2017 IEEE 33rd International Conference on Data Engineering

# A Human-and-Machine Cooperative Framework for Entity Resolution with Quality Guarantees

Zhaoqiang Chen*, Qun Chen* and Zhanhuai Li*
*School of Computer Science
Northwestern Polytechnical University, Xi'an, 710072
{chenzhaoqiang@mail., chenbenben@, lizhh@}nwpu.edu.cn

***Abstract*—For entity resolution, it remains very challenging to find the solution with quality guarantees as measured by both precision and recall. In this demo, we propose a HUman-and-Machine cOoperative framework, denoted by HUMO, for entity resolution. Compared with the existing approaches, HUMO enables a flexible mechanism for quality control that can enforce both precision and recall levels. We also introduce the problem of minimizing human cost given a quality requirement and present corresponding optimization techniques. Finally, we demo that HUMO achieves high-quality results with reasonable return on investment (ROI) in terms of human cost on real datasets.**

**Video: http://www.wowbigdata.com.cn/HUMO/video.html.**

## I. Introduction

Entity resolution (ER) usually refers to identifying the relational records that correspond to the same real-world entity. It has been extensively studied in literature [1]. However, most of the existing approaches do not have the mechanism for quality control. Even though there exists some work [2] (based on active learning) that can optimize recall while guaranteeing a user-specified precision level, it is usually desirable in practice that the results have quality guarantees on both precision and recall fronts.

To this end, we propose a human-and-machine cooperative framework (HUMO) with a flexible mechanism for quality control. Its primary idea is to divide the pair instances in an ER task into easy ones, which can be labelled by machine with high accuracy, and more challenging ones, which require human intervention. HUMO is, to some extent, motivated by the success of human and machine cooperation in problem solving as demonstrated by crowdsourcing applications. We note that crowdsourcing for ER [3] mainly focused on how to make human work effectively and efficiently given a task. HUMO instead investigates the problem of how to divide the workload in a task between human and machine such that a quality requirement can be met. Since the workload assigned to human can usually be performed by crowdsourcing, HUMO can be considered to be a preprocessing step before a crowdsourcing task can be invoked. This demo makes the following contributions:

- We propose a human-and-machine cooperative framework (HUMO) for entity resolution that can enforce quality control on both precision and recall fronts;
- We introduce the problem of minimizing human cost given a quality requirement in HUMO and propose corresponding optimization techniques;
- We demo that HUMO achieves high-quality results with reasonable ROI in terms of human cost on real datasets;

## II. Framework & Solutions

### A. Framework

Given a set of record pair instances $D$, an ER task is to label each instance in $D$ as *matched* or *unmatched*. The purpose of HUMO is to divide $D$ into two disjoint subsets $D_m$ and $D_h$, which are then labelled by machine and human respectively, such that user-specified precision and recall levels can be met with minimal human effort. We suppose that each instance in $D$ has a machine-computed similarity value, which measures the record similarity between the corresponding pair. As in [2], we suppose that $D$ statistically satisfies the property of monotonicity of match proportion, which is defined as follows:

*Assumption 1 (Monotonicity of Match Proportion):* A value interval $I_1$ is dominated by another interval $I_2$, denoted by $I_1 \preceq I_2$, if every value in $I_1$ is less than every value in $I_2$. We say that match proportion is monotonic with respect to similarity if for any two similarity value intervals $I_1 \preceq I_2$ in [0,1], $r(I_1) \leq r(I_2)$, in which $r(I_i)$ denotes the match proportion of the instances whose similarity values are located in $I_i$.

The underlying intuition of Assumption 1 is that the more similar two records are, the more likely it is that they refer to the same real-world entity. Based on the monotonicity assumption, HUMO divides the similarity interval [0,1] into three disjoint intervals, $\{I_1{=}[0,v_l),I_2{=}[v_l,v_u],I_3{=}(v_u,1]\}$, and correspondingly $D$ into three disjoint subsets, as shown in Figure 1, in which $D_i$ represents the set of instances whose similarity values are located in $I_i$. HUMO automatically labels the instances in $D_1$ as *unmatched*, the instances in $D_3$ as *matched*, and assigns the instances in $D_2$ for human verification.

