A Unified Probabilistic Framework for Name Disambiguation in Digital Library

TKDE@2012 by Jie Tang et al.

http://www.aminer.cn/

lina
2018.03.16

CONTENTS

1 Introduction

2 Problem Formalization

3 Framework

4 Parameter Estimation

5 Experiments and Conclusion

CONTENTS

1 Introduction

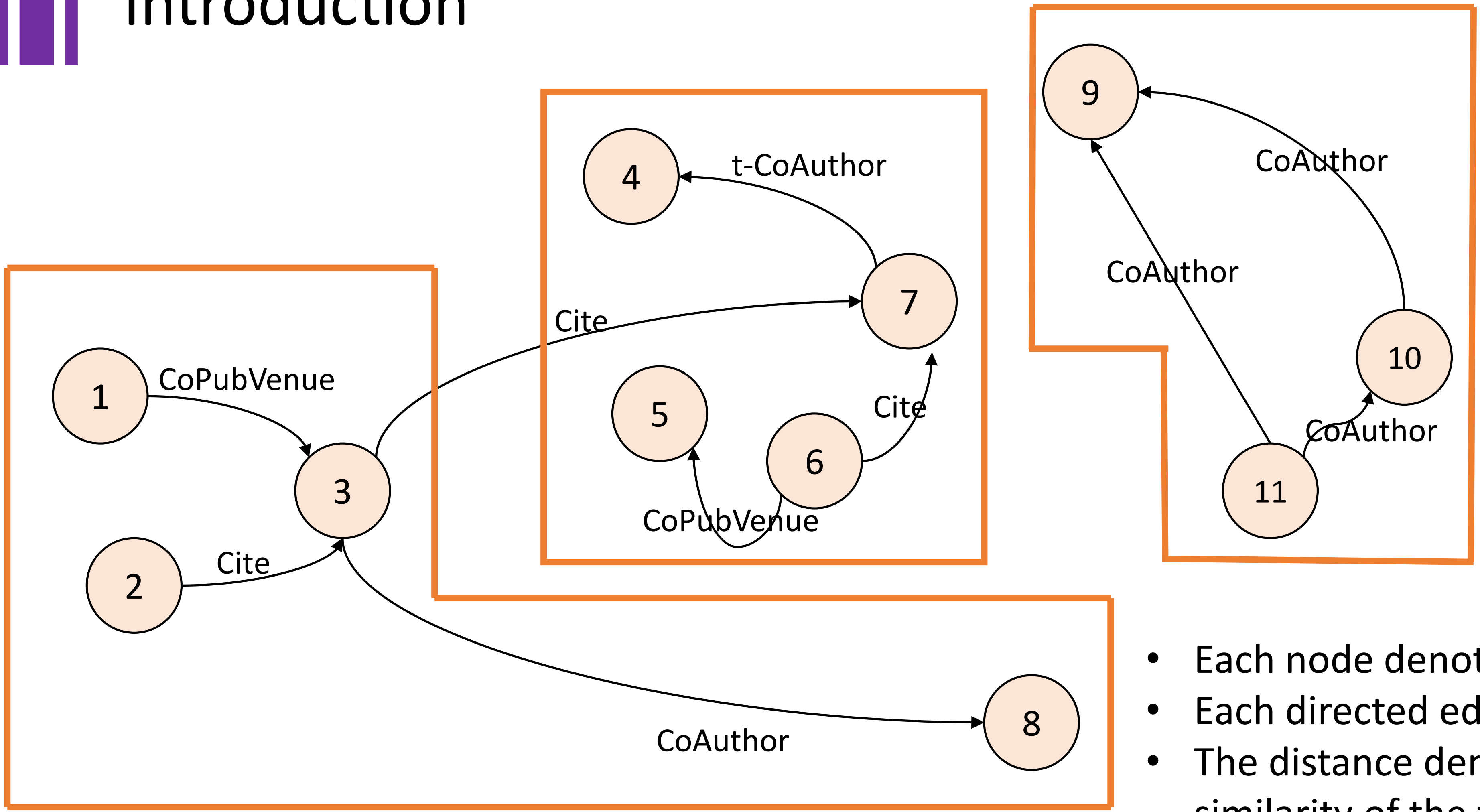
2 Problem Formalization

3 Framework

4 Parameter Estimation

5 Experiments and Conclusion

Introduction



- Each node denotes a paper.
- Each directed edge denotes a relationship.
- The distance denotes the content-based similarity of the two papers.

