A Unified Probabilistic Framework for Name Disambiguation in Digital Library

Jie Tang, A.C.M. Fong, Bo Wang, and Jing Zhang

Abstract—Despite years of research, the name ambiguity problem remains largely unresolved. Outstanding issues include how to capture all information for name disambiguation in a unified approach, and how to determine the number of people K in the disambiguation process. In this paper, we formalize the problem in a unified probabilistic framework, which incorporates both attributes and relationships. Specifically, we define a disambiguation objective function for the problem and propose a two-step parameter estimation algorithm. We also investigate a dynamic approach for estimating the number of people K. Experiments show that our proposed framework significantly outperforms four baseline methods of using clustering algorithms and two other previous methods. Experiments also indicate that the number K automatically found by our method is close to the actual number.

Index Terms—Digital libraries, information search and retrieval, database applications, heterogeneous databases.

1 INTRODUCTION

DIFFERENT e e a , a e de ca a e , e ea d. I e a ed , a , e 300 c a e a e a e ed b e , a 114 e e (a ab 78.74 e ce) , e U ed S a e (:// a e . aba .c / a e_ a e ,). I a a ca , ç a ce f c e a e a a e e a d f a e a , , e e e a e a e ed a , e de f e e e e , e f a . Na e a b ea , e a f , e e e ed f a .

T de e , e e e f , e b e , e , a e e a ed 100 e a e , e b ca da a a d f d, f e a e, , e e a e 54 a e a , ed b 25 d ffe e J Z a , e DBLP da aba e. A , ee de a ed Y L , a e ad a ed f , e f a , ' ab.

1.1 Motivation

We be b a , e b e , a e a e d a f a ea - d e (:// a e e .) [40]. I , e , e e ac e ea ç e f e f , e eb a d e a e , e b ca da a f e da aba e ç a DBLP, ACM D a L b a , C eSee , a d SCI. I , e e a , e e ab , a e , e a e a b b e . F . 1 , a f ed e a e. I F . 1, eaç de de e a a e (, e ed). Eaç d ec ed ed e de e a e a , be ee a e , a abe e e e , e e f , e e a , (cf. Sec 2.1 f def f , e e a , e). T e d a ce be ee de de e , e a f , e a e e f e c e -ba ed a ea e e (e. ., c e a). T e d e , e dea d a b a e , , ç d ca e , a 11 a e , d be a ed , ee d ffe e a , . A ed a e b e a f F . 1 , a a e d ba ed c e a (, e d a ce) d be d ff c ac e e a fac e f a ce, a d , a d ffe e e f e a , ca be , e f , b , d ffe e de ee f c b . F e a e, , e e a C A , e a , be ee de #3 a d #8. A , , e a be ee , e de , , , be ef f , e C A , e a , , e ca a , e. de (a e) , e a e a , . O , e c a , a , , , e e a C a e a , be ee de #3 a d #7, , e a e a e a ed d ffe e a , . T , e ç a e e , e e , de a a , f , e a e d a b a b e b c de b , a b e f a f , e de a d , e e a , be ee de .

1.2 Prior Work

T e b e , a bee de e de e a ed d ffe e d a , a d a a e e [4], [5], [7], eb a ea a ce d a b a [3], [20], a e de f ca [26], a d Ob ec d c [49]. De e a a aç e ed, , e a e a b b e e- a a e e ed.

I e e a , e e , d f a e d a b a a fa , ee ca e e : *supervised based*, *unsupervised based*, a d *constraint based*. T e e ed-ba ed a aç (e. ., [17]) e ea a ec f c c a f ca de f eaç a , a e f , e , a - abe ed a da a. T e , , e ea ed de ed ed c , e a , a e f eaç a e . I , e e ed-ba ed a aç (e. ., [18], [36], [37], [49]), c e a , c de a e e ed f d a e

- *J. Tang and J. Zhang are with the Department of Computer Science and Technology, Tsinghua University, Rm 1-308, FIT Building, Beijing 100084, China. E-mail: jietang@tsinghua.edu.cn, zhangjing0544@gmail.com.*
- *A.C.M. Fong is with the School of Computing and Mathematical Sciences, Auckland University of Technology, AUT Tower Level 1, 2-14 Wakefield Street, Auckland 1142, New Zealand. E-mail: afong@aut.ac.nz.*
- *B. Wang is with the Department of Computer Science, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China. E-mail: bowang@nuaa.edu.cn.*

Manuscript received 1 July 2008; revised 5 Apr. 2010; accepted 16 Nov. 2010; published online 27 Dec. 2010.
Recommended for acceptance by B.C. Ooi.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2008-07-0335.
Digital Object Identifier no. 10.1109/TKDE.2011.13.

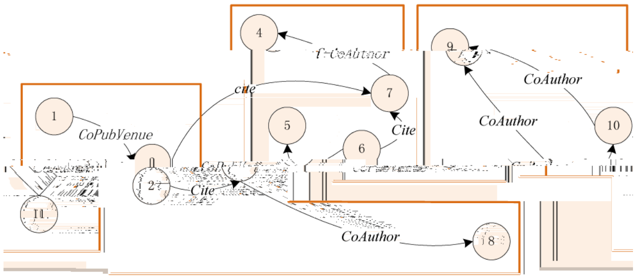

Fig. 1. An example of name disambiguation.

a , a d a e d ffe e a a e a ed d ffe e a . T e c a -ba ed a ac a e , e c e a . T e d ffe e ce , a e- ded c a a e ed de , e c e a , a d be e da a a (e. ., [2], [51]).

F , e e, e e a , e a ac e ba ed e, c a /a , a , a d c b a f , e d ffe e a ac e , a e bee d ed. F e a e, W a e a . [47] d ce a e a e e -ba ed a ac e e , e c e c e , e da aba e a d de e a de f a e e c ea e , e e . Da e a . [11] a e de e ed a e ac e e , e a e ca e , e cc e ce f a ed e e , a e e . T e e de f efe e ce a e a bec (e. ., a a c a b d) e e a ed a e f , a bec a d a c ec . McRae-S e ce a d S adb [28] e e a a -ba ed a ac a d a b a a e- ca e c a e b e f-c a , c a , e a . T e a ac ca ac e e a , ec b a e a e eca . Y e a . [50] a e de e ed e ed a ac e de f , e f f f a b abb e a , e c e , e a ea . M e ece , C e e a . [8] d c b e , e d ffe e d a b a a ac e a d e a e e e be f a e , c b e , e e f e ba e- e e e e a e , e acc ac f e e . W a e a . [46] e a e a e b c f a e , e e , e e e f b c a e ef ec ed b e e ce ed b c . O a d Lee [32] d e ca ab e f , e a e d a b a be . A , e , a bee ade, e e, d d ac e e a fac d a b a e de , e a :

1. S e e a c e e, d (e. ., [31], [35], [48]) f c a , e da a a , ba ed , e ca c e; e , e e, d (e. ., [18], [42]) a c e , e da a a , acc d de a . A fe e ea e (e. ., [38], [52]) c b e , e ece f f a . F e a e, Z e a . a e c b e f a ba ed b , e e a b e (.e., de a) a d a , ca c e b f c c a a b e a e ed a , , e c a e f ⟨a b e, a e⟩ a e ce , a d be e e a , e a e e ce ' c e e a e , b , d a d a de . T e a e c a ea , a e b e e d ca d ca f a . A , e a , e e abe de a e , a a b e a c ea e , e c e e f a e e ce , e d a ce ea e, a ba a ce , e c b f , e d ffe e f a a e be T e a e abe c c de , a add a b e a f a , e c e be c e de ade , e ac e c e e . F , e, [52], e e e e a da a e c a e fe a b e . T e f da a e (ca b) , a e (b a) a b e a d , e ec d da a e f DBLP b b a , ca da a , a a b e . We a e , a c e de a b e f a- e ed f ac , e a e d a b a- be effec e .
2. T e e f a ce f a , e af e e ed e, d de e d acc a e e a K. A , , e e a c e a , c a X- ea [33] ca a a ca f d , e be K ba ed e c e , c ea , e , e c a e, d ca be d ec a ed , e a e d a b a be .
3. I e e, d , , e da a a c a , e e de a d e a , ; , e be e , , e e a be e d ffe e e a (e. ., C A , a d C a) be ee de . T e e f d ffe e e a , a , a e d ffe e a ce f , e a e d a b a be . H a a ca de , e de ee f c b f d ffe e e a , a c a e be .

