ACTIVELY DISAMBIGUATING PERSON NAMES WITH USER INTERACTION

Xuezhi Wang*, Jie Tang*, Hong Cheng‡, Philip S. Yu†

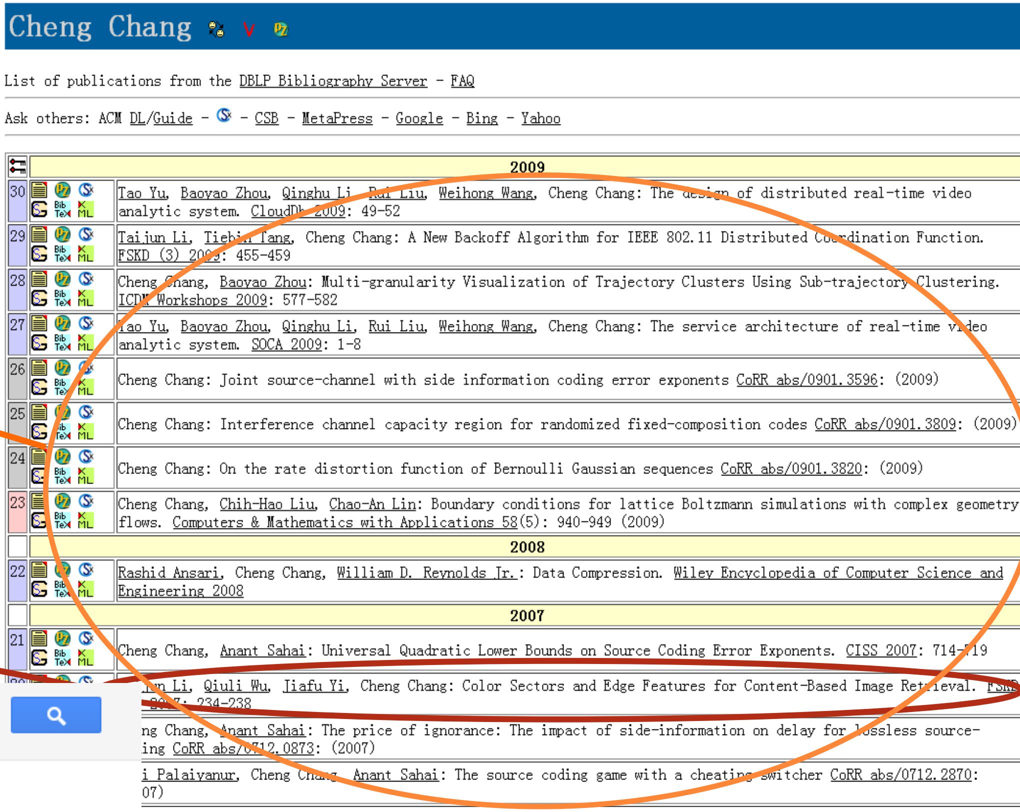
*Tsinghua University, ‡The Chinese University of Hong Kong,
†University of Illinois at Chicago

MOTIVATION

Search an author in DBLP

Do these papers really belong to Cheng Chang, student from Tsinghua and later went to Berkeley?

This paper actually belongs to Cheng Chang, from Hainan University.


Cheng Chang

List of publications from the DBLP Bibliography Server - FAQ

Ask others: ACM DL/Guide - CSB - MetaPress - Google - Bing - Yahoo

2009

30 Tao Yu, Baoyao Zhou, Qinghu Li, Rui Liu, Weihong Wang, Cheng Chang: The design of distributed real-time video analytic system. CloudDb 2009: 49-52

29 Taijun Li, Tiehin Tang, Cheng Chang: A New Backoff Algorithm for IEEE 802.11 Distributed Coordination Function. FSKD (3) 2009: 455-459

28 Cheng Chang, Baoyao Zhou: Multi-granularity Visualization of Trajectory Clusters Using Sub-trajectory Clustering. ICDM Workshops 2009: 577-582

27 Tao Yu, Baoyao Zhou, Qinghu Li, Rui Liu, Weihong Wang, Cheng Chang: The service architecture of real-time video analytic system. SOCA 2009: 1-8

26 Cheng Chang: Joint source-channel with side information coding error exponents CoRR abs/0901.3596: (2009)

25 Cheng Chang: Interference channel capacity region for randomized fixed-composition codes CoRR abs/0901.3809: (2009)

24 Cheng Chang: On the rate distortion function of Bernoulli Gaussian sequences CoRR abs/0901.3820: (2009)

23 Cheng Chang, Chih-Hao Liu, Chao-An Lin: Boundary conditions for lattice Boltzmann simulations with complex geometry flows. Computers & Mathematics with Applications 58(5): 940-949 (2009)

