Truth Discovery from Conflicting Multi-Valued Objects

Xiu Susie Fang
Supervised by Prof. Michael Sheng
School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia
xiu.fang@adelaide. edu.au

ABSTRACT

Truth discovery is a fundamental research topic, which aims at identifying the true value(s) of objects of interest given the conflicting multi-sourced data. Although considerable research efforts have been conducted on this topic, we can still point out two significant issues unsolved: i) *single-valued assumption*, i.e., current methods assume only one true value for each object, while in reality objects with multiple true values widely exist; ii) *sparse ground truth*, i.e., current works evaluate and compare existing truth discovery methods based on datasets with limited ground truth. Therefore, the empirical studies might be biased and cannot legitimately validate the existing methods. In this PhD project, we propose a full-fledged graph-based model, *SmartMTD* (Smart Multi-valued Truth Discovery), which incorporates four important implications to conduct truth discovery for multi-valued objects. Two graphs are constructed and further used to derive two aspects of source reliability via random walk computations. We also present a general approach, which utilizes Markov chain models with Bayesian inference, for comparing the existing truth discovery methods and validate our approach without ground truth. Initial empirical studies on two real-world datasets show the effectiveness of SmartMTD.

Keywords

Truth Discovery, Multi-Valued Objects, Source Relations, Object Popularity, Performance Evaluation, Ground Truth

1. PROBLEM

In the *Big Data* era, it is easy to observe that multiple sources provide *conflicting* descriptions on the same objects of interest, due to typos, out-of-date data, missing records or erroneous entries. To conduct *truth discovery*, considerable research efforts have been proposed under the single-valued assumption [4, 15, 9, 3]. However, in real world, multi-valued objects—such as the children of a person, the authors of a book—widely exist. One may argue that previous methods can deal with multi-valued objects by simply regarding a value set, which may contain several values, claimed by each source as a joint single value, and determining the most confident value set as the truth. However, the value sets provided by different sources are generally correlated. There may be some overlap between two sources' claimed value sets, indicating that they are not totally voting against each other. Neglecting this implication could degrade the accuracy of truth discovery. Another drawback of the previous single-valued truth discovery methods is that they overlook the important distinction between two aspects of quality, namely, false negatives and false positives, by measuring source quality using a single parameter, such as precision or accuracy. For multi-valued objects, some sources may provide erroneous values, making false positives, while some other sources may provide partial true values without erroneous values, making false negatives. Regarding these two types of errors as equivalent, the previous methods cannot distinguish the quality of those two types of sources. However, measuring source reliability by considering these two different types of errors is crucial to identify the complete true values for multi-valued objects.

The first problem to be solved in this PhD project is truth discovery for multi-valued objects, i.e., MTD. Formally, given a set of multi-valued objects ($\mathcal{O}$), conflicting values $\mathcal{V}$ can be collected from a set of sources ($\mathcal{S}$). We denote the set of all values of an object o provided by all sources in $\mathcal{S}$ as $\mathcal{U}_o$, the set of values provided by a source s on o as $\mathcal{V}_{s_o}$ (i.e., *positive claims*). By incorporating the *mutual exclusion assumption*, s is believed to implicitly invalidate all the other values on o, the disclaimed values are denoted as $\tilde{\mathcal{V}}_{s_o}$ (i.e., *negative claims*), which is calculated as $\mathcal{U}_o - \mathcal{V}_{s_o}$. The first goal of this PhD project is to identify a set of true values ($\mathcal{V}_o{}^*$) from $\mathcal{V}$, for each object o, satisfying that $\mathcal{V}_o{}^*$ is as close to the ground truth $\mathcal{V}_o{}^g$ as possible, while estimating two aspects of source reliability, namely *positive precision* ($\tau(s)$, the probability of the positive claims of a source being true), and *negative precision* ($\tilde{\tau}(s)$, the probability of the negative claims of a source being false). The perfect truth discovery result satisfies $\mathcal{V}_o{}^* = \mathcal{V}_o{}^g$.

