A Graph-Theoretic Fusion Framework for Unsupervised Entity Resolution

Presented by: Dongxiang Zhang

NUS National University of Singapore

未来媒体研究中心 CENTER FOR FUTURE MEDIA

电子科技大学 University of Electronic Science and Technology of China

Entity Resolution

Text Records	Identical Entity
Les Celebrites 160 Central Park S New York French	√
Les Celebrites 155 W. 58th St. New York City French (Classic)	
Palm 837 Second Ave. New York City Steakhouses	×
Palm Too 840 Second Ave. New York City Steakhouses	

Two examples from the restaurant dataset.

Previous Work

- **Distance-based Methods**
  - Edit Distance, TF-IDF
  - Simple and scalable, but not effective enough
- **Learning-based Methods**
  - Learn a distance metric
  - Model ER as a classification task and apply SVM
  - Require considerable amount of training data
- **Crowd-based Methods**
  - CrowdER, TransM, TransNode, GCER, ADC, Power+
  - Achieve state-of-the-art accuracy but require human intervention

Our Objective

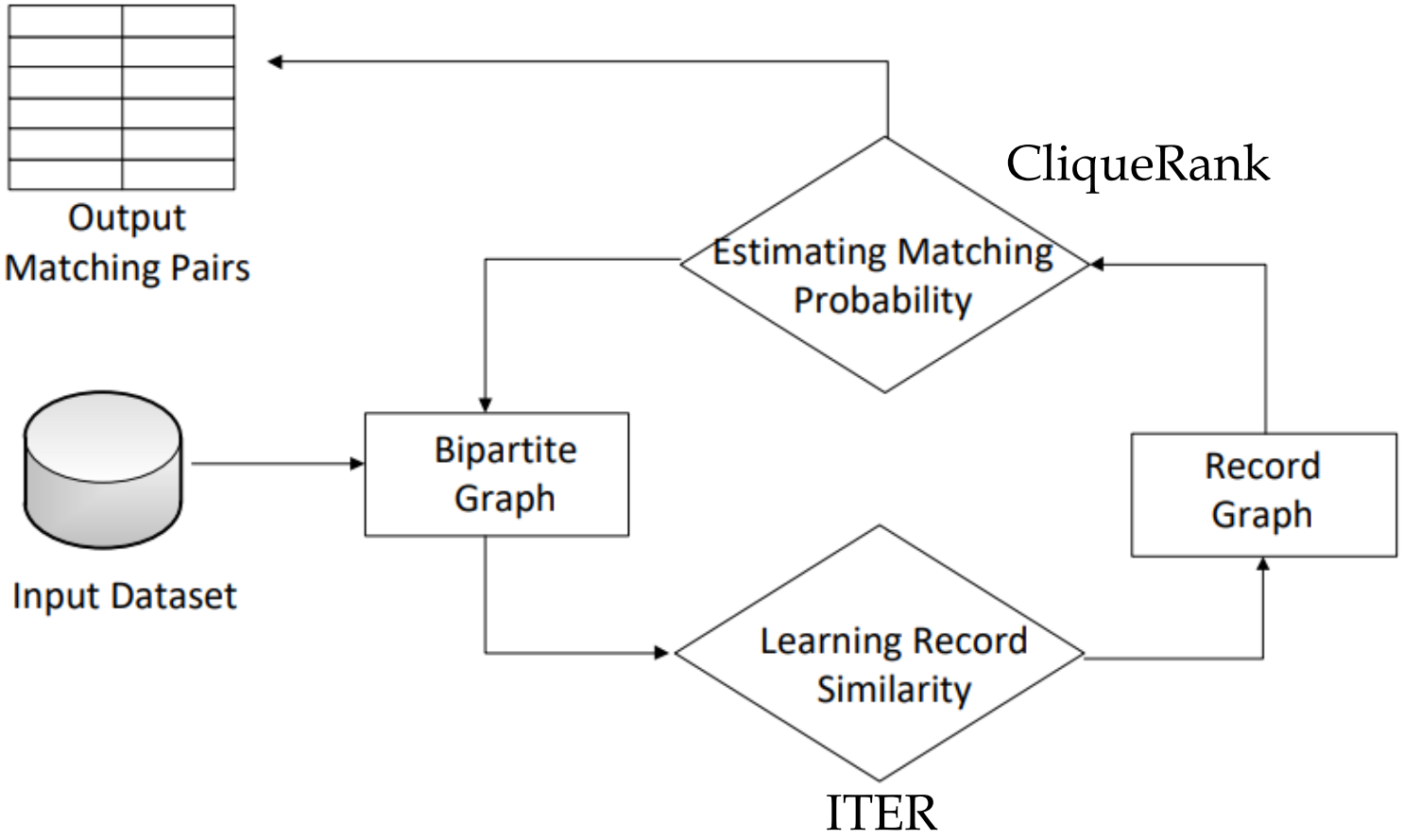
- **Propose an unsupervised approach**
  - More accurate when compared with distance-based methods
  - Require no training/labeling efforts when compared with learning-based methods
  - Require no human intervention and financial cost when compared with crowd-based methods