2008

22 Rashid Ansari, Cheng Chang, William D. Reynolds Jr.: Data Compression. Wiley Encyclopedia of Computer Science and Engineering 2008

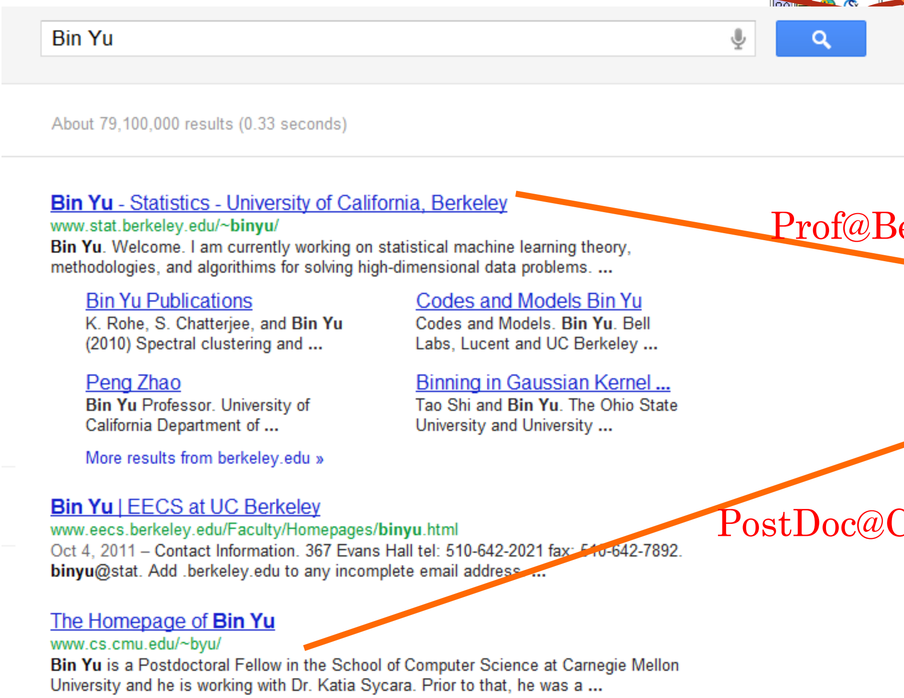
2007

21 Cheng Chang, Anant Sahai: Universal Quadratic Lower Bounds on Source Coding Error Exponents. CISS 2007: 714-719

...n Li, Qiuli Wu, Jiafu Yi, Cheng Chang: Color Sectors and Edge Features for Content-Based Image Retrieval. FSKD 2007: 234-238

...ng Chang, Anant Sahai: The price of ignorance: The impact of side-information on delay for lossless source-...ing CoRR abs/0712.0873: (2007)

...i Palaiyanur, Cheng Chang, Anant Sahai: The source coding game with a cheating switcher CoRR abs/0712.2870: ...07)


Bin Yu

About 79,100,000 results (0.33 seconds)

Bin Yu - Statistics - University of California, Berkeley
www.stat.berkeley.edu/~binyu/
Bin Yu. Welcome. I am currently working on statistical machine learning theory, methodologies, and algorithms for solving high-dimensional data problems. ...

Bin Yu Publications
K. Rohe, S. Chatterjee, and Bin Yu (2010) Spectral clustering and ...

Codes and Models Bin Yu
Codes and Models. Bin Yu. Bell Labs, Lucent and UC Berkeley ...

Peng Zhao
Bin Yu Professor. University of California Department of ...

Binning in Gaussian Kernel ...
Tao Shi and Bin Yu. The Ohio State University and University ...

More results from berkeley.edu »

Bin Yu | EECS at UC Berkeley
www.eecs.berkeley.edu/Faculty/Homepages/binyu.html
Oct 4, 2011 – Contact Information. 367 Evans Hall tel: 510-642-2021 fax: 510-642-7892. binyu@stat. Add .berkeley.edu to any incomplete email address ...

The Homepage of Bin Yu
www.cs.cmu.edu/~byu/
Bin Yu is a Postdoctoral Fellow in the School of Computer Science at Carnegie Mellon University and he is working with Dr. Katia Sycara. Prior to that, he was a ...

Prof@Berkeley

Search a name in a search engine

Which Bin Yu do you want to find?