1.3 Our Solution

Ha c d c ed a , , e a , e e a fed bab c f a e add e , e ab e c a e e . S ec f ca , e f a e , e d a b a be a Ma Ra d F e d (MRF) [16], [24], , c , e da a a e c , e e b , ca a b e a d e a , . We e e a d a c a ac f e a- , e be f e e K a d a - e a , f a a e e e a . T e ed a ac ca ac e e be e e f a ce a e d a b a , a e e , d beca e , e a ac a e ad a a e f e- de e de c e be ee a e a e . T , e be f ed e, , e f f a e a , e be f a e d a b a a fed f a e a d ac e , e be e, e .

T e ed f a e e e e a . O e ca c a e a e a a fea e ca fea e , e f a e , e. ., a fea e ba ed , e eb ea c e e ed. T e f a e ca be a e e ded dea , a , e be c a e e a e a a da aba e [4].

O c b , a e c de: 1) f a a f , e a e d a b a be a fed bab - c f a e ; 2) a f a a , e , e a a e e e a , e f a e ; a d 3) a e ca e f ca f , e effec e e f , e ed f a e .

TABLE 1
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	authors name set of p_i $\{a_i^{(0)}, a_i^{(1)}, \ldots a_i^{(u)}\}$
p_i.references	references of p_i

2 Problem Formalization

2.1 Definitions

I e d c a f , e a a b e eac a e p_i a Tab e 1. S c b ca da a ca be e ac ed f ce c a DBLP, L b a. a.c , A e e. , a d C e ee . . .ed .

Definition 1 (Principle Author and Secondary Author). *Each paper p_i has one or more authors $A_{pi} = \{a_i^{(0)}, a_i^{(1)}, \ldots a_i^{(u)}\}$. We describe the author name that we are going to disambiguate as the principle author $a_i^{(0)}$ and the rest (if any) as secondary authors.*

We def e f e e f d ec ed e a be ee a e (Tab e 2). S ec f ca ,

- C P bVe e (r_1) e ee a e b ed a e a e e e. F e a e, f b a e a e b ed a KDD, e c ea e a d ec ed C P bVe e e a be ee e a e . I e , e ea c e e a e a e a d ffe e e ea c f e d , d b a e a d ffe e e e .
- C A (r_2) e ee a a e p_1 a d p_2 a e a ec da a , e a e a e, .e., $A'_{p1} \cap A'_{p2} \neq \emptyset$, e e A'_{p1} de e e e f a f a e p_1 e c d e c e a $a_i^{(0)}$, .e., $A'_{p1} = A_{p1} \setminus a_i^{(0)}$. T ca , a e a a e a c c a d be e a e e .
- C a (r_3) e ee e a e c a e a e . I e a a a c e , e . F e , e c a e a e c a- f a a f : If a e p_1 c e a e $p_2, p_3, \ldots, p_n$, e e e ab d ec ed a e e a , a a c ed a e , add d ec ed a e e a be ee p_1 a d e c ed a e .
- C a (r_4) de e c a ed a e feedbac . F a ce, e e ca ec f a a e , d be d a b a ed e a e e , d be d ffe e e .
- τ-C A (r_5) e ee τ-e e C A e a . We e a e a e e a e a . S e a e p_i, a a Da d M c e a d A d e Ma , a d p_j, a a Da d M c e a d Fe a d M f d. We a e d a b a e Da d M c e . A d f A d e Ma a d Fe a d M f d a c a a e a e , e e a p_i a d p_j, a e a 2-C A e a .

TABLE 2
Relationships between 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	Constraint	feedback supplied by users
r_5	w_5	τ-CoAuthor	τ-extension co-authorship (τ>1

T a e c ea , e e a f e ab , de e e e e a e a e a τ-C A e a - . F e e e a e da a e , e ca c c a c a e , e e eac de de e a a a e a d eac ed e de e a c a e a . F a a e p_1 a d p_2, e ca b a e c e d- e A'_{p1} a d A'_{p2} b e c a . If a d f $A'_{p1} \cap A'_{p2} \neq \emptyset$, e a e a e a e a C A e a . F de e a 2-e e C A e a , e c c c a e A^2_{p1} a d A^2_{p2} acc d e c a e . S ec f ca , A^2_{p1} e e f a b e e d A'_{p1} a e b f e a A'_{p1}, .e., $A^2_{p1} = A'_{p1} \cup \{NB(a)\}_{a \in A'_{p1}}$, e e $NB(a)$ e e f e b f de a. T e , e a e a e p_1 a d p_2 a e a 2-C A e a , f a d f $A^2_{p1} \cap A^2_{p2} \neq \emptyset$. F de e e e a e a e a 3-e e C A e a , e f e e e d A^2_{p1} f d a a e A^3_p f eac a e a d f e e a e a e ec , e a e a e a e a 3-C A e a . T e e f eac e f e a r_i de ed b w_i. E a f e a e f d ffe e e be de c bed Sec 4.

I e a e d a b a be , e a e a ea be c e ed e e a be a ed e e b e e. T e e a e be a ed e d a b a a . We de c be c f a e a *cluster atom.*

Definition 2 (Cluster Atom). *A cluster atom is a cluster in which papers are closely connected (e.g., the similarity $K(x_i, x_j) >$ threshold). Papers with similarity less than the threshold will be assigned to disjoint cluster atoms.*

F d c e a d be ea e f a e d a b a . F e a e, e ca a e ec e a a e a f e d a b a a . F f d e c e a , e ca e a c a ed-ba ed c e a e e c a . I add , e def e e c ce f *cluster centroid.* De ed f e c e a a , e e a e ca e d f d e ce d f a c e , e da a a ea e e ce e f e c e e ce d a ca c a ed a e a e c ea f a da a a ed e c e .

2.2 Name Disambiguation

G e a e a e a, e de e b ca c a e a a e a a $P = \{p_1, p_2, \ldots, p_n\}$. T e b ca da a e a ca be de ed b e c de a d ed e . We e a ada e e f e -ca ed f a e a [13] e e e e b ca da a. P b ca a d e a a e a - f ed a d ec ed a , eac de e e e a a e a d eac ed e a e a . A b e f a a e a e a ac ed e c e d de a a fea e ec . F e ec , e e d (af e d f e a d e) e a b e f a a e a fea e a d e e be f e cc e ce a e a e . F a , e ca def e e b ca f a e a a f :

Definition 3 (Publication Informative Graph). *Given a set of papers $P = \{p_1, p_2, \ldots, p_n\}$, let $r_k(p_i, p_j)$ be a relationship r_k between p_i and p_j. A publication informative graph is a graph $G = (P, R, V_P, W_R)$, where each $v(p_i) \in V_P$ corresponds to the feature vector of paper p_i and $w_k \in W_R$ denotes the weight of relationship r_k. Let $r_k(p_i, p_j) = 1$ iff there is a relationship r_k between p_i and p_j; otherwise, $r_k(p_i, p_j) = 0$.*

S e e e a e K e $\{y_1, \ldots, y_K\}$ e a e a, a d a b a e e n b ca e ea e ea e y_i, $i \in [1, K]$. M e ec f ca , e a a f a e d a b a ca be def ed a :

1. F e d a b a b e . T e f - a eed c de b ca a b e fea e a c a ed eac a e a d e a - be ee a e .
2. S e b e a c ed a ac . Ba ed e f a , e a c ed a ac a d e a eff c e a .
3. De e e be f e e K. G e a d a b a a (a f a-), de e e e ac a K.

I a e f e e a . F , ed a e c ea , f e e e e d a b- a b e a f ed f a e . Sec d, e a de , e. ., Ma Ra d F e d [16], a e a a ed de e a a da a. H e e , e b ca f a e a , e a e be a b a c ec ed b d ffe e e f e a . I c ea , e f fe e ce (a a e e e a) a a , a b a c e. I add , e a e be f e e K a a a e a .