The General Idea

- **In the traditional unsupervised methods**
  - Step 1: Craft a distance measure between two records
  - Step 2: Tune a threshold such that two records with similarity score higher than the threshold are considered as the same entity

- **We are motivated to improve these two steps by**
  - Proposing ITER algorithm to learn record similarity
  - Proposing CliqueRank to estimate the likelihood of two records referring to the same entity
  - Iteratively Reinforcing these two components

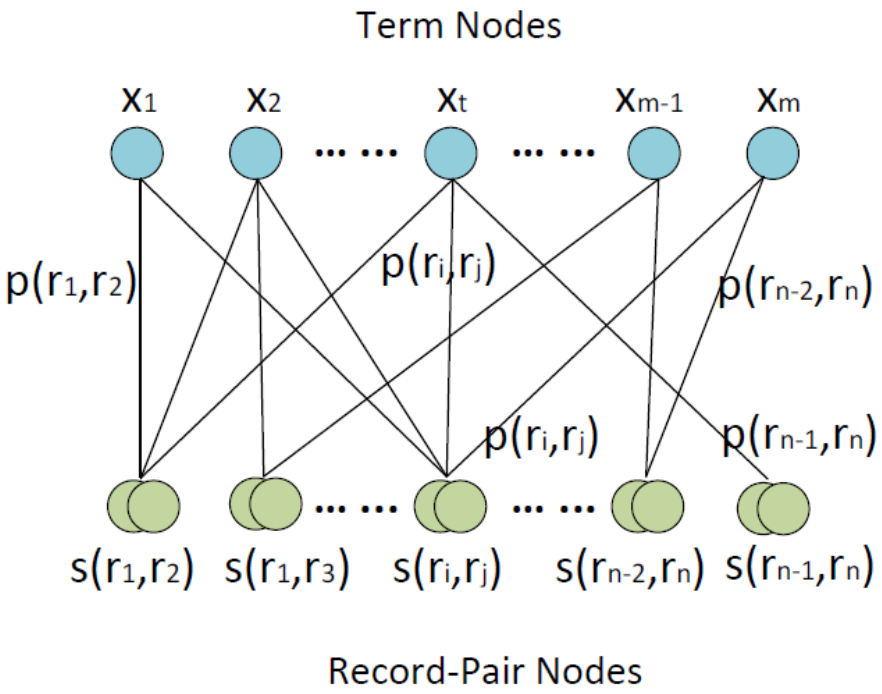
Unsupervised Fusion Framework



ITER Algorithm

- **If a term only occurs in a group of matching records, then we consider the term as highly discriminative**
  - Examples include product models for electronic devices or telephone numbers for restaurant.
  - These terms have low term frequency and may not be emphasized by TF-IDF
- **If a term is shared by many non-matching records, its weight will be punished**

