Enabling Quality Control for Entity Resolution: A Human and Machine Cooperation Framework

Zhaoqiang Chen, Qun Chen, Fengfeng Fan, Yanyan Wang, Zhuo Wang, Youcef Nafa, Zhanhuai Li, Hailong Liu, Wei Pan

Northwestern Polytechnical University

April 19

Outline

- Background
- Motivation
- The HUMO Framework
- Optimization Approaches
- Experiments
- Conclusion

Background

Entity Resolution (**ER**): Identify the relational records that correspond to the same real-world entity.

Data source 1:

id	name	···	price
30134	Apple Mac Mini 1.83GHz Intel Core 2 Duo Computer - MB138LLA	···	$599

Data source 2:

id	name	···	price
206362873	Apple Mac mini Desktop - MB138LL/A	···	$574

A same product.

Background

Measurement on the ***Quality*** of an ER solution:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

True Label \ Predicted Label	*match*	*unmatch*
match	True Positive (TP)	False Negative (FN)
unmatch	False Positive (FP)	True Negative (TN)

Outline

- Background
- Motivation
- The HUMO Framework
- Optimization Approaches
- Experiments
- Conclusion

Motivation

- Pure machine-based ER solutions usually struggle in ensuring desired quality guarantees specified at both ***precision*** and ***recall*** fronts.

Precision ≥ The requirement ?
and
Recall ≥ The requirement ?

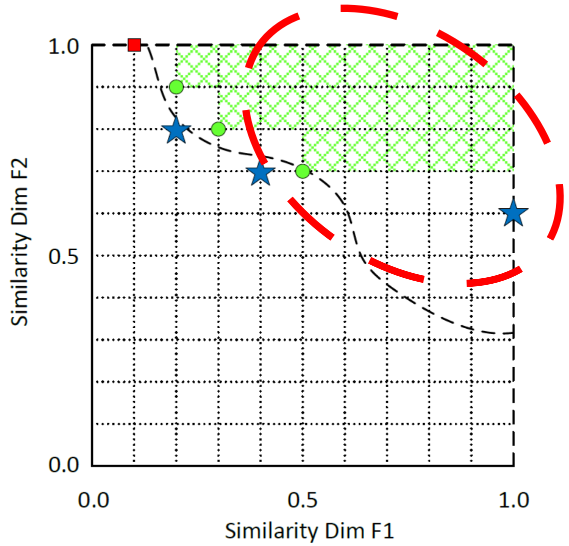
Motivation

ER Techniques	Quality Guarantees	
	Precision	Recall
Rules, Probabilistic Theory or Machine Learning based	✗	✗
Active-learning based [1][2]	✓	✗
HUMO	✓	✓

Difference: cannot enforce comprehensive quality guarantees specified by both precision and recall metrics as HUMO does.

[1] Learns record matching packages such that

$$Precision \geq Threshold$$



EnumerateBoundary[1]

[1] A. Arasu, M. Gotz, et al. On active learning of record matching packages. SIGMOD 2010.
[2] K. Bellare, S. Iyengar, et al. Active Sampling for entity matching. SIGKDD 2012.

Motivation

- Humans usually perform better than machines in terms of quality, but human labor is much more expensive.

- Therefore, HUMO has been designed with the purpose of minimizing human cost given a particular quality requirement.

Outline

- Background
- Motivation
- The HUMO Framework
- Optimization Approaches
- Experiments
- Conclusion

HUMO Framework

- Suppose that each instance pair can be evaluated by a machine metric.
  - *Pair similarity*
  - *Classification metrics, e.g., match probability and Support Vector Machine distance.*
- For simplicity of presentation, we use pair similarity as a machine metric example in this work. *However, HUMO is similarly effective with other machine metrics.*

HUMO Framework

Assumption [Monotonicity of Precision*]:

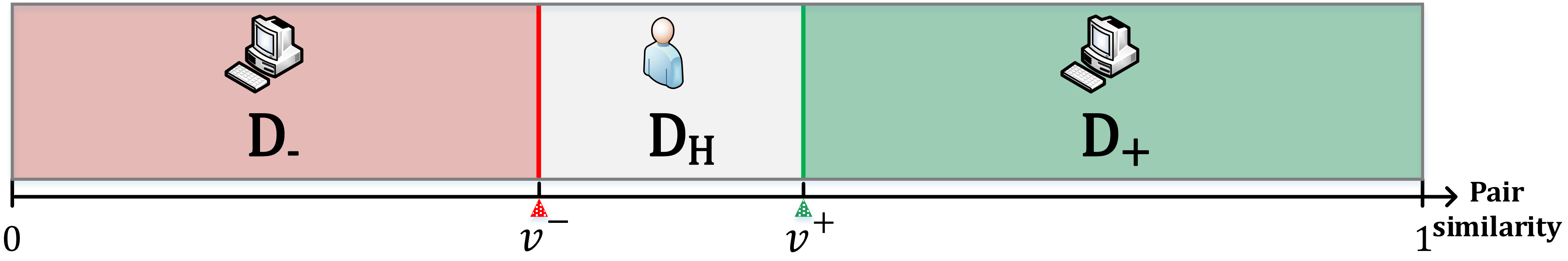
For any two value intervals $I_i \preccurlyeq I_j$ in [0, 1], we have $R(I_i) \leq R(I_j)$, in which $R(I_i)$ denotes the precision of the set of instance pairs whose metric values are located in I_i.

The higher (resp. lower) metric values a set of pairs have, the more probably they are matching pairs (resp. unmatching pairs).

* *It was first proposed by A. Arasu, M. Gotz, et al. On active learning of record matching packages. SIGMOD 2010.*

HUMO Framework

Automatically labeled with high accuracy + Challenging, Manual Verification + Automatically labeled with high accuracy ⇒ High Quality



D_- : labeled as unmatch D_H : manually labeled D_+ : labeled as match

Fig.1 The HUMO framework.

HUMO Framework

Given a HUMO solution S, the lower bound of its achieved precision and recall can be represented by,

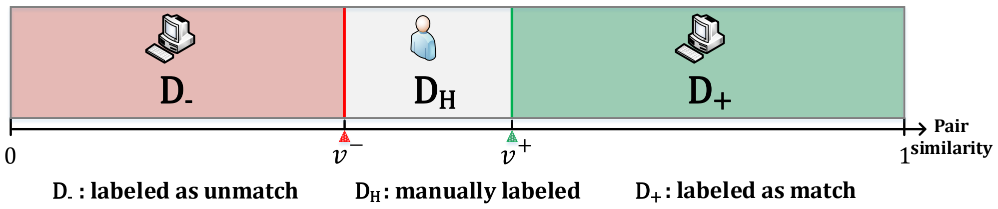


Fig.1 The HUMO framework.

\# of matches

$$Precision_l(S) = \frac{N_l^+(D_+) + N_l^+(D_H)}{N(D_+) + N(D_H)}$$

Lower bound

$$Recall_l(S) = \frac{N_l^+(D_+) + N_l^+(D_H)}{N_l^+(D_+) + N_l^+(D_H) + N_u^+(D_-)}$$

Upper bound

Note: In the case that human errors are introduced in D_H, we can adjust the estimated bounds accordingly.

In this paper, we assume that the pairs in D_H can be manually labeled accurately.

HUMO Framework

Optimization Problem:

A HUMO solution.

$$\underset{S_i}{argmin}(|D_H(S_i)|)$$

The number of manually inspected instance pairs.

Precision level.

$$subjet\ to\ P(precision(D,S_i) \geq \alpha) \geq \theta,$$

Confidence level.

$$P(recall(D,S_i) \geq \beta) \geq \theta.$$

Recall level.

HUMO Framework

The problem of searching for the minimum size D_H is challenging due to the fact that the ground-truth match proportions of D_- and D_+ are unknown.

