Crowdsourced Truth Discovery in the Presence of Hierarchies for Knowledge Fusion

Woohwan Jung
Seoul National University
whjung@kdd.snu.ac.kr

Younghoon Kim
Hanyang University
yhkim7951@gmail.com

Kyuseok Shim*
Seoul National University
shim@kdd.snu.ac.kr

ABSTRACT

Existing works for truth discovery in categorical data usually assume that claimed values are mutually exclusive and only one among them is correct. However, many claimed values are not mutually exclusive even for functional predicates due to their hierarchical structures. Thus, we need to consider the hierarchical structure to effectively estimate the trustworthiness of the sources and infer the truths. We propose a probabilistic model to utilize the hierarchical structures and an inference algorithm to find the truths. In addition, in the knowledge fusion, the step of automatically extracting information from unstructured data (e.g., text) generates a lot of false claims. To take advantages of the human cognitive abilities in understanding unstructured data, we utilize crowdsourcing to refine the result of the truth discovery. We propose a task assignment algorithm to maximize the accuracy of the inferred truths. The performance study with real-life datasets confirms the effectiveness of our truth inference and task assignment algorithms.

1 INTRODUCTION

Automatic construction of large-scale knowledge bases is very important for the communities of database and knowledge management. Knowledge fusion (KF) [8] is one of the methods used to automatically construct knowledge bases (a.k.a. knowledge harvesting). It collects the possibly conflicting values of objects from data sources and applies *truth discovery* techniques for resolving the conflicts in the collected values. Since the values are extracted from unstructured or semi-structured data, the collected information exhibits error-prone behavior. The goal of the *truth discovery* used in knowledge fusion is to infer the true value of each object from the noisy observed values retrieved from multiple information sources while simultaneously estimating the reliabilities of the sources. Two potential applications of knowledge fusion are web source trustworthiness estimation and data cleaning [10]. By utilizing truth discovery algorithms, we can evaluate the quality of web sources and find systematic errors in data curation by analyzing the identified wrong values.

Truth discovery with hierarchies: As pointed out in [6, 8, 25], the extracted values can be hierarchically structured. In this case, there may be multiple correct values in the hierarchy for an object even for functional predicates and we can utilize them to find the most specific correct value among the candidate values. For example, consider the three claimed values of 'NY', 'Liberty Island' and 'LA' about the location of the Statue of Liberty in Table 1. Because Liberty Island is an island in NY, 'NY' and 'Liberty Island' do not conflict with each other. Thus, we can conclude that the Statue of Liberty stands on Liberty Island in NY.

Table 1: Locations of tourist attractions

Object	Source	Claimed value
Statue of Liberty	UNESCO	NY
Statue of Liberty	Wikipedia	Liberty Island
Statue of Liberty	Arrangy	LA
Big Ben	Quora	Manchester
Big Ben	tripadvisor	London

We also observed that many sources provide generalized values in the real-life. Figure 1 shows the graph of the generalized accuracy against the accuracy of the sources in the real-life datasets *BirthPlaces* and *Heritages* used for experiments in Section 5. The accuracy and the generalized accuracy of a source are the proportions of the exactly correct values and hierarchically-correct values among all claimed values, respectively. If a source claims exactly correct values without generalization, it is located at the dotted diagonal line in the graph. This graph shows that many sources in real-life datasets claim with generalized values and each source has its own tendency of generalization when claiming values.

Most of the existing methods [7, 9, 30, 38, 39] simply regard the generalized values of a correct value as incorrect. Thus, it causes a problem in estimating the reliabilities of sources. According to [8], 35% of the false negatives in the data fusion task are produced by ignoring such hierarchical structures. Note that there are many publicly available hierarchies such as WordNet [32] and DBpedia [1]. Thus, a truth discovery algorithm to incorporate hierarchies is proposed in [2]. However, it does not consider the different tendencies of generalization and may lead to the degradation of the accuracy. Another drawback is that it needs a threshold to control the granularity of the estimated truth.

We propose a novel probabilistic model to capture the different generalization tendencies shown in Figure 1. Existing probabilistic models [7, 9, 30, 39] basically assume two interpretations of a claimed value (i.e., correct and incorrect). By introducing three interpretations of a claimed value (i.e., exactly correct, hierarchically correct, and incorrect), our proposed model represents the generalization tendency and reliability of the sources.

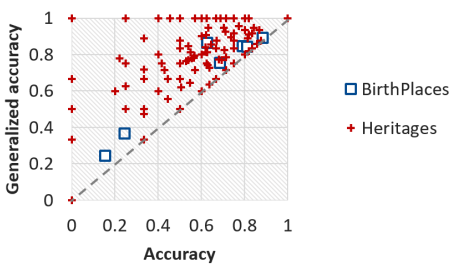


Figure 1: Generalization tendencies of the sources

*Corresponding author

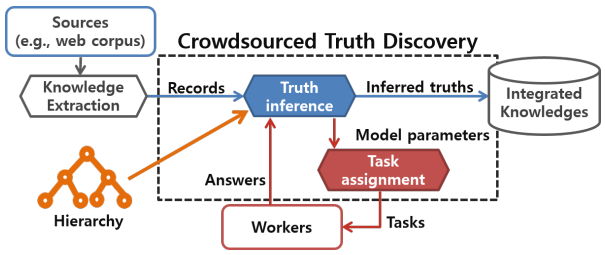


Figure 2: Crowdsourced truth discovery in KF

Crowdsourced truth discovery: It is reported in [8] that upto 96% of the false claims are made by extraction errors rather than by the sources themselves. Since crowdsourcing is an efficient way to utilize human intelligence with low cost, it has been successfully applied in various areas of data integration such as schema matching [12], entity resolution [34], graph alignment [17] and truth discovery [39, 41]. Thus, we utilize crowdsourcing to improve the accuracy of the truth discovery.

It is essential in practice to minimize the cost of crowdsourcing by assigning proper tasks to workers. A popular approach for selecting queries in active learning is *uncertainty sampling* [3, 18, 19, 39]. It asks a query to reduce the uncertainty of the confidences on the candidate values the most. However, it considers only the uncertainty regardless of the accuracy improvement. QASCA algorithm [41] asks a query with the highest accuracy improvement, but measures the improvement without considering the number of collected claimed values. It can be inaccurate since an additional answer may be less informative for an object which already has many records and answers.

Assume that there are two candidate values of an object with equal confidences. If only a few sources provide the claimed values for the object, an additional answer from a crowd worker will significantly change the confidence distribution. Meanwhile, if hundreds of sources already provide the claimed values for the object, the influence of an additional answer is likely to be very little. Thus, we need to consider the number of collected answers as well as the current confidence distribution. Based on the observation, we develop a new method to estimate the increase of accuracy more precisely by considering the number of collected records and answers. We also present an incremental EM algorithm to quickly measure the accuracy improvement and propose a pruning technique to efficiently assign the tasks to workers.

An overview of our truth discovery algorithm: By combining the proposed task assignment and truth inference algorithms, we develop a novel *crowdsourced truth discovery algorithm using hierarchies*. As illustrated in Figure 2, our algorithm consists of two components: *hierarchical truth inference* and *task assignment*. The hierarchical truth inference algorithm finds the correct values from the conflicting values, which are collected from different sources and crowd workers, using hierarchies. The task assignment algorithm distributes objects to the workers who are likely to increase the accuracy of the truth discovery the most. The proposed *crowdsourced truth discovery algorithm* repeatedly alternates the truth inference and task assignment until the budget of crowdsourcing runs out. As discussed in [20], some workers answer slower than others and increase the latency. However, we do not investigate how to reduce the latency in this work since we can utilize the techniques proposed in [13].

Our contributions: The contributions of this paper are summarized below.

- We propose a truth inference algorithm utilizing the hierarchical structures in claimed values. To the best of our knowledge, it is the first work which considers both the reliabilities and the generalization tendencies of the sources.
- To assign a task which will most improve the accuracy, we develop an incremental EM algorithm to estimate the accuracy improvement for a task by considering the number of claimed values as well as the confidence distribution. We also devise an efficient task assignment algorithm for multiple crowd workers based on the quality measure.
- We empirically show that the proposed algorithm outperforms the existing works with extensive experiments on real-life datasets.

2 PRELIMINARIES

In this section, we provide the definitions and the problem formulation of *crowdsourced truth discovery in the presence of hierarchy*.

2.1 Definitions

For the ease of presentation, we assume that we are interested in a single attribute of objects although our algorithms can be easily generalized to find the truths of multiple attributes. Thus, we use 'the target attribute value of an object' and 'the value of an object' interchangeably.

A *source* is a structured or unstructured database which contains the information on target attribute values for a set of objects. In this paper, a *source* is a certain web page or website and a *worker* represents a human worker in crowdsourcing platforms. The information of an object provided by a source or a worker is called a *claimed value*.

Definition 2.1. A *record* is a data describing the information about an object from a source. A record on an object o from a source s is represented as a triple (o, s, v_o^s) where v_o^s is the claimed value of an object o collected from s. Similarly, if a worker w answers that the truth on an object o is v_o^w, the *answer* is represented as (o, w, v_o^w).

Let S_o be the set of the sources which claimed a value on the object o and V_o be the set of candidate values collected from S_o. Each worker in W_o answers a question about the object o by selecting a value from V_o.

In our problem setting, we assume that we have a hierarchy tree H of the claimed values. If we are interested in an attribute related to locations (e.g., birthplace), H would be a geographical hierarchy with different levels of granularity (e.g., continent, country, city, etc.). We also assume that there is no answer with the value of the root in the hierarchy since it provides no information at all (e.g., Earth as a birthplace). We summarize the notations to be used in the paper in Table 2.

Example 2.2. Consider the records in Table 1. Since the source Wikipedia claims that the location of the Statue of Liberty is Liberty Island, it is represented by v_o^s ='Liberty Island' where o ='Statue of Liberty' and s ='Wikipedia'. If a human worker 'Emma Stone' answered Big Ben is in London, it is represented by v_o^w ='London' where o ='Big Ben' and w ='Emma Stone'.

2.2 Problem Definition

Given a set of objects O and a hierarchy tree H, we define the two subproblems of the crowdsourced truth discovery.