For the purpose of performance evaluation of various truth discovery methods, their effectiveness is measured in terms of accuracy (or error rate), F_1-measure, and recall for categorical data, Mean of Absolute Error (MAE) and Root of Mean Square Error (RMSE) for continuous data. All these metrics are measured based on the assumption that complete ground truth is available. However, in reality, ground truth is always limited or even out-of-reach, which is gen-

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053374

erally less than 10% of the original dataset's size [12]. The sparsity of ground truth may bring bias to the performance measurement of the methods. The incomplete ground truth is not statistically significant to be legitimately used for evaluating and comparing existing methods in a systematic way. The second problem to be solved in this PhD project is to compare the existing truth discovery methods and validate our approach without using ground truth.

2. STATE OF THE ART

Significant research efforts have been conducted for truth discovery in various application scenarios (see [7, 12, 8] for surveys). The *primitive* methods are typically *rule-based*, such as the methods that take the *majority voting* (for categorical data) or the *mean* (for continuous data) as the true values. These methods do not distinguish the reliability of sources and therefore have low accuracy when many sources provide low quality data. Yin et al. [15] first formulated the truth discovery problem in 2008. Since then, many advanced solutions have been proposed by additionally considering various implications of multi-source data. They generally fall into five categories. *Web-link* based methods [9, 4, 9, 10, 16] conduct random walks on the bipartite graph between sources and values of objects. They measure webpage authority based on their links to the claimed values and estimate source reliability and value correctness based on the bipartite graph. *Iterative* methods [15, 9, 3] iteratively calculate value veracity and source reliability from each other until certain convergence condition is met. *Bayesian point estimation* methods [1, 2, 14] adopt *Bayesian analysis* to compute the maximum posteriori probability or *MAP* value for each object. *Probabilistic graphical model* based methods [18, 11, 17] apply probabilistic graphical models to jointly reason about source trustworthiness and value correctness. Finally, *optimization* based methods [6, 13, 5] formulate the truth discovery problem as an optimization problem.

Despite active research in the field, MTD is rarely studied by the previous work. To the best of our knowledge, only two related works exist. LTM (Latent Truth Model) [17], a probabilistic graphical model based method, is the first solution. LTM makes strong assumptions about prior distributions for nine latent variables, rendering the model inhibitive and intractable to incorporating various implications to improve its performance. Moreover, Waguih et al. [12] conclude with extensive experiments that this type of methods cannot scale well. To relax unnecessary assumptions, Wang et al. [14] analyze the unique features of MTD and propose a MBM (Multi-truth Bayesian Model), which incorporates *source confidence* and finer-grained copy detection techniques into a Bayesian framework. Compared to those two methods, SmartMTD is a graph-based method, which incorporates four important implications to pursue better truth discovery.

To conduct performance evaluation and comparison with the state-of-the-art truth discovery methods, previous comparative studies, such as [6] and [7], conduct experiments on real-world datasets and sparse gold standards. Waguih et al. [12] point out that the sparse ground truth is not statistically significant to be legitimately leveraged for method accuracy evaluation and comparison. To the best of our knowledge, they are the first to implement a dataset generator to generate synthetic datasets with the control of the complete ground truth distribution, for the sake of comparing twelve existing methods. Different from their work, our idea is to propose an approach for comparing the existing methods without ground truth.

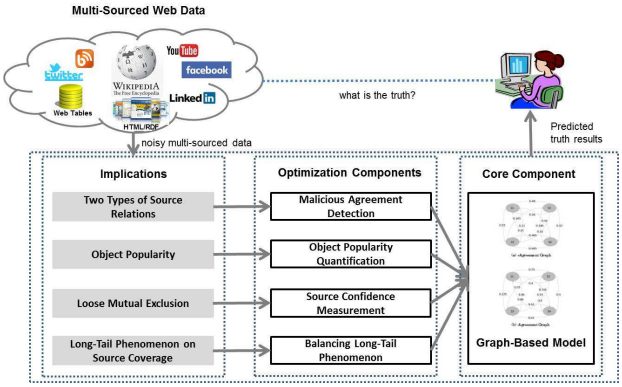

Figure 1: The framework of SmartMTD.