PostDoc@CMU

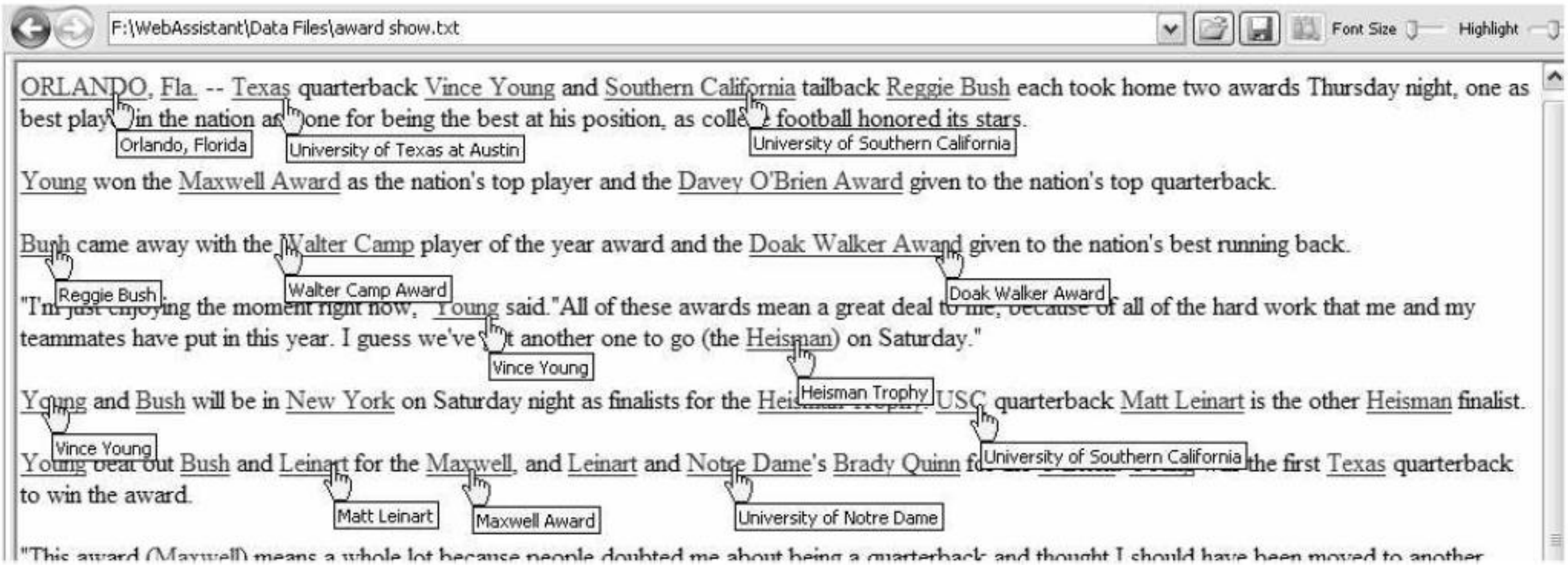
Existing Methods For Name Disambiguation

- Supervised-based approach:
  - Learn a specific classification model from training data
  - Use model to predict the assignment of each paper

- Unsupervised-based approach:
  - Clustering algorithms to find paper partitions.
  - Papers in different partitions are assigned to different persons.

- Constraint-based approach:
  - Utilizes the clustering algorithms.
  - User-provided constraints are used to guide the clustering towards better data partitioning.

Existing Methods with Interaction



- Several problems:
  - User has to check every result to see if it is correct
  - No propagation, correction only based on user input

ALGORITHM DESIGN

- How to combine features, relations and user feedback?

- Feature, between document pair and label
- Relation, between label and label
- User Feedback, constraint on partial labels

