Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop

Yutao Zhang, Fanjin Zhang, Peiran Yao, Jie Tang

Department of Computer Science and Technology, Tsinghua University

Summary

- Name disambiguation is an important and challenging problem:
  - The 300 most common male names are used by over 115 million people (taking about 78.74%) in the US.
  - Challenges include: measuring similarity of documents, determining the number of persons with the same name, integrating data continuously…
- We propose a novel representation learning method by incorporating both global and local information, and present an end-to-end cluster size estimation method:
  - A global metric learning model for all documents.
  - A local linkage model within a candidate set.
  - An end-to-end RNN-based model to estimate the number of persons associated with the same name.
  - +7-35% in terms of F1-score compared with baselines.

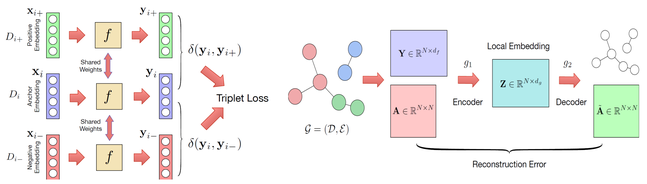
Global Metric Learning

First, embed each documents D_i as a IDF-weighted Word2Vec vectors x_i.
Idea: enforce positive pairs to be close in the embedding space and negative pairs to be far away.
Input: X (input document embedding matrix)
Output: Y (learned global document embedding matrix)

$$\mathbf{y_i} = f(\mathbf{x}_i)$$

Triplet Loss: $\mathcal{L}_f = \sum_{(D_i, D_{i+}, D_{i-}) \in \mathcal{T}} \max\{0, \delta(\mathbf{y}_i, \mathbf{y}_{i+}) - \delta(\mathbf{y}_i, \mathbf{y}_{i-}) + m\}.$

where $\mathcal{T}$ is the set of all triplets in training set, m is a margin enforced between positive pairs and negative pairs and $\delta(v_1, v_2) = ||v_1 - v_2||$ is the Euclidean distance.



global metric learning local linkage learning

Local Linkage Learning

Definition: (Local Linkage Graph)
For a given name reference a, we construct a local linkage graph $\mathcal{G}^a = (\mathcal{D}^a, \mathcal{E}^a)$, where $\mathcal{D}^a = \{D_i^a\}$ is the set of documents authored by a person named a, $\mathcal{E}^a = \{D_i^a, D_j^a\}$ is a set of edges capturing the similarity between the documents.

Graph Auto-Encoder
Input: A (the adjacency matrix of $\mathcal{G}^a$), Y (Output of global model)
Output: Z (learned latent local embedding matrix)
Encoder: Two-layer graph convolution network (GCN)

$$g_1(\mathrm{Y}, \mathrm{A}) = \grave{\mathrm{A}}\mathrm{ReLU}(\grave{\mathrm{A}}\mathrm{Y}\mathbf{W}_0)\mathbf{W}_1$$

Decoder: Inner product decoder

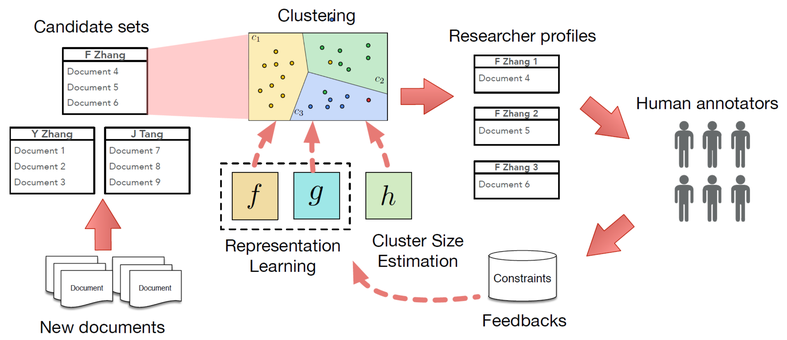
$$g_2(\mathrm{Z}) = \mathrm{sigmoid}(\mathbf{Z}^\top \mathbf{Z})$$

Loss Function: reconstruction adjacency matrix

$$p(\tilde{\mathrm{A}}_{ij} = 1 | \mathbf{z}_i, \mathbf{z}_j) = \mathrm{sigmoid}(\mathbf{z}_i^\top \mathbf{z}_j)$$

$$\mathcal{L}_g = -\sum_{D_i, D_j \in \mathcal{D}} \mathrm{A}_{ij} \log p(\tilde{\mathrm{A}}_{ij})$$

Framework Overview



Cluster Size Estimation

First, construct a pseudo-training set as right.