3. PROPOSED APPROACH

3.1 The SmartMTD

Fig. 1 shows our SmartMTD framework. Our model incorporates four implications, including two types of source relations, object popularity, loose mutual exclusion, and long-tail phenomenon on source coverage by integrating four optimization components into one graph-based core component.

The core component applies the following principle for truth discovery [8]: sources providing more true values are assigned with higher reliability; meanwhile, values provided by higher-quality sources are more likely to be true. Value confidence scores and source reliability are iteratively calculated from each other until convergence.

Intuitively, if the positive (resp., negative) claims of a source are agreed by the majority of other sources, this source is likely to have high positive (resp., negative) precision. This means that the inter-source agreements (i.e., the common values claimed or disclaimed by two sources) indicate source reliability endorsement. This intuition motivates us to measure the source positive (resp., negative) precision by quantifying the +agreements (resp., –agreements) among sources, i.e., the agreements among sources regarding their positive (resp., negative) claims. In reality, sources might not only support one another by providing the same true claims, but also may maliciously copy from others to provide the same false claims, which sometimes mislead the audience. Therefore, we identify two types of source relations. Specifically, sharing the same true values means one source supports/endorses the other source, indicating a *supportive relation* between two sources. We define the common values between these two sources as *supportive agreements*. We measure source reliability by quantifying source supportive agreements in the core component. On the contrary, if two sources share a significant amount of false values, they are likely to copy from each other, indicating a *copying relation* between them. We define these common false values as *malicious agreements*. We measure source independence scores by quantifying source malicious agreements in the *Malicious agreement detection component*.

To derive source positive and negative precision, the core component constructs $\pm$*supportive agreement graph*s. In

each graph, the vertices denote sources, each directed edge represents that one source agrees with the other source, and the weight on each edge depicts to what extent one source endorses the other source. The constructions of ±supportive agreement graphs incorporate the outputs of the four optimization components. In particular, we calculate the endorsement degree from s to s' on positive claims by:

$$\mathcal{A}(s,s') = \mathcal{L}(s,s') + \sum_{o\in\mathcal{O}_s\cap\mathcal{O}_{s'}} \frac{|A_o(s,s')|}{|\mathcal{V}_{s'o}|} \cdot(1-\prod_{v\in A_o(s,s')} \mathcal{C}_{\tilde{v}})\cdot\mathcal{P}_o\cdot I(s,o)\cdot\mu(s,o) \quad (1)$$

$$A_o(s,s') = \mathcal{V}_{s_o}\cap\mathcal{V}_{s'_o} \quad (2)$$

where $A_o(s,s')$ is the agreement between the positive claims of s and s' on o, $\mathcal{O}_s$ is the set of objects covered by s, $I(s,o)$ is the independence score of s providing positive claims on o, $\mathcal{P}_o$ is the popularity degree of o, $\mu(s,o)$ is the confidence score of s providing positive claims on o, and $\mathcal{L}(s,s')$ is the long-tail phenomenon compensation of edge from s to s'.

We calculate the weight on each edge of +supportive agreement graph using:

$$\omega(s\to s') = \beta + (1-\beta)\cdot\frac{\mathcal{A}(s,s')}{|\mathcal{O}_{s'}|} \quad (3)$$

where β is a smoothing factor, which guarantees that the graph is always connected and source reliability calculation can converge.

We define the calculation of edge weights of –supportive agreement graph in the similar way.

We apply the *Fixed Point Computation Model* (FPC) random walk to those two graphs, after normalization steps, we obtain $\tau(s)$ and $\tilde{\tau}(s)$ as positive and negative precision for each source.