3 Our Framework

3.1 Basic Idea

We a e ba c b e a f e a e d a b a- be : 1) a e , a c e e d a e e a e abe (be e a e a); a d 2) a e a ea , e d a e e a e abe , f e a e, a e c a , a a a e a e . A dea d a b a e e a e b e e a b c e a a d a e e a . T a a b e , beca e e c e e d ca e ba a ce e ece f f a .

I a e , e ea f ed f a e ba ed Ma Ra d F e d [16], [24]. M e acc a e , e f e b c e -ba ed f a a d c e- ba ed f a a H dde Ma Ra d F e d (HMRF) de a fea e f c . T e c b de ee f e e f f a a e f ed a e f e fea e f c . T e a ce f d ffe e e f e a , a de ed a e f c e d fea e f c . S e HMRF de c de b e a e e f fea e f c a d a a e d ffe e e . S a f a e a ffe add a ad a a e : f , e ed ea , e ed ea , a d e - e ed ea . I a e , e f c e ed ea f a e d a b a , b ea c a e e / e ed f a e de . Sec d, a a d de e ec e HMRF de . T e b ec e f c e HMRF de a e bab d b f dde a ab e e b e a , ac e f de e ec a e .

3.2 Hidden Markov Random Fields

A Ma Ra d F e d a c d a bab d b f abe (dde a ab e) a be e Ma e [16]. Ma ec a ca e f MRF ca be de e ed. A H dde Ma Ra d F e d a e be f e fa f MRF a d c ce de ed f H dde Ma M de (HMM) [15]. A HMRF a c ed f ee c e : a b e ab e e f a d a ab e $X = \{x_i\}_{i=1}^n$, a dde f e d f a d a ab e $Y = \{y_i\}_{i=1}^n$, a d e b d be ee eac a f a ab e e dde f e d.

We f e e d a b a b e a a f ea a a e d ffe e c e . Le e dde a ab e Y be e c e abe e a e . E e dde a ab e y_i a e a a e f e e $\{1, \ldots, K\}$, a e e de e f e c e . T e b e a a ab e X c e d a e , e e e e a d a ab e x_i e e a ed f a c d a bab d b $P(x_i|y_i)$ de e ed b e c e- d dde a ab e y_i. F e , e a d a ab e X a e a ed be e e a ed c d a de e de f e dde a ab e Y, .e.,

$$P(X|Y) = \prod_{x_i \in X} P(x_i|y_i). \tag{1}$$

F . 2 e a ca c e f e HMRF f e e a e F . 1. We ee a de e de ed e a e ded be ee e dde a ab e c e d e ea F . 1. T e a e f eac dde a ab e (e. ., $y_1 = 1$) de e e a e e . We d de e d ec ea be ee e b , b e de ca a a e e de e de c e a e ea .

A HMRF a ec a ca e f MRF, e bab d b f e dde a ab e be e Ma e . T , e bab d b f e a e f y_i f e b e a a ab e x_i de e d e c e abe f b e a a a e e a x_i [24]. B e f da e a e e f a d f e d [16], e bab d b f e abe c f a Y a e f

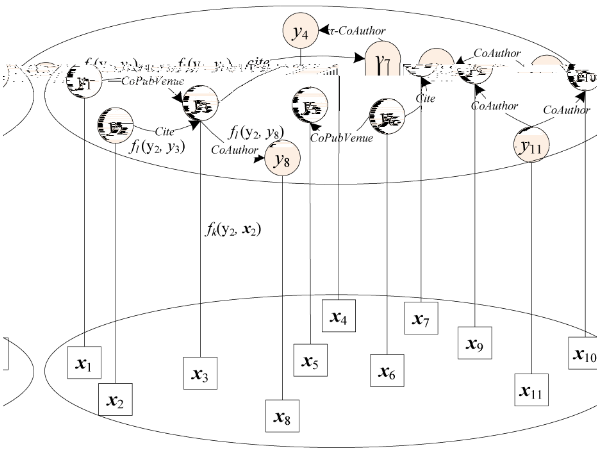

Fig. 2. Graphical representation of the HMRF model. $f(y_i, y_j)$ and $f(y_i, \mathbf{x}_i)$ are edge feature and node feature, respectively, and will be described in the next section.

$$P(Y) = \frac{1}{Z_1}\exp\left(\sum_{(y_i,y_j)\in E,k} \boldsymbol{\lambda}_k f_k(y_i, y_j)\right),$$
$$Z_1 = \sum_{y_i,y_j}\sum_{(y_i,y_j)\in E,k} \boldsymbol{\lambda}_k f_k(y_i, y_j) \tag{2}$$

a d b f e e c e b ca da a be e e a ed de e e ca Ga a d b , e a e

$$P(X|Y) = \frac{1}{Z_2}\exp\left(\sum_{x_i\in X,l} \boldsymbol{\alpha}_l f_l(y_i, x_i)\right),$$
$$Z_2 = \sum_{y_i}\sum_{x_i\in X,l} \boldsymbol{\alpha}_l f_l(y_i, x_i), \tag{3}$$

e e $f_k(y_i, y_j)$ a e a e e a f c (a ca ed e fea e f c) def ed ed e (y_i, y_j) a d E e e e a ed e e a ; $f_l(y_i, x_i)$ a e a f c def ed de x_i; λ_k a d α_l a e e , f e ed e fea e f c a d e de fea e f c , e ec e ; Z_1 a d Z_2 a e a fac .

T fac a e f e d c , e e eaf e e X de e e b ca e P a d e x_i de e e ec $v(p_i)$ f e a e p_i.

3.3 Disambiguation Objective Function

We def e a bec e f c a e Ma a- P e c f a f e HMRF, .e., b a $P(Y|X)$. $P(X)$ a a e a c a . T e ef e, acc d e Ba e e $P(Y|X) \propto P(Y)P(X|Y)$, bec e f c ca be def ed a

$$L_{\max} = \log(P(Y|X)) = \log(P(Y)P(X|Y)). \tag{4}$$

B b (2) a d (3) (4), e b a

$$L_{\max} = \log\left(\frac{1}{Z_1 Z_2}\exp\left(\sum_{(y_i,y_j)\in E,k} \boldsymbol{\lambda}_k f_k(y_i, y_j) + \sum_{x_i\in X,l} \boldsymbol{\alpha}_l f_l(y_i, x_i)\right)\right). \tag{5}$$

E e a , e ab e bec e f c , e e e d f fea e f c , de fea e f c $f_l(y_i, x_i)$ a d ed e fea e f c $f_k(y_i, y_j)$, e e e e a b e f a a c a ed eac a e a d e e a , f a be ee a e , e ec e .

T e ed e fea e f c $f_k(y_i, y_j)$ ed c a ac e- e e e a , be ee a e . I e , f a e , a e a e a , a d a a e a eac e , e e e e a e a e be a ed e a e c e . S ec f ca , e ed e fea e f c ca e e a e e a , c a C P bVe e a d C A (a Tab e 2) a d a ea e f a . T , e def e e ed e fea e f c a

$$f_k(y_i, y_j) = K(x_i, x_j)\sum_{r_m\in R_{ij}}[w_m r_m(x_i, x_j)]. \tag{6}$$

He e, $K(x_i, x_j)$ a a f c be ee a e x_i a d x_j; w_m e e f e a , r_m; R_{ij} de e e e f e a , be ee x_i a d x_j; a d $r(x_i, x_j)$ de e a f c f e e a , be ee x_i a d x_j. T e e a def e e e a f c $r(x_i, x_j)$ def e b a a e a de c bed Def 3. He e, e f e c de a def , c c b e e e f a , .e., $r_1(x_i, x_j) = exp\{-|x_i.year - x_j.year|\}$. T def de ed f a b e a e a e a b be : e C A a d C P bVe e e a- a e f e e-de e de , e. ., a e d. b , a e e a f c ed c fe e ce / a e e a ec f c e d a d c a a e d c ab a e eac e a ec f c e d.