ITER Algorithm



$$x_t = \sum_{i \neq j} \frac{p(r_i, r_j) s(r_i, r_j)}{P_t}$$

$$s(r_i, r_j) = \sum_{t \in r_i \wedge t \in r_j} norm(x_t)$$

ITER Algorithm

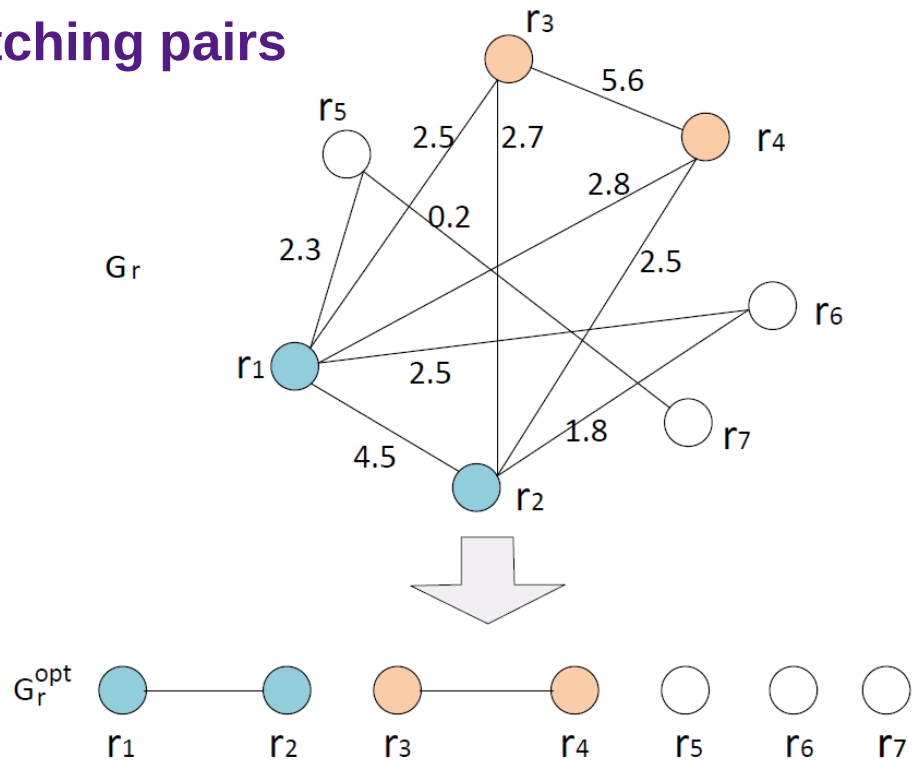
Algorithm 1: ITER Algorithm

Input: Bipartite graph structure with edge weight $p(r_i, r_j)$;
Output: Node salience x_t and record pair similarity score $s(r_i, r_j)$;

1. Randomly initialize x_t in $(0, 1)$;
2. **while** x_t does not converge **do**
3. **for** each record pair (r_i, r_j) **do**
4. Set its weight $s(r_i, r_j) \leftarrow \sum_{t \in r_i \wedge t \in r_j} x_t$;
5. **for** each term t **do**
6. Set its weight $x_t \leftarrow \sum_{i \neq j} \frac{p(r_i, r_j) s(r_i, r_j)}{P_t}$;
7. Set $x_t = 1/(1 + \frac{1}{x_t})$
8. **return** x_t and $s(r_i, r_j)$

CliqueRank Algorithm

- Given Gr, our goal is to identify matching probability.
- Ideally, the probability should be 1 for matching pairs and 0 for non-matching pairs



CliqueRank Algorithm

- **Random-Walk based interpretation**
  - **Ideally, if r_i and r_j refer to different entities, they should be located in different cliques and not reachable from each other**
  - **Otherwise, if we start a random walk from one record r_i, it will be very likely to visit the other record r_j within certain number of steps**

Random-Surfer Sampling

Algorithm 2: Random-Surfer Sampling (RSS)

