sequence of combining the existing methods is a permutation problem. In this paper, we randomly choose the methods one by one, and use the source reliability predicted by a method to initialize its direct successor method. We call the serial S-ensemble methods "*SS-#*" (e.g., SS-3).

Serial M-ensemble. We adapt the single methods, when necessary, by using the same operations designed for parallel M-ensemble. Then, we run the revised methods in the same order as applied for serial S-ensemble. Similarly, we call the serial M-ensemble methods "*SM-#*" (e.g., SM-3).

5 Experiments

5.1 Experimental Setup

We compared our approaches with three groups of truth discovery methods.

Original Single-Truth Discovery Methods (STD). We chose five typical and competitive algorithms from this category for the comparison. Note that Sums was revised by incorporating complementary vote.

– *Voting.* For each item, it predicts the most frequently provided claim as the estimated truth(s) without iteration.
– *Sums, Avg-Log, TruthFinder, 2-Estimates.* All these methods iteratively evaluate source reliability and claims alternately from each other using different calculation methods.

Multi-Truth Discovery Methods (MTD). There are two existing multi-truth discovery methods:

– *LTM.* Based on a probabilistic graphical model, it recognizes a value as true if its veracity score exceeds 0.5.
– *MBM.* This method incorporates a new mutual exclusion definition for multi-truth discovery from the reformatted claims.

Modified Single-Truth Discovery Methods (MMTD). We adapted four representative single-truth discovery methods for the multi-truth scenario by applying the operations described in Sect. 4.2, resulting in four new methods, namely *Voting**, *Sums**, *Average-Log**, *TruthFinder**, and *2-Estimates**.

Based on the above representative methods, we derived methods following our ensemble approaches as follows:

– *Parallel S-Ensemble Group.* It contains five methods, i.e., *PS-Voting*, *PS-Sums*, *PS-AvgLog*, *PS-TruthFinder*, and *PS-Estimates*.
– *Parallel M-Ensemble Group.* It consists of seven methods, i.e., *PM-LTM*, *PM-MBM*, *PM-Voting**, *PM-Sums**, *PM-AvgLog**, *PM-TruthFinder**, and *PM-2Estimates**.
– *Serial S-Ensemble Group.* As Voting does not consider source reliability, we combined the other four single-truth discovery methods and implemented *SS-4*. We combined the four methods in the following order: Sums, Avg-Log, TruthFinder, and 2-Estimates[4], and compared *SS-1* through *SS-4* by gradually adding one method each time in Sect. 5.4.
– *Serial M-Ensemble Group.* We combined six methods in the following order: Sum*, Avg-Log*, TruthFinder*, 2-Estimates*, LTM, and MBM, to implement *SM-6*. We chose this order for the same reason as SS-4). We compared *SM-1* through *SM-6* in Sect. 5.4.

We implemented all the above methods in Java 7 and ran experiments on 3 PCs with Intel Core i7-5600 processor (3.20 GH × 8) and 16 GB RAM. The methods were evaluated in terms of three metrics, including *precision*, which is the average percentage of the true positives returned by the methods in the set of all predicted true values on all values of all data items, *recall*, which is the average percentage of the true positives returned by the methods in the set of ground truths on all values of all data items, and F_1 *score*, which is the harmonic mean of precision and recall, from which we can see the comprehensive performance of all the compared methods.

[4] We chose this order because it is the increasing order of precision of these four methods performed on three real-world datasets in [15].

5.2 Experiments on Real-World Datasets

In this section, we present the evaluation of our ensemble approaches with respect to the existing methods on three real-world datasets (namely *Book dataset* [17], *Biography dataset* [12], and *Movie dataset* [15], described in Table 1), where we have removed the duplicated and invalid records to clean the original datasets.