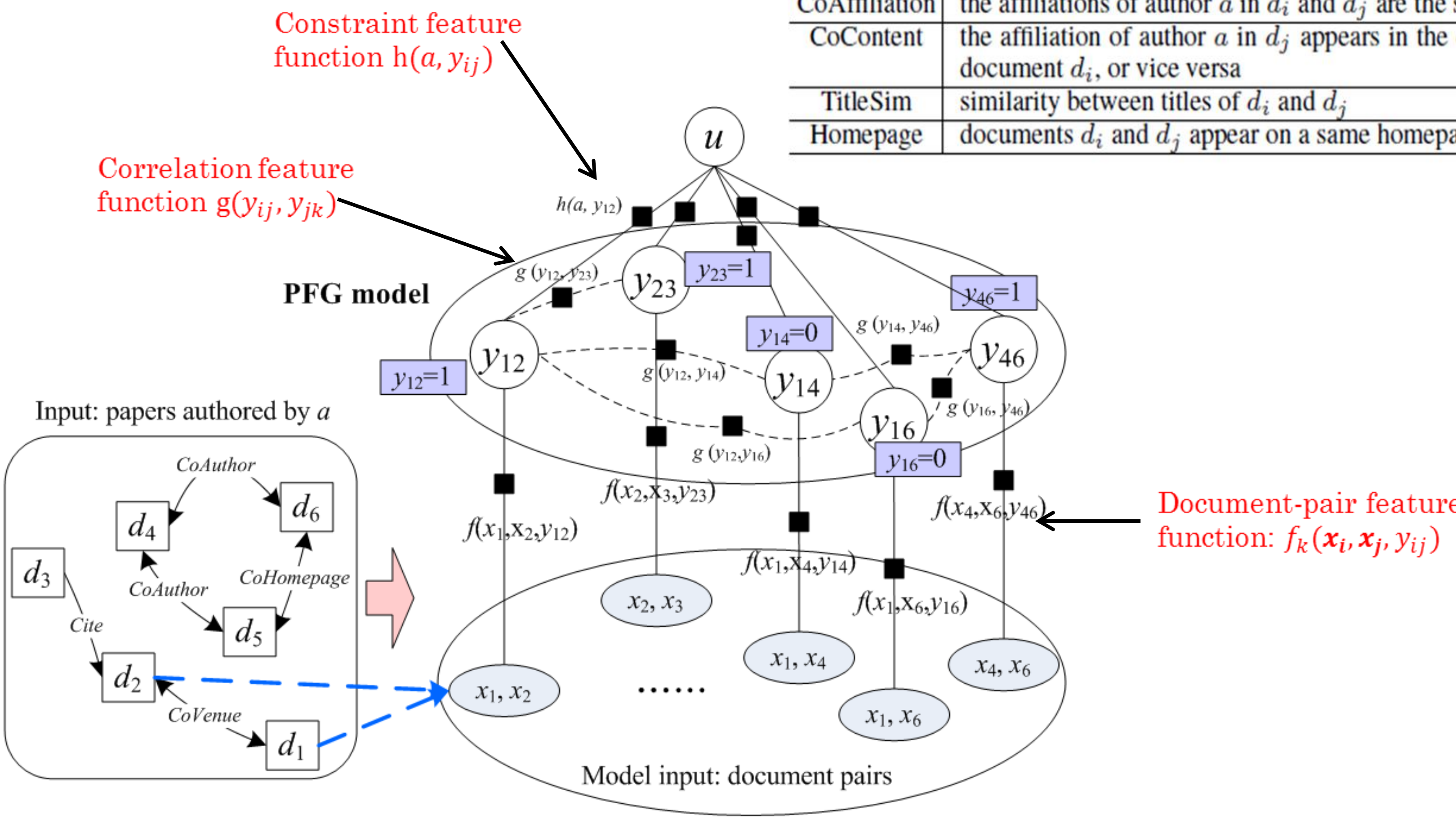
- We need a model to elegantly combine these altogether
- Inference on the model can give us the answer to paper assignment

Algorithm Design

—Pairwise Factor Graph Model

Feature Description

Name	Description
Citation	document d_i cites d_j in the reference, or vice versa.
CoAuthor	d_i and d_j share at least one coauthor (except author a)
CoVenue	d_i and d_j are published at the same venue (journal or conference)
CoAffiliation	the affiliations of author a in d_i and d_j are the same
CoContent	the affiliation of author a in d_j appears in the content of document d_i, or vice versa
TitleSim	similarity between titles of d_i and d_j
Homepage	documents d_i and d_j appear on a same homepage



$$p(Y|X) = \frac{1}{Z}\exp\{\sum_{i\neq j}\sum_k w_k f_k(\boldsymbol{x_i}, \boldsymbol{x_j}, y_{ij}) + \sum_{e(ij,jk)\in E}\mu g(y_{ij}, y_{jk}) + \sum_l \alpha_l h_l(a, y_{ij})\}$$

Learning Algorithm for PFG

Input: number of iterations;
Output: learned configuration for Y;

Initialize all $\theta = (\{w_k\}, \{\mu\}, \{\alpha_l\})$ as 1;
Initialize all hidden variables $Y = \{y_{ij}\}$ with $y_{ij} = 0$;
repeat
    % sample a new configuration Y' based on $q(Y'|Y)$;
    $Y' \leftarrow q(Y'|Y)$;
    $\tau \sim \min(\frac{p(Y',X|\theta)}{p(Y,X|\theta)}, 1)$;
    toss a coin s according to a $Bernoulli(\tau, (1-\tau))$;
    if $(s = 1)$ **then**
        % accept the new configuration Y';
        $Y \leftarrow Y'$;
    end
until *convergence*;
return Y;

2: The MH-based learning algorithm for PFG.

Metropolis-Hasting Algorithm for Approximate Inference

Why Active Name Disambiguation?