Then, adopt RNN as an encoder and try to map a set of embedding vectors to the true number of clusters in the set.

Optimize the Mean Squared Logarithmic Error (MSLE)

$$\mathcal{L}_h = \frac{1}{N}\sum_{t=1}^{a}[\log(1 + h(\mathcal{D}_t)) - \log(1 + K_t)]^2$$

ALGORITHM 1: Pseudo-training data generation strategy for cluster size estimation. $(\mathcal{D}_t, K_t)$ is a training example for RNN model $h(\mathcal{D}) \to \mathbb{R}$.

Input: Clean clusters C, $K_{\min}$, $K_{\max}$, sample size z, step t;
Output: Pseudo-training example $(\mathcal{D}_t, K_t)$;
$K_t \leftarrow$ Sample from $[K_{\min}, K_{\max}]$;
$\mathrm{C}_t \leftarrow$ Sample K_t clusters from C;
$\mathcal{D}_t \leftarrow$ Sample z documents from $\bigcup_{C_i \in \mathrm{C}_t}\{D \in C_i\}$ with replacement;
return $(\mathcal{D}_t, K_t)$;

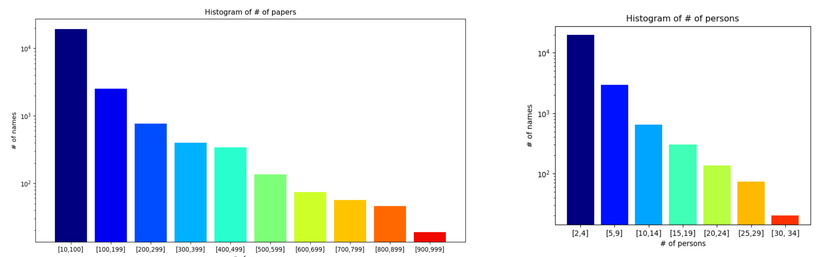
Datasets

- Extracted from AMiner

Statistics

Dataset	small (in paper)	large
# of names	100	23,823
# of persons	12,798	83,980
# of papers	70,258	1,203,482

Large Dataset Analysis



- Data Schema
  - Relation file: name → list of person id → list of paper id
  - Paper file: paper id, title, published year, abstract, author names, author organizations, keywords, venue…

URL: https://aminer.org/disambiguation/

Experimental Results

Table 1: Results of Author Name Disambiguation

	AMiner			Zhang et al.			GHOST			Louppe et al.			Rule		
Name	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Xu Xu	74.18	45.86	56.68	48.16	41.87	44.80	61.34	21.79	32.15	22.55	64.40	33.40	10.75	97.23	19.35
Rong Yu	89.13	46.51	61.12	65.48	40.85	50.32	92.00	36.41	52.17	38.85	91.43	54.53	30.81	97.79	46.86
Yong Tian	76.32	51.95	61.82	70.74	56.85	63.04	86.94	54.58	67.06	32.08	63.71	42.67	10.37	93.79	18.67
Lu Han	51.78	28.05	36.39	47.88	20.62	28.82	69.72	17.39	27.84	30.25	46.65	36.70	13.66	89.16	23.69
Lin Huang	77.10	32.87	46.09	71.84	34.17	46.31	86.15	17.25	28.74	24.86	71.32	36.87	13.86	99.46	24.33
Kexin Xu	91.37	98.64	94.87	90.02	82.47	86.08	92.90	28.52	43.64	91.26	98.35	94.67	91.45	99.60	95.35
Wei Quan	53.88	39.02	45.26	64.45	47.66	54.77	86.42	27.80	42.07	37.86	63.41	47.41	28.16	93.80	43.32
Tao Deng	81.63	43.62	56.86	53.04	29.89	38.23	73.33	24.50	36.73	40.46	51.38	45.27	16.30	95.16	27.84
Hongbin Li	77.20	69.21	72.99	54.66	53.05	53.84	56.29	29.12	38.39	19.48	85.96	31.77	13.25	96.41	23.30
Hua Bai	71.49	39.73	51.08	58.58	35.90	44.52	83.06	29.54	43.58	36.39	41.33	38.70	25.47	98.51	40.47
Meiling Chen	74.93	44.70	55.99	59.36	28.80	38.79	86.11	23.85	37.35	58.32	47.14	52.14	59.55	82.07	69.02
Yanqing Wang	71.52	75.33	73.37	60.40	51.97	55.87	80.79	40.39	53.86	29.64	79.08	43.11	25.72	62.47	36.44
Xudong Zhang	62.40	22.54	33.12	70.20	23.35	35.04	85.75	7.23	13.34	72.38	79.83	75.92	63.22	17.94	27.95
Qiang Shi	52.20	36.15	42.72	43.84	36.94	40.10	53.72	26.80	35.76	35.31	47.18	40.39	28.79	93.89	44.06
Min Zheng	57.65	22.35	32.21	54.76	19.70	28.98	80.50	15.21	25.58	25.86	32.67	28.87	15.41	98.72	26.66
Avg.	77.96	63.03	**67.79**	70.63	59.53	62.81	81.62	40.43	50.23	57.09	77.22	63.10	44.94	89.30	53.42