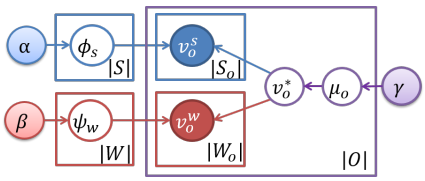


Figure 3: A graphical model for truth inference

Definition 2.3 (Hierarchical truth inference problem). For a set of records R collected from the sources and a set of answers A from the workers, we find the most specific true value v_o^* of each object $o \in O$ among the candidate values in V_o by using the hierarchy H.

Definition 2.4 (Task assignment problem). For each worker w in a set of workers W, we select the top-k objects from O which are likely to increase the overall accuracy of the inferred truths the most by using the hierarchy H.

We present a hierarchical truth inference algorithm in Section 3 and a task assignment algorithm in Section 4.

3 HIERARCHICAL TRUTH INFERENCE

For the hierarchical truth inference, we first model the trustworthiness of sources and workers for a given hierarchy. Then, we propose a probabilistic model to describe the process of generating the set of records and the set of answers based on the trustworthiness modeling. We next develop an inference algorithm to estimate the model parameters and determine the truths.

3.1 Our Generative Model

Our probabilistic graphical model in Figure 3 expresses the conditional dependence (represented by edges) between random variables (represented by nodes). While the previous works [5, 15, 31, 35] assume that all sources and workers have their own reliabilities only, we assume that each source or worker has its generalization tendency as well as reliability. We first describe how sources and workers generate the claimed values based on their trustworthiness. We next present the model for generating the true value. Finally, we provide the detailed generative process of our probabilistic model.

Model for source trustworthiness: For an object o, let v_o^* be the truth and v_o^s be the claimed value reported by a source s.

Table 2: Notations

Symbol	Description
s	A data source
w	A crowd worker
v_o^s	Claimed value from s about o
v_o^w	Claimed value from w about o
R	Set of all records collected from the set of sources S
A	Set of all answers collected from the set of workers W
V_o	Set of candidate values about o
S_o	Set of sources which post information about o
W_o	Set of workers who answered about o
O_s	Set of objects that source s provided a value
O_w	Set of objects that worker w answered to
$G_o(v)$	Set of values in V_o which are ancestors of a value v except the root in the hierarchy H
$D_o(v)$	Set of values in V_o which are descendants of v

Recall that V_o is the set of candidate values for an object o. Furthermore, we let $G_o(v)$ denote the set of candidate values which are ancestors of a value v except for the root in the hierarchy H.

There are three relationships between a claimed value v_o^s and the truth v_o^*: (1) $v_o^s = v_o^*$, (2) $v_o^s \in G_o(v_o^*)$ and (3) otherwise. Let $\phi_s = (\phi_{s,1}, \phi_{s,2}, \phi_{s,3})$ be the *trustworthiness distribution* of a source s where $\phi_{s,i}$ is the probability that a claimed value of the source s corresponds to the i-th relationship. In each relationship, a claimed value is generated as follows:

- **Case 1 ($v_o^s = v_o^*$):** The source s provides the exact true value with a probability $\phi_{s,1}$.
- **Case 2 ($v_o^s \in G_o(v_o^*)$):** The source s provides a *generalized true value* v_o^s with a probability $\phi_{s,2}$. In this case, the claimed value is an ancestor of the truth v_o^* in H. We assume that the claimed value is uniformly selected from $G_o(v_o^*)$.
- **Case 3 (*otherwise*):** The source s provides a wrong value v_o^s not even in $G_o(v_o^*)$. The claimed value is uniformly selected among the rest of the candidate values in V_o.

The probability distribution ϕ_s is an initially-unknown model parameter to be estimated in our inference algorithm. Accordingly, the probability of selecting an answer v_o^s among the values in V_o for an object o is represented by

$$P(v_o^s|v_o^*, \phi_s) = \begin{cases} \phi_{s,1} & \text{if } v_o^s = v_o^*, \\ \phi_{s,2}/|G_o(v_o^*)| & \text{if } v_o^s \in G_o(v_o^*), \\ \phi_{s,3}/(|V_o| - |G_o(v_o^*)| - 1) & \text{otherwise.} \end{cases} \quad (1)$$

For the prior of the distribution ϕ_s, we assume that it follows a Dirichlet distribution $Dir(\alpha)$, with a hyperparameter $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, which is the conjugate prior of categorical distributions.

Let O_H be the set of objects who have an ancestor-descendant relationship in their candidate set. In practice, there may exist some objects whose candidate values do not have an ancestor-descendant relationship. In this case, the probability of the second case (i.e., $\phi_{s,2}$) may be underestimated. Thus, if there is no ancestor-descendant relationship between the claimed values about o (i.e., $o \notin O_H$), we assume that a source generates its claimed value v_o^s with the following probability

$$P(v_o^s|v_o^*, \phi_s) = \begin{cases} \phi_{s,1} + \phi_{s,2} & \text{if } v_o^s = v_o^*, \\ \phi_{s,3}/(|V_o| - 1) & \text{otherwise.} \end{cases} \quad (2)$$

Model for worker trustworthiness: Let v_o^w be the claimed value chosen by a worker w among the candidates in V_o for an object o. Similar to the model for source trustworthiness, we also assume the three relationships between a claimed value v_o^w and the truth v_o^*: (1) $v_o^w = v_o^*$, (2) $v_o^w \in G_o(v_o^*)$ and (3) otherwise. Each worker w has its *trustworthiness distribution* $\psi_w = (\psi_{w,1}, \psi_{w,2}, \psi_{w,3})$ where $\psi_{w,i}$ is the probability that an answer of the worker w corresponds to the i-th relationship. We assume that the trustworthiness distribution is generated from $Dir(\beta)$ with a hyperparameter $\beta = (\beta_1, \beta_2, \beta_3)$.

Since it is difficult for the workers to be aware of the correct answer for every object, a worker can refer to web sites to answer the question. In such a case, if there is a widespread misinformation across multiple sources, the worker is also likely to respond with the incorrect information. Similar to [9, 30], we thus exploit the *popularity* of a value in Cases 2 and 3 to consider such dependency between sources and workers.

- **Case 1 ($v_o^w = v_o^*$):** The worker w provides the exact true value with a probability $\psi_{w,1}$.
- **Case 2 ($v_o^w \in G_o(v_o^*)$):** The worker w provides a generalized true value with a probability $\psi_{w,2}$. We assume that

$$f^v_{o,s} = \frac{P(v^s_o|v^*_o=v,\phi_s)\cdot\mu_{o,v}}{\sum_{v'\in V_o} P(v^s_o|v^*_o=v',\phi_s)\cdot\mu_{o,v'}} \qquad g^1_{o,s} = \frac{\phi_{s,1}\cdot\mu_{o,v^s_o}}{\sum_{v\in V_o} P(v^s_o|v^*_o=v,\phi_s)\cdot\mu_{o,v}} \qquad g^1_{o,w} = \frac{\psi_{w,1}\cdot\mu_{o,v^w_o}}{\sum_{v\in V_o} P(v^w_o|v^*_o=v,\psi_w)\cdot\mu_{o,v}}$$

$$f^v_{o,w} = \frac{P(v^w_o|v^*_o=v,\psi_w)\cdot\mu_{o,v}}{\sum_{v'\in V_o} P(v^w_o|v^*_o=v',\psi_w)\cdot\mu_{o,v'}} \qquad g^2_{o,s} = \begin{cases} \frac{\sum_{v\in D_o(v^s_o)} \frac{\phi_{s,2}}{|G_o(v)|}\cdot\mu_{o,v}}{\sum_{v\in V_o} P(v^s_o|v^*_o=v,\phi_s)\cdot\mu_{o,v}} & \text{if } o\in O_H \\ \frac{\phi_{s,2}\cdot\mu_{o,v^s_o}}{\sum_{v\in V_o} P(v^s_o|v^*_o=v,\phi_s)\cdot\mu_{o,v}} & \text{otherwise} \end{cases} \qquad g^2_{o,w} = \begin{cases} \frac{\sum_{v\in D_o(v^w_o)} \psi_{w,2}\cdot Pop_2(v^w_o|v^*_o=v)\cdot\mu_{o,v}}{\sum_{v\in V_o} P(v^w_o|v^*_o=v,\psi_w)\cdot\mu_{o,v}} & \text{if } o\in O_H \\ \frac{\psi_{w,2}\cdot\mu_{o,v^w_o}}{\sum_{v\in V_o} P(v^w_o|v^*_o=v,\psi_w)\cdot\mu_{o,v}} & \text{otherwise} \end{cases}$$

$$g^3_{o,s} = \frac{\sum_{v\in\neg D_o(v^s_o)} \frac{\phi_{s,3}}{|V_o - G_o(v)|-1}\cdot\mu_{o,v}}{\sum_{v\in V_o} P(v^s_o|v^*_o=v,\phi_s)\cdot\mu_{o,v}} \qquad g^3_{o,w} = \frac{\sum_{v\in\neg D_o(v^w_o)} \psi_{w,3}\cdot Pop_3(v^w_o|v^*_o=v)\cdot\mu_{o,v}}{\sum_{v\in V_o} P(v^w_o|v^*_o=v,\psi_w)\cdot\mu_{o,v}}$$

Figure 4: E-step for the proposed truth inference algorithm

the claimed value v^w_o is selected according to the popularity $Pop_2(v^w_o|v^*_o) = \frac{|\{s|s\in S_o, v^s_o=v\}|}{|\{s|s\in S_o, v^s_o\in G_o(v^*_o)\}|}$ which is the proportion of the records whose claimed value is v^w_o out of the records with generalized values of v^*_o.

- **Case 3 (*otherwise*):** The claimed value is selected from the wrong values according to the popularity $Pop_3(v^w_o|v^*_o) = \frac{|\{s|s\in S_o, v^s_o=v\}|}{|\{s|s\in S_o, v^s_o\notin G_o(v^*_o), v^s_o\neq v^*_o\}|}$.