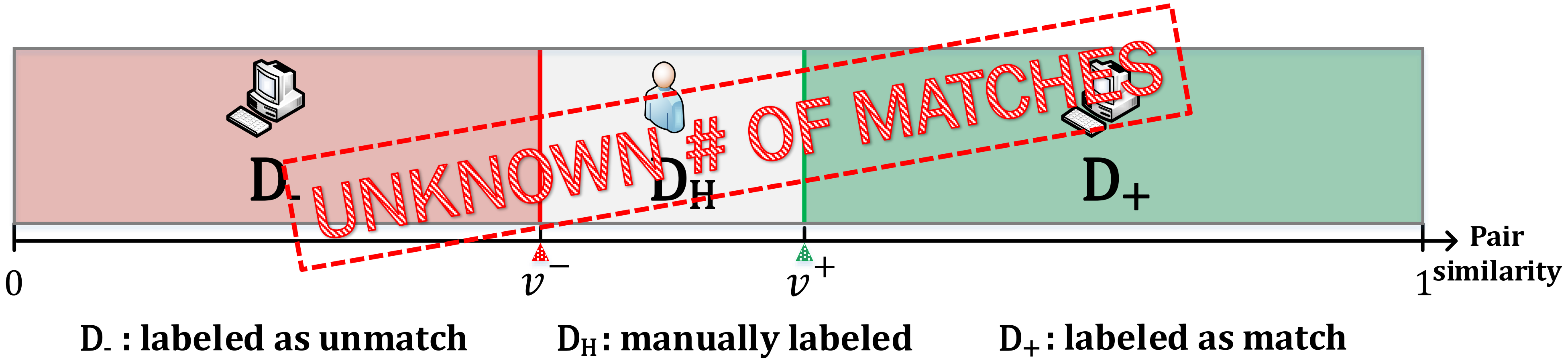
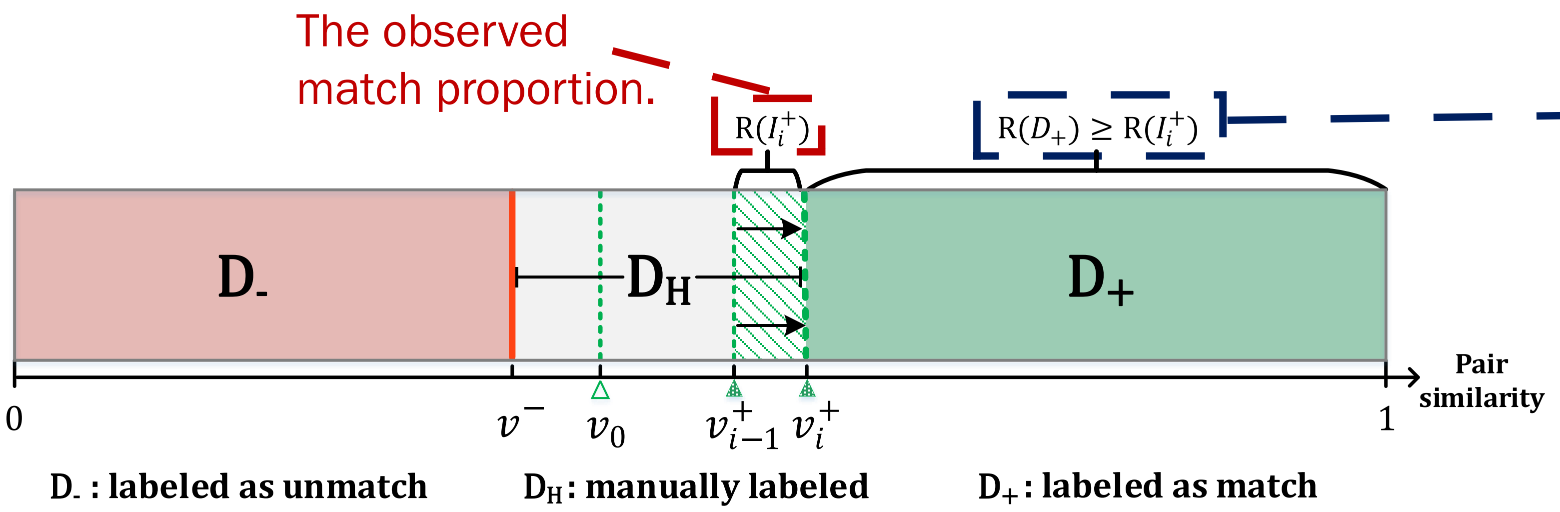


Fig.1 The HUMO framework.

Outline

- Background
- Motivation
- The HUMO Framework
- Optimization Approaches
  - *Baseline approach*
  - *Sampling-based approach*
  - *Hybrid approach*
- Experiments
- Conclusion

Baseline Approach



Monotonicity of Precision: the more similar two records are, the more likely they refer to the same real-world entity.

Fig.2 Incrementally moving the upper bound of D_H right.