To jointly determine value veracity from source reliability, we compute the confidence scores of each value v being true and false by:

$$\mathcal{C}_v = \frac{\sum_{s\in\mathcal{S}_v}\tau(s) + \sum_{s\in\mathcal{S}_{\bar{v}}}(1-\tilde{\tau}(s))}{|\mathcal{S}_o|} \quad (4)$$

$$\mathcal{C}_{\bar{v}} = \frac{\sum_{s\in\mathcal{S}_v}(1-\tau(s)) + \sum_{s\in\mathcal{S}_{\bar{v}}}\tilde{\tau}(s)}{|\mathcal{S}_o|} \quad (5)$$

$\mathcal{S}_o$ is the set of sources provide values on o, and $\mathcal{S}_v$ (resp., $\mathcal{S}_{\bar{v}}$) is the set of sources claim (resp., disclaim) v on o.

The four optimization components compute the parameters regarding the four implications required by the core component.

Malicious agreement detection component derives the independence score of each source providing claims on each object by constructing ±malicious agreement graphs for sources in $\mathcal{S}_o$, for each object $o\in\mathcal{O}$. The edge weight of each graph is calculated by:

$$\omega_{c_o}(s\to s') = \beta+(1-\beta)\cdot\frac{|A_o(s,s')|}{|\mathcal{V}_{s'o}|}\cdot(1-\prod_{v\in A_o(s,s')}\mathcal{C}_v)\cdot\mu(s,o) \quad (6)$$

$$\tilde{\omega}_{c_o}(s\to s') = \beta+(1-\beta)\cdot\frac{|\tilde{A}_o(s,s')|}{|\tilde{\mathcal{V}}_{s'o}|}\cdot(1-\prod_{v\in\tilde{A}_o(s,s')}\mathcal{C}_{\tilde{v}})\cdot\tilde{\mu}(s,o) \quad (7)$$

After random walk computations and normalization, we obtain the independence scores of each source on each object.

Intuitively, sources tend to publish more popular information to gain more attentions from the public, and the objects with more occurrences in the sources' claims indicate that they are more popular. Since the number of potential audiences of popular objects is usually bigger than that of less popular objects, if a source provides false values on a popular object, it will mislead more people than on a less popular object. For example, the phone number of a restaurant is more popular and has bigger impact than the year when it is opened because customers need to contact the restaurant. We therefore propose to distinguish source reliability by differentiating the popularities of objects, to minimize the number of people misguided by false values. *Object popularity quantification component* calculates the popularity of each object by applying the following equation, which comprehensively incorporates the occurrence of the object and the coverage of each source that covers the object:

$$\mathcal{P}_o^u = \sum_{s\in\mathcal{S}_o}\frac{1}{Cov(s)} \quad (8)$$

$\mathcal{P}_o^u$ is then normalized to obtain $\mathcal{P}_o$.

For multi-valued object, since sources may cautiously provide partial true values and omit the values they are not sure about, or audaciously provide all potential values, even if the veracity of the claimed values is uncertain, the mutual exclusion among values is not as strict as that of the single-valued object, i.e., the loose mutual exclusion. For this reason, SmartMTD uses *source confidence measurement component* to calculate the source confidence scores of providing positive (resp., negative) claims on each object, and reconcile sources' belief in their positive and negative claims. In particular, $\mu(s,o)$ and $\tilde{\mu}(s,o)$ are calculated as:

$$\mu(s,o) = \frac{1}{|\mathcal{V}_{s_o}|}\cdot(1-\frac{1}{|\mathcal{U}_o|}) \quad (9)$$

$$\tilde{\mu}(s,o) = \frac{1}{|\tilde{\mathcal{V}}_{s_o}|}\cdot\frac{1}{|\mathcal{U}_o|} \quad (10)$$

Finally, in reality, various datasets show the long-tail phenomenon on source coverage, which refers to the fact that very few sources provide extensive coverage for the objects of interest and most of the source only provide values for very few objects. The *balancing long-tail phenomenon component* calculates the compensation of long-tail phenomenon on source coverage for each link in the ±supportive agreement graphs to avoid small sources with very few claims from being assigned with extreme reliability.

$$\mathcal{L}(s,s') = \beta_{\mathcal{L}}\sum_{o\in\mathcal{O}_{s'}-(\mathcal{O}_s\cap\mathcal{O}_{s'})}\frac{1}{8}\cdot\mathcal{I}_o\cdot\frac{1}{|\mathcal{U}_o|}(1-\frac{1}{|\mathcal{U}_o|}) \quad (11)$$

$$\tilde{\mathcal{L}}(s,s') = \beta_{\mathcal{L}}\sum_{o\in\mathcal{O}_{s'}-(\mathcal{O}_s\cap\mathcal{O}_{s'})}\frac{1}{8}\cdot\mathcal{I}_o\cdot\frac{1}{|\mathcal{U}_o|}^2 \quad (12)$$

where $\beta_{\mathcal{L}}$ is an uncertainty factor of the compensation.

