TD-AC: Efficient Data Partitioning based Truth Discovery

Osias Noël Nicodème Finagnon TOSSOU
African Institute for Mathematical Sciences
Mbour, Senegal
osias.tossou@aims-senegal.org

Mouhamadou Lamine Ba
Université Alioune Diop de Bambey
Bambey, Senegal
mouhamadoulamine.ba@uadb.edu.sn

ABSTRACT
This paper introduces an effective algorithm, called TD-AC, for the truth discovery problem in scenarios where data attributes are correlated by distinct levels of reliability of the sources. TD-AC is built on an abstract representation of the truth in the data to automatically find an optimal partitioning of the input data using the k-means clustering technique and the silhouette measure. Such a data partitioning strategy ensures to maximize the accuracy of any base truth discovery process when executed on each partition. The intensive experiments conducted on synthetic and real datasets show that TD-AC outperforms baseline approaches with a more reasonable running time. It improves on synthetic datasets the accuracy of standard truth discovery algorithms by 1% at least and by 14% at most and also significantly when the data coverage rate is high for the other types of datasets.

KEYWORDS
Truth discovery, attribute, data partitioning, clustering, attribute truth vector, k-means, silhouette index, performance evaluation

1 INTRODUCTION
Dealing with contradictory claims about the same facts is a real concern in many real-world applications such as Web data integration systems [3], online crowdsourcing platforms, online news Websites, social media, etc. Truth discovery resolves such an issue by predicting which of the values provided by conflicting sources is true with no prior knowledge about the level of reliability of the sources. Many approaches [1, 2, 5, 7, 12] have been proposed for the truth discovery problem in the setting of the reliability of sources by corroborating their claims under various settings. As in [2], we investigate in this work the truth discovery with attribute partitioning problem that may occur in cases where the attributes over data are structurally correlated so that sources exhibit different levels of reliability on distinct subsets of facts. For instance, Source 1 is good on Q1 and Q3 and bad on Q2. Meanwhile Source 2 is good on Q2 and bad on Q1 and Q3. We say that Q1 and Q3 are about data attributes that are correlated according to the sources’ reliability levels; capturing these unknown groups of correlated attributes may help to avoid having a biased truth discovery process.

The approach in [2] finds the set of correlated data attributes for truth discovery as an optimal partitioning of the set of input data attributes using various weighting functions over sources’ reliability levels themselves estimated by the truth discovery algorithm. However, the different exploration strategies introduced in [2] are time-consuming and error-prone. In addition, its different weighting functions do not give any guarantees about the correctness of the returned optimal partition.

This paper revisits [2] and proposes a new more effective and efficient approach to the problem of truth discovery with attribute partitioning. The presented approach, called TD-AC, is based on an abstract representation of the truth in the data using the new concept of attribute truth vector. Given the set of attribute truth vectors, we rely on k-means clustering technique from machine learning domain to find the optimal partitioning of the data attributes. To determine the optimal number of clusters, we assess the homogeneity of the individuals in a clustering result with the help of the silhouette measure. This methodology guarantees to find an optimal partition or a near-optimal one maximizing the accuracy of any base truth discovery process, without an exploration of all the possible partitions. The results of our intensive experiments on synthetic, semi-synthetic and real datasets show that TD-AC outperforms approaches in [2], with a more reasonable time cost. On synthetic data, it improves the accuracy of standard algorithms at least by 1% and at most 14% at least and by 14% at most and also significantly when the data coverage rate is high for the other types of datasets.

Table 1: Example with sources having different levels of reliability with respect to distinct groups of data attributes

<table>
<thead>
<tr>
<th>Sources</th>
<th>Topic</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source 1</td>
<td>FB</td>
<td>Algeria</td>
<td>2000</td>
<td>12</td>
</tr>
<tr>
<td>Source 2</td>
<td>FB</td>
<td>Senegal</td>
<td>2019</td>
<td>11</td>
</tr>
<tr>
<td>Source 3</td>
<td>FB</td>
<td>Algeria</td>
<td>1994</td>
<td>12</td>
</tr>
<tr>
<td>Source 1</td>
<td>CS</td>
<td>Linux Torvalds</td>
<td>1830</td>
<td>7</td>
</tr>
<tr>
<td>Source 2</td>
<td>CS</td>
<td>Bill Gate</td>
<td>1991</td>
<td>8</td>
</tr>
<tr>
<td>Source 3</td>
<td>CS</td>
<td>Steve Jobs</td>
<td>1991</td>
<td>10</td>
</tr>
</tbody>
</table>
by 14% and also significantly when the data coverage is high for the other types of datasets.