1. Construct a record graph G_r based on $s(r_i, r_j)$;
2. **for** each edge $(r_i, r_j) \in G_r$ **do**
3. $\quad c_1 \leftarrow 0; \quad c_2 \leftarrow 0$
4. $\quad$ **for** $m \leftarrow 1; m \leq M/2; m++$ **do**
5. $\quad\quad c_1 \leftarrow c_1 + RandomWalk(r_i, r_j)$;
6. $\quad$ **for** $m \leftarrow 1; m \leq M/2; m++$ **do**
7. $\quad\quad c_2 \leftarrow c_2 + RandomWalk(r_j, r_i)$;
8. $\quad p(r_i, r_j) \leftarrow (c_1 + c_2)/M$;
9. **return** $p(r_i, r_j)$;

Random Walk Algorithm

Algorithm 3: RandomWalk($start$, $target$)

1. $cur \leftarrow start$;
2. **for** $s \leftarrow 1; s \leq S; s++$ **do**
3. pick a random value $b \in (0, 1)$;
4. $s'(cur, target) \leftarrow (1 + b) \cdot s(cur, target)$;
5. $next \leftarrow$ pick a node from neighbors of cur based on the new transition probability $p_b(r_i \rightarrow r_j)$;
6. **if** $next == target$ **then**
7. **return** 1;
8. **if** edge $(next, target) \notin G_r$ **then**
9. **return** 0;
10. $cur \leftarrow next$;
11. **return** 0;

To handle large cliques

$$p(r_i \rightarrow r_j) = \frac{s(r_i, r_j)^\alpha}{\sum_{r_j \in O(r_i)} s(r_i, r_j)^\alpha},$$

To champion edge with high score

For early termination

CliqueRank Algorithm

- Iterative sampling is slow, and we switch to matrix operation

- M_t be the matrix with reaching probability from r_i to r_j with 1 step

$$M_t[i,j] = p(r_i \rightarrow r_j)$$

- M_t^S be the matrix with reaching probability from r_i to r_j with S steps

$$M_t^S = \underbrace{M_t \times M_t \times \ldots \times M_t}_{S-1 \text{ times of operation } \times}$$

- The random surfer algorithm essentially estimates such probability

CliqueRank Algorithm

- We make customizations to the RSS algorithm
- $M_b[i,j]$ be the initial transition probability matrix

$$p_b(r_i \rightarrow r_j) = \frac{(1+b)^{\alpha} s(r_i,r_j)^{\alpha}}{norm(r_i,r_j)}.$$

- $M_n[i,j]$ is set to 1 if r_i to r_j are connected in Gr
- Finally, we can define the reaching probability with S steps

$$M_t^S = \begin{cases} M_b & \text{if } S = 1 \\ M_t \times (M_t^{S-1} \odot M_n) & \text{if } S > 1 \end{cases}$$

Benchmark Datasets

- **Restaurant**
  - 858 non-identical restaurant records.
  - Each record contains the information of restaurant name and address.
- **Product**
  - 1081 records from the abt website and the other 1092 records from the buy website.
  - Each product record contains its name and descriptive information.
- **Paper**
  - 1865 non-identical publication records.
  - Each record has a cluster id and its textual information consists of authors, title, publication venue and year.

Experimental Setup

- **For the three datasets, we use the same setting of parameters**
  - $\alpha=20$
  - $S=20$
  - $\eta=0.98$
  - 5 iterations between the reinforcement of ITER and CliqueRank
- **Eigen library is used to boost matrix multiplication**

  http://eigen.tuxfamily.org/index.php?title=Main Page

Experiment & Analysis

- **Accuracy**

	Restaurant	Product	Paper
String-distance based approaches			
Jaccard	0.836	0.332	0.792
TF-IDF	0.871	0.658	0.821
Machine-learning based approaches			
Gaussian Mixture Model [5]	0.704	-	-
HGM+Bootstrap [5]	0.844	-	-
MLE [5]	0.904	-	-
SVM [6]	0.922	-	0.824
Crowd-sourcing based approaches			
CrowdER [8]	0.934	0.800	0.824
TransM [10]	0.930	0.792	0.740
GCER [9]	0.930	0.760	0.785
ACD [12]	0.934	0.805	0.820
Power+ [13]	0.934	-	0.820
Graph-theoretic baselines			
SimRank	0.645	0.376	0.730
PageRank	0.905	0.564	0.316
Hybrid	0.946	0.593	0.748
Proposed graph-theoretic fusion framework			
ITER+CliqueRank	0.927	0.764	**0.890**