By the above model, the probability of selecting an answer v^w_o for the truth v^*_o of an object o is formulated as

$$P(v^w_o|v^*_o,\psi_w) = \begin{cases} \psi_{w,1} & \text{if } v^w_o = v^*_o, \\ \psi_{w,2}\cdot Pop_2(v^w_o|v^*_o) & \text{if } v^w_o \in G_o(v^*_o), \\ \psi_{w,3}\cdot Pop_3(v^w_o|v^*_o) & \text{otherwise.} \end{cases} \quad (3)$$

Similar to the model for source trustworthiness, if there is no ancestor-descendant relationship in the candidate values of an object o, the probability of selecting a claimed value v^w_o is

$$P(v^w_o|v^*_o,\psi_w) = \begin{cases} \psi_{w,1}+\psi_{w,2} & \text{if } v^w_o = v^*_o, \\ \psi_{w,3}\cdot Pop_3(v^w_o|v^*_o) & \text{otherwise.} \end{cases} \quad (4)$$

Model for truth: We introduce the probability distribution over the candidate answers to determine the truth, called *confidence distribution*. Each object o has a confidence distribution $\mu_o = \{\mu_{o,v}\}_{v\in V_o}$ where $\mu_{o,v}$ is the probability that the value $v\in V_o$ is the true answer for o. We also use a dirichlet prior $Dir(\gamma_o)$ for the confidence distribution μ_o where $\gamma_o = \{\gamma_{o,v}\}_{v\in V_o}$ is a hyperparameter.

Based on the above three models, the generative process of our model works as follows.

Generative process: Given a set of objects O, a set of sources S and a set of workers W, our proposed model assumes the following generative process for the set of records R and the set of answers A:

(1) Draw $\phi_s \sim Dir(\alpha)$ for each source $s\in S$
(2) Draw $\psi_w \sim Dir(\beta)$ for each worker $w\in W$
(3) For each object $o\in O$
  (a) Draw $\mu_o \sim Dir(\gamma_o)$
  (b) Draw a true value $v^*_o \sim Categorical(\mu_o)$
  (c) For each source $s\in S_o$
    (i) Draw a value v^s_o following $P(v^s_o|v^*_o,\phi_s)$
  (d) For each worker $w\in W_o$
    (i) Draw a value v^w_o following $P(v^w_o|v^*_o,\psi_w)$

3.2 Estimation of Model Parameters

We now develop an inference algorithm for the generative model. Let $\Theta = \boldsymbol{\phi}\cup\boldsymbol{\psi}\cup\boldsymbol{\mu}$ be the set of all model parameters where $\boldsymbol{\phi} = \{\phi_s|s\in S\}$, $\boldsymbol{\psi} = \{\psi_w|w\in W\}$ and $\boldsymbol{\mu} = \{\mu_o|o\in O\}$. We propose an EM algorithm to find the maximum a posteriori (MAP) estimate of the parameters in our model.

The maximum a posteriori (MAP) estimator: Recall that $R = \{(o,s,v^s_o)\}$ is the set of records from the sources and $A = \{(o,w,v^w_o)\}$ is the set of answers from the workers. For every object o, each source $s\in S_o$ and each worker $w\in W_o$ generates its claimed values independently. Then, the likelihood of R and A based on our generative model is

$$P(R,A|\Theta) = \prod_{o\in O}\prod_{s\in S_o} P(v^s_o|\phi_s,\mu_o)\cdot\prod_{o\in O}\prod_{w\in W_o} P(v^w_o|\psi_w,\mu_o)$$

where the probability of generating a claimed value by a source or a worker becomes

$$P(v^s_o|\phi_s,\mu_o) = \sum_{v\in V_o} P(v^s_o|\phi_s, v^*_o=v)\cdot\mu_{o,v} \quad (5)$$

$$P(v^w_o|\psi_w,\mu_o) = \sum_{v\in V_o} P(v^w_o|\psi_w, v^*_o=v)\cdot\mu_{o,v}. \quad (6)$$

Consequently, the MAP point estimator is obtained by maximizing the log-posterior as

$$\hat{\Theta} = \arg\max_{\Theta}\{\log P(R,A|\Theta) + \log P(\Theta)\} = \arg\max_{\Theta}\mathbb{F} \quad (7)$$

where the objective function $\mathbb{F}$ is

$$\begin{aligned}\mathbb{F} = &\sum_{o\in O}\sum_{s\in S_o}\log\sum_{v\in V_o} P(v^s_o|\phi_s,v^*_o=v)\cdot\mu_{o,v} \\ &+\sum_{o\in O}\sum_{w\in W_o}\log\sum_{v\in V_o} P(v^w_o|\psi_w,v^*_o=v)\cdot\mu_{o,v} \\ &+\sum_{s\in S}\log p(\phi_s|\alpha)+\sum_{w\in W}\log p(\psi_w|\beta)+\sum_{o\in O}\log p(\mu_o|\gamma_o).\end{aligned} \quad (8)$$

Note that although we assumed that each claimed value is generated independently according to its probability distribution defined in Eq. (5) and (6), the dependencies between sources and workers are already considered in $Pop_2(v^w_o|v^*_o)$ and $Pop_3(v^w_o|v^*_o)$.

The EM algorithm: We introduce a random variable C_v to represent the type of the relationship between the claimed value v and the truth v^*_o. It is defined as follows:

$$C_v = \begin{cases} 1 & \text{if } v = v^*_o, \\ 2 & \text{if } v\in G_o(v^*_o), \\ 3 & \text{otherwise.} \end{cases}$$

In the **E-step**, we compute the conditional distributions of the hidden variables $C_{v^s_o}$, $C_{v^w_o}$ and v^*_o under our current estimate of the parameters Θ. Let $f^v_{o,s}$, $f^v_{o,w}$, $g^t_{o,s}$ and $g^t_{o,w}$ denote the conditional probabilities $P(v^*_o=v|v^s_o,\mu_o,\phi_s)$, $P(v^*_o=v|v^w_o,\mu_o,\psi_w)$, $P(C_{v^s_o}=t|\mu_o,\phi_s)$ and $P(C_{v^w_o}=t|\ \mu_o,\psi_w)$, respectively. Using Bayes' rule, we can update the conditional probabilities as shown in Figure 4 where $D_o(v) = \{v'|v\in G_o(v')\wedge v'\in V_o\}$ is the set of descendants of v among the candidate values and $\neg D_o(v) =$

$V_o - D_o(v) - \{v\}$ is the set of candidate values each of which is neither a descendant of the value v nor the v itself.

In the **M-step**, we find the model parameters Θ that maximize our objective function $\mathbb{F}$. We first add Lagrange multipliers to enforce the constraints of model parameters.

$$\mathbb{L} = \mathbb{F} + \sum_{s \in S} \lambda_{\phi,s} \left(1 - \sum_{t=1}^{3} \phi_{s,t}\right) + \sum_{w \in W} \lambda_{\psi,w} \left(1 - \sum_{t=1}^{3} \psi_{w,t}\right) + \sum_{o \in O} \lambda_{\mu,o} \left(1 - \sum_{v \in V_o} \mu_{o,v}\right)$$

We obtain the following equations for updating the model parameters Θ by taking the partial derivative of the Lagrangian $\mathbb{L}$ with respect to each model parameter and setting it to zero:

$$\mu_{o,v} = \frac{\sum_{s \in S_o} f^v_{o,s} + \sum_{w \in W_o} f^v_{o,w} + \gamma_{o,v} - 1}{|S_o| + |W_o| + \sum_{v' \in V_o} (\gamma_{o,v'} - 1)} \tag{9}$$

$$\phi_{s,t} = \frac{\sum_{o \in O_s} g^t_{o,s} + \alpha_t - 1}{|O_s| + \sum_{t'=1}^{3} (\alpha_{t'} - 1)} \tag{10}$$

$$\psi_{w,t} = \frac{\sum_{o \in O_w} g^t_{o,w} + \beta_t - 1}{|O_w| + \sum_{t'=1}^{3} (\beta_{t'} - 1)} \tag{11}$$

where O_s and O_w are the sets of objects claimed by s and w, respectively. We infer the truth by choosing the value with the maximum confidence among the candidate values as

$$v^*_o = \arg\max_{v \in V_o} \mu_{o,v}. \tag{12}$$

Extension to numerical data: In the world wide web, numerical data also have an implicit hierarchy due to the significant digits which carry meaning contributing to its measurement resolution. For example, even though the area of Seoul is $605.196km^2$, different websites may represent the area in various forms depending on the significant figures (e.g., $605.2km^2$, $605km^2$). An existing algorithm [21] to handle numerical data utilizes a weighted sum of the claimed values to consider the distribution of the claimed values. However, such method is sensitive to outliers and thus need a proper preprocessing to remove the outliers. To overcome the drawbacks, we generate the underlying hierarchy in the numerical data by assuming that v_d is a descendant of v_a if a value v_a can be obtained by rounding off a value v_d. Then, we can use our TDH algorithm to find the truths in numerical data by taking into account the relationship between the values in the implicit hierarchy. Our algorithm is also robust to the outliers with extremely small or large value since we estimate the truth by selecting the most probable value from the candidate values rather than computing a weighted average of the claimed values.

4 TASK ASSIGNMENT TO WORKERS

In this section, we propose a task assignment method to select the best objects to be assigned to the workers in crowdsourcing systems. We first define a quality measure of tasks called *Expected Accuracy Increase (EAI)* and develop an incremental EM algorithm to quickly estimate the quality measure. Finally, we present an efficient algorithm for assigning the k questions to each worker w in a set of workers W based on the measure.

4.1 The Quality Measure

Given a worker w, our goal is to choose an object to be assigned to the worker w which is likely to increase the accuracy of the estimated truths the most. Thus, we define a quality measure for a pair of worker and an object based on the improvement of the accuracy. As discussed in [41], the improvement of the accuracy by a task can be estimated by using the difference between the highest confidence as follows:

$$(Accuracy\ improvement) = \{\max_v \mu_{o,v|w} - \max_v \mu_{o,v}\}/|O| \tag{13}$$

where $\mu_{o,v|w}$ is the estimated confidence on v if the worker w answers about an object o.

The quality measure used by QASCA: The QASCA[41] algorithm calculates the estimated confidence by using the current confidence distribution and the likelihood of the answer v^w_o given the truth $v^*_o = v$ as

$$\mu_{o,v|w} \propto \mu_{o,v} \cdot p(v^w_o = v' | v^*_o = v)$$

where v' is a sampled claimed value. There are two drawbacks in the quality measure of QASCA. First, since it computes the estimated confidence $\mu_{o,v|w}$ based on a sampled answer $v^w_o = v$, the value of the quality measure is very sensitive to the sampled answer. In addition, QASCA does not consider the number of claimed values collected so far and the estimated confidence $\mu_{o,v|w}$ may not be accurate. For instance, assume that there exist two objects which have identical confidence distributions. If one of the objects already has many collected claimed values, an additional answer is not likely to change the confidence significantly. Thus, task assignment algorithms should select another object who has a smaller number of collected records and answers.