1 PART

Introduction

- Prior work:
  - Only consider topological structure of graph or node similarity.
  - Few methods can find the number K automatically.
- Solution:
  - Formalize the disambiguation using a Markov Random Fields(MRF).
  - Explore a dynamic approach for estimating the number of people K.
  - Present a two-step algorithm for parameter estimation.

CONTENTS

1 Introduction

2 Problem Formalization

3 Framework

4 Parameter Estimation

5 Experiments and Conclusion

Definitions

- Attributes of Each Publication p_i

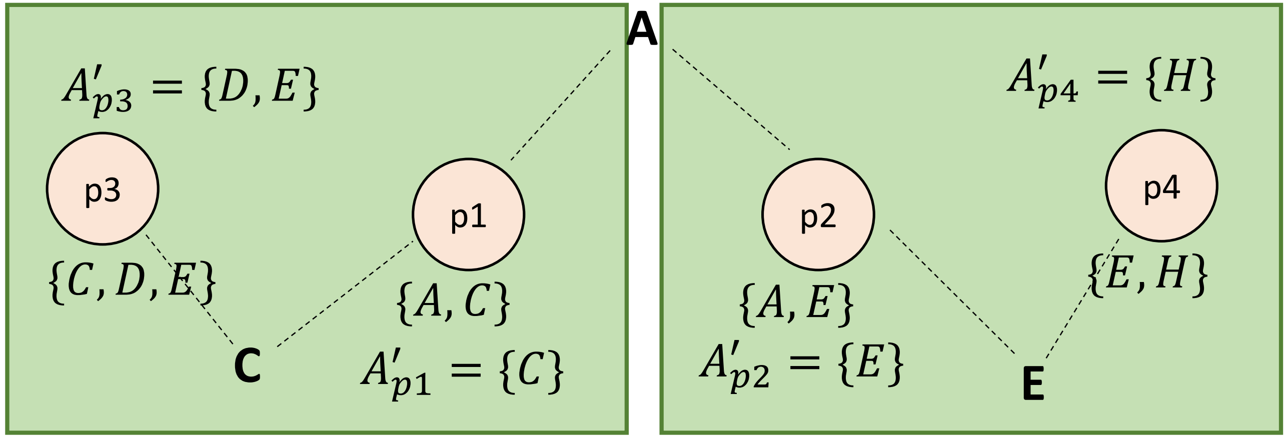
Attribute	Description
$p_i.title$	title of p_i
$p_i.pubvenue$	published conference/journal of p_i
$p_i.year$	published year of p_i
$p_i.abstract$	abstract of p_i
$p_i.authors$	author name set of $p_i\{a_i^{(0)}, a_i^{(1)}, \dots, a_i^{u}\}$
$p_i.references$	References of p_i

- Principle Author and Secondary Author

Each paper p_i has one or more authors $A_{pi} = \{a_i^{(0)}, a_i^{(1)}, \dots, a_i^{u}\}$, we describe the author that we are going to disambiguate as the principle author $a_i^{(0)}$ and the rest as the secondary authors denoted as A'_{pi}.

- Five types of undirected relationships between two papers.

R	W	Relation Name	Description
r_1	w_1	CoPubVenue	$p_i.pubvenue = p_j.pubvenue$
r_2	w_2	CoAuthor	$\exists r, s > 0, a_i^{(r)} = a_j^{(s)}$
r_3	w_3	Citation	p_i cites p_j or p_j cites p_i
r_4	w_4	Constrait	Feedback supplied by users
r_5	w_5	τ-CoAuthor	τ-extension co-authorship(τ>1)



$A^2_{p1} = \{C, D, E\}$ $\quad$ $A^2_{p2} = \{E, H\}$

$A^2_{p1} \cap A^2_{p2} = \{E\}$

2 PART

Name Disambiguation

Publication Informative Graph:

$$G = (P, R, V_p, W_R)$$

- $P = \{p_1, p_2, \dots, p_n\}$ denotes the publications containing the author name a.
- $r_k(p_i, p_j)$ is a relationship r_k between p_i and p_j,

  $r_k(p_i, p_j) = 1$ if there is a relationship r_k between p_i and p_j; otherwise, $r_k(p_i, p_j) = 0$
- Each $v(p_i) \in V_P$ corresponds to the feature vector of paper p_i
- $w_k \in W_R$ denotes the weight of relationship r_k