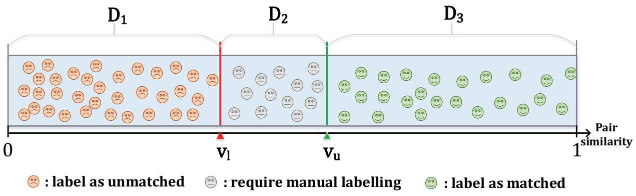


Fig. 1: The HUMO Framework

For simplicity of presentation, this demo assumes that the instances in $D_2$ can be manually labelled with 100% accuracy. The effectiveness of HUMO however does not depend on the 100%-accuracy assumption. In fact, HUMO can work properly provided that a quality guarantee can be enforced on $D_2$, but the best quality guarantee it can achieve is no better than that of $D_2$. The optimization purpose of HUMO is to minimize the required human effort given a quality guarantee.

2375-026X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDE.2017.197

By quantifying human effort by the number of instances in $D_2$, we can present the optimization problem as

$$\begin{aligned} & minimize \quad |D_2| \\ & subject \quad to \quad \frac{|D_2| \cdot r(D_2) + |D_3| \cdot r(D_3)}{|D_2| \cdot r(D_2) + |D_3|} \geq \beta, \\ & \frac{|D_2| \cdot r(D_2) + |D_3| \cdot r(D_3)}{|D_1| \cdot r(D_1) + |D_2| \cdot r(D_2) + |D_3| \cdot r(D_3)} \geq \gamma, \end{aligned} \tag{1}$$

in which $\beta$ and $\gamma$ denote the user-specified precision and recall levels, and $r(D_i)$ the ground-truth match proportion of $D_i$. Note that HUMO achieves the 100% precision and recall levels in the extreme case of all the instances being assigned to human (i.e. $D_2$=$D$). In general, its achieved precision and recall levels tend to decrease as $D_2$ becomes smaller.

### B. Solutions

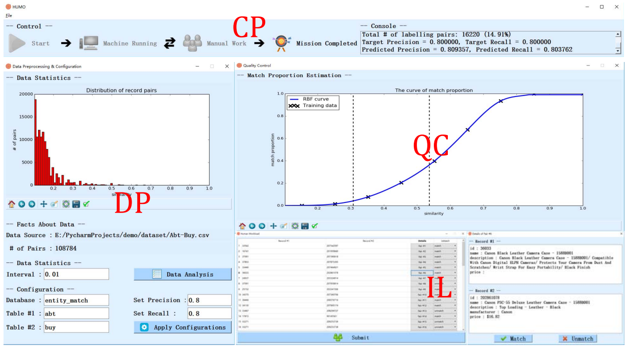
The challenge of solving the optimization problem as presented in Equation 1 mainly results from the observation that the ground-truth match proportions of $D_1$ and $D_3$ are unknown and have to be estimated. Here we briefly sketch our solutions, whose technical details can be found in our technical report [4].

**Baseline.** The baseline solution begins with an initial middle-valued similarity (e.g. the boundary value of a learned classifier) and then incrementally identifies the lower and upper bounds of the similarity interval of $D_2$, $v_l$ and $v_u$. To enforce precision, it iteratively moves $v_u$ from $v_u^i$ to a *higher* value of $v_u^{i+1}$. With the monotonicity assumption, it can be proved that the precision requirement of $\beta$ on $D$ would be met once the match proportion of $I_i$=[$v_u^i$,$v_u^{i+1}$] exceeds a corresponding threshold. Similarly, recall can be enforced by incrementally moving $v_l$ to a *lower* value.

**Sampling-based.** The drawback to the baseline solution is that it may overestimate (sometimes greatly) the match proportion of the instances with low similarities and also underestimate (but to a lesser extent) the match proportion of the instances with high similarities. Therefore, we have also proposed an improved approach based on match proportion estimation. It divides $D$ into many disjoint subsets and estimates their match proportions by sampling. To save human cost, *not* all the subsets are required to be sampled. With the sampled estimates, it then approximates the match proportions of all the subsets by either a singular function (e.g. logistic function) or a combination of radial basis functions (RBFs). Finally, it computes the lower and upper similarity bounds of $D_2$ based on the estimated match proportion function on $D$.