Our quality measure: To avoid the sensitiveness caused by sampling answers, we develop a new quality measure *Expected Accuracy Improvement (EAI)* which is obtained by taking the expectation to Eq. (13). That is,

$$EAI(w, o) = \{E[\max_v \mu_{o,v|w}] - \max_v \mu_{o,v}\}/|O|. \tag{14}$$

By the definition of expectation, $E[\max_v \mu_{o,v|w}]$ becomes

$$E[\max_v \mu_{o,v|w}] = \sum_{v' \in V_o} P(v^w_o = v' | \psi_w, \mu_o) \cdot \max_v \mu_{o,v|v^w_o = v'}. \tag{15}$$

where $\mu_{o,v|v^w_o = v'}$ is the conditional confidence when a worker w answers with v' about the object o.

Since $P(v^w_o = v' | \psi_w, \mu_o)$ can be computed by Eq. (6), to compute $E[\max_v \mu_{o,v|w}]$ by Eq. (15), we need the estimation of the conditional confidence $\mu_{o,v|v^w_o = v'}$ with an additional answer $v^w_o = v'$. Recall that the estimated confidence computed by QASCA may not be accurate because it does not consider the collected records and answers so far. To reduce the error, we use them to compute the conditional confidence $\mu_{o,v|v^w_o = v'}$. We can compute the conditional confidence $\mu_{o,v|v^w_o = v'}$ by applying the EM algorithm in Section 3.2 with the collected records and answers including $v^w_o = v'$. However, since it is computationally expensive, we next develop an *incremental EM algorithm*.

4.2 The Incremental EM Algorithm

Let $\mathbb{F}_{v^w_o = v'}$ be the objective function in Eq. (7) after obtaining an additional answer (o, w, v'). Then, we have

$$\mathbb{F}_{v^w_o = v'} = \mathbb{F} + \log \sum_{v \in V_o} P(v^w_o = v' | \psi_w, v^*_o = v) \cdot \mu_{o,v}$$

by adding the related term of the additional answer (log likelihood of the additional answer) to Eq. (8). Instead of running the iterative EM algorithm in Section 3.2, we incrementally perform a *single EM-step* to speed up for only the additional answer with the current model parameters and the above objective function.

E-step: Since we use the current model parameters, the probabilities of the hidden variables for collected records and answers are not changed. Thus, we only need to compute the conditional probabilities of the hidden variable given the additional answer as

$$f^v_{o,w|v^w_o = v'} = \frac{P(v^w_o = v' | v^*_o = v, \psi_w) \cdot \mu_{o,v}}{\sum_{v'' \in V_o} P(v^w_o = v' | v^*_o = v'', \psi_w) \cdot \mu_{o,v''}} \tag{16}$$

based on the equation for $f^{v}_{o,w}$ used at the E-step in Figure 4.

M-step: For the objective function $\mathbb{F}_{v^w_o=v'}$, we obtain the following equation of the M-step for the confidence distribution μ_o with the additional answer $v^w_o = v'$

$$\mu_{o,v|v^w_o=v'} = \frac{\sum_{s\in S_o} f^{v}_{o,s} + \sum_{w'\in W_o} f^{v}_{o,w'} + f^{v}_{o,w|v^w_o=v'} + \gamma_{o,v} - 1}{|S_o| + |W_o| + 1 + \sum_{v''\in V_o}(\gamma_{o,v''} - 1)}$$

by adding the related terms $f^{v}_{o,w|v^w_o=v'}$ and 1 to the numerator and the denominator of the update equation in Eq. (9), respectively. Let $N_{o,v}$ and D_o be the numerator and the denominator in Eq. (9), respectively. Then, the above equation can be rewritten as

$$\mu_{o,v|v^w_o=v'} = \frac{N_{o,v} + f^{v}_{o,w|v^w_o=v'}}{D_o + 1}. \tag{17}$$

By substituting $f^{v}_{o,w|v^w_o=v'}$ in Eq. (17) with Eq. (16), the conditional confidence becomes

$$\mu_{o,v|v^w_o=v'} = \frac{N_{o,v} + \frac{P(v^w_o=v'|v^*_o=v,\psi_w)\cdot\mu_{o,v}}{\sum_{v''\in V_o} P(v^w_o=v'|v^*_o=v'',\psi_w)\cdot\mu_{o,v''}}}{D_o + 1}. \tag{18}$$

Since $N_{o,v}$ and D_o are proportional to the number of the existing claimed values, the confidence will be changed very little if there are many claimed values already. Thus, we can overcome the second drawback of QASCA. Since $N_{o,v}$s and D_os are repeatedly used to compute $\mu_{o,v|v^w_o=v'}$, our truth inference algorithm keeps $N_{o,v}$s and D_os in main memory to reduce the computation time.

Time complexity analysis: To calculate $E[\max_v \mu_{o,v|w}]$ by Eq. (15), $P(v^w_o = v'|\psi_w, \mu_o)$ is computed $|V_o|$ times and $\mu_{o,v|v^w_o=v'}$ is calculated for every pair of v and v' (i.e., $O(|V_o|^2)$ times). Moreover, computing $P(v^w_o = v'|\psi_w, \mu_o)$ and $\mu_{o,v|v^w_o=v'}$ take $O(|V_o|)$ time. Thus, it takes $O(|V_o|^3)$ time to compute $EAI(w, o)$ by Eq. (14). In reality, $|V_o|$ is very small compared to $|O|$,$|S|$ and $|W|$. In addition, by utilizing the pruning technique in the next section, we can significantly reduce the computation time. Therefore, the task assignment step can be performed within a short time compared to the truth inference. The execution time for each step will be presented in the experiment section.

4.3 The Task Assignment Algorithm

To find the k objects to be assigned to each worker, we need to compute $EAI(w, o)$ for all pairs of w and o. To reduce the number of computing $EAI(w, o)$, we develop a pruning technique by utilizing an upper bound of $EAI(w, o)$.

An upper bound of EAI: We provide the following lemma which allows us to compute an upper bound $U_{EAI}(o)$.

Lemma 4.1. (Upper Bound of Expected Accuracy Increase) *For an object o and a worker w, we have*

$$EAI(w, o) \le U_{EAI}(o) = \frac{1 - \max_v \mu_{o,v}}{|O| \cdot (D_o + 1)}. \tag{19}$$

Proof. From Eq. (18), since $\sum_{v'} P(v^w_o=v'|\psi_w, \mu_o) = 1$, we get

$$\begin{aligned} E[\max_v \mu_{o,v|w}] &= \sum_{v'\in V_o} P(v^w_o=v'|\psi_w, \mu_o) \cdot \max_v \mu_{o,v|v^w_o=v'} \\ &\le \max_{v,v'} \mu_{o,v|v^w_o=v'} \cdot \sum_{v'\in V_o} P(v^w_o=v'|\psi_w, \mu_o) \\ &= \max_{v,v'} \mu_{o,v|v^w_o=v'}. \end{aligned} \tag{20}$$

Moreover, from Eq. (17), we obtain

$$\mu_{o,v|v^w_o=v'} = \frac{N_{o,v} + f^{v}_{o,w|v^w_o=v'}}{D_o + 1} \le \frac{N_{o,v} + 1}{D_o + 1}. \tag{21}$$

Algorithm 1 Task Assignment

Input: set of workers W, number of questions k

1: Compute the upper bound $U_{EMCI}(o)$ for $o \in O$
2: $h_{UB} \leftarrow$ BuildMaxHeap($\{\langle U_{EAI}(o), o\rangle | o \in O\}$)
3: Sort workers in the decreasing order of $\psi_{w,1}$ (i.e., $\psi_{1,1} \ge \psi_{2,1} \ge \cdots \ge \psi_{|W|,1}$).
4: **for** w = 1 to $|W|$ **do**
5: $\quad h_{EAI}[w] \leftarrow$ BuildMinHeap($\{\}$)
6: **while** True **do**
7: $\quad \langle U_{EAI}(o), o\rangle \leftarrow h_{UB}$.extractMax()
8: $\quad$ **if** $h_{EAI}[|W|].size = k$ **and** $h_{EAI}[w].min > U_{EAI}(o)$ for all w **then**
9: $\quad\quad$ **break**
10: $\quad$ **for** w = 1 to $|W|$ **do**
11: $\quad\quad$ **if** w already answered on o **or** $h_{EAI}[w].min > U_{EAI}(o)$ **then**
12: $\quad\quad\quad$ **continue**
13: $\quad\quad$ Compute $EAI(w, o)$
14: $\quad\quad$ $h_{EAI}[w]$.insert($\langle EAI(w, o), o\rangle$)
15: $\quad\quad$ **if** $h_{EAI}[w].size \le k$ **then**
16: $\quad\quad\quad$ **break**
17: $\quad\quad$ $o \leftarrow h_{EAI}[w]$.extractMin().value()

By substituting Eq. (21) for $\mu_{o,v|v^w_o=v'}$ in Eq. (20), we derive

$$E[\max_v \mu_{o,v|w}] \le \max_{v,v'} \mu_{o,v|v^w_o=v'} \le \frac{\max_v N_{o,v} + 1}{D_o + 1}. \tag{22}$$

In addition, by applying Eq. (22) to Eq. (14), we get

$$EAI(w, o) \le \left(\frac{\max_v N_{o,v}+1}{D_o+1} - \max_v \mu_{o,v}\right)/|O|.$$

Since $\mu_{o,v} = \frac{N_{o,v}}{D_o}$, we finally obtain the upper bound of $EAI(w, o)$.

$$\begin{aligned} EAI(w, o) &\le \left(\frac{\max_v N_{o,v}+1}{D_o+1} - \frac{\max_v N_{o,v}}{D_o}\right)/|O| \\ &= \frac{1-\frac{\max_v N_{o,v}}{D_o}}{|O|\cdot(D_o+1)} = \frac{1-\max_v \mu_{o,v}}{|O|\cdot(D_o+1)} = U_{EAI}(o). \end{aligned}$$

□

We devise an algorithm to assign the best k objects to each available worker in crowdsourcing systems. Since a single answer is sufficient to find the correct value for some objects, we assign an object to only a single worker in each round. If the answer is not sufficient to find the correct value of the object, we assign the object to another worker in the next round.