CONTENTS

1 Introduction

2 Problem Formalization

3 Framework

4 Parameter Estimation

5 Experiments and Conclusion

3.1 PART

Basic Idea

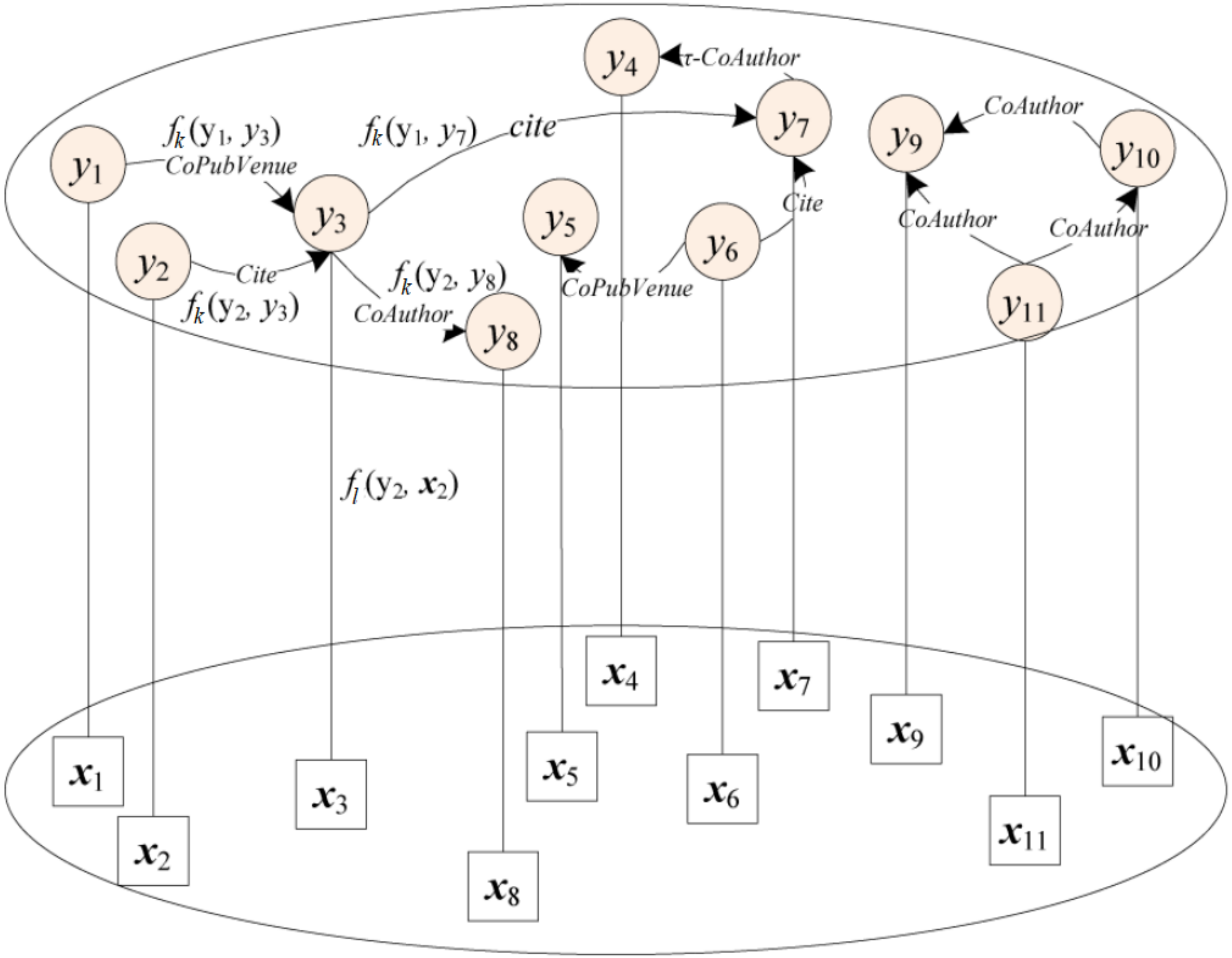
papers with similar content tend to have the same label

papers having strong relationship tend to have the same label

leveraging both content similarity and paper relationships

formalize both content-based information and structure-based information into a Hidden Markov Random Field(HMRF) model

Hidden Markov Random Fields



Two components:

- An observable set of random variables $X = \{x_i\}_{i=1}^{n}$
- A hidden field of random variables $Y = \{y_i\}_{i=1}^{n}$

$$P(Y) = \frac{1}{Z_1}\exp\left(\sum_{(y_i,y_j)\in E,k} \lambda_k f_k(y_i, y_j)\right)$$

$$Z_1 = \sum_{y_i,y_j} \sum_{(y_i,y_j)\in E,k} \lambda_k f_k(y_i, y_j)$$

$$P(X|Y) = \frac{1}{Z_2}\exp\left(\sum_{x_i\in X,l} \alpha_l f_l(y_i, x_i)\right)$$

$$Z_2 = \sum_{y_i} \sum_{x_i\in X,l} \alpha_l f_l(y_i, x_i)$$

3.3 PART

Disambiguation Objective Function

We define an objective function as the Maximum a Posteriori configuration of the HMRF.

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} \propto \mathrm{P(Y)P(X|Y)}$$

$$L_{max} = \log\big(P(Y|X)\big) \Rightarrow \log(P(Y)P(X|Y))$$

$$= \log\left(\frac{1}{Z_1 Z_2}\exp(\Sigma_{(y_i,y_j)\in E,k}\lambda_k f_k(y_i,y_j)+\Sigma_{x_i\in X,l}\alpha_l f_l(y_i,x_i))\right)$$

$$f_k(y_i, y_j) = K(x_i, x_j)\sum_{r_k\in R_{ij}}[w_k r_k(x_i, x_j)]$$

$$f_l(y_i, x_i) = K(y_i, x_i) = K(\mu_{(i)}, x_i)$$

$$\Longrightarrow L_{max} = \sum_{(y_i,y_j)\in E,k}\lambda_k K(x_i, x_j) r_k(y_i, y_j) + \sum_{x_i\in X,l}\alpha_l K(\mu_{(i)}, x_i) - logZ\ (\text{其中} Z = Z_1 Z_2)$$

CONTENTS

1 Introduction

2 Problem Formalization

3 Framework

4 Parameter Estimation

5 Experiments and Conclusion

4.1 PART

Algorithm

Algorithm 1. Parameter estimation

Input: $P=\{p_1, p_2, \ldots, p_n\}$

Output: model parameters Θ and $Y=\{y_1, y_2, \ldots, y_n\}$, where $y_i \in [1, K]$

1. Initialization

1.1 randomly initialize parameters Θ;

1.2 for each paper x_i, choose an initial value y_i, with $y_i \in [1, K]$;

1.3 calculate each paper cluster centroid $\mu_{(i)}$;

1.4 for each paper x_i and each relationship (x_i, x_j), calculate $f_l(y_i, x_i)$ and $f_k(y_i, y_j)$.

2. Assignment

2.1 assign each paper to its closest cluster centroid;

3. Update

3.1 update of each cluster centroid;

3.2 update of the weight for each feature function.

$$\Theta = \{\lambda_1, \lambda_2, \ldots; \alpha_1, \alpha_2, \ldots\}$$

2. Assignment

$$\log P(y_i|x_i) \propto L_{x_i}(\mu_{(h)}, x_i)$$
$$= \sum_{(x_i, x_j)\in E_i, R_i, k} \lambda_k K(x_i, x_j) r_k(y_i, y_j) + \sum_l \alpha_l K(x_i, \mu_{(i)}) - \log Z$$

$$K(x_i, x_j) = \frac{x_i^T x_j}{||x_i|| \cdot ||x_j||}, \text{where } ||x_i|| = \sqrt{x_i^T x_i}$$

4.1 PART Algorithm

2. Assignment

Maximizing the log-likelihood is equalient to minizing the KL divergence.

$$maxL = max\ \log\left(\prod_{y_i} p(y_i|x_i)\right)$$

$$= max\ \sum_{y_i} logp(y_i|x_i)$$

$$\Rightarrow maxE_{q(y_i)}logp(y_i|x_i)$$

$$=< logP(y_i|x_i) >_{q(y_i)}$$

$$KL(q||P) = \sum_{y_i} q(y_i|x_i)log\frac{q(y_i|x_i)}{p(y_i|x_i)}$$

$$= \sum_{y_i} q(y_i|x_i)logq(y_i|x_i) - \sum_{y_i} q(y_i|x_i)logP(y_i|x_i)$$

$$= -H(q) -< logP(y_i|x_i) >_{q(y_i)}$$

q is an approximation P

$$L^{KL} = KL(q^0||P) \approx KL(q^0|P) - KL(q^l|P)$$

$$=< logP(y_i|x_i) >_{q^0(y_i)} -< logP(y_i|x_i) >_{q^l(y_i)}$$ [1]

$$\Rightarrow KL(q^0||q^1)$$ [1]

So, we can simply consider one Gibbs sampling iteration to minimize the $KL(q^0||q^1)$.