3.2 The Comparison Approach - Ongoing Work

To evaluate the existing truth discovery methods without ground truth, we propose an approach that utilizes Markov chain models with Bayesian inference. In particular, given a dataset for truth discovery, we first utilize a Markov chain

model for modeling the statistical relations among sources as an adjacency matrix (we name it as *raw matrix*), with each element as the transition probability between the corresponding two sources. We then consider the output of each truth discovery method as the hypothesis about the raw matrix. To mathematically express the hypothesis, we also construct an adjacency matrix by leveraging the source reliability and value confidence scores output by the truth discovery methods, as those evaluations are the key elements for identifying the truth of each object. The hypothesis matrices are then used for eliciting informative Dirichlet priors using an adapted version of the (trial) roulette method. Finally, we leverage the sensitivity of Bayes factors on the priors for comparing hypotheses with each other.

4. METHODOLOGY

The methodology used in the development of this work comprises five tasks:

i) Extensive literature review: this includes the study of the literature about existing truth discovery methods, how current works compare their works with others, and the datasets and ground truth current works utilize for their empirical studies.

ii) Formalization of the MTD problem.

iii) Development of a graph-based approach as overall solution incorporating four important implications. We need to quantify the implications, and then implement SmartMTD.

iv) Evaluation and validation of SmartMTD to measure its performance with respect to the other state-of-the-art truth discovery methods. The experiments need to be conducted as follows:

(1) we first collect known datasets to run our experiments, and choose several typical and competitive existing methods for comparison, and exclude the methods that are inapplicable to the MTD scenario. To ensure the fair comparison, we need to run a series of experiments to determine the optimal parameter settings for each baseline method.

(2) execution of experiments and statistical studies of the obtained results to deduce conclusions about the proposed solution. All methods should be evaluated in terms of precision (P), recall (R), F_1 score (F1), and execution time (T). Since we introduce a new concept of object popularity, we use object popularity weighted precision (WP), recall (WR) and F_1 score (WF1) as additional accuracy metrics.

(3) To evaluate the impact of different implications, we should implement several variants of SmartMTD, and analyze the experimental results.

v) We first model the truth discovery datasets as source transition matrices and the existing truth discovery methods as hypothesis matrices. Then we implement our approach for comparing those matrices.

vi) Evaluation and validation of our comparison approach. We utilize both real-world datasets with limited ground truth and the synthetic datasets with complete ground truth generated by the dataset generator proposed by Waguih et al. [12]. The performance evaluation output by our comparison approach need to be compared with the P, R, F1 of each method. The experimental results are then finally analyzed and concluded.

5. RESULTS

This PhD project is at the beginning of the third year, we have done an extensive literature review, and implemented SmartMTD. We have also done some preliminary experiments on the approach for comparing truth discovery methods without using ground truth. The initial results show the potential of our approach. In this section, due to the limited space, we briefly report our progress of the implementation of SmartMTD, and omit the empirical studies of our second research problem.

We compared SmartMTD with three types of baseline methods: i) methods under single-valued assumption (i.e., STD methods), including *Voting*, *Sums* [4], *Average-Log* [9], *TruthFinder* [15], and *2-Estimates* [3]; ii) existing MTD methods, *LTM* [17] and *MBM* [14]; iii) modified STD methods, we modified the above five STD methods by incorporating truth number prediction. In particular, for each method, we treated the values in each claimed value set of each source individually, and ran the original method to output source reliability and value confidence scores. Then, we computed $|\mathcal{V}_{s_o}|$ for each source on each object, based on which we predicted the number of true values for each object by:

$$P_o(n) = \sqrt[|\mathcal{S}_o|]{\prod_{|\mathcal{V}_{s_o}|=n, s\in\mathcal{S}_o} A(s) \cdot \prod_{|\mathcal{V}_{s_o}|\neq n, s\in\mathcal{S}_o} (1 - A(s))} \tag{13}$$

where $P_o(n)$ is the unnormalized probability[1] of the number of values of an object o to be n, and $A(s)$ is the reliability of s calculated by each method.