T e de fea e f c $f_l(y_i, x_i)$ a ca e e a b e f a a c a ed a e x_i. T e ba c dea e e f e a e a a e e a e a c e e e e a e a e be a ed e c e . F c e a e e a , e def e e de fea e f c a

$$f_l(y_i, x_i) = K(y_i, x_i) = K(\boldsymbol{\mu}_{(i)}, x_i), \tag{7}$$

e e $\mu_{(i)}$ e c e ce d a e a e x_i a ed . N a $K(x_i, \mu_{(i)})$ e e e e a be ee a e x_i a d a ed c e ce e $\mu_{(i)}$.

T e , (6) a d (7) (5), e b a

$$L_{\max} = \sum_{(x_i,x_j)\in E,k}\lambda_k K(x_i, x_j) r_k(x_i, x_j) + \sum_{x_i\in X,l}\boldsymbol{\alpha}_l K(x_i, \boldsymbol{\mu}_{(i)}) - \log Z, \tag{8}$$

e e $Z = Z_1 Z_2$. W a f e e a , e c b e e e , f ed e fea e f c λ_k a d e e f e e a , w_m, a d e a λ f c .

3.4 Criteria for Model Selection

We e Ba e a I f a C e (BIC) a e c e e a e e be f e K. We def e a bec e f c f e d a b a a . O a e a a a e e e , a a e e ca bec e f c , e e K a d f d a be K a a e e ba bec e f c .

S ec f ca , e f c de $K = 1$, a , ee e e , e e a e a a. T e , e e a ea e e de e e , e e , e a e c e , d be bc e . Ne , f eac bc e , e a a e , e ea e e de e e , e , e . T e e a e ea e c d- a f ed (e. ., bc e ca be). I e ce , e ca M_h , e de c e d , e , , e e be h. We , e ef e , a e a fa f a e a e de M_h , , e e h a e f 1 n, c e .

N , a ç e , e be de f M_h . Ma ea e e ca be ed f de e ec , ç a S , e e C eff c e [23], M De c Le , (MDL) [34], A a e I f a C e (AIC) [1], a d e bab e a [22]. We ç e BIC a , e c e , beca e BIC c e f da e- a a , e c e a ç a MDL a d , a a e e a , a , e , e c e a ç a AIC, , ç de ab e b e . Ba ed , e e c de a , e e a a a f , e BIC ea e e [22] a , e c e

$$BIC^v(M_h) = \log(P(M_h|P)) - \frac{|\lambda|}{2}\cdot\log(n), \tag{9}$$

, e e $P(M_h|P)$, e e bab f de M_h e , e b e a P. $|\lambda|$, e be f a a e e M_h (, ç ca be def ed d ffe e a , e. ., , e be f z-- e a a e e , e de M_h , e f , e bab e f $P(Y)$). n , e a e be . T e ec d a a e a de c e .

I e e ce, a BIC c e a a e , a - a e , e de M_h f , e , e da a e . We e , c e f , e de e ec beca e ca be ea e e ded d ffe e a . F e a e, c e- a c e a , e K- ea [27] X- ea [33] e a d , e da a a de e de a d , , e e bab $P(M_h|P)$ ca be f ed $P(P|M_h)$ acc d , e Ba e a e $P(M_h|P) \propto P(P|M_h)P(M_h)$ b a , e $P(M_h)$ a f . H e e , e e d a e ad a a e f de e de c e be ee , e c e e . T , e $P(M_h)$ a f a a e. O def (2) c de , e de e de c e a Ma f e d.

4 PARAMETER ESTIMATION

4.1 Algorithm

T e a a e e e e a be de e e , e a e f , e a a e e $\Theta = \{\lambda_1, \lambda_2, \ldots; \alpha_1, \alpha_2, \ldots\}$ a d de e e a e f a a e . M e acc a e , e z-- e , e - e , d bec e f c (8) , e ec a c d a de $P(Y|X,\Theta)$.

A a , e e , , e ea a , (cf. A , 1) f a a e e e a a c f e a e e : *Assignment* f a e , a d *Update* f a a e e Θ. T e ba c dea , a e f a d ç e a a a e e e Θ a d e ec a ce d f eac c e . Ne , e a eac a e c e c e a d , e ca c a e , e ce d f eac a e -c e ba ed , e a e . Af e , a , e da e , e e , f eac fea e f c b a z-- , e bec e f c .

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.

F a-- a , e a d a , e a e f eac a a e e (λ a d α). F a-- a f , e c e ce d, e f e a a , c e e , d de f , e c e a . Ba ca , a e , a e , a a , e , d be a ed d c e a . We eed a a e , e de c bed fa , b a a ç , e a e , a , a , e , , e a , e c e ce d u. I , a , e e γ c e a . If γ e a , e be f e e K, , e , e e γ a e ed a a a e . If $\gamma < K$, e a d ç e a , e (K-γ) a e a , e c e ce d . If $\gamma > K$, e , e ea e c e a , e e a e K ef . We d ce de a , e e a a e e e a a , .

Assignments. I *Assignments*, eac a e x_i a ed $\mu_{(h)}$ a z-- e $\log P(y_i|x_i)$

$$\begin{aligned}\log P(y_i|x_i) &\propto L_{x_i}(\mu_{(h)}, x_i)\\ &= \sum_{(x_i,x_j)\in E_i, R_i, k} \lambda_k K(x_i,x_j) r_k(x_i,x_j)\\ &\quad + \sum_l \alpha_l K(x_i, \mu_{(h)}) - \log Z,\end{aligned} \tag{10}$$

, e e Z de ade a a-- a fac x_i a d ca be e ed a e ca e ab , e e a e c e f f c ,e. ., d e e a d ad e e. H e e, , e a e e , e f c e a .

N , , e a ca c a e a a a e c e (10). T e f e (10) a e a a c b a f , e a f c $K(x_i, \mu_{(h)})$ a d , e e a a - a f c $K(x_i, x_j)$, , ç ca be ca c a ed. H e e, ac ab e b a a e ac f , e a f c , .e., (Z), beca e , e a a d a e ace , , e a $(Z = Z_1 Z_2)$. A fe a - , a e bee ed f a a e fe e ce, e. ., be ef a a [30] a d c a e d e e ce (CD) [19]. We e a e, a a e , e a f c a c a e d e e ce d a b a bec e f c .

Ba ed Je e ′ e a [21], e ca b a a e b d f , e e a e - e, d (L) a K - bac -L eb e (KL) d e e ce

$$
\begin{aligned}
L^{KL} &= KL(q\|P) \\
&= \sum_{y_i} q(y_i|x_i)\log(q(y_i|x_i)) - \sum_{y_i} q(y_i|x_i)\log(P(y_i|x_i)) \\
&= -H(q) - \langle \log(P(y_i|x_i))\rangle_{q(y_i)},
\end{aligned}
\tag{12}
$$

, e e $q(y_i|x_i)$ a a a f , e d b $P(y_i|x_i)$. $\langle . \rangle_q$, e e ec a de , e d b q.

Ma , e - e, d f , e da a (5) e a- e , e KL d e e ce (12) be ee , e da a d b q^0 a d , e e b d b e , e be a ab e, q^∞, , e e , e f e ca be ca c a ed b , e b e a , , e c e a ed abe a d , e ec d e , e bab , e e e de d b , a be abe . A a , , e b d ff c , e ab e e a , de e , e ec d e . A Ma ç a M e Ca (MCMC) e d ca be ed e a e , e a a d b $q^\infty(y_i|x_i)$, , e a f MCMC be e ed a $q^0(y_i|x_i)$. T a e , e ce e eff c e , e ca e , e c a e d e e ce a [19], , ç a a e , e d - b b e e a G bb a e (e e). T , , e bec e f c bec e

$$
\begin{aligned}
L^{KL} &= KL(q^0\|P) \approx KL(q^0\|P) - KL(q^l\|P) \\
&= \langle \log(P(y_i|x_i))\rangle_{q^0(y_i)} - \langle \log(q^l(y_i|x_i))\rangle_{q^l(y_i)}.
\end{aligned}
\tag{13}
$$

I c a e d e e ce ea , ead f $KL(q^0\|q^\infty)$, e e , e d ffe e ce be ee $KL(q^0\|q^l)$ a d $KL(q^l\|q^\infty)$, , e e q^l , e d b e , e l- e ec c f , e da a ec (.e., b e a) , a a e e e a ed af e l- e G bb a . A d ca ed [19], , e e l ca be e a 1 ca e . (T a , e ca c de e G bb a e a e , e $KL(q^0\|q^1)$). T e ced e f ec c , e da a ec (.e., q^1) f , e d b q^0 de c bed A , 2.