Table 2 shows the evaluation results. For each single method group (i.e., single-truth discovery method group and multi-truth discovery method group, including the modified single-truth methods), no methods consistently outperformed the others on all the real-world datasets, which is consistent with the previous survey studies [11]. Among those single methods, Voting almost always achieved the best precision. As the data items in all the three real-world datasets

Table 1. Characteristics of three real-world datasets

Book dataset	Biography dataset	Movie dataset
# sources (Websites): 649	# sources (users): 55,259	sources (Websites): 16
# claims: 13,659	# claims: 227,584	# claims: 33,194
attribute: author names	attribute: children	attribute: director names
# objects (books): 664	# objects (person): 2,579	# objects (movies): 6,402
ground truths count (GT):	ground truths count (GT):	ground truths count (GT):
86 books (12.95 %)	2,578 person (99.9 %)	200 movies(3.12 %)
Avg. Coverage per source: 0.0317	Avg. Coverage per source:0.0016	Avg. Coverage per source: 0.0625
Avg. # distinct values per data item	Avg. # distinct values per data item	Avg. # distinct values per data item
(conf): 3.2	(conf): 2.45	(conf): 1.2
Avg. # claims per source: 21.05	Avg. # claims per source: 4.12	Avg. # claims per source: 2074.62

Table 2. Method comparison on real-world datasets and synthetic datasets (The best performance values in each method group are in bold. We consider multi-truth discovery methods and modified single-truth methods as one group. The best performance values among our ensemble approaches are highlighted in the gray background).

Group	Method	Book			Biography			Movie			Syn.(R)	Syn.(80P)
		Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Corr. Rate	Corr. Rate
STD	Voting	**0.837**	0.328	0.471	0.876	0.855	0.865	**0.91**	0.292	0.442	**0.321**	0.581
	Sums	**0.837**	0.54	0.656	0.859	0.881	0.87	0.847	0.591	0.696	0.319	0.623
	AvgL.	0.826	0.605	0.698	0.904	0.886	0.895	0.847	0.643	0.731	0.317	0.58
	TruthF.	**0.837**	0.605	0.702	0.905	0.886	0.895	0.847	**0.71**	**0.772**	**0.32**	0.62
	Est.	**0.837**	**0.621**	**0.713**	**0.908**	**0.888**	**0.898**	0.863	0.692	0.768	0.319	**0.626**
MTD	LTM	**0.826**	0.651	0.728	0.91	0.88	0.895	0.812	0.813	0.812	0.225	0.223
	MBM	**0.826**	**0.744**	**0.783**	**0.915**	**0.89**	**0.902**	0.852	**0.833**	**0.842**	**0.32**	0.533
MMTD	Voting*	0.756	0.638	0.692	0.873	0.851	0.862	0.864	0.523	0.652	0.318	0.586
	Sums*	**0.826**	0.644	0.724	0.905	0.887	0.896	0.81	0.534	0.644	0.319	0.623
	AvgL*	0.663	0.709	0.685	0.88	**0.89**	0.885	0.812	0.65	0.722	0.317	0.58
	TruthF.*	0.698	0.709	0.703	0.876	0.88	0.878	0.853	0.723	0.783	**0.32**	0.623
	Est.*	**0.826**	0.734	0.777	0.89	0.88	0.885	**0.865**	0.722	0.787	0.319	**0.626**
PS-ens.	PS-Voting	**0.837**	0.63	0.719	**0.905**	**0.886**	**0.895**	0.915	0.75	0.824	**0.323**	**0.632**
	PS-Sums	**0.837**	**0.64**	**0.725**	**0.905**	**0.886**	**0.895**	0.92	0.78	0.844	0.322	0.631
	PS-AvgL.	**0.837**	0.638	0.724	**0.905**	**0.886**	**0.895**	0.92	0.78	0.844	0.322	**0.632**
	PS-TruthF.	**0.837**	**0.64**	**0.725**	**0.905**	**0.886**	**0.895**	**0.927**	0.792	0.854	0.322	0.631
	PS-Est.	**0.837**	**0.64**	**0.725**	**0.905**	**0.886**	**0.895**	0.925	**0.816**	**0.867**	0.322	0.631
PM-ens.	PM-Voting*	**0.86**	0.754	0.804	0.91	0.9	0.905	0.899	0.821	0.858	0.321	0.627
	PM-Sums*	0.827	0.751	0.787	0.91	0.89	0.9	0.883	0.833	0.857	0.32	0.627
	PM-AvgLog*	0.829	0.763	0.795	0.915	0.897	0.906	0.886	0.833	0.859	**0.325**	0.623
	PM-TruthF.*	0.834	0.791	0.812	0.91	0.9	0.905	0.886	0.854	0.87	0.322	0.626
	PM-Est.*	0.842	0.766	0.802	0.92	0.89	0.905	0.904	0.846	0.874	0.32	0.626
	PM-LTM	0.837	0.808	0.822	0.93	0.91	0.92	0.91	**0.86**	0.884	0.322	0.623
	PM-MBM	**0.86**	**0.812**	**0.836**	**0.93**	**0.92**	**0.925**	**0.922**	0.85	**0.885**	0.32	**0.628**
SS-ens.	SS-4	0.837	0.721	0.775	0.91	0.9	0.905	0.87	0.753	0.807	0.325	0.628
SM-ens.	SM-6	0.836	0.764	0.798	0.93	0.92	0.925	0.913	0.866	0.889	0.321	0.563