Wei Wang
(University of North Carolina at Chapel Hill)

2011

[512] Xiuhua Guo, Xiangye Liu, Huan Wang, Zhigang Liang, Wei Wu, Qian He, Kuncheng Li, Wei Wang. Enhanced CT Images by the Wavelet Transform Improving Diagnostic Accuracy of Chest Nodules.. - J. Digital Imaging, 2011

[511] Yisha Liu, Zidong Wang, Wei Wang. Reliable H filtering for discrete time-delay systems with randomly occurred nonlinearities via delay-partitioning method.. - Signal Processing, 2011

[510] Jun Zhao, Quanli Liu, Wei Wang, Zhuoqun Wei, Peng Shi. A parallel immune algorithm for traveling salesman problem and its application on cold rolling scheduling.. - Inf. Sci., 2011

[509] Jianping Zeng, Jiangjiao Duan, Wei Wang, Chengrong Wu. Semantic multi-grain mixture topic model for text analysis.. - Expert Syst. Appl., 2011

[508] Xiaoping Zhang, Jun Zhao, Wei Wang, Liqun Cong, Weimin Feng. An optimal method for prediction and adjustment on byproduct gas holder in steel industry.. - Expert Syst. Appl., 2011

2010

[507] Wei Wang, Jiong Yang. Mining High-Dimensional Data.. - , 2010

[506] Ning Jin, Calvin Young, Wei Wang. Discriminative Subgraph Mining for Protein Classification.. - IJKDB, 2010

[505] Xiangliang Zhang, Wei Wang. Self-adaptive Change Detection in Streaming Data with Non-stationary Distribution.. - Database Technologies: Concepts, Methodologies, Tools, and Applications, 2010

[504] Decong Li, Sujian Li, Wenjie Li, Wei Wang, Weiguang Qu. A Semi-Supervised Key Phrase Extraction Approach: Learning from Title Phrases through a Document Semantic Network.. - ACL (Short Papers), 2010

[503] Wei Wang, Pengtao Zhang, Ying Tan. An Immune Concentration Based Virus Detection Approach Using Particle Swarm Optimization.. - ICSI (1), 2010

[502] Yong Wang, Huihui Zhao, Jianxin Chen, Chun Li, Wenjing Chuo, Shuzhen Guo, Junda Yu, Wei Wang. Classification and Diagnosis of Syndromes in Chinese Medicine in the Context of Coronary Heart Disease Model Based on Data Mining Methods.. - LSMS/ICSEE, 2010

[501] Jiaqi Zhang, Zhenying He, Yue Tao, Xiansheng Wang, Hao Hu, Wei Wang. Efficient SLCA-Based Keyword Search on XML Databases: An Iterative-Skip Approach.. - DEXA (1), 2010

[500] Wei Wang, X. Z. Gao. A single neuron PID controller based on

Order by "Wei Wang".
Wei Wang
(Others)

Dubious

[352] Wei Wang, Hong Zhao, Qiang Li, Zhixiong Liu. A Novel Hybrid Intelligent Model for Financial Time Series Forecasting and Its Application.. - BIFE, 2009

[351] Wei Wang, Hong Zhao, Qiang Li, Zhixiong Liu. Engineering Signals' Blind Source Separation in Frequency Domain and Its Application.. - ICNC (2), 2009

[350] Wei Wang, Hong Zhao, Chunhong Zhu, Qiang Li. Study of a Novel Algorithm for Incipient Fault Diagnosis and Its Application.. - WKDD, 2010

[349] Wei Wang, Tinglei Huang, Hui Liu, Fei Pang. Localization Algorithm Based on SVM-Data Fusion in Wireless Sensor Networks.. - WGEC, 2009

[348] Chuanqi Li, Wei Wang. Projection Pursuit Model Based on Complex Algorithm and its Application in Water Quality Evaluation of Yuqing Lake Reservoir.. - FSKD (2), 2009

2011

[347] Kaiquan S. J. Xu, Wei Wang, Jimmy Ren, Jin S. Y. Xu, Long Liu, Stephen Liao. Classifying Consumer Comparison Opinions to Uncover Product Strengths and Weaknesses.. - IJIIT, 2011

[346] Wei Wang, Xiao Long Xin. On fuzzy filters of pseudo BL-algebras.. - Fuzzy Sets and Systems, 2011

[345] Changyou Wang, Shu Wang, Wei Wang. Global asymptotic stability of equilibrium point for a family of rational difference equations.. - Appl. Math. Lett., 2011

[344] Ting Zhou, Hamid Sharif, Michael Hempel, Puttipong Mahasukhon, Wei Wang, Tao Ma. A Novel Adaptive Distributed Cooperative Relaying MAC Protocol for Vehicular Networks.. - IEEE Journal on Selected Areas in Communications, 2011

2010

[343] Ke Zhu, Ying Zhang, Xuemin Lin, Gaoping Zhu, Wei Wang. NOVA: A Novel and Efficient Framework for Finding Subgraph Isomorphism Mappings in Large Graphs.. - DASFAA (1), 2010

[342] Wenjie Zhang, Xuemin Lin, Ying Zhang, Jian Pei, Wei Wang. Threshold-based probabilistic top- dominating queries.. - VLDB J., 2010