For each object, we chose the number with the highest probability (denoted as N) as the number of true values and output the top-N values instead of choosing the value set with the biggest confidence score as the outputs. Finally, we obtained five new methods, namely *Voting**[2], *Sums**, *Average-Log**, *TruthFinder**, and *2-Estimates**. We ran the above methods on two real-world datasets, including *book-author dataset* [15] and *Parent-Children dataset* (extracted from *Biography dataset* [9]).

Table 1 shows the comparison results. For all the accuracy evaluation metrics except precision, SmartMTD consistently achieved the highest value. Even in terms of precision, SmartMTD still achieved the second best performance on Parent-Children dataset. Though SmartMTD sacrificed precision for recall due to the limited size of the Book-Author dataset, SmartMTD achieved the best F_1 score as the overall performance. Among the three methods specially designed for the MTD problem, our approach is the most efficient one with the lowest execution time. This is due to the reasons that LTM includes complicated Bayesian inference over the complex probabilistic graphical model, and MBM conducts time-consuming copy detection, while our approach is based on a relatively simple graph model. The modified STD methods performed even worse than their original versions. This depicts that in reality the majority of the sources tend to be cautious and only provide values they are sure to be true, thus the predicted numbers of true values were generally smaller than the real ones, leading to lower precision and recall of the modified STD methods.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we present two significant problems we aim to solve in this PhD project. Firstly, we have solved the problem of discovering true values for multi-valued objects

[1]Such values are then normalized to represent probabilities.

[2]For Voting*, we predict the number of true values as the number with the highest vote counts.

Table 1: Comparison of different methods: the best and second best performance values are in bold.

Method	Book-Author Dataset							Parent-Children Dataset						
	P	R	F1	WP	WR	WF1	T(s)	P	R	F1	WP	WR	WF1	T(s)
Voting	**0.84**	0.63	0.72	**0.83**	0.64	0.72	**0.07**	0.88	0.85	0.87	0.69	0.68	0.69	**0.56**
Sums	**0.84**	0.64	0.73	**0.83**	0.64	0.72	0.85	**0.90**	0.89	0.90	**0.88**	0.86	0.87	1.13
Avg-Log	**0.83**	0.60	0.70	**0.83**	0.64	0.72	0.61	**0.90**	0.89	0.89	**0.88**	0.86	0.87	**0.75**
TruthFinder	**0.84**	0.60	0.70	**0.83**	0.60	0.70	0.74	**0.90**	0.89	0.90	**0.88**	0.85	0.86	1.24
2-Estimates	0.81	0.70	0.75	0.80	0.68	0.74	0.38	**0.91**	0.89	0.90	**0.88**	0.86	0.87	1.34
Voting*	0.77	0.42	0.54	0.80	0.39	0.53	0.13	0.87	0.85	0.86	0.71	0.68	0.69	0.89
Sums*	**0.83**	0.24	0.38	**0.85**	0.21	0.34	0.99	0.86	0.88	0.87	0.67	0.84	0.75	1.45
Avg-Log*	0.74	0.49	0.59	0.80	0.53	0.64	**0.08**	0.89	0.87	0.88	0.77	0.82	0.79	0.92
TruthFinder*	0.70	0.71	0.70	0.75	**0.72**	0.73	0.99	0.85	0.91	0.88	0.69	0.88	0.77	1.16
2-Estimates*	**0.83**	0.24	0.38	0.81	0.21	0.34	0.79	0.86	0.89	0.87	0.66	0.83	0.74	1.47
LTM	0.82	0.65	0.73	0.82	0.62	0.71	0.98	0.87	0.90	0.88	0.86	0.89	0.87	0.99
MBM	**0.83**	**0.74**	**0.78**	0.82	0.71	**0.76**	0.67	**0.90**	**0.92**	**0.91**	0.87	**0.90**	**0.88**	2.17
SmartMTD	0.81	**0.79**	**0.80**	**0.83**	**0.81**	**0.82**	0.45	**0.90**	**0.94**	**0.92**	**0.92**	**0.95**	**0.93**	0.92

