Entity Resolution with Weighted Constraints

Zeyu Shen and Qing Wang

Research School of Computer Science
Australian National University
Australia
qing.wang@anu.edu.au

Entity Resolution

- *Entity resolution* (ER) is to determine whether or not different entity representations (e.g., records) correspond to the same real-world entity.

Entity Resolution

- *Entity resolution* (ER) is to determine whether or not different entity representations (e.g., records) correspond to the same real-world entity.
- Consider the following relation AUTHORS:

ID	Name	Department	University
i_1	Peter Lee	Department of Philosophy	University of Otago
i_2	Peter Norrish	Science Centre	University of Otago
i_3	Peter Lee	School of Philosophy	Massey University
i_4	Peter Lee	Science Centre	University of Otago

- Questions:
  - Are Peter Lee (i_1) and Peter Lee (i_3) the same person?
  - Are Peter Norrish (i_2) and Peter Lee (i_4) not the same person?
  - ...

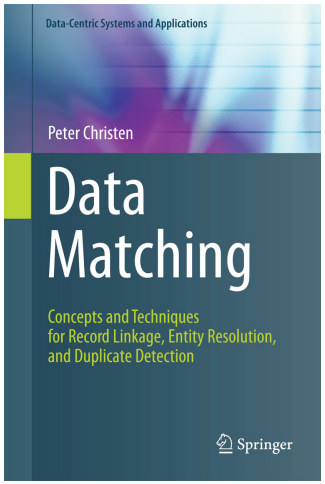
State of The Art

- State-of-the-art approaches to entity resolution favor similarity-based methods.

State of The Art

- State-of-the-art approaches to entity resolution favor similarity-based methods.
- Numerous techniques under a variety of perspectives:

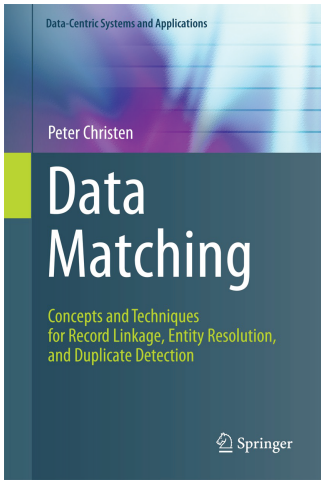
  a. threshold-based
  b. cost-based
  c. rule-based
  d. supervised
  e. active learning
  f. clustering-based
  g. . . .



State of The Art

- State-of-the-art approaches to entity resolution favor similarity-based methods.
- Numerous techniques under a variety of perspectives:

  a. threshold-based
  b. cost-based
  c. rule-based
  d. supervised
  e. active learning
  f. clustering-based
  g. . . .



- The central idea is

> "The more similar two entity representations are, the more likely they refer to the same real-world entity."

Goal of this Paper

- To study entity resolution in the presence of constraints, i.e., ER constraints.

Goal of this Paper

- To study entity resolution in the presence of constraints, i.e., ER constraints.
  - ER constraints ubiquitously exist in real-life applications.

  (1) " ICDM" refers to "IEEE International Conference on Data Mining" and vice versa (Instance level).

  (2) Two paper records refer to different papers if they do not have the same page numbers (Schema level).

Goal of this Paper

- To study entity resolution in the presence of constraints, i.e., ER constraints.
  - ER constraints ubiquitously exist in real-life applications.

  (1) " ICDM" refers to "IEEE International Conference on Data Mining" and vice versa (Instance level).

  (2) Two paper records refer to different papers if they do not have the same page numbers (Schema level).

  - They allow us to leverage rich domain semantics for improved ER quality.
  - Such constraints can be obtained from a variety of sources:

    a. background knowledge,
    b. external data sources,
    c. domain experts,
    d. ...

Research Questions

- We study *two questions* on ER constraints:

  (1) How to effectively specify ER constraints?

  (2) How to efficiently use ER constraints?