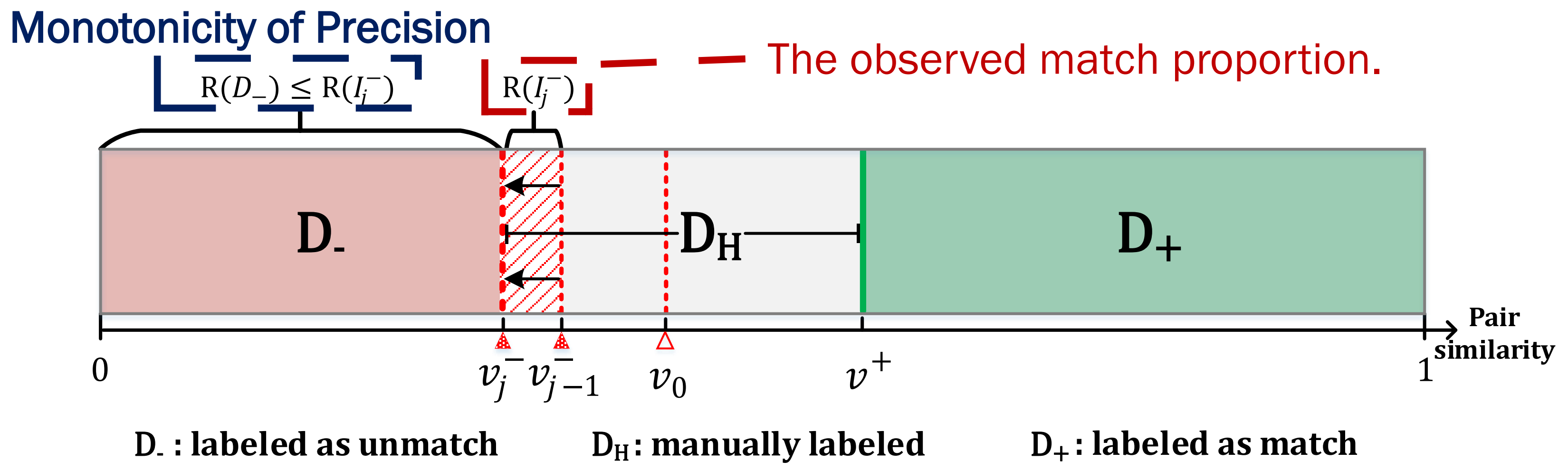


Fig.3 Incrementally moving the lower bound of D_H left.

Baseline Approach

The **precision requirement** $\boldsymbol{\alpha}$ and **recall requirement** $\boldsymbol{\beta}$ would be satisfied once:

$$R(I_i^+) \geq \frac{\alpha \cdot |D_+| - (1-\alpha) \cdot R(D_H) \cdot |D_H|}{|D_+|}$$

$$R(I_j^-) \leq \frac{(1-\beta) \cdot (|D_H| \cdot R(D_H) + |D_+| \cdot R(I_i^+))}{\beta \cdot |D_-|}$$

However:

- It may ***underestimate*** the match proportion of D_+.
- It may ***overestimate*** the match proportion of D_-.