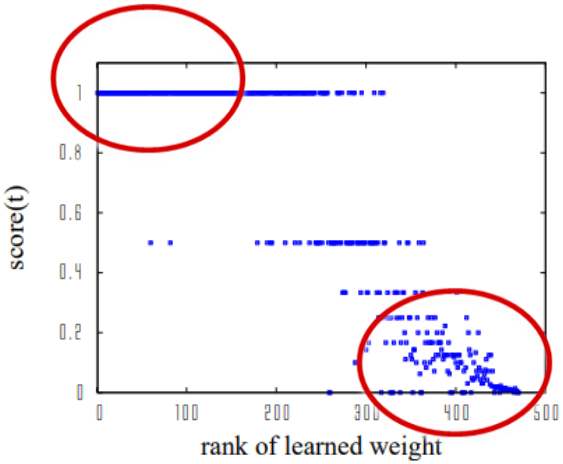
Experiment & Analysis

- **Efficiency**

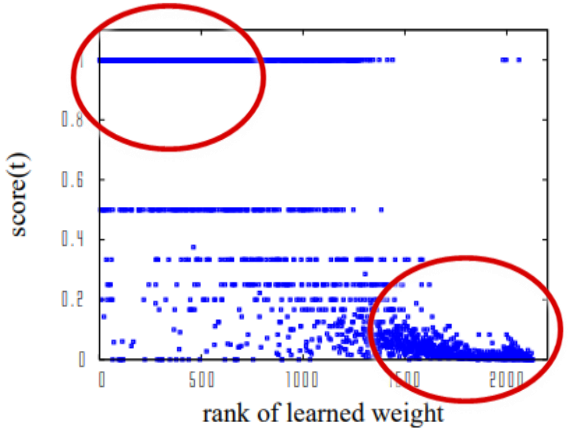
	Restaurant	Product	Paper
Number of nodes in G_r	858	2173	1865
Number of edges in G_r	5,320	151,939	980,780
Total running time	1.1min	21.6min	24.2min
Running time for ITER	3sec	20sec	58sec
Speedup compared to RSS	1.3x	1.5x	60x

Experiment & Analysis

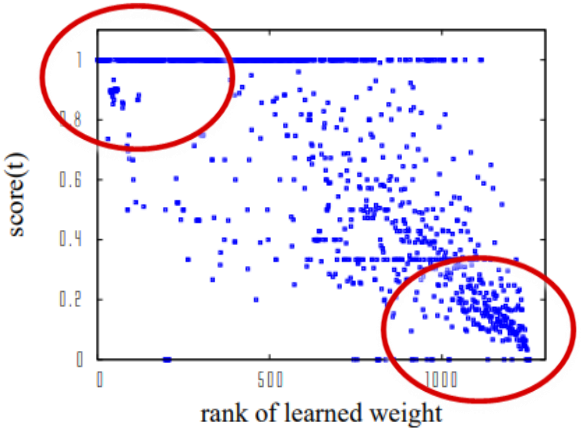
- **Effectiveness of Learned Term Weights**



(a) Restaurant



(b) Product



(c) Paper

ground-truth score: $score(t) = \frac{\sum_{t \in r_i \wedge t \in r_j} I(r_i, r_j)}{P_t}$