involve multiple true values, LTM and MBM generally achieved better performance than the original single-truth discovery methods, esp. in recall and F_1 score. The modified single-truth discovery methods also achieved relatively higher precision and recall than their original methods. The original single-truth discovery methods showed higher precision but achieved lower recall than multi-truth discovery methods. This indicates that the original single-truth discovery methods tend to underestimate the number of true values.

Both our parallel ensemble methods, i.e., PM and PS, returned better results than the element methods. The serial ensemble methods, i.e., SS-4 and SM-6, also showed relatively better performance. In particular, both PM and SM-6 (resp., PS and SS-4) outperformed the original multiple (resp., single) truth discovery methods they combined in terms of precision, recall and F_1 score on all the three real-world datasets. In our experiments, five single-truth discovery methods are combined for PS and seven multi-truth discovery methods are combined for PM. The obtained 3-dimensional matrices are not significantly different from each other, which resulted in the outcome that all PM and PS methods show similar performance. Due to the existence of multiple true values in the datasets, PM and SM-6 methods performed better than PS and SS-4 methods. However, neither the SM-6 nor the PM methods could consistently dominate the other, and the results are different among different datasets. Similar situations occurred when we compared SS-4 with PS. Further performance studies of SS and SM will be presented in Sect. 5.4.

5.3 Experiments on Synthetic Datasets

Due to the limited ground truths of real-world datasets, the performance evaluation may be biased by the available ground truth. In this section, we present the comparison of our approaches with the element methods on synthetic datasets with a wide spectrum of distribution settings and complete ground truths. We first generated synthetic datasets by applying the dataset generator proposed by Waguih et al. [14]. This generator contains six parameters that can be configured to simulate a wide spectrum of truth discovery scenarios. Three parameters, namely the number of sources (M), the number of data items (N), and the number of distinct values per data item (V), determine the scale of the generated dataset, while the other three parameters, source coverage (cov), ground truth distribution per source (GT), and distinct value distribution per data item ($conf$), determine the characteristics of the generated dataset.

We fixed the scale parameters by setting $M = 50$, $N = 1,000$, and $V = 20$, configured both cov and $conf$ to follow exponential distributions. In particular, we chose two distributions (i.e., the random[5] and 80-pessimistic[6] distributions) for GT. We chose these distributions as they are closest to the real world scenarios. Specifically, for the exponential distribution of $conf$, the majority of data

[5] Random ground truth distribution per source means the number of true positive claims per source is random.

[6] 80-pessimistic ground truth distribution per source means 80 % of the sources provide 20 % true positive claims, while 20 % of the sources provide 80 % true positive claims.