Algorithm 2: One-step sampling

Input: current observation x^0 and labels y^0

Output: sampling results of y^1 and x^1

1: Draw an observation x, from the distribution of $q^0(x_i)$ ($q(x)$ can be obtained by summing over all possible labels);

2: Compute $P(y_i|x)$, the posterior probability distribution over the label variable given the observation x;

3: [illegible]

4: Draw a new label y^1_i for each observation from the probabi distribution $P(y_i|x)P(y_i|y_{-i})$;

5: Given the chosen label, compute the conditional distribution of $P(x_i|y_i$

6: Draw each feature of the new observation x^1_i from the conditi distribution $P(x_i|y_i)$.

F a , ba ed , e ec c ed da a ec , e ca ca c a e (13). T e ç a c a e e e de a d . T a e e eff c e , e ca e , e de e c ea f e d a , [44] e ace , e a ced e.

Af e , e , d e (10), e ca c e , e f , e , e bec e f c . F a , a eed a , ed e e a da e , e a e f eac a e . A a e f a a e e f ed , e ee , e , e a e f ed. T e ce e ea ed a e ç a e a e be ee cce e e a .

Update. I U da e, eac c e ce d f da ed b , e a , e c ea f , e a e c a ed

$$
\mu_{(h)} = \frac{\sum_{i:y_i=h} x_i}{\|\sum_{i:y_i=h} x_i\|_{\mathbf{A}}}. \tag{14}
$$

T e , b d ffe e a , e bec e f c , e ec eac a a e e λ_k, e , a e

$$
\frac{\partial L}{\partial \lambda_k} = -\sum_{(x_i,x_j)\in E} K(x_i,x_j) r(x_i,x_j) - \frac{\partial \log Z}{\partial \lambda_k}. \tag{15}
$$

We ee , a , e ec d e ac ab e, beca e ca c a f Z eed a b e f a e f eac a e . A a , e a f , e KL d e e ce bec e f c (13) a d e , e CD a , ca c a e , e de a e f L^{KL} , e ec λ_k

$$
\begin{aligned}
\frac{\partial L^{KL}}{\partial \lambda_k} &= \left\langle \frac{\partial \log(P(y_i|x_i))}{\partial \lambda_k} \right\rangle_{q^0(y_i)} - \left\langle \frac{\partial \log(q(y_i|x_i))}{\partial \lambda_k} \right\rangle_{q^1(y_i)} \\
&= -\sum_{(x_i,x_j)\in E} K(x_i,x_j) r(x_i,x_j) - \left\langle \frac{\partial \log(q(y_i|x_i))}{\partial \boldsymbol{\lambda}_k} \right\rangle_{q^1(y_i)}.
\end{aligned}
\tag{16}
$$

T e f e , e a c b a f , e a f c a d , e ec d e ca be ca c a ed af e , e 1- e a (A , 2).

F a , eac a a e e da ed b

$$
\lambda_k^{new} = \lambda_k^{old} + \Delta \frac{\partial L}{\partial \lambda_k}, \tag{17}
$$

, e e Δ , e ea a e. We d , e a e f α.

4.2 Estimation of K

O ae f e a K (ee A , 2) a b e a 1 a d e e e e e BIC c e ea e , e e , e c e c e . T e a , e a e . I eac e a , e e e c e C bc e . We ca c a e a ca BIC c e f e e b de M_2. If BIC(M_2) $>$ BIC(M_1), e e e c e . We ca c a e a ba BIC c e f e e de . T e ce c e b de e f b e f e . F a , e de , e , e ba BIC c e ç e .

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;

O e d ff c e a , be f d e be bc e de f e c e C (L e 4). W d ffe e a a , e e bc e , be d ffe e . F a e , b e a e a ed f a e , be ef f e c e a de f ca . I d a b a , a c e ca c f e e a c e a . T f e , e e e c e a a ce d a d a , e d e ab e e .

F e a a e e $|\lambda|$ (9), e def e a e f e K c e bab e , a a e e , a d c e ce d , .e.,

$$\sum_{i=1}^{K}(P(y_i)+\mu_{(i)})+\sum_{\lambda\in\Theta}\lambda. \tag{18}$$

5 Experimental Results

5.1 Experimental Setting

Data Sets. We e a a ed e ed e d e c e f A e M e . [40]. We c ea ed a da a e , ç c de 32 ea a a e a d 2,074 a e . I e e a e , e a e a e a c a ed a fe e , f e a e C e C a , e a e f ee e a d We Ga f ; e e a e ee be a . F e a e, e e a e 25 e e a e J Z a a d 40 e a ed Le Wa . S a c f da a e a e , Tab e 3. F e P D de f CS c d c ed a a d a b a a a e f e 32 a a e . A ec a c ea ed de e a a ce . Eaç a e a abe ed a be d ca e ac a e . T e abe a ca ed ba ed e b ca e a , ’, e- a e a d ba ed e aff a , e- a add e e e eb da aba e (e. ., ACM D a L b a). We ca c a ed e Ka a c eff c e f e a a ed da a. T e a e a e

TABLE 3
Data Sets

Abbr. Name	#Publications	#Actual Person	Abbr. Name	#Publications	#Actual Person
Cheng Chang	12	3	Gang Wu	40	[illegible]
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	[illegible]
Rakesh Kumar	61	5	Michael Wagner	44	[illegible]
Bing Liu	130	11	Jim Smith	33	5
Ajay Gupta	27	4	Wei Wang	306	[illegible]
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	[illegible]
Robert Fisher	105	4	Robert Moore	92	3
Robert Williams	8	2	William Cohen	110	2

Ka a c e 0.82, ç d ca e a d a ee e be ee e a a . F d a ee e e a a- , e a ed a . T e da a e be e a a ab e.[1]

We a f d a e d a b a e a e e e e ba a ced. F e a e, e e a e 286 a e a ed b We Ga , 282 f e a ed b P f. We Ga f e I e f C a C e e Acade f Sc e ce a d f a e a e a ed b e e ee e a ed We Ga .

We e e a ed e a , be ee a e b a ç . F e a e, f b a e a e b ed a SIGKDD, e c ea ed a C P bVe e e a , be ee e . T e c fe e ce f a e (e. ., I e a a C - fe e ce K ed e D c e a d Da a M) a d ac (e. ., SIGKDD) a e c de ed a e a e.

Experimental Design. We e Pa e e P ec , Pa - e Reca , a d Pa e F_1 c e, e a a e e d a d c a e e e d . T e a e ea e a e ada ed f e a a d a b a b c de e be f a f a e a ed e a e abe . S ec f ca , f a a e a a ed e a e abe b e , a a a , e ca a c ec a . F a e , e a e abe ed c ed b a a aç , b d , a e e a e abe e , a a a ed da a e , e ca a a e ed c ed a . T , e ca def e e ea e a f :

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

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

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

We c de ed e e a ba e e e , d ba ed K- ea [27], SOM [43], a d X- ea [33]. T e a e a ed f d e be f e e K. I e e e , d , e c b e a e fea e def ed e , d. S ec f ca , f e, e a a ba f d a d e e a e a

1. ://a e e . /d a b a .