Sampling-based Approach

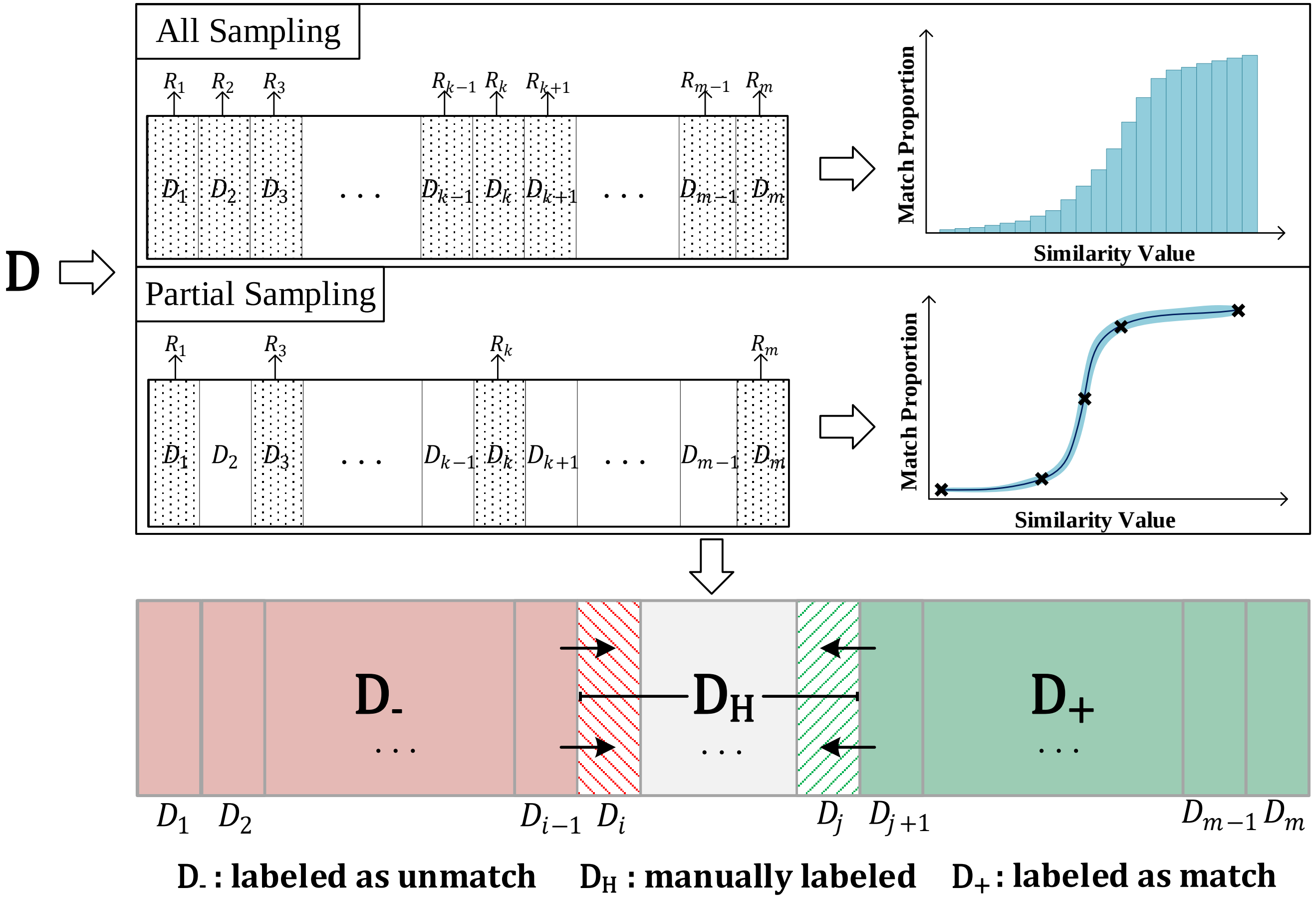


Fig.4 The demonstration of sampling-based solution.

Sampling-based Approach

All-Sampling Solution:

- Stratified Random Sampling.
- Sample every subset → human cost consumed on labeling samples is usually prohibitive.

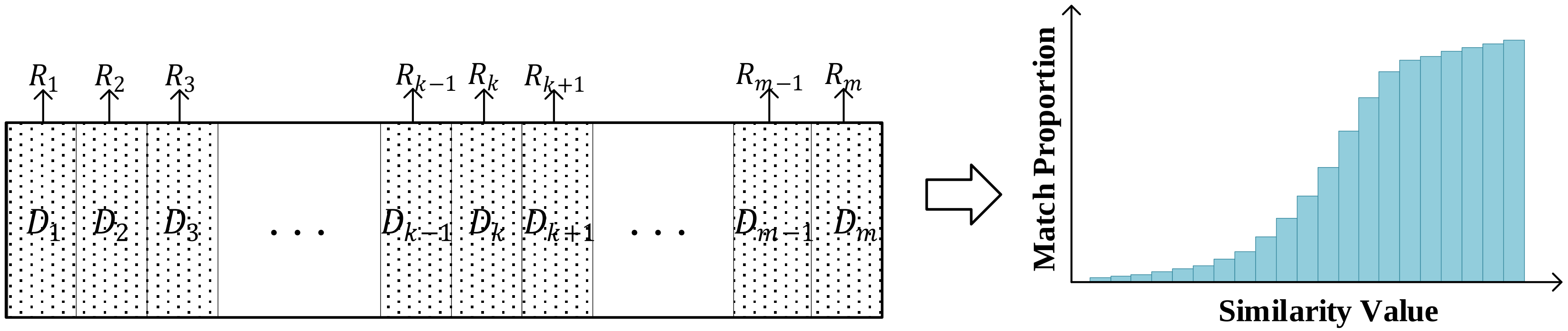


Fig.5 All-sampling solution.

Sampling-based Approach

Partial-Sampling Solution:

- Gaussian Process Regression.
- The match proportions of subsets have a joint Gaussian distribution.