Our task assignment algorithm sequentially assigns each object to a worker by scanning the objects o with non-increasing order of the upper bound $U_{EAI}(o)$. To allocate an object to a worker, since $\psi_{w,1}$ is the probability of answering the truth, we consider the workers w with non-increasing order of $\psi_{w,1}$. After assigning an object to a worker w, if the number of assigned objects to the worker w exceeds k, we remove the object o with the minimum $EAI(w, o)$ and assign the deleted object to the next worker and perform the same step. While scanning the objects, we stop the assignment if the upperbound $U_{EAI}(o)$ is smaller than the minimum $EAI(w, o')$ among the $EAI(w, o')$s of all assigned objects and each worker has k assigned objects. The reason is that the $EAI(w, o)$ of the remaining objects o can be larger than that of any assigned object.

The pseudocode: It is shown in Algorithm 1. We first compute the upper bound $U_{EAI}(o)$ for every object $o \in O$ by Lemma 4.1 and build a maxheap h_{UB} of all objects by using $U_{EAI}(o)$ as the key to assign the objects to workers in the decreasing order of $U_{EAI}(o)$ (in lines 1-2). The workers are sorted in the decreasing order of $\psi_{w,1}$ to give a higher priority to reliable workers (in line 3). We next initialize a minheap $h_{EAI}[w]$ for each worker w to contain the k assigned objects (in lines 4-6). Then, we repeatedly extract an object from h_{UB} and assign the object to a worker in the sorted order of $\psi_{w,1}$ (in lines 12-18). Before assigning an object o, if the heaps $h_{EAI}[w]$s of all workers are full and the

minimum value of $EAI(w, o')$ of the objects o' in all $h_{EAI}[w]$s is larger than the upper bound $U_{EAI}(o)$, we stop immediately.

5 EXPERIMENTS

The experiments are conducted on a computer with Intel i5-7500 CPU and 16GB of main memory.

Datasets: We collected the two real-life datasets publicly available at http://kdd.snu.ac.kr/home/datasets/tdh.php.

BirthPlaces: We crawled 13,510 records about the birthplaces of 6,005 celebrities from 7 websites (sources). For the gold standard data to evaluate the correctness of discovered birthplaces, we used IMDb biography which is available at *http://www.imdb.com*. Moreover, the geographical hierarchy was created by using the IMDb data. For example, if there is a person who was born in 'LA, California, USA', we assigned 'LA' as a child of 'California' and 'California' as a child of 'USA'. The hierarchy contains 4,999 nodes (e.g., countries, cities and etc.) and its height is 5.

Heritages: This is a dataset of the locations of World Heritage Sites provided by UNESCO World Heritage Centre, available at *http://whc.unesco.org*. We queried about the locations of 785 World Heritage Sites with Bing Search API and obtained 4,424 claimed values from 1,577 distinct websites. The hierarchy was created in the same way as we did for *BirthPlaces* and it has 1,027 nodes. The height of this hierarchy tree is 6.

Quality Measures: We use *Accuracy*, *GenAccuracy* and *AvgDistance* to evaluate the truth discovery algorithms. Let t_o be the truth of the object o in the gold standard and v_o^* be the estimated truth by an algorithm. Note that t_o may not exist in the set of candidate values. In this case, the most specific candidate value among the ancestors of the truth is assumed to be t_o. *Accuracy* is the proportion of objects that the algorithm discovers the truth exactly. It is actually used in [10, 39–41] to evaluate truth discovery algorithms.

$$(Accuracy) = \Sigma_{o \in O} I(v_o^* = t_o)/|O|$$

The ancestors of t_o are less informative but still correct values. Thus, we develop an evaluation measure named *GenAccuracy* which is the proportion of objects o whose estimated truth v_o^* is either the truth t_o or an ancestor of the truth.

$$(GenAccuracy) = \Sigma_{o \in O} I(v_o^* \in G_H(t_o) \cup \{t_o\})/|O|$$

Ancestors of the truth have a different level of informativeness depending on the distance to the truth. For example, 'New York' is more informative than 'USA' as the location of the Statue of Liberty. Thus, we utilize another evaluation measure named *AvgDistance* which weights the estimated truth based on the distance from the ground truth. More specifically, it is the average number of edges $d(v_o^*, t_o)$ between the truth t_o and the estimated truth v_o^* in the hierarchy H.

$$(AvgDistance) = \Sigma_{o \in O} d(v_o^*, t_o)/|O|$$

AvgDistance is robust to the case where the ground truth is less specific than the estimated truth. The estimated truth is regarded as a wrong value when we compute *Accuracy* and *GenAccuracy* even though the estimate truth is correct and more specific. Since the distance between the less specific ground truth and the estimated truth is generally small, *AvgDistance* compensates the drawback of *Accuracy* and *GenAccuracy*.

Settings for simulated crowdsourcing: To evaluate the truth discovery algorithms with varying the quality of the answers from workers, we conducted experiments with simulated crowd workers. In our simulation, we assumed that each simulated worker answers a question correctly with its own probability p_w and randomly selects an answer from the candidate values with probability $1-p_w$. We sampled the probability p_w from a uniform distribution ranging from $\pi_p - 0.05$ to $\pi_p + 0.05$ where the default value of π_p is 0.75. In the experiments, each of 10 worker answers 5 questions for each round.

Table 3: Performance of truth inference algorithms

	Dataset					
	BirthPlaces			*Heritages*		
Algorithm	*Accuracy*	*GenAccuracy*	*AvgDistance*	*Accuracy*	*GenAccuracy*	*AvgDistance*
TDH	**0.8913**	**0.8988**	**0.3151**	**0.7414**	0.8726	**0.5210**
VOTE	0.7900	0.8924	0.4961	0.6892	**0.8994**	0.6382
LCA	0.8834	0.8923	0.3414	0.6930	0.8866	0.6611
DOCS	0.8828	0.8916	0.3409	0.6904	0.8866	0.6599
ASUMS	0.8543	0.8571	0.4573	0.6229	0.7414	1.2000
MDC	0.8263	0.8432	0.5320	0.7254	0.8087	0.6869
ACCU	0.8137	0.8296	0.6063	0.5834	0.7656	1.0637
POPACCU	0.8133	0.8300	0.6070	0.6561	0.8586	0.7554
LFC	0.8085	0.8743	0.4669	0.6803	0.8076	0.8076
CRH	0.8083	0.8271	0.6120	0.6841	0.8828	0.6688

5.1 Implemented Algorithms

We implemented 10 truth inference algorithms and 4 task assignment algorithms in Python for comparative experiments. The truth inference algorithms are referred to as follows:

- TDH: This is our algorithm proposed in Section 3. For the prior distribution $Dir(\alpha)$, we set the hyperparameter $\alpha = (3, 3, 2)$ since correct values are more frequent than wrong values for most of the sources. For the other hyperparameters β and γ, we set every dimension of β and γ to 2.
- ACCU: It is the algorithm proposed in [7] which considers the dependencies between sources to find the truths. The algorithm exploits Bayesian analysis to find the dependencies.
- POPACCU: This denotes the algorithm in [9] which extends ACCU. It computes the distribution of the false values from the records while ACCU assumes that it is uniform.
- LFC: This algorithm is proposed in [31] and utilizes a confusion matrix to model a source's quality.
- CRH: It is proposed in [22] to resolve conflicts in heterogeneous data containing categorical and numerical attributes.
- LCA: It is a probabilistic model proposed in [30]. We select GuessLCA to be compared in this paper which is one of the best performers among the 7 algorithms proposed in [30].
- ASUMS: This is proposed in [2] by adapting an existing method SUMS [29] to hierarchical truth discovery.
- MDC: This denotes the truth discovery method designed for medical diagnose from non-expert crowdsourcing in [24].
- DOCS: This is the state-of-the-art technique presented in [39] that suggests the domain-sensitive worker model.
- VOTE: This is a baseline that selects a value with the highest frequency in the claimed values.

We implemented the following task assignment algorithms.

- *EAI*: This is our proposed algorithm in Section 4.
- *MB*: It is the task assignment algorithm used by DOCS [39].
- *QASCA*: It is a task assignment algorithm proposed in [41].
- *ME*: This is our baseline algorithm which utilizes an uncertainty sampling. It selects an object o^* whose confidence distribution has the maximum entropy. (i.e., $o^* = argmax_{o \in O}$ $(-\sum_{v \in V_o} \mu_{o,v} \cdot \log \mu_{o,v})$)

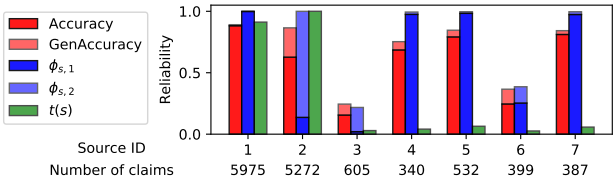
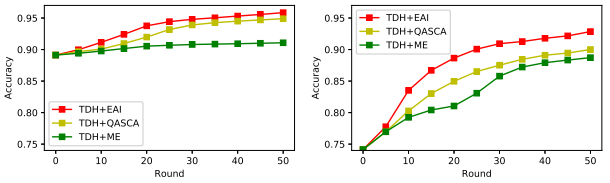


Figure 5: Source reliability distribution in *BirthPlaces*



(a) *BirthPlaces* **(b) *Heritages***

Figure 6: Evaluation of task assignment algorithms

Note that *EAI* and *MB* are the task assignment algorithms specially designed to work with *TDH* and *DOCS*, respectively. *QASCA* can work with truth inference algorithms based on probabilistic models such as *TDH*, *DOCS*, *LCA*, *ACCU* and *POPACCU*. All the truth inference algorithms can be combined with *ME*.

5.2 Truth Inference

We first provide the performances of the truth inference algorithms without using crowdsourcing in Table. 3.

***BirthPlaces*:** Our TDH outperforms all other algorithms in terms of all quality measures since TDH finds the exact truths by utilizing the hierarchical relationships. Since TDH estimates the reliabilities of the sources and workers by considering the hierarchies, it does not underestimate the reliabilities of the sources and workers. Thus, TDH also finds more correct values including the generalized truths. We will discuss the reliability estimation in detail at the end of this section by comparing TDH with ASUMS. LCA is the second-best performer and VOTE shows the lowest *Accuracy* among all compared algorithms. However, in terms of *GenAccuracy*, VOTE performs the second-best. It is because many websites claim the generalized values rather than the most specific value.