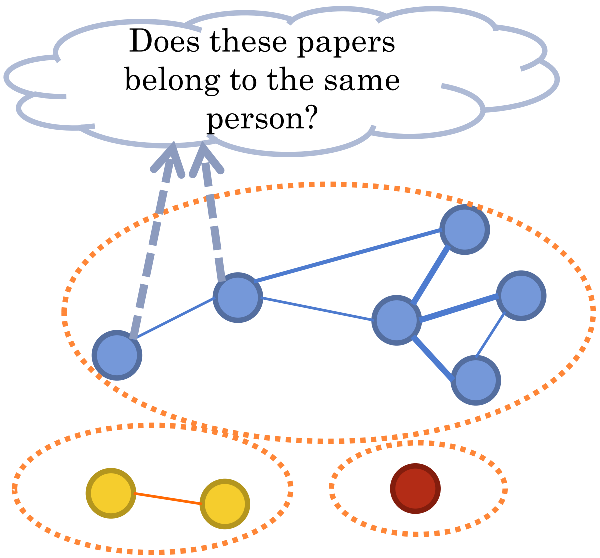
[341] Muhammad Aamir Cheema, Xuemin Lin, Wei Wang, Wenjie Zhang, Jian Pei. Probabilistic Reverse Nearest Neighbor Queries on Uncertain Data.. - IEEE Trans. Knowl. Data Eng., 2010

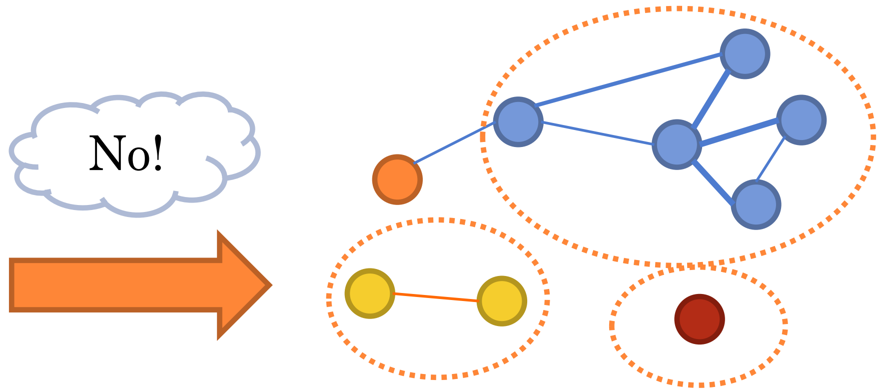
Are they correct?

How to find document pairs that are most likely to be wrongly classified?

Uncertainty-based Active Selection

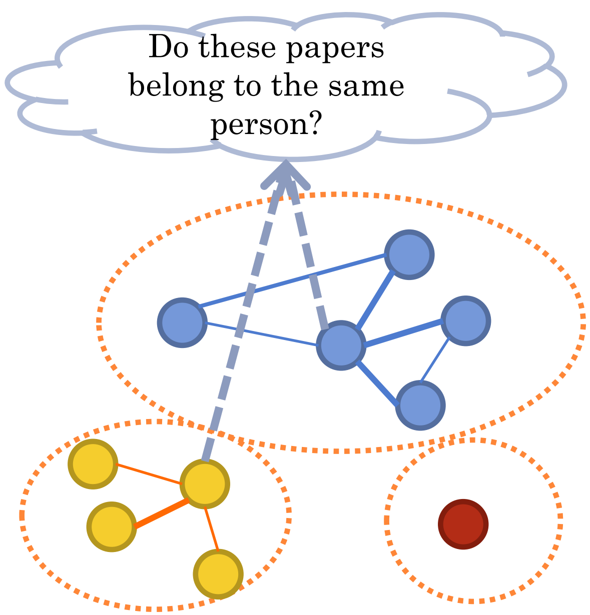
- $p(y_{ij} = 1 | \boldsymbol{x_i}, \boldsymbol{x_j}, \theta)$