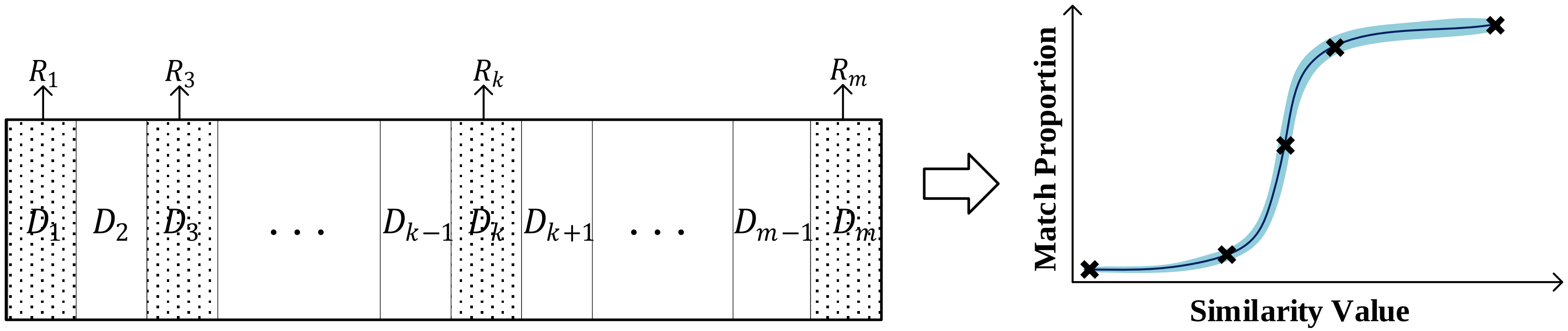


Fig.6 Partial-sampling solution.

Sampling-based Approach

Given the confidence level θ and the ***recall level*** β, the HUMO solution meets the recall requirement if:

Lower bound of True Positives.

$$\beta \le \frac{\boxed{lb(n^{+}_{[i,m]}, \sqrt{\theta})}}{\underbrace{\boxed{ub\left(n^{+}_{[1,i-1]}, \sqrt{\theta}\right)} + \boxed{lb(n^{+}_{[i,m]}, \sqrt{\theta})}}}$$

Upper bound of False Negatives.

Lower bound of True Positives.

Lower bound of the estimated recall.

Sampling-based Approach

Given the confidence level θ and the ***precision level*** α, the HUMO solution meets the precision requirement if:

$$\alpha \le \underbrace{\frac{\boxed{lb\left(n^{+}_{[i,j]}, \sqrt{\theta}\right) + lb\left(n^{+}_{[j+1,m]}, \sqrt{\theta}\right)}}{\boxed{lb\left(n^{+}_{[i,j]}, \sqrt{\theta}\right) + n_{[j+1,m]}}}}_{\text{Lower bound of the estimated precision.}}$$

Lower bound of True Positives.

Lower bound of True Positives + Upper bound of False Positives.

Hybrid Approach

- The baseline approach

  -- *overestimates the match proportion of* D_-;

  -- *underestimates the match proportion of* D_+.

- The sampling-based approach

  -- *has to consider confidence margins in the estimations of* D_- *and* D_+.

  -- *has large error margins when sample size is small.*

Hybrid Approach

✓ Takes advantage of both estimations and uses the better of both worlds in the process of bound computation.

- *Begins with an initial solution of the partial-sampling approach, S_0, and its lower and upper bounds of D_H;*
- *Incrementally redefines D_H's bounds using the better between the baseline and sampling-based estimates.*

Outline

- Background
- Motivation
- The HUMO Framework
- Optimization Approaches
- Experiments
- Conclusion

Experiments

- Datasets: DBLP–Scholar[1] (abbr. DS); Abt–Buy[2] (abbr. AB); Synthetic Datasets.