***Heritages*:** In terms of *AvgDistance* and *Accuracy*, TDH performs the best among those of the compared algorithms. VOTE shows the highest *GenAccuracy* because many sources provide the generalized truths. In fact, a high *GenAccuracy* with low *Accuracy* and *AvgDistance* can be easily obtained by providing the most general values for the truths. However, such values usually are not informative. Since our algorithm shows much higher *Accuracy* and much lower *AvgDistance* than VOTE, we can see that the estimated truth by TDH is more accurate and precise than the result from VOTE. *Heritages* contains many sources and most of the sources have a few claims. Thus, it is very hard to estimate the reliability of each source accurately. Therefore, most of the compared algorithms show worse performance than VOTE in terms of *AvgDistance*. In particular, ACCU has the lowest *Accuracy*. The reason is that ACCU requires many shared objects between two sources in order to accurately determine the dependency between the sources. The average accuracy of the sources in *Heritages* is 58.0% while that of the sources in *BirthPlaces* is 72.1%. Thus, every algorithm shows a lower *Accuracy* in this dataset than in *BirthPlaces*.

Comparison with ASUMS: Since ASUMS [2] is the only existing algorithm which utilizes hierarchies for truth inference, we show the statistics related to the reliability distributions estimated by TDH and ASUMS for *BirthPlaces* dataset in Figure 5. *Accuracy* and *GenAccuracy* represent the actual reliabilities of each source computed from the ground truths. Recall that $\phi_{s,1}$ and $\phi_{s,2}$ are the estimated probabilities of providing a correct value and a generalized correct value respectively for a source s by our TDH, as defined in Section 3. In addition, $t(s)$ is the estimated reliability of a source s by ASUMS which ignores the generalization level of each source. The reliabilities of the sources 4, 5 and 7 computed by ASUMS (i.e. $t(s)$) are quite different from the actual reliabilities (i.e., *Accuracy*). As we discussed in Section 1, for a pair of sources that provide different claimed values with an ancestor-descendant relationship in a hierarchy, existing methods may assume that one of the claimed values is incorrect. Thus, the reliability of the source with the assumed wrong value tends to become lower by the existing methods. ASUMS suffers from the same problem and underestimates the reliabilities of the sources 4, 5 and 7 which provide a small number of claimed values. Meanwhile, our proposed algorithm TDH accurately estimates the reliabilities of the sources by introducing another class of the claimed values (generalized truth).

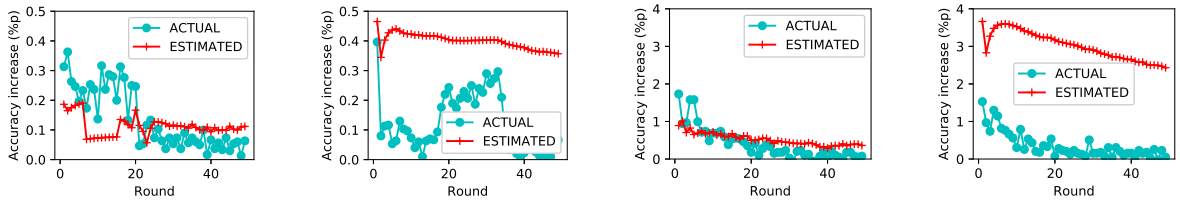
5.3 Task Assignment

Before providing the full comparison of all possible combinations of truth inference algorithms and task assignment algorithms, we first evaluate the task assignment algorithms with our proposed truth inference algorithm. We plotted the average *Accuracy* of the truth discovery algorithms with different task assignment algorithms for every 5 round in Figure 6. The points at the 0-th round represent the *Accuracy* of the algorithms without crowdsourcing. All algorithms show the same *Accuracy* at the beginning since they use the same truth inference algorithm TDH. As the round progresses, the *Accuracy* of TDH+EAI increases faster than those of all other algorithms. The *Accuracy* of TDH+ME is the lowest since ME selects a task based only on the uncertainty without estimating the accuracy improvement by the task.

As discussed in Section 4.1, our task assignment algorithm EAI estimates the accuracy improvement by considering the number of existing claimed values and the confidence distribution whereas QASCA considers the confidence distribution only. We plotted the actual and estimated accuracy improvements by EAI and QASCA in Figure 7. The graphs show that the estimated accuracy improvement by EAI is similar to the actual accuracy improvement while QASCA overestimates the accuracy improvement at every round. On average, the absolute estimation errors

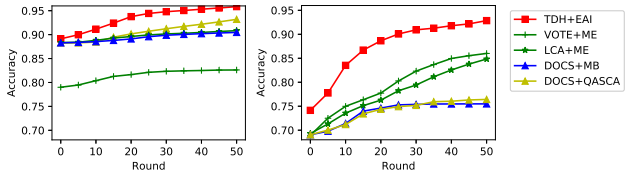
Table 4: *Accuracy* of the algorithms after the 50th round

	BirthPlaces				*Heritages*			
	EAI	MB	QASCA	ME	EAI	MB	QASCA	ME
TDH	**0.9601**	-	**0.9500**	**0.9109**	**0.9304**	-	**0.8999**	**0.8884**
DOCS	-	**0.9052**	0.9341	0.8842	-	**0.7546**	0.7661	0.7631
LCA	-	-	0.8823	0.9089	-	-	0.7136	0.8507
POPACCU	-	-	0.9295	0.8987	-	-	0.7512	0.8336
ACCU	-	-	0.8468	0.8257	-	-	0.5796	0.5896
ASUMS	-	-	-	0.8700	-	-	-	0.7427
CRH	-	-	-	0.9000	-	-	-	0.8459
MDC	-	-	-	0.8254	-	-	-	0.7241
LFC	-	-	-	0.8287	-	-	-	0.7327
VOTE	-	-	-	0.8261	-	-	-	0.8634



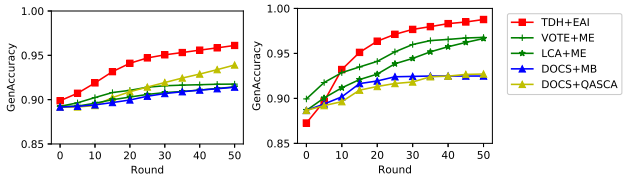
(a) *BirthPlaces*-EAI (b) *BirthPlaces*-QASCA (c) *Heritages*-EAI (d) *Heritages*-QASCA

Figure 7: Actual and estimated accuracy improvement by EAI and QASCA



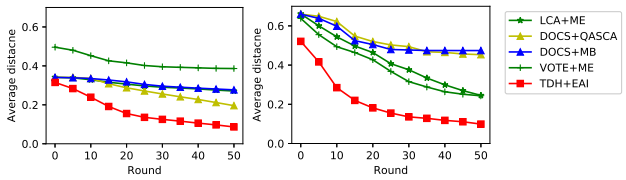
(a) *BirthPlaces* (b) *Heritages*

Figure 8: *Accuracy* with crowdsourced truth discovery



(a) *BirthPlaces* (b) *Heritages*

Figure 9: *GenAccuracy* with crowdsourced truth discovery



(a) *BirthPlaces* (b) *Heritages*

Figure 10: *AvgDistance* with crowdsourced truth discovery

from EAI are 0.08 and 0.26 percentage points (pps) while those errors from QASCA are 0.28 and 2.66 pps in *BirthPlaces* and *Heritages* datasets, respectively. This result confirms that EAI outperforms QASCA by effectively estimating the accuracy improvement. In terms of the other quality measures *GenAccuracy* and *AvgDistance*, our proposed EAI also outperforms the other task assignment algorithms in both datasets. Due to the lack of space, we omit the results with the other quality measures.

5.4 Simulated Crowdsourcing

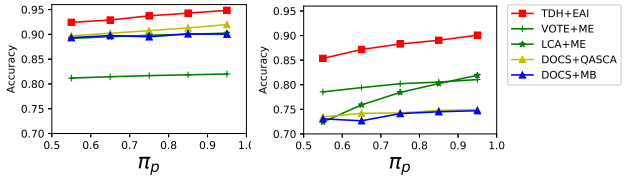
We evaluate the performance of crowdsourced truth discovery algorithms with the simulated crowdsourcing.

For all possible combinations of the implemented truth inference and task assignment algorithms, we show the *Accuracy* after 50 rounds of crowdsourcing in Table 4 where the impossible combinations are denoted by '-'. As expected, TDH+EAI has the highest *Accuracy* in both datasets for all possible combinations. The result also shows that both TDH and EAI contribute to increasing *Accuracy*. The improvement obtained by EAI can be estimated by comparing the result of TDH+EAI to that of the second performer TDH+QASCA. The accuracies of TDH+EAI in *BirthPlaces* and *Heritages* datasets are 1 and 3 percentage points (pps) higher than those of TDH+QASCA, respectively. In addition, for each combined task assignment algorithm, the improvement by TDH can be inferred by comparing the results with those of other truth inference algorithms. In both datasets, TDH shows the highest *Accuracy* among the applicable truth inference algorithms for each task assignment algorithm. For example, TDH+QASCA shows 2.6 and 13 pps higher *Accuracy* in *BirthPlaces* and *Heritages* datasets, respectively, than the second performer DOCS+QASCA among the combinations with QASCA. In the rest of the paper, we report *Accuracy*, *GenAccuracy* and *AvgDistance* of TDH+EAI, DOCS+MB, DOCS+QASCA, LCA+ME and VOTE+ME only since these combinations are the best or the second-best for each task assignment algorithm.

Cost efficiency: We plotted the average *Accuracy* of the tested algorithms for every 5 rounds in Figure 8. TDH+EAI shows the highest *Accuracy* for every round in both datasets. For the *BirthPlaces* dataset, DOCS+QASCA was the next best performer which achieved 0.9341 of *Accuracy* at the 50-th round. Meanwhile, TDH+EAI only needs 17 rounds of crowdsourcing to achieve the same *Accuracy*. Thus, TDH+EAI saved 66% of crowdsourcing cost compared to the second-best performer DOCS+ QASCA. Likewise, TDH+EAI reduced the crowdsourcing cost 74% in *Heritages* dataset compared to the next performer. In terms of *GenAccuracy* and *AvgDistance*, TDH+EAI also outperforms all the other algorithms as plotted in Figure 9 and Figure 10. The results confirm that TDH+EAI is the most efficient as it achieves the best qualities in terms of both *Accuracy* and *GenAccuracy*.