Research Questions

- We study *two questions* on ER constraints:

  (1) How to effectively specify ER constraints?

  (2) How to efficiently use ER constraints?

- Our task is

  > incorporate semantic capabilities (in form of ER constraints) into existing ER algorithms to improve the quality, while still being computationally efficient.

Research Questions

- We study *two questions* on ER constraints:

  (1) How to effectively specify ER constraints?

  (2) How to efficiently use ER constraints?

- Our task is

  > incorporate semantic capabilities (in form of ER constraints) into existing ER algorithms to improve the quality, while still being computationally efficient.

- A key ingredient is to associate each constraint with a weight that indicates the confidence on the robustness of semantic knowledge it represents.

Not all constraints are equally important.

An Example

A database schema
$\text{PAPER} := \{pid, authors, title, journal, volume, pages, tech, booktitle, year\}$
$\text{AUTHOR} := \{aid, pid, name, order\}$
$\text{VENUE} := \{vid, pid, name\}$

Views	
$\text{TITLE} := \pi_{pid,title}\text{PAPER}$	$\text{HASVENUE} := \pi_{pid,vid}\text{VENUE}$
$\text{PAGES} := \pi_{pid,pages}\text{PAPER}$	$\text{VNAME} := \pi_{vid,name}\text{VENUE}$
$\text{PUBLISH} := \pi_{aid,pid,order}\text{AUTHOR}$	$\text{ANAME} := \pi_{aid,name}\text{AUTHOR}$

	Constraints	Weights
r_1 :	$\text{PAPER}^*(x,y) \leftarrow \text{TITLE}(x,t), \text{TITLE}(y,t'), t \approx_{0.8} t'$	0.88
r_2 :	$\text{PAPER}^*(x,y) \leftarrow \text{TITLE}(x,t), \text{TITLE}(y,t'), t \approx_{0.6} t', \text{SAMEAUTHORS}(x,y)$	0.85
r_3 :	$\text{PAPER}^*(x,y) \leftarrow \text{TITLE}(x,t), \text{TITLE}(y,t'), t \approx_{0.7} t', \text{HASVENUE}(x,z),$ $\text{HASVENUE}(y,z'), \text{VENUE}^*(z,z')$	0.95
r_4 :	$\neg\text{PAPER}^*(x,y) \leftarrow \text{PAGES}(x,z), \text{PAGES}(y,z'), \neg z \approx_{0.5} z'$	1.00
r_5 :	$\text{VENUE}^*(x,y) \leftarrow \text{HASVENUE}(z,x), \text{HASVENUE}(z',y), \text{PAPER}^*(z,z')$	0.75
r_6 :	$\text{VENUE}^*(x,y) \leftarrow \text{VNAME}(x,n_1), \text{VNAME}(y,n_2), n_1 \approx_{0.8} n_2$	0.70
r_7 :	$\neg\text{AUTHOR}^*(x,y) \leftarrow \text{PUBLISH}(x,z,o), \text{PUBLISH}(y,z',o'), \text{PAPER}^*(z,z'), o \neq o'$	0.90
r_8 :	$\text{AUTHOR}^*(x,y) \leftarrow \text{COAUTHORML}(x,y), \neg\text{CANNOT}(x,y)$	0.80

Learning Constraints

- Two-step process:
  - Specify ground rules to capture the semantic relationships, which may have different interpretations for similarity atoms in different applications.

$$g_1: \quad \text{PAPER}^*(x, y) \leftarrow \text{TITLE}(x, t), \text{TITLE}(y, t'), t \approx_\lambda t'$$

Learning Constraints

- Two-step process:
  - Specify ground rules to capture the semantic relationship, which may have different interpretations for similarity atoms in different applications.

$$g_1: \quad \text{PAPER}^*(x, y) \leftarrow \text{TITLE}(x, t), \text{TITLE}(y, t'), t \approx_\lambda t'$$

  - Refine ground rules into the "best" ones for specific applications by learning.