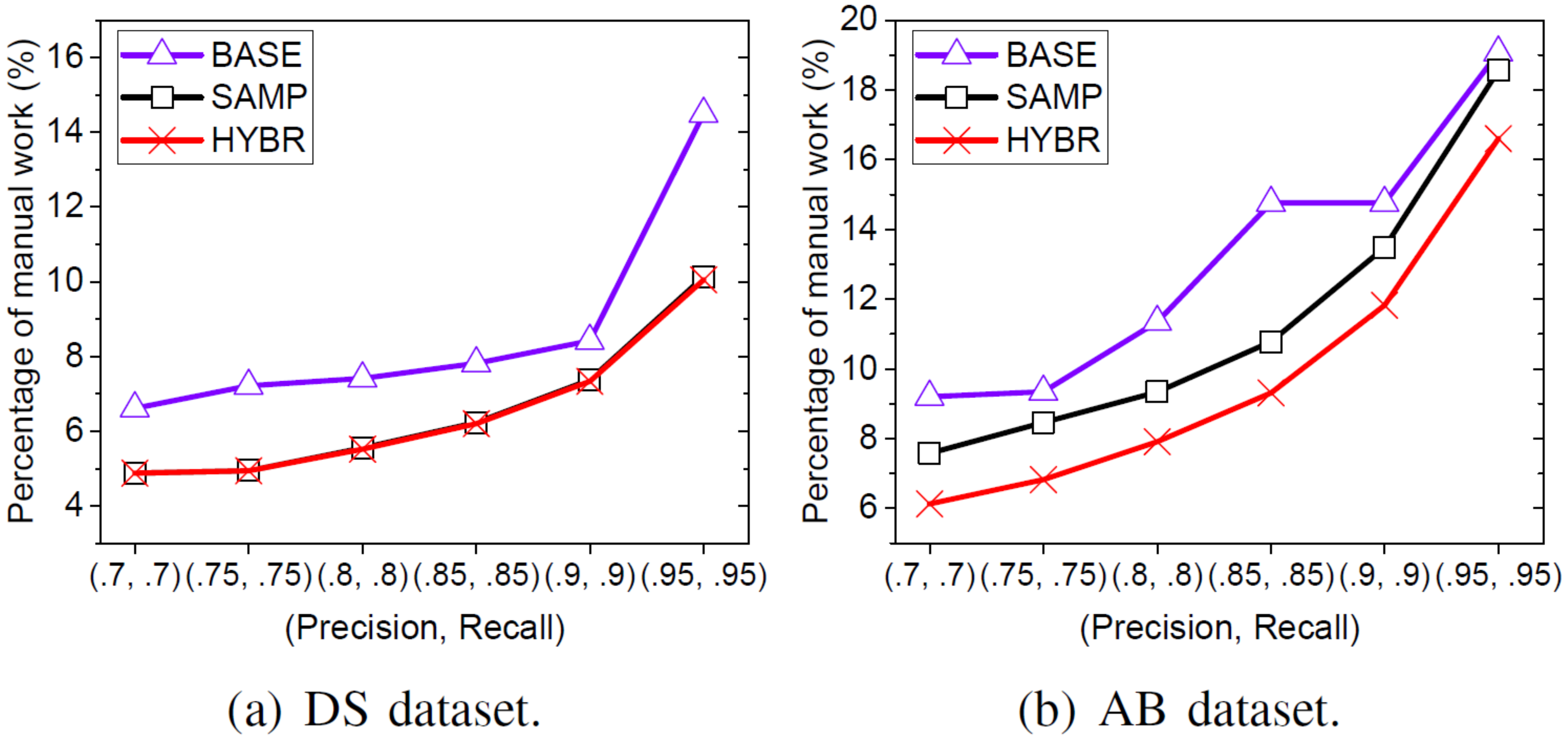


Fig.7 Comparison of human cost on two real datasets (with confidence set to 0.9).

[1] https://dbs.uni-leipzig.de/file/DBLP-Scholar.zip [2] https://dbs.uni-leipzig.de/file/Abt-Buy.zip

Baseline approach requires lesser manual work than Sampling-based one.

Hybrid approach can effectively use the better of both BASE and SAMP estimates.