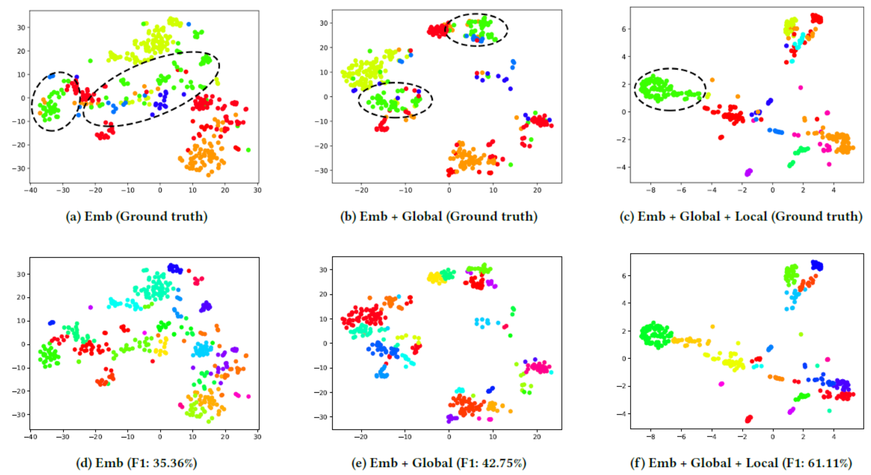
Table 2: Contribution of Each Component.

	Pre.	Rec.	F1
Embedding	66.85	42.04	49.79
Global	68.40	47.42	54.56
Local	68.97	67.68	66.55
Overall	77.96	63.03	67.79

Table 3: Results of Clustering Size Estimation.

	Actual	RNN	Regression	X-means
RMSLE	-	**0.2493**	1.6006	2.1065
Song Chen	125	101.39	173.80	10
Jian Du	87	62.89	110.21	5
Fosong Wang	4	5.71	184.75	5
J Yu	346	74.06	24.92	7
Yang Shen	157	153.77	89.52	7
Xiaobing Luo	13	11.01	143.44	3
Jian Feng	102	149.73	113.88	8
Lu Han	129	114.51	173.16	7

Embedding Analysis



(a) Emb (Ground truth) (b) Emb + Global (Ground truth) (c) Emb + Global + Local (Ground truth)

(d) Emb (F1: 35.36%) (e) Emb + Global (F1: 42.75%) (f) Emb + Global + Local (F1: 61.11%)

Figure 4: t-SNE Visualization of embedding spaces on a candidate set. Each color in (a), (b), (c) denotes an individual ground truth cluster, while each color in (d), (e), (f) denotes a predicted cluster by hierarchical agglomerative clustering. Emb indicates the original feature embedding. Global and Local represent the use of global metric learning and local linkage learning respectively. The dashed black ellipses in (a), (b), (c) circle the points of the same ground truth cluster.

Acknowledgement

The work is supported by the National High-tech R&D Program (2015AA124102), Development Program of China (2016QY01W0200), National Basic Research Program of China (2014CB340506), National Natural Science Foundation of China (61631013,61561130160), National Social Science Foundation of China (13&ZD190), a research fund supported by MSRA, and the Royal Society-Newton Advanced Fellowship Award.

Contact: Jie Tang, jietang@tsinghua.edu.cn **GitHub**: https://github.com/neozhangthe1/disambiguation/