(1). $\text{PAPER}^*(x, y) \leftarrow \text{TITLE}(x, t), \text{TITLE}(y, t'), t \approx_{0.8} t'$;

(2). $\text{PAPER}^*(x, y) \leftarrow \text{TITLE}(x, t), \text{TITLE}(y, t'), t \approx_{0.7} t'$;

(3). ...

Learning Constraints

- Positive and negative rules have different metrics α and β:

	Positive rules	Negative rules
α	$\frac{tp}{tp+fp}$	$\frac{tn}{tn+fn}$
β	$\frac{tp}{tp+fn}$	$\frac{tn}{fp+tn}$

Learning Constraints

- Positive and negative rules have different metrics α and β:

	Positive rules	Negative rules
α	$\frac{tp}{tp+fp}$	$\frac{tn}{tn+fn}$
β	$\frac{tp}{tp+fn}$	$\frac{tn}{fp+tn}$

- Objective functions must be *deterministic* and *monotonic*.
  - Soft rules

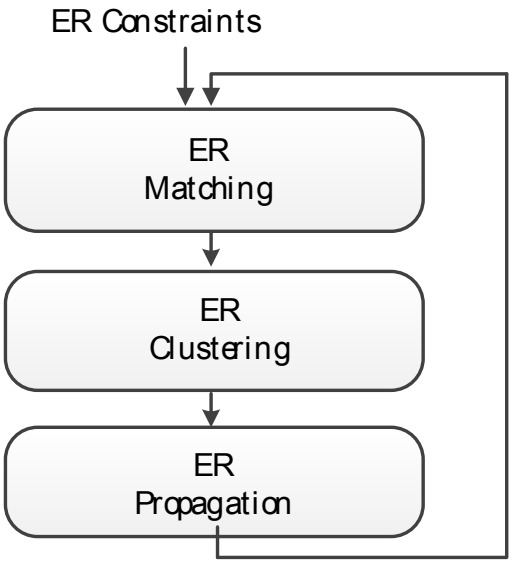
$$max_\lambda \xi(\alpha, \beta)$$
$$\text{subject to } \alpha \geq \alpha_{min} \text{ and } \beta \geq \beta_{min}.$$

  - Hard rules

$$max_\lambda \xi(\alpha, \beta)$$
$$\text{subject to } w \geq 1 - \varepsilon.$$

Using Constraints

- ER matching: To obtain ER graphs $G = (V, E, \ell)$, each having a set of *matches* $(u, v)^{=}$ or *non-matches* $(u, v)^{\neq}$, together with a *weight* $\ell(u, v)$ for each match and non-match.
- ER clustering: Given an ER graph $G = (V, E, \ell)$, to find *a valid clustering* over G such that vertices are grouped into one cluster iff their records represent the same real-world entity.



Using Constraints - ER Matching

- Soft rules with one hard rule: $\ell(u, v) = \top$ or $\ell(u, v) = \bot$

rule	match/non-match	weight
r_1	$(u, v)^{=}$	$\omega(r_1) = 0.88$
r_3	$(u, v)^{=}$	$\omega(r_3) = 0.95$
r_4	$(u, v)^{\neq}$	$\omega(r_4) = 1$

$\dashrightarrow$

$\ell(u, v) = \bot$

(a hard edge between u and v)

Using Constraints - ER Matching

- Soft rules with one hard rule: $\ell(u, v) = \top$ or $\ell(u, v) = \bot$

rule	match/non-match	weight
r_1	$(u, v)^{=}$	$\omega(r_1) = 0.88$
r_3	$(u, v)^{=}$	$\omega(r_3) = 0.95$
r_4	$(u, v)^{\neq}$	$\omega(r_4) = 1$

--→ $\ell(u, v) = \bot$

(a hard edge between u and v)