The remaining of this paper is organized as follow. First, we give some preliminaries and define the studied problem in Section 2. Then, we detail our proposed approach by providing its different building blocks in Section 3. In order to validate our approach, we present in Section 4 the results of our intensive experiments conducted on various types of datasets and a thorough analysis of the obtained results. We briefly review the state-of-the-art truth discovery algorithms in Section 5 before concluding in Section 6 with some research perspectives.

2 CONCEPTS AND STUDIED PROBLEM

This section resumes the key concepts of the truth discovery problem and informally introduces the studied problem.

2.1 DEFINITION OF CONCEPTS

A typical truth discovery process usually assumes a structured world where input data consist of a set O of objects corresponding to real world entities. Each object is characterized by a set A of attributes (or properties) with values in V coming from a collection S of data sources. In a one-truth setting, every attribute for each object has one true value and several possible false values. Thus, the notion of value confidence C is used to assess the level of veracity of every value v. Meanwhile, the level of reliability T of a source s (or source accuracy) models its ability to provide true values for given real-world object attributes. In real applications, the confidence scores over provided values and the reliability levels of sources are both often unknown and initialized to default values depending on the setting before being updated during the execution of the truth discovery algorithm.

This work considers groups of attributes over data to be structurally correlated if every source has the same reliability level on these latter.

2.2 PROBLEM STATEMENT

Given the triplet (S, A, O) in a one-truth setting in which a given source may not cover all the objects or attributes, the truth discovery problem is commonly defined as follows.

**Problem 1.** Find, for each object o in O, the true value of every attribute a in A, amongst its set $V_o$, of possible values by corroborating claims from sources in S, where $A_o$ and $S_o$ are the set of attributes of o and the set of sources providing values for o.

We informally introduce the truth discovery with attribute partitioning problem as follows.

**Problem 2.** Find an optimal partitioning P of A that maximizes the accuracy of any solution for Problem 1 where each partition in P contains correlated data attributes according to sources' reliability levels.

In next, we propose an efficient clustering based approach to solve Problem 2 when data attributes are structurally correlated.

3 TRUTH DISCOVERY WITH CLUSTERING

This section presents our proposed algorithm, called TD-AC, that discovers the truth by data partitioning. TD-AC, that stands for Truth Discovery with Attribute Clustering, applies k-means to find optimal clusters of structurally correlated data attributes based on sources’ reliability level by relying on attribute truth vectors and the silhouette index, as we detail it below.

### 3.1 DATA ATTRIBUTE TRUTH VECTORS

We define and use the concept of data attribute truth vectors as an abstract representation of the precision (or quality) of a given truth discovery algorithm using attributes as dimensions. To build such vectors, we firstly apply a base truth discovery algorithm (e.g. majority voting) on input data to obtain a reference truth. Then, for each attribute of an object and every source we verify whether or not the value given by the source is true regarding the reference truth; we set the value for each rank of any attribute vector according to Equation 1.

$$V_a \in A, V_o \in O, V_s \in S; x (a, o, s) = \begin{cases} 1 & \text{if } p \text{ is true} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $p = (v(a, o, s) \in v_p(a, o))$ with $v(a, o, s)$ representing the value given by s about a of o, $v_p(a, o)$ is the true value of a of o predicted by the base algorithm and $x(a, o, s)$ is a binary value of our truth vector. Table 2 sketches the matrix of attribute truth vectors obtained on our running example in Table 1 by applying the procedure described above and Equation 1.