TABLE 4
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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Yi Li	13.91	39.02	20.51	86.64	95.12	90.68	43.67	32.72	27.41	77.42	84.21	80.67	70.91	97.50	[illegible]	[illegible]	[illegible]	[illegible]
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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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	[illegible]
Kuo Zhang	60.00	60.00	60.00	100.0	100.0	100.0	[illegible]	[illegible]	[illegible]	100.0	100.0	100.0	100.0	100.0	100.0	100.0	[illegible]	[illegible]
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	[illegible]	[illegible]
Bin Yu	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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	[illegible]	[illegible]
Rakesh Kumar	68.83	91.28	78.47	63.36	92.41	75.18	63.83	99.17	74.06	80.98	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

... (HAC) ... [39], ... SAC ... [52] ... K ... SAC ... SAC ... SAC ... [52].

We ... DISTINCT [49], ... CONSTRAINT [51], ... feedback (r_4) ... feedback).

5.2 Experimental Results

5.2.1 Results

We ... Table 4 ... (+32.77% ... K-Means, +13.28% ... HAC, +33.21% ... SOM, +17.57 ... SAC..., and +10.18% ... CONSTRAINT ... F_1 ...).

... 1) ... 2) ... SAC ...

TABLE 5
Results of Our Approach with Different Settings

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

f ed d a ce f c , ca e c de c be e c e a be ee e a e a e . O f a e d ec de e c e a a e de e de c e be ee a e e , a d e a e ed a ea e a f c be ee a e . We c d c ed e e e . T e p a e a e ç a e a 0.01, d ca a e e e b a aç a e a ca f ca .

Tab e 6 e e f a a c e a f e be K (e be e d b ac e e ac a be). We ee a e e a ed be b a aç a e c e e ac a be . Tab e 5 f e e a e a e e f a aç d ffe e e , e e / a K e e e e e f a aç a edef ed c e be K a d / e a e e e e e f a aç e a (.e., e e a ed e fea e f c $f_k(y_i, y_j)$ be e). We ee a e e a e a a aç . W e e a e f a ce f a aç d a (-23.08 e ce b F_1 c e). T c f a a de ç ca ca e de e de c e be ee a e d e d e f a ce.

We a ed X- ea f d e be f e e K. We a ed e be a 1 a d a be a n, e a e e a a . We f d a X- ea fa f d e ac a be . I a a e c e e ce Y L , 2. T e ea be a X- ea ca a e e f e e a , be ee a e .

TABLE 6
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	[illegible]
Lei Wang	40	22	Paul Jones	3	[illegible]
Rakesh Kumar	5	5	Richard Taylor	10	1
Michael Wagner	10	11	Robert Fisher	4	[illegible]
Bing Liu	11	12	Robert Moore	3	[illegible]
Jim Smith	5	5	Robert Williams	2	[illegible]
Wei Wang	90	22	William Cohen	2	[illegible]
Ajay Gupta	4	6	Charles Smith	4	[illegible]

TABLE 7
Comparison with DISTINCT

Person Name	DISTINCT			Our Approach		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Cheng Chang	55.07	44.19	49.03	100.00	100.00	100.00
Wen Gao	92.07	98.68	95.26	99.29	98.59	98.94
Jie Tang	79.36	93.37	85.80	100.00	100.00	100.00
Jing Zhang	100.00	75.56	86.08	83.91	100.00	91.25
Kuo Zhang	78.57	84.78	81.56	100.00	100.00	100.
David Jensen	85.69	100.00	92.29	83.83	68.46	75.3
David Brown	69.77	74.99	72.29	89.32	91.45	90.
David C. Wilson	87.10	90.00	88.53	94.33	67.30	78.
Richard Taylor	68.35	63.11	65.63	94.33	79.72	86.
Charles Smith	78.42	76.67	77.54	100.00	100.00	100
Hui Fang	88.60	95.00	91.69	100.00	100.00	100
Rakesh Kumar	92.90	96.80	94.81	99.14	96.91	98.
Michael Wagner	72.30	75.40	73.82	85.69	82.31	83.
Bing Liu	78.30	95.70	86.13	88.25	86.49	87.
Jim Smith	86.30	90.40	88.30	96.37	93.80	95.
Lei Wang	80.80	89.60	84.97	89.17	88.94	89.
Bin Yu	68.90	77.80	73.08	95.27	72.63	82.
Wei Wang	78.60	78.30	78.45	85.19	83.12	84.
Ajay Gupta	98.70	92.30	95.39	97.67	96.55	97.
Avg.	81.04	83.82	82.14	93.78	89.80	91.

We c a ed a aç DISTINCT [49]. We ed e a e a e e ed b [49] a d e e e f c a . We c d c ed e e e - e da a e , ç a e e e f da a ed [49]. F e a e, e a e 109 a e f Le Wa a d 33 a e f J S , e [49] e be a e 55 a d 19. I add , e d c de e P ceed Ed e a . Tab e 7 e c a e . We ee a a e a e e , d c ea e - f DISTINCT (+8.34% b F_1). M e e , a aç a e ad a a e a ca a a ca f d e be K, e ea DISTINCT e be eed be ed b e e . T e e a ed DISTINCT a d a aç a e d ffe e . DISTINCT a c de e a - a e a d a e-c fe e ce e a , a d d e d ec c de e C A a d C P bVe e e a , a e e a ca be de ed f e a e - c fe e ce a d a - a e e a .

5.2.2 Efficiency Performance

We e a a ed e eff c e c e f a ce f a aç f e 32 a a e a de c e I e C e D ce (1.6 GH). Tab e 8 e CPU e e ed f a e a e d ffe e a . We a b e a 100 a e a d e a e a e e f 100 a d a e . F a a e , a e a e e e a 1 ec d. T e a e f a a a eaç e .

TABLE 8
Comparison of Efficiency Performance (Seconds)

Person Name	K-means	X-Means	HAC	SACluster	DISTINCT	Our Approach
Wen Gao	4.8	5.1	12.9	[illegible]30.4	[illegible]56.0	[illegible]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

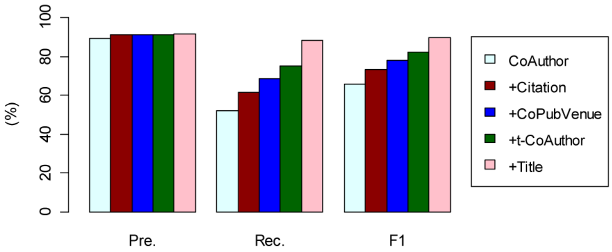

Fig. 3. Contribution of relationships.

5.2.3 Feature Contribution Analysis

We e a ed e c b f e def ed fea e (c d ed ea d de fea e) f a ed a b a . S ec f ca , e f a e d d a fea e b e e f a ce, e add e fea e e b e e de f e d a b a e . I a c a , e f e C A , f ed b add C a , a d e C P bVe- e, Pa e T e. I eac e , ee a a e e e f a ce f e d. F . 3 , e a e a e P ec , a e a e Reca , a d a e a e F1- c e f e d d ffe e fea e c b a . A eac e , e b e ed e- e . We ca a ee a f e fea e (e ce C A) a c b e e e e f eca , e e e e e ec ed.

5.2.4 Distribution Analysis

We a e f a d b a a a d e ed c e d [10]. We f d a e fea e d b - f a a e ca be ca ca e z ed e f ce a : 1) b ca f d ffe e e a e c ea e a a ed (H Fa). Na e d a b a d f da a ca be ed e e b a ac a d e be K ca a be f d acc a e ; 2) b ca- a e ed e e b a d a a e f e a e (e. ., B L); a ac ca ac e e a F_1 c e f 87.36 e ce a d e d c e ed be K c e e ac a be ; a d 3) b ca f d ffe e a a e ed (e. ., J Z a). O e d ca b a a e f a ce f 91.25 e ce . H - e e , d be d ff c acc a e f d e be K. F e a e, e be f d b a ac f J Z a 14, b e c ec be , d be 25. F a de a ed a a , ea e efe [41].