- Only soft rules: $\ell(u, v) \in [0, 1]$

rule	match/non-match	weight
r_1	$(u, v)^{=}$	$\omega(r_1) = 0.88$
r_3	$(u, v)^{=}$	$\omega(r_3) = 0.95$
r_9	$(u, v)^{\neq}$	$\omega(r_4) = 0.70$

--→ $\ell(u, v) = 0.215$

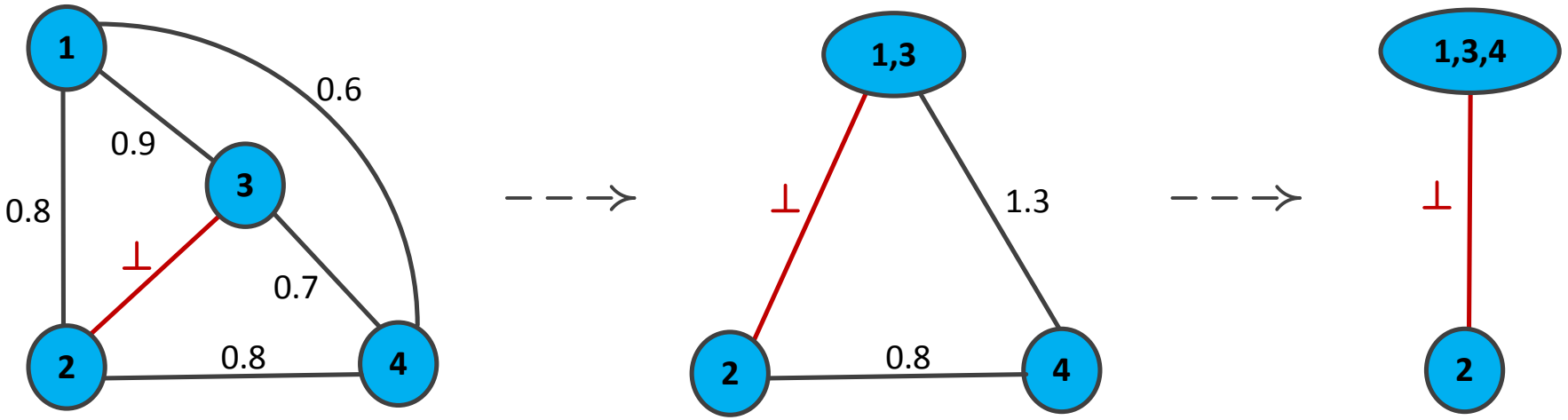
(a soft edge between u and v)

Using Constraints – ER Clustering

- A natural view is to use correlation clustering techniques.
  - Clustering objectives are often defined as minimizing disagreements or maximizing agreements.
  - However, it is known that correlation clustering is a NP-hard problem.
- Two approaches we will explore:
  - *Pairwise nearest neighbour* (PNN)
  - *Relative constrained neighbour*(RCN)