<table>
<thead>
<tr>
<th></th>
<th>FB</th>
<th>CS</th>
<th>FB</th>
<th>CS</th>
<th>FB</th>
<th>CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Q2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Q3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: Matrix of attribute truth vectors with data in Table 1 using TruthFinder as base algorithm

### 3.2 GROUPING CORRELATED ATTRIBUTES

We find and group correlated data attributes by assessing the similarity distance of their corresponding truth vectors. Given two distinct attributes $a_1, a_2$ and their truth vectors $(a_1^1, a_1^2, \ldots, a_1^l)$ and $(a_2^1, a_2^2, \ldots, a_2^l)$, we define the similarity between $a_1$ and $a_2$ using the Hamming distance as:

$$d(a_1, a_2) = \sum_{i=1}^{l} |a_1^i - a_2^i|$$

To automatically devise the threshold value for grouping the attributes based on our similarity measure, we rely on k-means and its optimization strategy in order to provide a domain-independent clustering process in practical cases. The k-means clustering approach [8] uses a similarity distance metric between data points to group them in k clusters. Given a set of observations $(a_1, a_2, \ldots, a_n)$, where every observation is an attribute truth vector having l dimensions, we define the partitioning of these attributes using k-means algorithm as the clustering of the n observations in k $(k \leq n)$ disjoint sets (or clusters) $C = \{g_1, g_2, \ldots, g_k\}$ in such a way that the sum of the squares (i.e. the Inertia) within each cluster is minimized. Formally, the goal is to find:

$$\text{argmin}_C \sum_{i=1}^{k} \sum_{a \in g_i} ||a - \mu_i||^2 = \text{argmin}_C \sum_{i=1}^{k} |g_i|\text{Inertia}(q_i)$$

where $\mu_i$ is the centroid of the points in $g_i$. This corresponds to minimize the squared deviations of the points in the same cluster:

$$\text{argmin}_C \sum_{i=1}^{k} \sum_{a \in g_i} \frac{1}{|g_i|} \sum_{a \in g_i, a \neq a'} ||a_1 - a_2||^2$$

K-means requires to specify the value of k in input. We find the optimal k using the silhouette index as described next.

### 3.3 ESTIMATION OF k WITH SILHOUETTE

The silhouette index [11] evaluates the quality of a clustering result with the help of the separation criteria $\beta$ and the cohesion criteria $\alpha$. Consider two attributes $a_1$ and $a_2$ that belong to clusters $g(1)$ and $g(2)$, respectively. Formally, the silhouette coefficient $CS(a_1)$ of the attribute $a_1$ is defined as:

$$CS(a_1) = \frac{\max(\alpha(a_1), \beta(a_1))}{\min(\alpha(a_1), \beta(a_1))}$$

with $\alpha(a_1) = \frac{1}{|g(1)|} \sum_{a \in g(1), a \neq a_1} d(a_1, a)$ and $\beta(a_1) = \min_{a_1, a_2} \frac{1}{|g(2)|} \sum_{a_2 \in g(2)} d(a_1, a_2)$ (5). If $CS(a_1) <
0, \( a_1 \) is badly classified. Conversely, if \( CS(a_1) > 0 \) \( a_1 \) is well classified. Finally, if \( CS(a_1) = 0 \) then \( a_1 \) is between two clusters. The silhouette coefficient \( CS(g) \) of a cluster \( g \) is thus given by:

\[
CS(g) = \frac{1}{|g|} \sum_{a \in g} CS(a)
\]

The silhouette value of a partition \( P \) is the average of the silhouette coefficients of all its clusters:

\[
CS(P) = \frac{1}{|P|} \sum_{g \in P} CS(g) \tag{7}
\]

The optimal \( K \) is the one associated to the partition having the highest silhouette coefficient amongst all the possible partitions.

### 3.4 TD-AC TRUTH DISCOVERY APPROACH

As depicted by Algorithm 1, our proposed algorithm TD-AC runs as follows: (i) considers a base truth discovery algorithm and input data (\( A, O, S \)); (ii) computes the matrix of attribute truth vectors from input data using the base algorithm and Equation 1; (iii) efficiently clusters the data attributes by applying k-means combined with the silhouette index; and (iv) executes the input base truth discovery algorithm on each data partition, and then aggregates the partial results to generate the entire result.