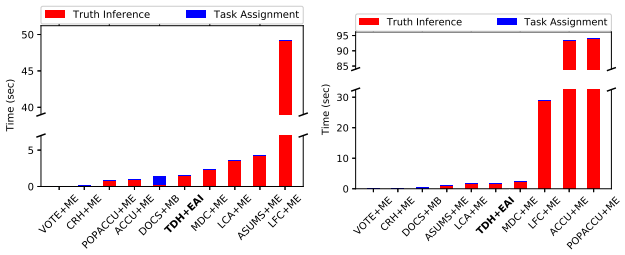
Varying π_p: We plotted the average *Accuracy* of all algorithms with varying the probability of correct answer π_p of simulated workers for *BirthPlaces* and *Heritages* datasets in Figure 11(a) and Figure 11(b), respectively. As we can easily expect, the accuracies increase with growing π_p for most of the algorithms. For both datasets, TDH+EAI achieves the best accuracy with all values of π_p. In Heritages dataset, a source provided less than 10 claims on average and it makes difficult for truth discovery algorithms to estimate the reliabilities of sources. Therefore, the baseline VOTE+ME shows good performance on Heritages dataset. Meanwhile, the performance of the state-of-the-art DOCS is significantly degraded on the Heritages dataset.

Execution times: We plotted the average execution times of the tested algorithms over every round in Figure 12. VOTE, CRH+ME, DOCS+MB and TDH+EAI run in less than 2.0 seconds per round on average for both datasets. Other algorithms except for ACCU+ME, POPACCU+ME and LFC+ME also take less than 5 seconds, which is acceptable for crowdsourcing. Since LFC builds the confusion matrix whose size is the square of the number of candidate values, LFC is the slowest with *BirthPlaces* data. On the other hand, for *Heritages* dataset which is collected



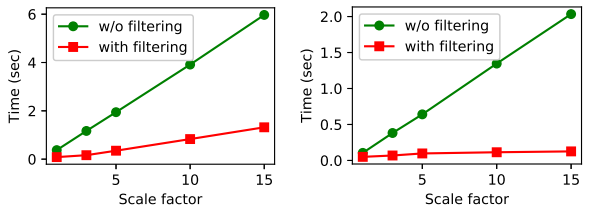
(a) *BirthPlaces* (b) *Heritages*

Figure 11: Varying π_p



(a) *BirthPlaces* (b) *Heritages*

Figure 12: Execution time per round



(a) *BirthPlaces* (b) *Heritages*

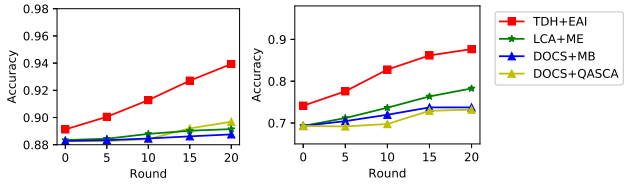
Figure 13: Execution time for task assignment per round

from much more sources than *BirthPlaces* dataset, ACCU and POPACCU take longer time for truth inference to calculate the dependencies between sources.

Effects of the filtering for task assignments: To test the scalability of our algorithm, we increase the size of both datasets by duplicating the data by upto 15 times. In Figure 13, with increasing data size, we plotted the execution times of our task assignment algorithm EAI with and without exploiting the upper bound proposed in Section 4.3. The filtering technique saved 78% and 94% of the computation time for the task assignment at the scale factor 15. The graphs show that the proposed upper bound enables us to scale for large data effectively. For the total execution time, including the truth inference, the filtering reduced 21% and 6% of the execution time on *BirthPlaces* and *Heritages* respectively at the scale factor 15.

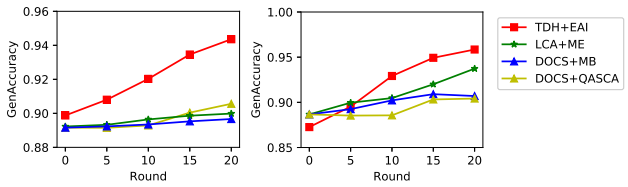
5.5 Crowdsourcing with Human Annotators

We evaluated the performance of the truth discovery algorithm by crowdsourcing real human annotations. For this experiment, we selected DOCS+QASCA, DOCS+MB and LCA+ME for comparison with the proposed algorithm TDH+EAI. This is because they are the best existing algorithms for each task assignment algorithm. We conducted this experiment with 10 human annotators for 20 rounds on our own crowdsourcing system. For each worker, we assigned 5 tasks in each round. Figure 14,15 and 16 show the performances of the algorithms against the rounds.



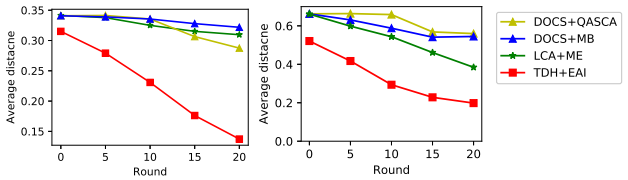
(a) *BirthPlaces* (b) *Heritages*

Figure 14: *Accuracy* with human annotations



(a) *BirthPlaces* (b) *Heritages*

Figure 15: *GenAccuracy* with human annotations



(a) *BirthPlaces* (b) *Heritages*

Figure 16: *AvgDistance* with human annotations

For both of the datasets, the results confirm that the proposed TDH+EAI algorithm outperforms the compared algorithms as in the previous simulations. Without crowdsourcing, the other algorithms show a higher *GenAccuracy* than TDH for *Heritages* dataset, because these algorithms tend to estimate the truths with more generalized form than TDH does. However, TDH+EAI shows the highest *GenAccuracy* after the 3rd round because it correctly estimates the reliabilities and the generalization levels of the sources by using the hierarchy. For *BirthPlaces* dataset, *Accuracies* of the algorithms increase a little bit faster than those in the experiment with simulated crowdsourcing. However, for *Heritages* dataset, *Accuracies* of the algorithms increase much slower than in the experiment with simulated crowdsourcing. It seems that finding the locations of a world heritages is a quite harder task than finding the birthplaces of celebrities because the birthplaces are often big cities (such as LA), which are familiar to workers, but World Cultural Heritages and World Natural Heritages are often located in unfamiliar regions.

5.6 Crowdsourcing with AMT

We evaluate the performances of TDH+EAI, DOCS+QASCA, DOCS+MB and LCA+ME based on the answers collected from Amazon Mechanical Turk (AMT). We collected answers for all objects in *Heritages* dataset from 20 workers in AMT. We made the collected answers available at http://kdd.snu.ac.kr/home/datasets/tdh.php. To evaluate the algorithms based on the collected answers, we assign 5 tasks for each worker in a round. We plotted the performance of the algorithms in Figure 17. Since we use more workers than we did in Section 5.5, the performances

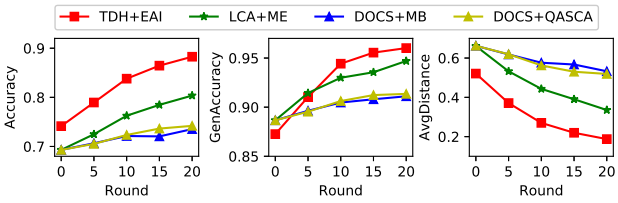


Figure 17: Crowdsourced truth discovery in *Heritages*

improve a little bit faster, but the trends are very similar to those with 10 human annotators in the previous section. We observe that our TDH+EAI outperforms all compared algorithms even with a commercial crowdsourcing platform.

5.7 Multi-truths Discovery Algorithms

Since there are multiple correct values including generalized values, we also implement multi-truth discovery algorithms such as DART[27], LFC[31] and LTM[38] to compare with our TDH algorithm. Since the multi-truths discovery algorithms independently generate the correct values, they may output the true values where there exist a pair of true values without ancestor-descendant relationship in the hierarchy. For example, from the given claimed values in Table 1, the multi-truth algorithms can answer that the 'Statue of Liberty' is located in LA and Liberty island. In this case, we cannot evaluate the result by our evaluation measures *Accuracy*, *GenAccuracy* and *AvgDistance*. Thus, to evaluate the performance of the tested algorithms, we utilize precision, recall and F1-score which are the evaluation measures typically used for multi-truths discovery. To use the multi-truths algorithms and the evaluation measures, we treat the ancestors of v and v itself as the multi-truths of v. LFC can work as either a single truth algorithm or a multi-truths algorithm. We refer to the multi-truth version of LFC as LFC-MT to avoid the confusion.

Table 5: Performance of truth discovery algorithms

		Dataset					
		BirthPlaces			*Heritages*		
Algorithm		Precision	Recall	F1	Precision	Recall	F1
Single truth	TDH	**0.899**	**0.921**	**0.910**	0.873	0.795	**0.832**
	VOTE	0.892	0.804	0.846	**0.899**	0.717	0.798
	LCA	0.892	0.913	0.903	0.878	0.711	0.786
	DOCS	0.892	0.913	0.902	0.887	0.722	0.796
	ASUMS	0.857	0.888	0.872	0.741	0.660	0.698
	POPACCU	0.847	0.858	0.852	0.859	0.694	0.768
	LFC	0.874	0.838	0.856	0.808	0.727	0.765
	MDC	0.844	0.853	0.848	0.807	0.792	0.800
	ACCU	0.830	0.842	0.836	0.766	0.631	0.692
	CRH	0.827	0.833	0.830	0.883	0.716	0.791
Multi -truths	LFC-MT	0.763	0.723	0.742	0.898	0.684	0.777
	DART	0.590	0.855	0.698	0.357	**0.994**	0.525
	LTM	0.780	0.472	0.588	0.871	0.672	0.759

Table 6: Performance evaluation for numerical data

	Change rate		Open price		EPS	
Algorithm	MAE	R/E	MAE	R/E	MAE	R/E
TDH	**0.0006**	**0.1011**	**0.0195**	**0.0354**	**0.0352**	**1.9513**
LCA	**0.0006**	**0.1011**	**0.0195**	**0.0354**	0.3831	16.2212
CRH	0.0020	1.6339	**0.0195**	**0.0354**	0.0610	1.9882
CATD	0.0104	2.3529	0.0211	0.0395	0.0803	3.2059
VOTE	**0.0006**	**0.1011**	**0.0195**	**0.0354**	0.0765	2.8402
MEAN	0.2837	30.8747	0.4047	0.5782	0.1762	7.3937

Table 5 shows the performance of the truth discovery algorithms in terms of precision, recall and F1-score. For both datasets, the TDH algorithm is the best in terms of F1-score. Recall that the VOTE algorithm tends to find a generalized value of the exact truth. Since a generalized truth generates a small number of multi-truths, the VOTE algorithm shows the highest precision in *Heritages* dataset. However, since its recall is much lower than that of our TDH algorithm, the F1-score of the VOTE algorithm is lower than that of the TDH algorithm. Similarly, although the DART algorithm has the highest recall in *Heritages* dataset, the precision of the DART algorithm is the smallest among the precisions of all compared algorithms.