Pairwise Nearest Neighbour

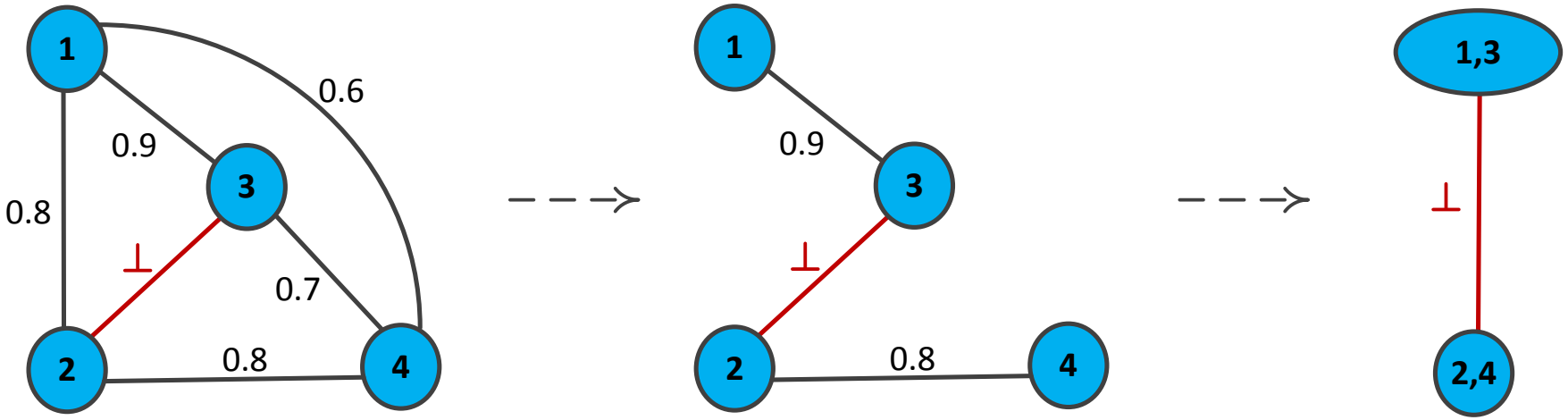
- Iteratively, a pair of two clusters that have the strongest positive evidence is merged, until the total weight of edges within clusters is maximized.



- Negative soft edges are "hardened" into negative hard edges under certain conditions.

Relative Constrained Neighbour

- Iteratively, a cluster that contains hard edges $\perp$ is split into two clusters based on the weights of relative constrained neighbours.



- Negative soft edges are "hardened" into negative hard edges under certain conditions.

Experimental Study

- We focused on three aspects:

  - ER models:

    How effectively can constraints and their weights be learned from domain knowledge for an ER model?

  - ER clustering:

    How useful can weighted constraints be for improving the ER quality?

  - ER scalability:

    How scalable can our method be over large data sets?

Experiments - ER Models

- We chose $\xi(\alpha, \beta) = (2 * \alpha * \beta)/(\alpha + \beta)$ for both data sets. The ER model over the Cora data set has 10 ground rules $g_1 - g_{10}$ for three entity types.

g_1: PAPER$^*(x, y) \leftarrow$ TITLE(x, t), TITLE$(y, t'), t \approx_{\lambda_1} t'$				
No	λ_1	Precision	Recall	F1-measure
1	0.8	0.879	0.815	0.846
2	0.7	0.818	0.926	0.869
3	0.6	0.725	0.985	0.835

g_2: PAPER$^*(x, y) \leftarrow$ TITLE(x, t), TITLE$(y, t'), t \approx_{\lambda_1} t'$, AUTHORS$(x, z)$, AUTHORS$(y, z'), z \approx_{\lambda_2} z'$, YEAR$(x, u)$, YEAR$(y, u'), u \approx_{\lambda_3} u'$						
No	λ_1	λ_2	λ_3	Precision	Recall	F1-measure
1	0.5	0.5	0.5	0.990	0.640	0.778
2	0.4	0.4	0.4	0.991	0.672	0.801
3	0.3	0.3	0.3	0.978	0.677	0.800

Experiments - ER Clustering

- We compared the quality of ER using three different methods:
  - Dedupalog[1],
  - ER-PNN,
  - ER-RCN,