#### Algorithm 1

**Input:** Set of observations (\( A, O, S \)), Base algorithm \( F \)

**Output:** Truth predicted by TD-AC

1. **results** ← \[\]
2. **truth_vector_matrix** ← buildTruthVectors\( (F,A,O,S) \)
3. // Find the optimal partition with k-mean and silhouette
4. **indice_silhouette** ← 0
5. opt_partition ← \[
6. for all \( k \in [2, |A| - 1] \) do
7. \( partition=\text{kmeansAttrClustering}(truth\text{-vector}\text{matrix}) \)
8. **silhouette\_index\_tmp** ← CS(partition)
9. if \( k == 2 \) then
10. **silhouette\_index** ← **silhouette\_index\_tmp**
11. else
12. **indice_silhouette** ← **indice_silhouette\_tmp**
13. if \( indice\_silhouette < indice\_silhouette\_tmp \) then
14. **silhouette\_index** ← **silhouette\_index\_tmp**
15. opt_partition ← partition
16. end if
17. end if
18. \( F, A, O, S \) ∈ results
19. // Truth discover on the optimal partition found
20. for each \( g \in opt\_partition \) do
21. \( A_g, O_g, S_g \leftarrow \text{getData}(g) \)
22. partial_result ← \( F(A_g, O_g, S_g) \)
23. Add partial_result in results
24. end for

### 4 EXPERIMENTS AND RESULTS

In this section, we demonstrate the efficiency of our approach on various datasets, proving that it outperforms approaches proposed in [2] and standard truth discovery algorithms in the literature in the presence of structurally correlated data attributes. We also show that its execution time is similar to that of standard algorithms unlike partitioning strategies in [2]. We start by presenting the experiment setting up and performed tests.

#### 4.1 EXPERIMENTATION SETTING UP

For the comparison purposes, we have implemented the different analyzed algorithms using Python programming language. The following standard truth discovery algorithms have been implemented: MajorityVote, TruthFinder [14], DEPEN, Accu and AccuSim [4]. We have compared ourselves to these algorithms because they are amongst the best in terms of efficiency and effectiveness for solving the truth discovery problem in various settings. In addition we have also implemented AccuGenPartition in [2] along with the different weighting functions to compute the optimal partition. The source codes of the tested algorithms are all available at https://github.com/osiastossou/ProjeTD-AC.git.

We have conducted all our experiments on an Intel Core i5 2.6GHz laptop computer with 8GB of RAM, 250GB of hard disk space, and 1.5GB of graphics memory. The implemented algorithms here require all hyper-parameters in input whose values have been fixed for the various tests according to [12]. At last, we have relied on usual metrics such as precision, recall, F1-measure, accuracy, and execution time to evaluate and compare the performance of our tested algorithms.

### 4.2 EXPERIMENTS ON SYNTHETIC DATA

We detail here the results of our experiments on synthetic data which simulate conditions where data attributes are structurally correlated.

We have used and re-implemented in Python the synthetic data generator in [2] to produce our synthetic data sets; we defer to [2] for the details. For the evaluation process, we have then generated three synthetic datasets (DS1, DS2 and DS3) of 6 attributes, 1000 objects, 10 sources and 60,000 observations with three different configurations as depicted in Table 3; DS1 meets the setting of this work while DS2 relaxes the assumptions to test the robustness of our approach. The partition selected for each configuration is given in Table 5.

Tables 4a, 4b and 4c respectively present the performances of each algorithm on DS1, DS2 and DS3. For the tests, we used Accu as our base algorithm similarly to the approaches in [2].

We observe that the attribute partitioning truth discovery algorithms perform better than the standard ones on all three synthetic datasets, proving the importance of partitioning when data attributes are structurally correlated. Specifically, TD-AC is the only partitioning strategy with a precision comparable to the real world (i.e. an Oracle) without a blowup of the running time. Table 5 reports the partitions returned by the different partitioning approaches.