[1] Hinton G E. Training products of experts by minimizing contrastive divergence.[M]. MIT Press, 2002.

Algorithm

3. Update

$$\mu_{(h)} = \frac{\sum_{i:y_i=h} x_i}{|| \sum_{i:y_i=h} x_i ||_A}$$

$$\frac{\partial L^{KL}}{\partial \lambda_k} = < \frac{\partial logP(y_i|x_i)}{\partial \lambda_k} >_{q^0(y_i)} - < \frac{\partial logq(y_i|x_i)}{\partial \lambda_k} >_{q^1(y_i)}$$
$$= -\sum_{(x_i,x_j)\in E_i} K(x_i,x_j) r_k(y_i,y_j) - < \frac{\partial logq(y_i|x_i)}{\partial \lambda_k} >_{q^1(y_i)}$$

$$\lambda_k^{new} = \lambda_k^{old} + \Delta \frac{\partial L}{\partial \lambda_k} (\Delta\ is\ learning\ rate.)$$

[1] Hinton G E. Training products of experts by minimizing contrastive divergence.[M]. MIT Press, 2002.

4.2 PART

Estimation of K

Algorithm 3. Estimation of K

Input: $P=\{p_1, p_2, \ldots, p_n\}$

Output: K, $Y=\{y_1, y_2, \ldots, y_n\}$, where $y_i \in [1,K]$

i=0, K=1, that is to view P as one cluster: $C^{(i)}=\{C_1\}$;
do{
  foreach cluster C in $C^{(i)}${
    find a best two sub-clusters model M_2 for C;
    if(BIC(M_2)>BIC(M_1))
      split cluster C into two sub clusters $C^{(i+1)}=\{C_1, C_2\}$;
    calculate BIC score for the obtained new model;
}while(existing split);
choose the model as output with the highest BIC score;

To seek the best balance between the model complexity and model's ability to describe the data set:

BIC measurement:

$$BIC = kln(n) - 2ln(L)$$

k: number of parameters.

n: number of samples.

L: likelihood function.

4.2 PART

Estimation of K

Algorithm 3. Estimation of K

Input: $P=\{p_1, p_2, \ldots, p_n\}$

Output: K, $Y=\{y_1, y_2, \ldots, y_n\}$, where $y_i \in [1,K]$

1: i=0, K=1, that is to view P as one cluster: $C^{(i)}=\{C_1\}$;

2: do{

3: foreach cluster C in $C^{(i)}${

4: find a best two sub-clusters model M_2 for C; ?

5: if(BIC(M_2)>BIC(M_1))

6: split cluster C into two sub clusters $C^{(i+1)}=\{C_1, C_2\}$;

7: calculate BIC score for the obtained new model;

8: }while(existing split);

9: choose the model as output with the highest BIC score;

$$BIC^v(M_h) = \log\big(P(M_h|P)\big) - \frac{|\lambda|}{2} \cdot \log(n)$$

$$|\lambda| = \sum_{i=1}^{K}\big(P(y_i) + \mu_{(i)}\big) + \sum_{\lambda\in\Theta} \lambda$$

M_h is the model corresponding to person number h.
$P(M_h|P)$ is the posterior probability if model M_h given the observations P.
$|\lambda|$ is the number of parameters in M_h.

Benefiting from the cluster atoms identification, this problem is alleviated in our framework.

CONTENTS

1 Introduction

2 Problem Formalization

3 Framework

4 Parameter Estimation

5 Experiments and Conclusion

Data Sets

Data Sets

Abbr. Name	#Public-ations	#Actual Person	Abbr. Name	#Public-ations	#Actual Person
Cheng Chang	12	3	Gang Wu	40	16
Wen Gao	286	4	Jing Zhang	54	25
Yi Li	42	21	Kuo Zhang	6	2
Jie Tang	21	2	Hui Fang	15	3
Bin Yu	66	12	Lei Wang	109	40
Rakesh Kumar	61	5	Michael Wagner	44	12
Bing Liu	130	11	Jim Smith	33	5
Ajay Gupta	27	4	Wei Wang	306	90
Dimitry Pavlov	16	2	David Jensen	43	3
Charles Smith	7	4	David Brown	53	7
David C. Wilson	52	5	George Miller	17	2
James H. Anderson	112	2	James Johnson	17	3
John Miller	74	2	Joseph Miller	10	2
Paul Jones	13	3	Richard Taylor	93	10
Robert Fisher	105	4	Robert Moore	92	3
Robert Williams	8	2	William Cohen	110	2

32 real author names and 2074 papers.