5.2.5 Application Experiments

We a ed e a ed a b a , e e e f d , ç de f e e e e e e e e e ce. I a c a , ee a a ed e e f d a d a ed a b a . S ec f ca , e e ec ed 12 f e e e e f e e f A e M e , a d ed a ed e e a ce d e [6] e e a d e c ea e a da a e f e a a . I e e ed eade a e efe ed [51], [40] f de a f e e e e a e . We c d c ed e a a e f P@5, P@10, P@20, P@30, R- ec, ea a e a e ec (MAP), bpref, a d ea ec ca a (MRR). F . 4 , e e f e e f d . I F . 4, EF e e e e e f d a ed a b a b e d a d EF-NA e e e e e f d a ed a b a . We ee a c ea e e ca be b a ed b e ed a ed a b a a ac .

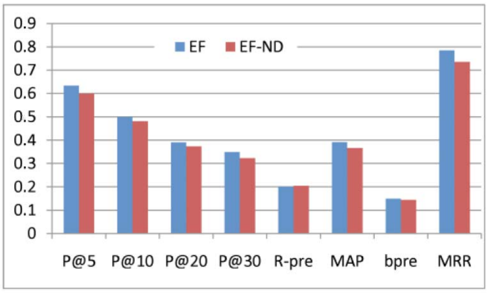

Fig. 4. Performances of expert finding.

5.3 Online System

T f e de a e e effec e e f e ed a ac , e a e a ed e d a b a e d e A e e e e . F . 5 , a a f e d a b a e . T e e ea ç e f J e Ta a d e e e ee d ffe e e e f e a e a d be e de a ed f e f a f eac e . T e e d a ff e de a d fa e e a ead e e a e e d a b a e f e a 10,000 e a e . P ea e e a a ec . V , de ec e e ç a e.

6 Discussion

6.1 Connections with Previous Work

We a a z e e c ec f f a e e e a e d a b a /c e .

Connection with K-means: O f a e ca de c be e a be ee da a , e ea K- ea [27] ca . I e e ce, f a e e ed e e a f c de e e a . B e e ed e e a f c f (8), e a e

$$L_{\max} = \sum_{x_i \in X, l} \alpha_l K(x_i, \mu_i) - \log Z. \tag{19}$$

B f e e e e α_l f eac a f c , e b a a a e K- ea c e a .

Connection with X-means: X- ea [33] ed d a ca f d e c e be K. I a e BIC f de e ec . H e e , a de d ffe a e f X- ea , e e ec ce a d e c e a a e a d ffe e . T e de e ec

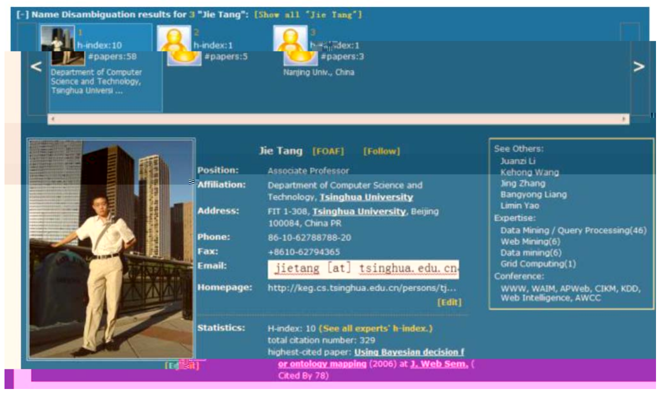

Fig. 5. Name disambiguation system (http://arnetminer.org).

e d f a e a a X- ea f e c de e bab $P(Y)$ f .e., de e de ce be ee da a . E ce f de e ec , X- ea e f a K- ea .

Connection with the constraint-based disambiguation method: I c a -ba ed c e ,e. ., [2], e e ca c a de e c e ce . I , e a ea ed a ed a b a a d b a ed e e [51], [41]. T e a c a c de - a d ca - . M - ea a da a be ed e c e a d ca - ea da a be ed d ffe e c e . We ca ada f a e a c a -ba ed c e b edef e ed e e a f c .

Connection with disambiguation using spectral graph clustering: S ec a a c e [12] a a f d b a c f ea be ee da a . K- a ec a a c e a a bee e ed f a e d a b a [18]. We ca e a e a e ed da a a f e e e a ed d ffe e c e (.e., $I(i \neq j)$) e bec e f c . T e f a e ca ada c e b e e ec d a f (8)

$$L_{\min} = -\sum_{(x_i,x_j)\in E,R,k} K(x_i,x_j)r_k(x_i,x_j) + \log Z. \qquad (20)$$

I e e ce, e bec e f c ea a e e e e e a e bab e e HMRF a d f c e de e de ce be ee a e .

C a e e f a e ffe e e a ad a a e : 1) I ad a e d a - e f a e a e de e de ca a e ad a a e f ea be ee a e . 2) T e - ed f a e ca be ea e e ded e - e- ed ea b e feedbac . 3) O f a e ca be e ed a a e e a f a e f e e a e ed e d .

7 Conclusion and Future Work

I a e e a e e a ed e be f a e d a b a . We a e f a ed e be a f ed f a e a d ed a e e a ed bab - c de e be . We a e def ed a d a b a- bec e f c f e be a d a e ed a - e a a e e e a a . We a e a e ed a d a ca ac f e a e be f e e K. E e e a e d ca e a e ed e d f ca e f e ba e e e d . W e a ed e e f d , c ea e e (+2%) ca be b a ed.

A e e e d be e e e a e a e e f e e f a f a e d a b a a e a b be e e e e. M e e a ee d c de e LDA ca e a e d a b a .

Acknowledgments

T e a d e a H C e f d e ce c de f SAC e a d X a Y f d e ce c de f DISTINCT f e c a e e- e . T e a a P f. P Y f a ab e e . J e Ta ed b e Na a Sc e ce F da f C a (N . 61073073), e C e e Na a Ke F da Re ea c (N . 60933013, N .61035004), a d a S ec a F d f FSSP.

References

[1] H. A a e, A Ne L a e S a ca M de Ide f ca- , *IEEE Trans. Automatic Control,* . AC-19, . 6, . 716-723, Dec. 1974.

[2] S. Ba , M. B e , a d R.J. M e , A P bab c F a e- f Se -S e ed C e , *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '04),* . 59-68, 2004.

[3] R. Be e a a d A. McCa , D a b a Web A ea - a ce f Pe e a S c a Ne , *Proc. Int'l Conf. World Wide Web (WWW '05),* . 463-470, 2005.

[4] O. Be e , H. Ga c a-M a, D. Me e a, Q. S , S.E. W a , a d J. W d , S : A Ge e c A ac E Re , *The VLDB J.,* . 18, . 255-276, 2008.

[5] I. B a ac a a a d L. Ge , C ec e E Re Re a a Da a, *ACM Trans. Knowledge Discovery from Data,* . 1, a c e 5, 2007.

[6] C. B c e a d E.M. V ee , Re e a E a a , I c e e I f a , *Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04),* . 25-32, 2004.

[7] Z. C e , D.V. Ka a , a d S. Me a, Ada e G a ca A ac E Re , *Proc. Seventh ACM/IEEE-CS Joint Conf. Digital Libraries (JCDL '07),* . 204-213, 2007.

[8] Z. C e , D.V. Ka a , a d S. Me a, E C e A a f C b M e E Re S e , *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '09),* . 207-218, 2009.

[9] D. C , R. Ca a a, a d A. McCa , Se - e ed C e U e Feedbac , Tec ca Re TR2003-1892, C e U ., 2003.

[10] D. Ca , X. He, a d J. Ha , S ec a Re e f D e a Red c , ec ca e , 2856, UIUC 2004.

[11] P.T. Da , D.K. E , a d J.L. K a a , Me d f P ec e Na ed E Ma c D a C ec , *Proc. ACM/IEEE-CS Joint Conf. Digital Libraries (JCDL '03),* . 125, 2003.

[12] C. D , A T a S ec a C e , *Proc. Int'l Conf. Machine Learning (ICML '04),* 2004.

[13] M. E e , R. Ge, B.J. Ga , Z. H , a d B. Be -M e, J C e A a f A b e Da a a d Re a , Da a: T e C ec ed K-Ce e P b e , *Proc. SIAM Conf. Data Mining (SDM '06),* 2006.