#### 4.3 TESTS ON SEMI-SYNTHETIC DATA

The semi-synthetic datasets have been generated from a real dataset called Exam. This real dataset comes from [2] and has been used in that paper to validate the proposed approaches. The Exam dataset has been obtained by aggregating the anonymous results of admission examinations. Unfortunately, it cannot be redistributed for privacy reasons. We had access to answers from 248 students (sources) to 124 questions (attributes) in total, from 9 different domains: Math 1A, Chemistry 1, Math 1B, Physics, Electrical Engineering, Computer Science, Chemistry 2, Science of life, and Math 2. We also know the correct answer to each question. Math 1A and Physics were only mandatory with the choice of an additional domain between Chemistry 1 and Math 1B. The five remaining domains were completely optional and wrong answers were penalized. As a result, all the attributes were not covered (missing data). For each unanswered
and TD-AC + TruthFinder. In general, we note that combining a base algorithm with TD-AC does not highly deteriorate the performance of the standard algorithm whatever the configuration considered, and even improves it in some cases, for example for the dataset with 124 attributes (see Figures 2 and 3).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1-measure</th>
<th>Time(s)</th>
<th>#Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>MajorityVote</td>
<td>0.613</td>
<td>0.688</td>
<td>0.751</td>
<td>0.666</td>
<td>437</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>TruthFinder</td>
<td>0.652</td>
<td>0.702</td>
<td>0.714</td>
<td>0.682</td>
<td>376</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>DEPEN</td>
<td>0.698</td>
<td>0.749</td>
<td>0.814</td>
<td>0.752</td>
<td>112</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>AccuGenPartition (Max)</td>
<td>0.782</td>
<td>0.909</td>
<td>0.860</td>
<td>0.845</td>
<td>152</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>TD-AC (F=Accu)</td>
<td>0.782</td>
<td>0.923</td>
<td>0.885</td>
<td>0.877</td>
<td>112</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4: Performance of all tested algorithms on the synthetic datasets DS1, DS2 and DS3

Figure 1: Comparison of the accuracy of all tested algorithms on DS1, DS2 and DS3

Table 5: Partitions chosen by the generator and returned by the different partitioning algorithms

Table 6: Performance of Accu, TruthFinder, TD-AC(F=Accu), and TD-AC(F=TruthFinder) on semi-synthetic datasets with 62 attributes

4.4 EXPERIMENTS ON REAL DATA

To end our performance evaluation, we present in this section the results of the experimentation of our approach and the existing algorithms on real data. The evaluation on real data sets enables to validate our approach against practical applications. For this purpose, we have considered and used the real datasets Exam [2], Stocks and Flights [9]. Real data contain missing values that

Table 7: Performance of Accu, TruthFinder, TD-AC(F=Accu), and TD-AC(F=TruthFinder) on semi-synthetic datasets with 124 attributes
4.5 Analysis of the results and discussion

The analysis of the presented intensive performance evaluation carried out on several datasets yields to three main observations.

Table 9 presents the performance measures of Accu, TD-AC+Accu, TruthFinder, and TD-AC+TruthFinder. We have also reported in Figures 4 and 5 the comparative study of the accuracy values of Accu and TD-AC+Accu on the one hand and TruthFinder and TD-AC+TruthFinder on the other hand on real datasets with data coverage greater than 66% and less than 55% respectively. We observe that Accu and TruthFinder outperform when used with our TD-AC approach, especially when the data coverage rate is greater than 66%. We also remark that the execution time of TD-AC is very close to that of standard algorithms on real data, unlike AccuGenPartition.

Figure 4: Study of the impact of TD-AC on Accu and TruthFinder by pairwise comparison of the accuracy values on real datasets Exam with 32 attributes, Stocks and Flights (DRC ≥ 66)

Table 9: Performance of Accu, TruthFinder, TD-AC+Accu, and TD-AC+TruthFinder on real datasets

Figure 5: Study of the impact of TD-AC on Accu and TruthFinder by pairwise comparison of the accuracy values on the real datasets Exam with 62 and 124 attributes (DRC ≤ 55)

TD-AC outperforms baseline partitioning approaches. TD-AC highly improves the accuracy of AccuGenPartition by at least and by 14% at most (see Figure 1) with a time complexity.
around 200 less significant (see Tables 4a, 4b and 4c). AccuGen-Partition is our baseline brute force approach proposed in [2] for the truth discovery with attribute partitioning problem with two weighting functions: Max and Avg. To discover the optimal partition, k-means combined with the silhouette index has been shown in Table 5 to be better than Max and Avg because: (i) k-means is a robust partitioning technique with a well-defined optimization strategy; and (ii) silhouette returns the most structurally homogeneous existing clusters. This explains the effectiveness of TD-AC. The drastic reduction of the running time with TD-AC is because it only requires one iteration to last without exploring all the possible partitions.