5.2 PART

Experimental Design

Measures:

$$PairwisePrecision = \frac{\#PairsCorrectlyPredictedToSameAuthor}{\#TotalPairsPredictedToSameAuthor}$$

$$PairwiseRecall = \frac{\#PairsCorrectlyPredictedToSameAuthor}{\#TotalPairsToSameAuthor}$$

$$PairwiseF_1 = \frac{2 \times PairwisePrecision \times PairwiseRecall}{PairwisePrecision + PairwiseRecall}、$$

Baselines:

K-means

SOM

X-means

HAC

SACluster

CONSTRAINT

5.3
PART

Experimental

Results of Name Disambiguation (Percent)

Person Name	K-means			HAC			SOM			SACluster			CONSTRAINT			Our Approach (Fixed K)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Cheng Chang	89.47	68.00	77.27	100.0	100.0	100.0	76.30	65.42	70.44	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Wen Gao	96.25	49.78	65.62	96.60	62.64	76.00	98.12	47.14	63.68	73.52	98.27	84.11	99.29	98.59	98.94	99.29	98.59	98.94
Yi Li	13.91	39.02	20.51	86.64	95.12	90.68	43.67	32.72	37.41	77.42	84.21	80.67	70.91	97.50	82.11	70.91	97.50	82.11
Jie Tang	95.38	72.09	82.12	100.0	100.0	100.0	84.92	70.65	77.13	90.14	82.04	85.90	100.0	100.0	100.0	100.0	100.0	100.0
Gang Wu	28.41	20.49	23.81	97.54	97.54	97.54	24.79	31.28	27.66	43.66	87.32	58.22	71.86	98.36	83.05	81.62	98.36	89.21
Jing Zhang	7.88	26.03	12.10	85.00	69.86	76.69	38.76	64.23	48.35	72.00	86.75	78.69	83.91	100.0	91.25	83.91	100.0	91.25
Kuo Zhang	60.00	60.00	60.00	100.0	100.0	100.0	82.50	70.20	75.85	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Hui Fang	60.87	90.32	72.73	100.0	100.0	100.0	40.60	80.60	54.00	92.21	54.20	68.27	100.0	100.0	100.0	100.0	100.0	100.0
Bin Yu	21.23	35.50	26.57	67.22	50.25	57.51	18.30	27.50	21.98	39.26	55.19	45.88	92.31	66.67	77.42	89.32	84.53	86.86
Lei Wang	11.98	21.87	15.48	68.45	41.12	51.38	21.52	57.34	31.29	44.40	75.59	55.94	91.58	92.59	92.08	88.64	89.06	88.85
Rakesh Kumar	68.82	91.28	78.47	63.36	92.41	75.18	62.83	90.17	74.06	80.98	82.43	81.70	92.37	99.18	95.65	99.14	96.91	98.01
Michael Wagner	57.66	52.32	54.86	18.35	60.26	28.13	52.18	46.39	49.11	42.20	64.04	50.87	26.25	77.78	39.25	85.19	76.16	80.42
Bing Liu	53.10	31.73	39.72	84.88	43.16	57.22	76.80	72.60	74.64	30.21	63.05	40.85	83.72	98.63	90.57	88.25	86.49	87.36
Jim Smith	62.59	44.16	51.78	92.43	86.80	89.53	43.10	40.50	41.76	83.14	80.87	81.99	70.91	97.50	82.11	95.81	93.56	94.67
Wei Wang	11.97	10.30	11.07	8.70	100.0	16.01	10.50	10.50	10.50	12.00	66.73	20.35	33.67	84.26	48.11	83.67	84.26	83.96
Ajay Gupta	67.33	58.62	62.67	41.88	100.0	59.04	61.82	43.59	51.13	51.16	77.65	61.68	90.67	96.55	93.52	97.67	96.55	97.11
Dimitry Pavlov	85.71	85.71	85.71	85.71	85.71	85.71	87.40	83.20	85.25	100.0	100.0	100.0	88.70	89.23	88.96	86.67	100.0	92.86
David Jensen	82.57	41.51	55.25	85.85	94.88	90.14	80.52	40.13	53.56	81.13	85.26	83.14	82.51	65.23	72.86	83.83	68.46	75.37