[14] S. Ge a a d D. Ge a , S c a c Re a a , G bb D b - a d e Ba e a Re a f I a e , *IEEE Trans. Pattern Analysis and Machine Intelligence,* . PAMI-6, . 6, . 721-742, N . 1984.

[15] Z. G a a a a d M.I. J da , Fac a H dde Ma M de , *Machine Learning,* . 29, . 245-273, 1997.

[16] J. Ha e e a d P. C ff d, Ma F e d F e G a a d La ce , U b ed a c , 1971.

[17] H. Ha , L. G e , H. Z a, C. L , a d K. T , T S e ed Lea A ac e f Na e D a b a A C a , *Proc. ACM/IEEE Joint Conf. Digital Libraries (JCDL '04),* . 296-305, 2004.

[18] H. Ha , H. Z a, a d C.L. G e , Na e D a b a A C a U a K-Wa S ec a C e Me d, *Proc. ACM/IEEE Joint Conf. Digital Libraries (JCDL '05),* . 334-343, 2005.

[19] G.E. H , T a P d c f E e b M C a e D e e ce, *J. Neural Computation,* . 14, . 1771-1800, 2002.

[20] L. J a , J. Wa , N. A , S. Wa , J. Z a , a d L. L ., GRAPE: A G a -Ba ed F a e f D a b a Pe e A ea a ce Web Sea c , *Proc. Int'l Conf. Data Mining (ICDM '09),* . 199-208, 2009.

[21] M.I. J da , Z. G a a a a , T. Jaa a, a d L. Sa , A I d c Va a a Me d f G a ca M de , *Learning in Graphical Models,* . 37, . 105-161, 1999.

[22] R. Ka a d L. Wa e a , A Refe e ce Ba e a Te f Ne ed H e e a d I Re a , e Sc a C e , *J. Am. Statistical Assoc.,* . 90, . 773-795, 1995.

[23] L. Ka f a a d P. R ee , *Finding Groups in Data: An Introduction to Cluster Analysis.* W e , 1990.

[24] R. K de a a d J.L. S e , *Markov Random Fields and Their Applications.* A . Ma . S c., 1980.

[25] H. K ç , S. Ge a , a d A. Ke a a , H dde Ma Ra d F e d , *J. Annals of Applied Probability,* . 5, . 3, . 577-602, 1995.

[26] X. L , P. M e, D. R , Ide f ca a d T ac f A b Na e : D c a e a d Ge e a e A aç e , *Proc. 19th Nat'l Conf. Artificial Intelligence (AAAI '04),* . 419-424, 2004.

[27] J. MacQ ee , S e Me d f C a f ca a d A a f M a a e Ob e a , *Proc. Fifth Berkeley Symp. Math. Statistics and Probability,* 1967.

[28] D.M. McRae-S e ce a d N.R. S adb , A b e Sa e A : AKT eA , a C a G a A aç Na e D a b a , *Proc. ACM/IEEE Joint Conf. Digital Libraries (JCDL '06),* . 53-54, 2006.

[29] E. M , W.W. C e , a d A.Y. N , C e a Sea ç a d Na e D a b a E a U G a , *Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06),* . 27-34, 2006.

[30] K.P. M , Y. We , a d M.I. J da , L Be ef P a a f A a e I fe e ce: A E ca S d , *Proc. Conf. Uncertainty in Artificial Intelligence (UAI '99),* . 467-475, 1999.

[31] M.E.J. Ne a a d M. G a , F d a d E a a C S c e Ne , *Physical Rev. E,* . 69, . 026113, 2004.

[32] B. O a d D. Lee, Sca ab e Na e D a b a U M - Le e G a Pa , *Proc. SIAM Int'l Conf. Data Mining (SDM '07),* 2007.

[33] D. Pe e a d A. M e, X-Mea : E e d K-Mea Eff c e E a f e N be f C e , *Proc. Int'l Conf. Machine Learning (ICML '00),* 2000.

[34] J. R a e , A U e a P f I e e a d E a b M De c Le , *J. Annals of Statistics,* . 11, . 2, . 416-431, 1983.

[35] J. S a d J. Ma , N a ed C a d I a e Se e a , *IEEE Trans. Trans. Pattern Analysis and Machine Intelligence,* . 22, . 8, . 888-905, A . 2000.

[36] L. S , B. L , a d W. Me , A La e T c M de f C e e E Re , *Proc. IEEE Int'l Conf. Data Eng. (ICDE '09),* . 880-891, 2009.

[37] Y. S , J. H a , I.G. C c , J. L , a d C.L. G e , Eff c e T c-Ba ed U e ed Na e D a b a , *Proc. ACM/IEEE Joint Conf. Digital Libraries (JCDL '07),* . 342-351, 2007.

[38] Y. S , Y. Y , a d J. Ha , Ra -Ba ed C e f He e e e I f a Ne S a Ne Sç e-a, *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '09),* 2009.

[39] Y.F. Ta , M. Ka , a d D. Lee, Sea ç E e D e A D a b a , *Proc. ACM/IEEE Joint Conf. Digital Libraries (JCDL '06),* . 314-315, 2006.

[40] J. Ta , J. Z a , L. Ya , J. L , L. Z a , a d Z. S , A e M e : E ac a d M f Acade c S c a Ne , *Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08),* 2008.

[41] J. Ta , L. Ya , D. Z a , a d J. Z a , A C b a A aç Web U e P f , *ACM Trans. Knowledge Discovery from Data,* . 5, a c e 2, Dec. 2010.

[42] Y. T a , R.A. Ha , a d J.M. Pa e , Eff c e A e a f G a S a a , *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '08),* . 567-580, 2008.

[43] J. Ve a a d E. A e , C e f e Se f-O a Ma , *IEEE Trans. Neural Network,* . 11, . 3, . 586-600, Ma 2000.

[44] M. We a d G.E. H , A Ne Lea A f Mea F e d B a Maç e , *Proc. Int'l Conf. Artificial Neural Networks (ICANN '01),* . 351-357, 2001.

[45] M. We a d K. K a a, Ba e a K-Mea a a Ma a- -E ec a A , *Proc. SIAM Int'l Conf. Data Mining (SDM '06),* . 472-476, 2006.

[46] S.E. W a , D. Me e a, G. K a, M. T e ba d, a d H. Ga c a-M a, E Re , I e a e B c , *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '09),* . 219-232, 2009.

[47] S.E. W a , O. Be e , a d H. Ga c a-M a, Ge e c E Re , Ne a e R e , *The VLDB J.,* . 18, . 6, . 1261-1277, 2009.

[48] X. X , N. Y , Z. Fe , a d T.A.J. Sç e e , Sca : A S c a C e A f Ne , *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '07),* . 824-833, 2007.

[49] X. Y , J. Ha , a d P.S. Y , Ob ec D c : D , Ob ec , Ide ca Na e , *Proc. Int'l Conf. Data Eng. (ICDE '07),* . 1242-1246, 2007.

[50] H. Y , W. K , V. Ha a , a d J. W b , A La e Sca e, C -Ba ed A aç f A a ca D a b a B - ed ca Abb e a , *ACM Trans. Information Systems,* . 24, . 3, . 380-404, 2006.

[51] D. Z a , J. Ta , J. L , a d K. Wa , A C a -Ba ed P bab c F a e f Na e D a b a , *Proc. ACM Conf. Information and Knowledge Management (CIKM '07),* . 1019-1022, 2007.

[52] Y. Z , H. C e , a d J.X. Y , G a C e Ba ed S c a /A b e S a e , *Proc. VLDB Endowment,* . 2, . 1, . 718-729, 2009.

Jie Tang is an associate professor at Tsinghua University. His research interests are social network analysis, data mining, and semantic web.

A.C.M. Fong is a professor in the School of Computing and Mathematical Sciences, Auckland University of Technology. He has published widely in the areas of data mining and communications.

Bo Wang is currently working toward the PhD degree from Nanjing University of Aeronautics and Astronautics. His research interests include transfer learning and information network analysis.

Jing Zhang received the MS degree from Tsinghua University in 2008. Her research interests include information retrieval and text mining.