## III. Demo and Evaluation

We have implemented a prototype system, as shown in Figure 2, which consists of four components of controller panel (CP), data preprocessing (DP), quality control (QC), and instance labelling (IL). CP provides with an interactive interface to carry out an ER task and print out the logs. DP analyzes raw data and measures the similarity between a pair of records by a user-specified metric. QC samples instances from an input $D$, estimates the match proportion function on $D$ and identifies the subset of $D$ requiring human verification. IL enables human to label instances from QC as matched or unmatched and returns results to QC.



Total # of labelling pairs: 16220 (14.91%)
Target Precision = 0.800000, Target Recall = 0.800000
Predicted Precision = 0.809357, Predicted Recall = 0.803762

-- Facts About Data --
Data Source : E:/PycharmProjects/demo/dataset/Abt-Buy.csv
# of Pairs : 108784

-- Data Statistics --
Interval : 0.01

-- Configuration --
Database : entity_match
Table #1 : abt
Table #2 : buy
Set Precision : 0.8
Set Recall : 0.8

-- Record #1 --
id : 36033
name : Canon Black Leather Camera Case - 15888001
description : Canon Black Leather Camera Case - 15888001/ Compatible With Canon Digital ELPH Cameras/ Protects Your Camera From Dust And Scratches/ Wrist Strap For Easy Portability/ Black Finish
price :

-- Record #2 --
id : 202861078
name : Canon PSC-55 Deluxe Leather Camera Case - 15888001
description : Top Loading - Leather - Black
manufacturer : Canon
price : $16.82

Fig. 2: The HUMO Demo System

TABLE I: Evaluation Results on the Abt-Buy Dataset.

| Quality Requirement | Precision & Recall | | | Pct. of Manual Work (%) | | |
|---|---|---|---|---|---|---|
| | Baseline | LOG (Avg.) | RBF (Avg.) | Baseline | LOG (Avg.) | RBF (Avg.) |
| $\beta = 0.80$<br>$\gamma = 0.80$ | $\hat{\beta} = 0.99$<br>$\hat{\gamma} = 0.98$ | $\hat{\beta} = 0.81$<br>$\hat{\gamma} = 0.92$ | $\hat{\beta} = 0.82$<br>$\hat{\gamma} = 0.84$ | 62.09 | 29.25 | 16.10 |
| $\beta = 0.90$<br>$\gamma = 0.90$ | $\hat{\beta} = 0.99$<br>$\hat{\gamma} = 0.99$ | $\hat{\beta} = 0.90$<br>$\hat{\gamma} = 0.98$ | $\hat{\beta} = 0.90$<br>$\hat{\gamma} = 0.92$ | 73.14 | 52.33 | 26.36 |

The evaluation results of HUMO on the real Abt-Buy dataset[1] are presented in Table I. The reported manual work percentages of the sampling-based solutions include sampling cost. Also note that for the sampling-based solutions, different runs may result in different subsets of $D_2$ and different matching qualities. Their reported results are therefore the averages over 100 runs. We have the following observations: (1) the baseline solution achieves the precision and recall levels well beyond what are required, but needs heavy human effort. In comparison, the sampling solutions achieve the precision and recall levels very close to what are required with much less human effort; (2) the sampling solution based on RBFs performs the best among them. It achieves good-quality results with reasonable ROI.

## Acknowledgements

This work is supported by the Ministry of Science and Technology of China, National Key Research and Development Program (Project Number: 2016YFB1000700), the National High Technology Research and Development Program of China under Grant No.2015AA015307, the National Natural Science Foundation of China under Grant No.61332006, No.61672432, No.61472321 and No.61502390.

## References

[1] P. Christen, *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[2] A. Arasu, M. Götz, and R. Kaushik, "On active learning of record matching packages," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 2010, pp. 783–794.

[3] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," vol. 5, no. 11. VLDB Endowment, 2012, pp. 1483–1494.

[4] Z. Chen, Q. Chen, and Z. Li, "A human-and-machine cooperative framework for entity resolution with quality guarantees (technical report)," Tech. Rep., 2016. [Online]. Available: http://www.wowbigdata.com.cn/HUMO/technical-report.pdf

[1] http://dbs.uni-leipzig.de/de/research/projects/object_matching/fever/benchmark_datasets_for_entity_resolution