David Brown	63.84	78.64	70.47	35.89	100.0	52.82	59.21	36.34	45.04	42.29	86.39	56.78	50.23	75.23	60.24	89.32	91.45	90.37
David C. Wilson	65.50	21.58	32.46	85.54	99.79	92.12	49.53	23.12	31.52	100.0	100.0	100.0	75.12	60.45	66.99	94.33	67.30	78.55
George Miller	85.19	65.71	74.19	85.87	75.24	80.20	68.90	67.85	68.37	50.97	79.94	62.25	72.37	74.56	73.45	85.87	75.24	80.20
James H. Anderson	80.23	96.05	87.43	89.15	99.27	93.94	76.50	76.50	76.50	98.08	51.52	67.55	85.99	80.12	82.95	88.51	85.80	87.13
James Johnson	69.23	81.82	75.00	73.77	100.0	84.91	81.76	53.82	64.91	88.11	69.52	77.72	78.32	75.67	76.97	100.0	100.0	100.0
John Miller	69.99	96.81	81.24	69.35	90.75	78.62	72.83	68.51	70.60	77.36	63.08	69.49	72.65	79.07	75.72	83.38	97.73	89.99
Joseph Miller	57.14	72.73	64.00	54.55	54.55	54.55	49.32	67.18	56.88	61.29	44.19	51.35	55.21	59.34	57.20	86.55	74.55	80.10
Paul Jones	51.61	64.00	57.14	36.36	80.00	50.00	48.19	59.31	53.17	16.79	63.49	26.56	38.64	63.45	48.03	84.00	84.00	84.00
Richard Taylor	68.85	19.91	30.89	80.17	99.93	88.97	72.31	34.56	46.77	53.80	94.69	68.62	68.23	64.54	66.33	94.33	79.72	86.41
Robert Fisher	92.87	61.17	73.76	96.14	100.0	98.03	73.16	48.57	58.38	81.02	86.57	83.70	85.21	74.54	79.52	92.82	79.13	85.43
Robert Moore	92.10	66.01	76.90	86.90	93.10	89.89	80.60	48.33	60.43	100.0	100.0	100.0	89.91	78.54	83.84	84.04	75.66	79.63
Robert Williams	63.64	46.67	53.85	66.67	66.67	66.67	57.83	33.96	42.79	73.90	90.69	81.44	65.12	58.23	61.48	86.67	60.00	70.91
William Cohen	82.25	90.12	86.01	81.53	97.98	89.00	80.45	52.60	63.61	100.0	100.0	100.0	86.01	85.23	61.48	80.37	83.34	81.83
Charles Smith	50.00	33.00	39.76	30.00	100.0	46.15	57.92	62.15	59.96	44.42	74.46	55.65	45.27	67.89	85.62	100.0	100.0	100.0
Avg.	61.49	56.03	56.21	73.58	85.53	75.52	60.41	53.34	54.59	68.80	79.63	71.23	76.47	83.09	78.62	**90.13**	**88.26**	**88.80**

5.3 PART

Experimental Results

Results of Our Approach with Different Settings

without auto K

Method	Precision	Recall	F1-Measure
Our Approach (Auto K)	83.01	79.54	80.05
Our Approach (w/o auto K)	90.13	88.26	88.80
Our Approach (w/o relation)	67.05	50.59	55.95

without relationship

Result of Automatically Discovered Person Number

Person Name	Actual Number	Auto Number	Person Name	Actual Number	Auto Number
Cheng Chang	3	3	Dimitry Pavlov	2	1
Wen Gao	4	5	David Jensen	3	6
Yi Li	21	13	David Brown	7	9
Jie Tang	2	2	David C. Wilson	5	5
Gang Wu	16	12	George Miller	2	6
Jing Zhang	25	16	James H. Anderson	2	7
Kuo Zhang	2	2	James Johnson	3	3
Hui Fang	3	3	John Miller	2	5
Bin Yu	12	10	Joseph Miller	2	3
Lei Wang	40	22	Paul Jones	3	5
Rakesh Kumar	5	5	Richard Taylor	10	14
Michael Wagner	10	11	Robert Fisher	4	7
Bing Liu	11	12	Robert Moore	3	6
Jim Smith	5	5	Robert Williams	2	5
Wei Wang	90	22	William Cohen	2	9
Ajay Gupta	4	6	Charles Smith	4	4