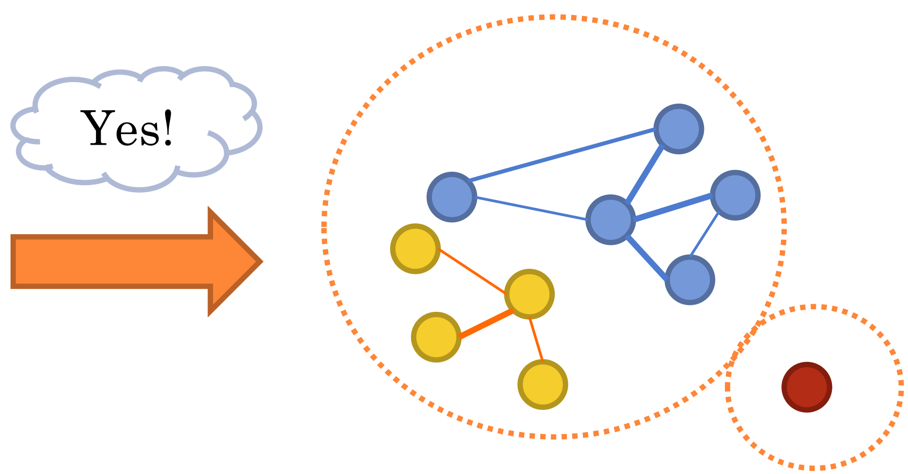




Influence Maximization-based Active Selection

- $max_{y_{ij}} \sum_{e(ij,jk)\in E} \mu g(y_{ij}, y_{jk})$





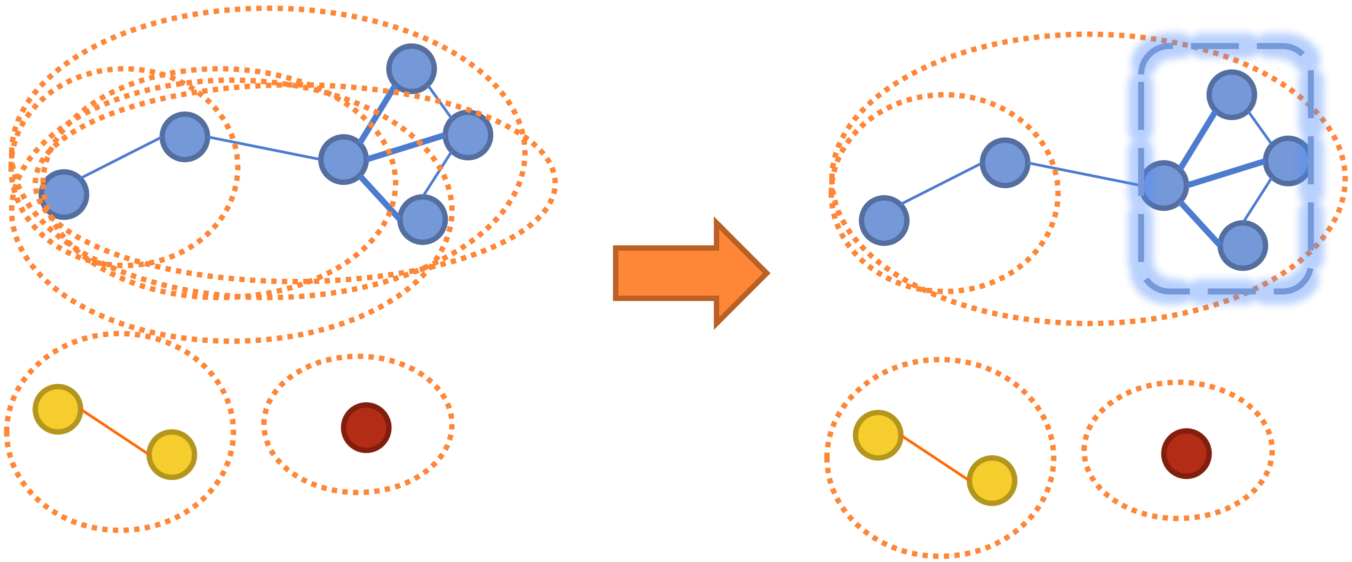
Model Refinement

- Maximizing the conditional probability P(Y|X)

- SampleRank algorithm
- for $\theta \in \{w_k, \mu, \alpha_l\}$, parameters in our PFG model
- y : original configuration; y': new configuration
- $\theta = \theta \begin{cases} -\eta \cdot \phi_{y',y} & if\ y \text{ is preferred } and\ M(y',y) > 0 \\ +\eta \cdot \phi_{y',y} & if\ y' \text{ is preferred } and\ M(y',y) \leq 0 \end{cases}$
- where η is the learning rate
- $M(y',y) = \theta \cdot \phi_{y',y}$ is the unnormalized log probability ratio according to the Metropolis-Hastings Model

Improving Efficiency by Atomic Cluster

- In practice, enumerating all possible document pairs can be really time-consuming and infeasible for an online system
- Atomic cluster-based method
  - Atomic cluster: in this cluster every paper has very high probability that they belong to the same person
  - Bias-classifier——AdaboostM1, aiming to minimize the number of false positives, thus obtaining very high precision