TD-AC improves the accuracy of base algorithms. When data attributes are structurally correlated, TD-AC significantly enhances the accuracy (from 5 to 35%) of standard algorithms (see Tables 4a, 4b, 4c, 9 and Figure 3). Standard algorithms alone do not capture the structural correlations between attributes leading to biased results. In the cases where the conditions do not match our working setting, TD-AC does not degrade the performances of the standard algorithms (see Tables 6 and 7). The impact of TD-AC is more important for Accu than TruthFinder because the former captures better the different levels of reliability of the sources. Such a impact introduces, however, a surplus in terms of execution time which is fortunately reasonable.

Correlation between coverage and TD-AC accuracy. The main observation is that TD-AC is more efficient when the data coverage is very high, i.e. DCR ≥ 66% (see Figure 4) because more one has in terms of information the better is the clustering with k-means. Lot of missing values, i.e. very sparse truth vectors affect both the quality of the clustering and the truth discovery process (see Figure 5).

5 RELATED WORK

A significant effort has been made in truth discovery area over the past years which has led to several approaches [12, 15]. The simplest approach is the majority vote which considers the truth said by the majority of sources. More elaborated approaches try to model the different levels of reliability of the sources and domain-specific aspects of the truth. For instance, TruthFinder[14], one of the first proposed standard algorithms, is a probabilistic model based on Bayesian analysis with similar value supporting each others in vote counts. Methods such as DEPEN, Accu and AccuSim[4] take into consideration copy relationships that may exist between sources by penalizing the vote of a source if it is detected as a copy of another source. DART (Domain-Aware Truth Discovery) [10] is both a probabilistic and a bayesian model which integrates the domain expertise level. Very recent methods such as [6, 15] capture the correlations between objects in the domain of Mobile Crowd Sensing.

The research works that are connexe to our studied problem are [2] and [13]. The proposal in [2] is a brute force approach that explores all the possible partitions of a given set of attributes in order to discover the one maximizing the precision of a standard truth discovery algorithm. The goodness of a partition in this case is based on a weighting function over sources’ reliability levels. The work in [13] focuses on object partitioning based on domain knowledge and some additional constraints.

6 CONCLUSION AND PERSPECTIVES

In this work, we have studied the truth discovery problem in a setting where the attributes of the data are structurally correlated.

As a solution, we have proposed a new approach, called TD-AC, built on an abstract representation of the truth in the data, the k-means clustering technique and the silhouette measure to automatically find an optimal partitioning of the input data (or a near-optimal) maximizing the accuracy of any base truth discovery process. Through an intensive experimental evaluation over various types of datasets, we have then shown the effectiveness and efficiency of TD-AC compared to existing partitioning strategies and its positive impact to the accuracy of any standard truth discovery process.

Despite of that, we have noticed that when the dataset contains lot of missing values, the impact of our approach is less significant. This can be explained by the use of sparse truth vectors in the clustering step, making the finding of the optimal partition hard. Moreover, even if the running time of our approach and standard algorithms is reasonable in the presence of small size datasets, it become important when the number of attributes, objects and sources is very large. As research perspectives, we plan to (i) improve our approach to better account for data with lot of missing values on the one hand; and (ii) on the other hand, to propose an optimization of the running time of our approach, in particular the optimal partition computation, by using parallel computation. We also plan to compare ourselves to a larger set of standard truth discovery algorithms and the partitioning approach in [13].

REFERENCES

[14] Xiaoxin Yin, Jiawei Han, and Philip Yu. 2008. Truth Discovery with Multiple Conflicting Information Providers on the Web. Knowledge and Data Engineering, IEEE Transactions on 20 (Jul 2008), 796 – 808.