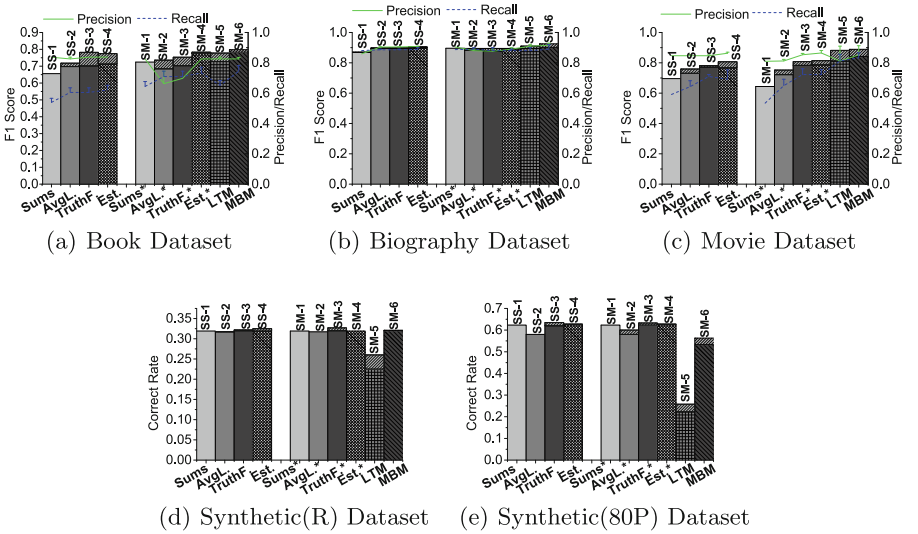
items have few distinct values while few data items have many conflicts. For the case of exponential source coverage, most sources claim values for few data items whereas few sources cover the majority of data items. When we face with the challenging task of information extraction and knowledge base construction, the majority of sources are always error-prone, and truths are maintained by the minority. Therefore, random and 80-pessimistic GT distributions are more representative. Based on the above configurations, we obtained two types of synthetic datasets, namely *Synthetic(R)* and *Synthetic(80P)*, each containing 10 datasets. The metrics of each method were measured as the average of 10 executions over the 10 datasets included by the same dataset type.

Table 2 shows the performance comparison of different methods on the synthetic datasets. As each data item in the synthetic datasets has only one single true value, every method predicted values for all the data items. In this case, we specially measure the methods in terms of *correct rate* by computing the percentage of matched values between each method's output and ground truths. Specifically, the experimental results show almost the same pattern with those on the real-world datasets, which confirms that the ensemble approaches indeed lead to more accurate truth discovery. As sources in Synthetic(R) claim random numbers of true positive values, all methods returned low-quality results for this dataset with correct rate kept around 0.32. Our ensemble methods only showed slightly better performance. The multi-truth discovery methods, especially LTM, failed to return good results on both datasets, where each data item has only one single true value. This is also the reason why SM-6 and PM methods performed worse than SS-4 and PS.

5.4 Impact of Method Numbers on Serial Ensemble Model

To analyze the impact of the number of methods (which are used to derive the ensemble approaches) on the two serial ensemble models (i.e., SS and SM), we conducted experiments on all the above datasets. In particular, we studied the performance of the serial ensemble methods by gradually adding one method each time. We combined the existing methods in the same order as described in Sect. 4.3, where SS-1 is the same as Sums, the source reliability output by Sums was used as the input of AverageLog to realize SS-2. Following a similar way, we further added TruthFinder and 2-Estimates to implement SS-3 and SS-4. Similarly, we gradually combined Sums*, AverageLog*, TruthFinder*, 2-Estimates*, LTM, and MBM to form SM-1 through SM-6. Through the above procedures, we finally obtained four SS methods (from SS-1 to SS-4) and six SM methods (from SM-1 to SM-6).

Figure 2 shows the performance of SS, SM, and the applied existing methods. In particular, the precision, recall and F_1 score of SS and SM fluctuated on all the real-world datasets, and the correct rate of them fluctuated on all the synthetic datasets, while we gradually combined more methods. Each serial ensemble method outperformed the last combined method except the special case of SS-1 (exactly Sums) and SM-1 (exactly Sums*), where the two methods are the same. This indicates that naively and serially combining more methods



(a) Book Dataset (b) Biography Dataset (c) Movie Dataset

(d) Synthetic(R) Dataset (e) Synthetic(80P) Dataset

Fig. 2. Impact of combining different numbers of single methods on SS and SM. The offsets on the precision and recall lines are the corresponding precision and recall of the corresponding SS and SM methods, while the upper bounds of the stack columns are the corresponding F_1 score of the corresponding SS and SM methods.

does not necessarily improve the effectiveness of the serial ensemble methods in a proportional manner. However, the accuracy of a single-truth discovery method could be improved by using the source reliability predicted by other methods as inputs. This indicates parallel ensemble model is generally better than serial ensemble model in obtaining the best ensemble performance.