Data Set

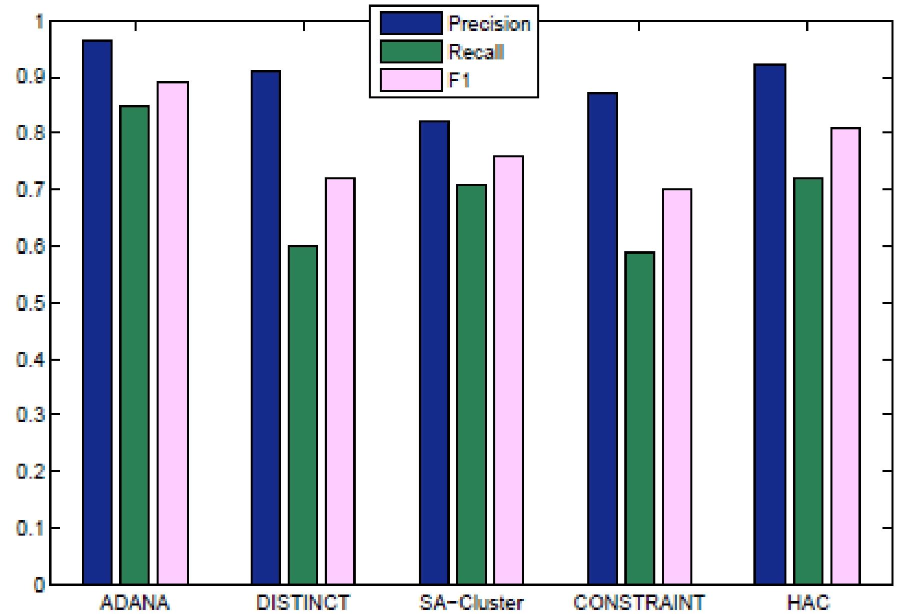
- Publication Data Set
  - From ArnetMiner.org, manually labeled 6,730 papers for 100 author names
- CALO Set
  - Email Directory, labeled data set of 1,085 webpages for 12 names
- News Stories
  - 755 ambiguous entities appearing in 20 web pages

Dataset	#Names	#Persons	#Documents
Publication	100	1,382	6,730
CALO	12	187	1,085
News Stories	380	755	20

EXPERIMENT

Publication Data Set (Average)

Precision	95.4%
Recall	85.6%
F1-score	89.2%



Performance(F1-score) of the comparison methods.

CALO Set

Method	Recall	Precision	F1-score
LS+A/CDC [5]	0.745	0.869	0.803
Our Approach	0.761	0.878	0.815

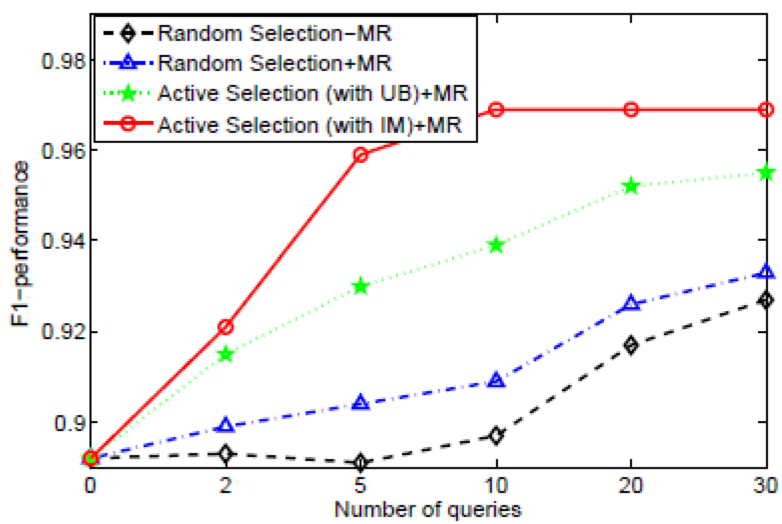
News Data Set

Method	Baseline in [23]	Approach in [23]	Our Approach
Accuracy	0.517	0.914	0.973

- Result of active name disambiguation (MR: the model refinement)
  - UB: Uncertainty-based active selection
  - IM: Influence Maximization-based active selection

Method	Random Selection-MR			Random Selection+MR			Active Selection (with UB)+MR			Active Selection (with IM)+MR		
#Query	Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score
0	0.856	0.954	0.892	0.856	0.954	0.892	0.856	0.954	0.892	0.856	0.954	0.892
2	0.857	0.954	0.893	0.867	0.953	0.899	0.896	0.953	0.915	0.892	0.955	0.921
5	0.855	0.954	0.891	0.873	0.953	0.904	0.922	0.952	0.930	0.976	0.953	0.959
10	0.863	0.956	0.897	0.885	0.951	0.909	0.937	0.953	0.939	0.994	0.952	0.969
20	0.889	0.963	0.917	0.905	0.959	0.926	0.958	0.953	0.952	0.996	0.951	0.969
30	0.903	0.964	0.927	0.915	0.961	0.933	0.965	0.953	0.955	0.997	0.951	0.969

- How F1-score varies with number of queries



Thank you!