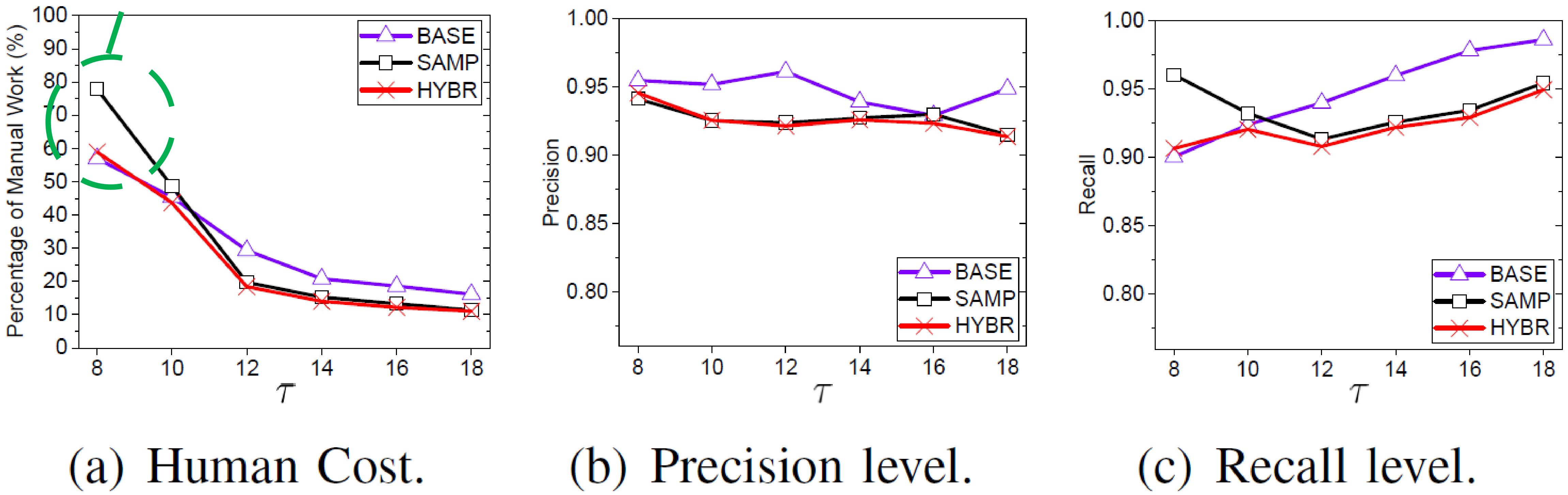


Fig.8 Varying τ (steepness) of the logistic curve on the synthetic datasets.

Note: The smaller the value of τ is, the more challenging the generated ER workload would be.

Active learning-based approaches [1], [2] have been proposed in order to satisfy the precision requirement for ER.

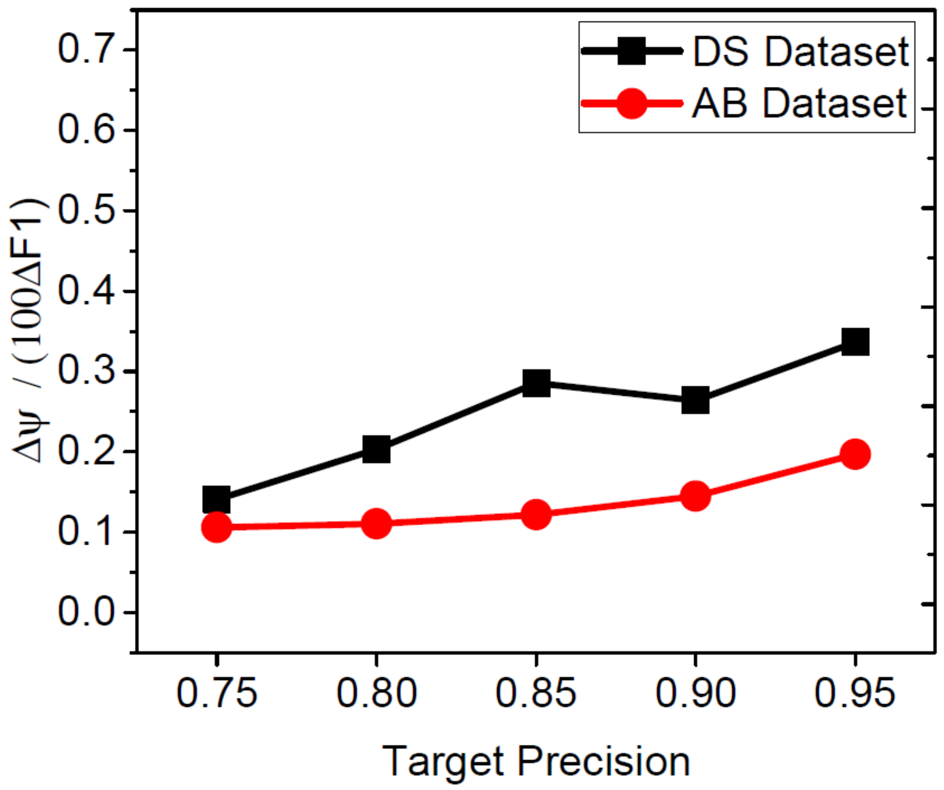


Fig.9 The percentage of manual work incurred by HUMO for 1% absolute improvement in F1 score over $ACTL$[1].

HUMO can effectively improve the resolution quality with reasonable return on investment in terms of human cost.

[1] A. Arasu, M. Gotz, et al. On active learning of record matching packages. SIGMOD 2010.
[2] K. Bellare, S. Iyengar, et al. Active Sampling for entity matching. SIGKDD 2012.

Outline

- Background
- Motivation
- The HUMO Framework
- Optimization Approaches
- Experiments
- Conclusion

Conclusion

- A human and machine cooperation framework for ER.
- It enables a flexible mechanism for comprehensive quality control at both precision and recall levels.
- Three optimization approaches to minimize human cost given a quality requirement.

Future Work

- Integrate HUMO into existing crowdsourcing platforms.

- As a general paradigm, HUMO can be potentially applied to other challenging classification tasks requiring high quality guarantees (e.g., financial fraud detection and malware detection).

Thank you !

Q & A