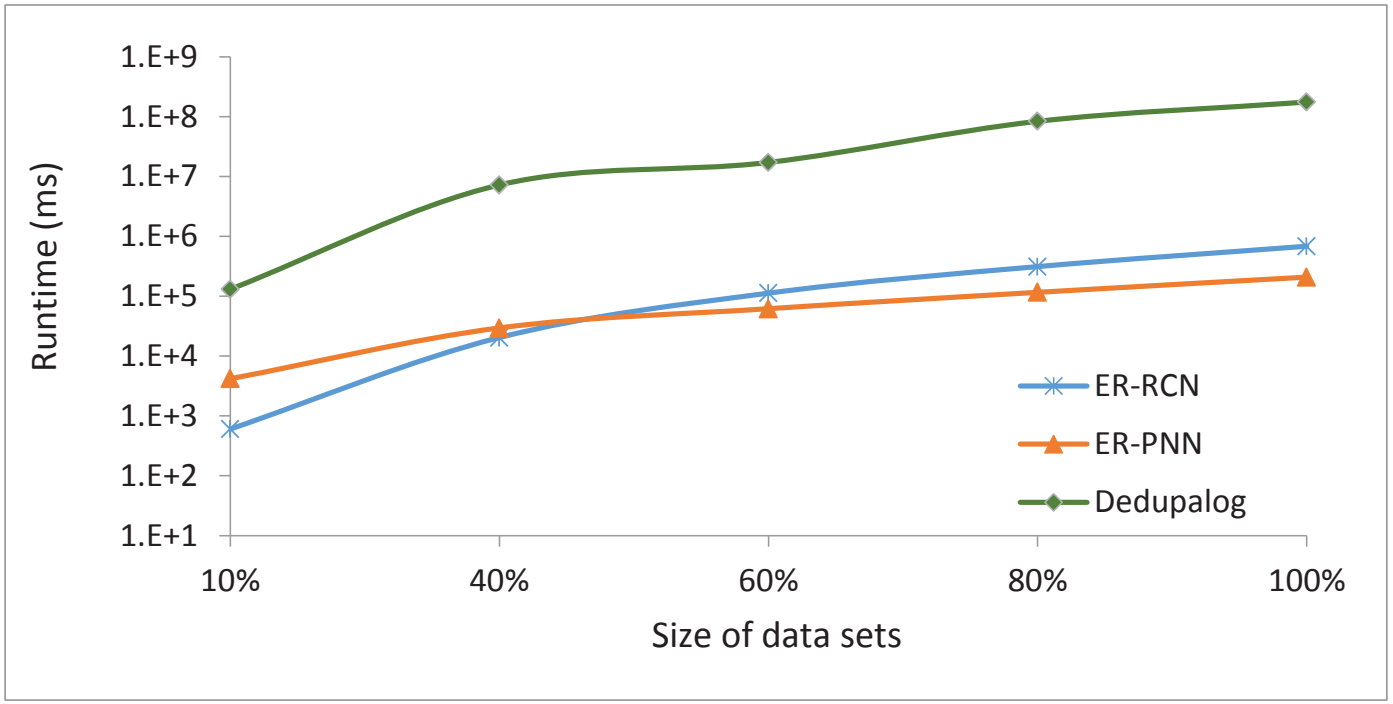
  where ER-PNN and ER-RCN only differ in the clustering algorithms.

Methods	Cora			Scopus		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure
Only positive rules	0.7324	0.9923	0.8428	0.9265	0.9195	0.9230
Dedupalog	0.7921	0.9845	0.8779	0.9266	0.9196	0.9231
ER-RCN	0.9752	0.9685	0.9719	0.9271	0.9192	0.9231
ER-PNN	0.9749	0.9660	0.9705	0.9271	0.9193	0.9232

[1] A. Arasu, C. Re, and D. Suciu. Large-scale deduplication with constraints using dedupalog. In ICDE, pages 952-963, 2009.

Experiments - ER scalability

- We conducted the scalability tests over Scopus (contains 47,333 author records).



Conclusions and Future Work

- We studied the questions of how to properly specify and how to efficiently use weighted constraints for performing ER tasks.
  - using a learning mechanism to "guide" the learning of constraints and their weights from domain knowledge
  - adding weights into constraints to leverage domain knowledge for resolving conflicts
- We plan to study knowledge reasoning for ER in the context of probabilistic modeling.
  - extending the current blocking techniques
  - identifying fragments of first-order logic for representing and reasoning ER constraints