(or MTD), which has rarely been studied in the truth discovery community. We propose a full-fledged graph-based model, *SmartMTD*, by incorporating four implications including *two types of source relations* (i.e., *supportive relations* and *copying relations*), *object popularity*, *loose mutual exclusion*, and *long-tail phenomenon on source coverage*. In particular, we construct ±*supportive agreement graph*s to model the endorsement among sources on their positive and negative claims, from which two-sided source reliability is derived. Copying relations among sources are captured by constructing the ±*malicious agreement graph*s based on the consideration that sources sharing the same false values are more likely to be dependent. We consider the popularities of objects and develop techniques to quantify object popularity based on object occurrences and source coverage. We apply source confidence scores to differentiate the extent to what a source believes its positive claims and negative claims. For the ubiquitous long-tail phenomenon on source coverage, we also add smoothing weights to the ±supportive agreement graphs to avoid the reliability of small sources from being over- or under-estimated. Empirical studies on two real-world datasets show the effectiveness of SmartMTD. Secondly, we propose an initial idea of comparing the truth discovery methods without ground truth. In the rest of this PhD project, we will further complete the design of our approach, including modeling the raw datasets as empirical source transition matrices, and modeling the methods as hypothesis matrices. Finally, validate the effectiveness of SmartMTD by applying our novel comparison approach. Our future work will focus on improving SmartMTD by exploring and incorporating more implications.

7. REFERENCES

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

[2] X. L. Dong, B. Saha, and D. Srivastava. Less is more: selecting sources wisely for integration. *Proc. the VLDB Endowment*, 6(2):37–48, 2012.

[3] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *Proc. ACM International Conference on Web Search and Data Mining (WSDM)*, pages 131–140, 2010.

[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[5] 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. the VLDB Endowment*, 8(4), 2014.

[6] 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 *Proc. ACM SIGMOD International Conference on Management of Data*, pages 1187–1198, 2014.

[7] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: is the problem solved? *Proc. the VLDB Endowment*, 6(2):97–108, 2012.

[8] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, 17(2):1–16, 2016.

[9] J. Pasternack et al. Knowing what to believe (when you already know something). In *Proc. Intl. Conference on Computational Linguistics (COLING)*, pages 877–885, 2010.

[10] J. Pasternack and D. Roth. Making better informed trust decisions with generalized fact-finding. In *Proc. International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 2324–2329, 2011.

[11] J. Pasternack and D. Roth. Latent credibility analysis. In *Proc. International World Wide Web Conference (WWW)*, pages 1009–1020, 2013.

[12] D. A. Waguih and L. Berti-Equille. Truth discovery algorithms: an experimental evaluation. *arXiv preprint arXiv:1409.6428*, 2014.

[13] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: a maximum likelihood estimation approach. In *Proc. ACM International Conference on Information Processing in Sensor Networks (Sensys)*, pages 233–244, 2012.

[14] X. Wang et al. An integrated bayesian approach for effective multi-truth discovery. In *Proc. the 24th ACM Intl. Conference on Information and Knowledge Management (CIKM)*, pages 493–502, 2015.

[15] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 20(6):796–808, 2008.

[16] X. Yin and W. Tan. Semi-supervised truth discovery. In *Proc. International World Wide Web Conference (WWW)*, pages 217–226, 2011.

[17] B. Zhao et al. A bayesian approach to discovering truth from conflicting sources for data integration. *Proc. the VLDB Endowment*, 5(6):550–561, 2012.

[18] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *Proc. International Workshop on Quality in DataBases (QDB), coheld with VLDB*, 2012.