5.4
PART

Efficiency Performance

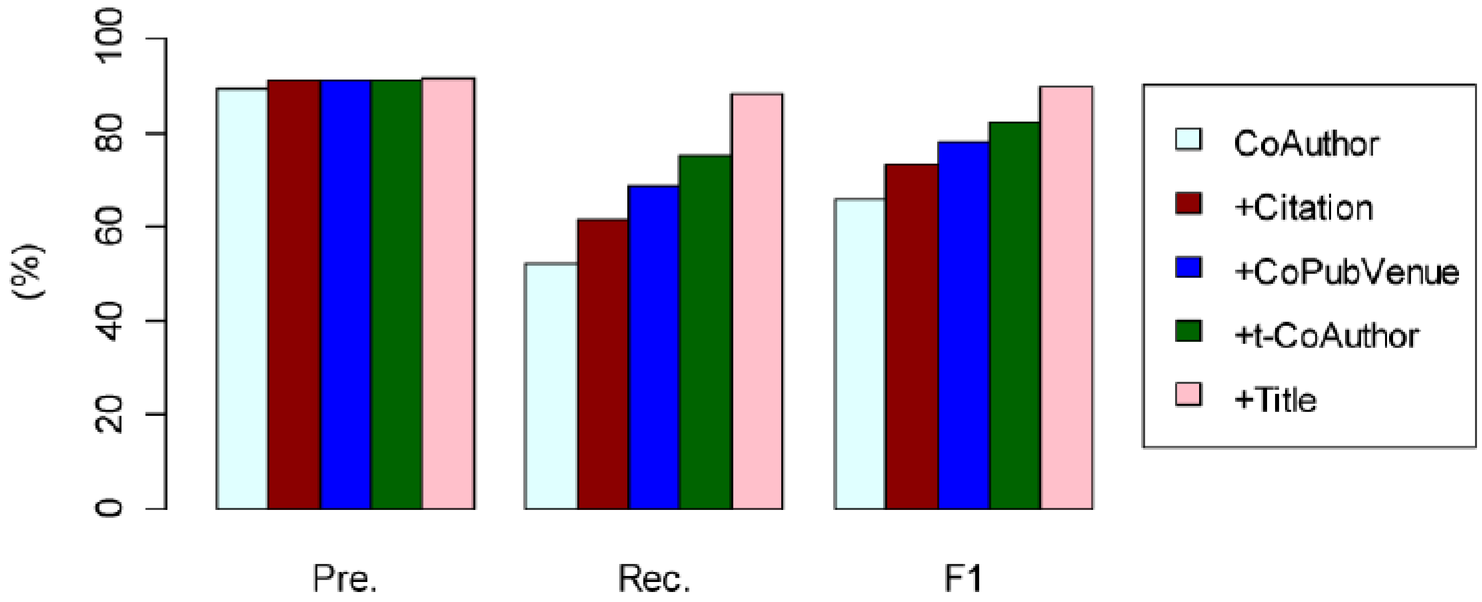
with Intel Core Duo processor(1.6 GHz)

only list six authors who publish more than 100 papers and the average for 100 random names.

Comparison of Efficiency Performance (Seconds)

Person Name	*K*-means	X-Means	HAC	SACluster	DISTINCT	Our Approach
Wen Gao	4.8	5.1	12.9	30.4	56.0	20.3
Lei Wang	3.7	2.4	6.8	4.1	12.1	4.6
Bing Liu	1.6	1.9	4.2	5.4	1.1	5.8
Wei Wang	28.7	5.1	73.1	46.9	83.3	100.2
Robert Fisher	2.8	1.3	5.6	0.2	0.2	0.8
William Cohen	0.8	1.2	3.0	0.06	0.6	0.9
Average over 100	0.52	0.26	1.14	0.96	0.87	1.42

Feature Contribution Analysis



5.6
PART

Conclusion

- Formalize the problems in a unified framework and proposed a generalized probabilistic model to the problem.
- Define a disambiguation objective function for the problem and have proposed a two-step parameter estimation algorithm.
- Explore a dynamic approach for estimating the number of people K.

THANKS!