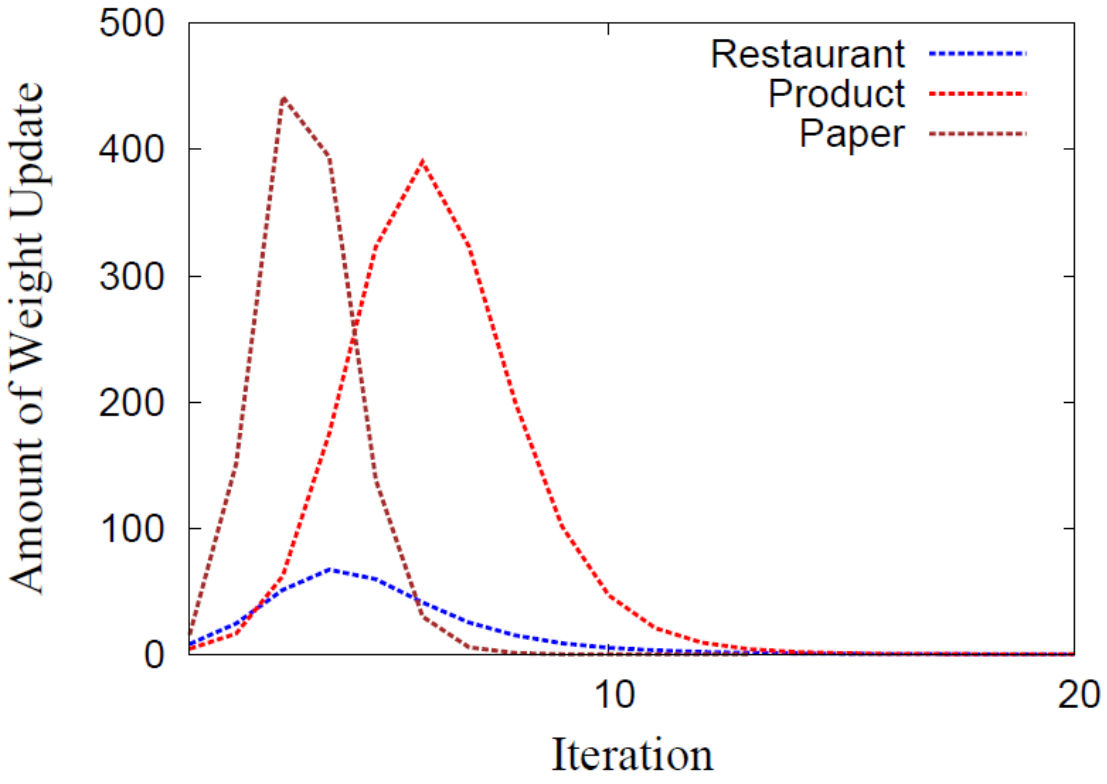
Experiment & Analysis

- **Top-Ranked Terms in the Benchmark Datasets**

Restaurant	coyote, chin's, 702/731-7547, 702/734-0410, 702/791-7111, 3645, 3400, gotham, 2880, arnie, seasons, 12, gramercy, chinois
Product	85w, trackpad, mirroring, magsafe, led-backlit, isight, dis-playport, 5400-rpm, spreadsheets, formulas, dramatically, compromising, multi-angle, 30p, 24mbps, diameter, s320, 7mm
Paper	thurn, wentzel, pachovicz, dzeroski, bloendorn, dze-roski, vafaic, kreusiger, kaufaman, pachowitz, re-ich, dzerowski, weldel, cmu-cd-91-197, jianping, jerzy, janusz, juergen, haleh, cmu-cs-91-197

Experiment & Analysis

- **Convergence of ITER**



Experiment & Analysis

- **Effect of Reinforcement**

	Restaurant		Product		Paper	
Iteration	F1-score	Time	F1-score	Time	F1-score	Time
1	0.916	13	0.543	253	0.844	207
2	0.935	25	0.712	514	0.888	515
3	0.931	39	0.747	768	0.889	819
4	0.931	52	0.754	1027	0.890	1135
5	0.927	64	0.764	1296	0.890	1453

Conclusion

- We propose an unsupervised graph-theoretic framework for entity resolution.
- Two novel algorithms ITER and CliqueRank are proposed, one for term-based similarity and the other for topological confidence. These two components can reinforce each other.
- Experimental results on three benchmark datasets show that our algorithm is accurate

Codes are available at:
https://github.com/uestc-db/Unsupervised-Entity-Resolution

Thank you!

Q&A