Robustness and Scalability of Truth Finding Algorithms

Dalia Attia Wagui Mouhamadou Lamine BA Laure Berti-Equille

ABSTRACT

Motivations. Truth finding is an important because conflicting , erroneous, and dirty information are everywhere. The truth must be tell when reconciling such a conflicting data from different sources. This has lead to much effort of the database community and well founded truth discovering algorithms. However, there is a lack of a comparative study of both the scalability and the robustness of these algorithms. The existing comparative studies only focus on accuracy aspects. We describe, reimplement, and compare the most prominent solutions, so far, for the truth finding problems.

To tackle this lack, we propose in this paper an experimental study of the robustness and the scalability of the most referenced truth finding algorithms. Our outline is as follows. First, we overview the truth finding problem by giving preliminary definitions and propose a classificartion of the literature in truth finding. Then, we restrict ourselves to the twelve more referenced truth finding algorithms, by detailing each of them. We finally proceed to the experimental study of the robustness and the scalability of the twelve algorithms.

1. INTRODUCTION

2. OVERVIEW ON TRUTH FINDING ALGORITHMS

2.1 Preliminaries

2.2 Classification

Cluster the algorithms in the three following classes.

- Iterative algorithms
- EM based algorithms
- Dependency detection based algorithms

3. TWELVE TRUTH FINDING ALGORITHMS

We details in the sections, the truth finding algorithms we have considered for the comparative study. Some points to put in this section.

- Details each technique by giving its pseudo-code
- Hierarchical graph based representation of the common characteristics of the different algorithms

4. EXPERIMENTAL STUDY

4.1 Robustness evaluation

In this section, we evaluate and compare the robustness of the algorithms in terms of the variation of their accuracy with respect to parameter initialization and the characteristics of used datasets.

4.1.1 Parameter Initialization

Evaluation of the robutness of the parameters of each algorithm on Book dataset. Validation of parameter setting inferred from experiments on the Book dataset by using other datasets, e.g., Flight dataset.

4.1.2 Dataset Characteristics

Evaluation of the robutness of each algorithm with respect to the type of dataset.

4.2 Scalability evaluation

We evaluate and compare here the scalability of the algorithms, i.e., their running time when the amount of data increases exponentially.

5. GUIDELINES AND DISCUSSIONS

Propose guidelines in terms of what algorithm to use in what contexts with respect to the results of the experimental study. Then, discuss about potential improvements of compared algorithms and maybe a more efficient way to perform truth finding.

6. CONCLUSION