6 Conclusion

In this paper, we focus on the problem of ensembling the existing truth discovery methods for more robust and consistent truth discovery. Several surveys have shown that a "one-fits-all" truth discovery method is not achievable due to the limitations of the existing methods. Therefore, combing various competing methods could be an effective alternative for conducting high-quality truth discovery. Given very few research efforts have been conducted on this issue, we analyze the feasibility of such an ensemble approach. We propose two novel models, namely *serial model* and *parallel model*, for combining the truth discovery methods. We further present several implementations based on the above models for both single-truth and multi-truth discovery problems. Extensive experiments over three real-world datasets and various synthetic datasets demonstrate the effectiveness of our ensemble approaches.

References

1. Berti-Equille, L.: Data veracity estimation with ensembling truth discovery methods. In: IEEE Big Data Workshop on Data Quality Issues in Big Data (2015)
2. Dietterich, T.G.: Ensemble methods in machine learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems (MCS 2000), Cagliari, Italy (2000)
3. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. **29**(2), 103–130 (1997)
4. Dong, X.L., et al.: From data fusion to knowledge fusion. In: Proceedings of the 40th International Conference on Very Large Data Bases (VLDB 2014), Hangzhou, China (2014)
5. Dong, X.L., et al.: Integrating conflicting data: the role of source dependence. VLDB Endowment (PVLDB) **2**(1), 550–561 (2009)
6. Galland, A., Abiteboul, S., Marian, A., Senellart, P.: Corroborating information from disagreeing views. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010), New York, NY, USA (2010)
7. Goasdoué, F., et al.: Fact checking and analyzing the web. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD 2013), New York, NY, USA (2013)
8. 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 (SIGMOD 2014), Snowbird, Utah, USA (2014)
9. Li, Q., et al.: A confidence-aware approach for truth discovery on long-tail data. VLDB Endowment (PVLDB) **8**(4), 425–436 (2015)
10. Li, X., et al.: Truth finding on the deep web: is the problem solved? VLDB Endowment (PVLDB) **6**(2), 97–108 (2013)
11. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., Han, J.: A survey on truth discovery. ACM SIGKDD Explor. Newsl. (2016)
12. Pasternack, J., Roth, D.: Knowing what to believe (when you already know something). In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Stroudsburg, PA, USA (2010)
13. Pasternack, J., Roth, D.: Latent credibility analysis. In: Proceedings of the 22nd International World Wide Web Conference (WWW 2013), Rio de Janeiro, Brazil (2013)
14. Waguih, D.A., Berti-Equille, L.: Truth discovery algorithms: an experimental evaluation. CoRR abs/1409.6428 (2014)
15. Wang, X., et al.: An integrated Bayesian approach for effective multi-truth discovery. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015), Melbourne, Australia (2015)
16. Yin, X., Tan, W.: Semi-supervised truth discovery. In: Proceedings of the 20th International World Wide Web Conference (WWW 2011), Hyderabad, India (2011)
17. Yin, X., et al.: Truth discovery with multiple conflicting information providers on the web. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007), San Jose, California, USA (2007)

18. Yu, D., et al.: The wisdom of minority: unsupervised slot filling validation based on multi-dimensional truth-finding. In: Proceedings of the International Conference on Computational Linguistics (COLING 2014), Dublin, Ireland (2014)
19. Zhao, B., et al.: A Bayesian approach to discovering truth from conflicting sources for data integration. VLDB Endowment (PVLDB) **5**(6), 550–561 (2012)
20. Zhao, B., Han, J.: A probabilistic model for estimating real-valued truth from conflicting sources. In: Proceedings of 10th International Workshop on Quality in Databases (QDB 2012), Instanbul, Turkey (2012)