5.8 Performance on a Numerical Dataset

To evaluate the extension to numerical data, we conducted an experiment on the stock datatset [23] which is trading data of 1000 stock symbols from 55 sources on every work day in July 2011. The detailed description of the data can be found in [23]. As we discussed at the end of Section 3.2, we can utilize our TDH algorithm for numeric dataset with implied hierarchy. We select three attributes 'change rate', 'open price' and 'EPS' of the dataset, and compared our TDH algorithm with the LCA, CRH, CATD[21], VOTE and MEAN algorithms. Among the second best performers DOCS and LCA in Table 4, we use only LCA for this experiment since DOCS requires the domain information while it is not available for this dataset. In addition to LCA, we implemented and tested the two algorithms CRH[22] and CATD[21] which are designed to find the truth in numerical data. Recall that VOTE is a baseline algorithm which selects the candidate value collected from majority sources. We also implemented a baseline algorithm, called MEAN, which estimates the correct value as the average of the claimed numeric values.

Table 6 shows the mean squared error (MAE) and the relative error (R/E) of the tested algorithms. The TDH algorithm performs the best for every attribute. The MEAN and CATD algorithms show worse performance than the other algorithms. Since they utilize an average or a weighted average of the claimed values, they are sensitive to outliers. The result confirms that our TDH algorithm is effective even for numerical data.

6 RELATED WORK

The problem of resolving conflicts from multiple sources (i.e., truth discovery) has been extensively studied [4, 5, 5, 7, 9, 16, 22, 24, 26, 27, 31, 33, 35–39, 42]. Truth discovery for categorical data has been addressed in [5, 7, 9, 22, 24, 31, 36, 39]. According to a recent survey [40], LFC[31] and CRH[22] perform the best in an extensive experiment with the truth discovery algorithms [4, 5, 16, 21, 35, 40, 42]. There exist other interesting algorithms [7, 9, 24, 39] which are not evaluated together in [40]. Accu[7] and PopAccu[9] combine the conflicting values extracted from different sources for the knowledge fusion [8]. They consider the dependencies between data sources to penalize the copiers' claims. DOCS[39] utilizes the domain information to consider the different levels of worker expertises on various domains. MDC[24] is a truth discovery algorithm devised for crowdsourcing-based medical diagnosis. The works in [26, 33, 37] studied how to resolve conflicts in numerical data from multiple sources.

The truth discovery algorithms in [30, 37–39] are based on probabilistic models. Resolving the conflicts in numerical data is addressed in [37] and discovering multiple truths for an object is studied in [38]. Probabilistic models for finding a single truth

for each object is proposed in [30, 39]. However, none of those algorithms exploit the hierarchical relationships of claimed values for truth discovery. The work in [2] adopts an existing algorithm to consider hierarchical relationships.

Task assignment algorithms[3, 11, 14, 28, 39, 41] in crowdsourcing have been studied widely in recent years. The works in [3, 39, 41] can be applied to our crowdsourced truth discovery. For task assignment, AskIt[3] selects the most uncertain object for a worker. Meanwhile, the task assignment algorithm in [39] selects the object which is expected to decrease the entropy of the confidence the most. QASCA [41] chooses an object which is likely to most increase the accuracy. Since QASCA outperforms AskIt in the experiments presented in [39, 41], we do not consider AskIt in our experiments. In [14], task assignment for binary classification was investigated but it is not applicable to our problem to find the correct value among multiple conflicting values. Meanwhile, the task assignment algorithm is proposed in [28] for the case when the required skills for each task and the skill set of every worker is available. However, it is not applicable to our problem. A task assignment algorithm proposed in [11] assigns every object to a fixed number of workers. However, since we already have claimed values from sources, we do not have to assign all objects to workers.

7 CONCLUSION

In this paper, we first proposed a probabilistic model for truth inference to utilize the hierarchical structures in claimed values and an inference algorithm for the model. Furthermore, we proposed an efficient algorithm to assign the tasks in crowdsourcing platforms. The performance study with real-life datasets confirms the effectiveness of the proposed algorithms.

ACKNOWLEDGMENTS

We appreciate the reviewers for providing their insightful comments. This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7063570). This research was also supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2016R1D1A1A02937186).

REFERENCES

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.
[2] Valentina Beretta, Sébastien Harispe, Sylvie Ranwez, and Isabelle Mougenot. 2016. How Can Ontologies Give You Clue for Truth-Discovery? An Exploratory Study. In *WIMS*. 15.
[3] Rubi Boim, Ohad Greenshpan, Tova Milo, Slava Novgorodov, Neoklis Polyzotis, and Wang-Chiew Tan. 2012. Asking the right questions in crowd data sourcing. In *ICDE*. 1261–1264.
[4] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* (1979), 20–28.
[5] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *WWW*. 469–478.
[6] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *SIGKDD*. 601–610.
[7] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. Integrating conflicting data: the role of source dependence. *PVLDB* 2, 1 (2009), 550–561.
[8] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. 2014. From data fusion to knowledge fusion. *PVLDB* 7, 10 (2014), 881–892.
[9] Xin Luna Dong, Barna Saha, and Divesh Srivastava. 2012. Less is more: Selecting sources wisely for integration. In *PVLDB*, Vol. 6. 37–48.
[10] Xin Luna Dong and Divesh Srivastava. 2015. Knowledge curation and knowledge fusion: challenges, models and applications. In *SIGMOD*. 2063–2066.
[11] Ju Fan, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, and Jianhua Feng. 2015. icrowd: An adaptive crowdsourcing framework. In *SIGMOD*. 1015–1030.
[12] Ju Fan, Meiyu Lu, Beng Chin Ooi, Wang-Chiew Tan, and Meihui Zhang. 2014. A hybrid machine-crowdsourcing system for matching web tables. In *ICDE*. 976–987.
[13] Daniel Haas, Jiannan Wang, Eugene Wu, and Michael J Franklin. 2015. Clamshell: Speeding up crowds for low-latency data labeling. *PVLDB* 9, 4 (2015), 372–383.
[14] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. 2013. Adaptive task assignment for crowdsourced classification. In *ICML*. 534–542.
[15] David R Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. In *NIPS*. 1953–1961.
[16] Hyun-Chul Kim and Zoubin Ghahramani. 2012. Bayesian classifier combination. In *AISTATS*. 619–627.
[17] Younghoon Kim, Woohwan Jung, and Kyuseok Shim. 2017. Integration of graphs from different data sources using crowdsourcing. *Information Sciences* 385 (2017), 438–456.
[18] Younghoon Kim, Wooyeol Kim, and Kyuseok Shim. 2017. Latent ranking analysis using pairwise comparisons in crowdsourcing platforms. *Inf. Syst.* 65 (2017), 7–21.
[19] David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR*. 3–12.
[20] Guoliang Li, Jiannan Wang, Yudian Zheng, and Michael J Franklin. 2016. Crowdsourced data management: A survey. *TKDE* 28, 9 (2016), 2296–2319.
[21] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. 2014. A confidence-aware approach for truth discovery on long-tail data. *PVLDB* 8, 4 (2014), 425–436.
[22] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *SIGMOD*. 1187–1198.
[23] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. 2012. Truth finding on the deep web: Is the problem solved?. In *PVLDB*, Vol. 6. 97–108.
[24] Yaliang Li, Nan Du, Chaochun Liu, Yusheng Xie, Wei Fan, Qi Li, Jing Gao, and Huan Sun. 2017. Reliable Medical Diagnosis from Crowdsourcing: Discover Trustworthy Answers from Non-Experts. In *WSDM*. 253–261.
[25] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A Survey on Truth Discovery. *SIGKDD Explor. Newsl.* 17, 2 (Feb. 2016), 1–16.
[26] Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. On the discovery of evolving truth. In *SIGKDD*. 675–684.
[27] Xueling Lin and Lei Chen. 2018. Domain-aware multi-truth discovery from conflicting sources. *PVLDB* 11, 5 (2018), 635–647.
[28] Panagiotis Mavridis, David Gross-Amblard, and Zoltán Miklós. 2016. Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing. In *WWW*. 843–853.
[29] Jeff Pasternack and Dan Roth. 2010. Knowing what to believe (when you already know something). In *COLING*. 877–885.
[30] Jeff Pasternack and Dan Roth. 2013. Latent Credibility Analysis. In *WWW*. 1009–1020.
[31] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *JMLR* 11, Apr (2010), 1297–1322.
[32] Princeton University. 2010. About WordNet. https://wordnet.princeton.edu/
[33] Mengting Wan, Xiangyu Chen, Lance Kaplan, Jiawei Han, Jing Gao, and Bo Zhao. 2016. From Truth Discovery to Trustworthy Opinion Discovery: An Uncertainty-Aware Quantitative Modeling Approach. In *SIGKDD*. 1885–1894.
[34] Jiannan Wang, Tim Kraska, Michael J Franklin, and Jianhua Feng. 2012. Crowder: Crowdsourcing entity resolution. *PVLDB* 5, 11 (2012), 1483–1494.
[35] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*. 2035–2043.
[36] Xiaoxin Yin, Jiawei Han, and S Yu Philip. 2008. Truth discovery with multiple conflicting information providers on the web. *TKDE* 20, 6 (2008), 796–808.
[37] Bo Zhao and Jiawei Han. 2012. A probabilistic model for estimating real-valued truth from conflicting sources. In *QDB*.
[38] Bo Zhao, Benjamin IP Rubinstein, Jim Gemmell, and Jiawei Han. 2012. A bayesian approach to discovering truth from conflicting sources for data integration. *PVLDB* 5, 6 (2012), 550–561.
[39] Yudian Zheng, Guoliang Li, and Reynold Cheng. 2016. DOCS: a domain-aware crowdsourcing system using knowledge bases. *PVLDB* 10, 4 (2016), 361–372.
[40] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: is the problem solved? *PVLDB* 10, 5 (2017), 541–552.
[41] Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, and Jianhua Feng. 2015. QASCA: A quality-aware task assignment system for crowdsourcing applications. In *SIGMOD*. 1031–1046.
[42] Denny Zhou, Sumit Basu, Yi Mao, and John C Platt. 2012. Learning from the wisdom of crowds by minimax entropy. In *NIPS*. 2195–2203.