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.

Search a name in a search engine

Which Bin Yu do you want to find?
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.
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

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

**Mathematical Model**

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

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

1. Initialize all $\theta = \{w_k\}, \{\mu\}, \{\alpha_j\}$ as 1;
2. Initialize all hidden variables $Y = \{y_{ij}\}$ with $y_{ij} = 0$;

3. repeat
   4. % sample a new configuration $Y'$ based on $q(Y'|Y)$;
   5. $Y' \leftarrow q(Y'|Y)$;
   6. $\tau \sim \min\left(\frac{p(Y', X|\theta)}{p(Y, X|\theta)}, 1\right)$;
   7. toss a coin $s$ according to a $Bernoulli(\tau, 1 - \tau)$;
   8. if ($s = 1$) then
      9. % accept the new configuration $Y'$;
      10. $Y \leftarrow Y'$;
   11. end
4. until convergence;
5. return $Y$;

2: The MH-based learning algorithm for PFG.
WHY ACTIVE NAME DISAMBIGUATION?

Are they correct?

How to find document pairs that are most likely to be wrongly classified?
Uncertainty-based Active Selection

- Does these papers belong to the same person?

\[ p(y_{ij} = 1| x_i, x_j, \theta) \]

Influence Maximization-based Active Selection

- Do these papers belong to the same person?

\[ \max_{y_{ij}} \sum_{e(i,j,k) \in E} \mu_g(y_{ij}, y_{jk}) \]
MODEL REFINEMENT

- Maximizing the conditional probability $P(Y \mid X)$

- SampleRank algorithm
  - for $\theta \in \{w_k, \mu, \alpha_l\}$, parameters in our PFG model
  - $y$: original configuration; $y'$: new configuration
  - $\theta = \begin{cases} 
  -\eta \cdot \phi_{y',y} & \text{if } y \text{ is preferred and } M(y',y) > 0 \\
  +\eta \cdot \phi_{y',y} & \text{if } y' \text{ is preferred and } M(y',y) \leq 0
  \end{cases}$

- where $\eta$ 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Names</th>
<th>#Persons</th>
<th>#Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication</td>
<td>100</td>
<td>1,382</td>
<td>6,730</td>
</tr>
<tr>
<td>CALO</td>
<td>12</td>
<td>187</td>
<td>1,085</td>
</tr>
<tr>
<td>News Stories</td>
<td>380</td>
<td>755</td>
<td>20</td>
</tr>
</tbody>
</table>
EXPERIMENT

Publication Data Set (Average)

- Precision: 95.4%
- Recall: 85.6%
- F1-score: 89.2%

CALO Set

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS+A/CDC$^{[5]}$</td>
<td>0.745</td>
<td>0.869</td>
<td>0.803</td>
</tr>
<tr>
<td>Our Approach</td>
<td>0.761</td>
<td>0.878</td>
<td>0.815</td>
</tr>
</tbody>
</table>

News Data Set

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.517</td>
<td>0.914</td>
<td>0.973</td>
</tr>
</tbody>
</table>
Result of active name disambiguation (MR: the model refinement)

- UB: Uncertainty-based active selection
- IM: Influence Maximization-based active selection

<table>
<thead>
<tr>
<th>Method</th>
<th>Random Selection-MR</th>
<th>Random Selection+MR</th>
<th>Active Selection (with UB)+MR</th>
<th>Active Selection (with IM)+MR</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Query</td>
<td>Recall</td>
<td>Precision</td>
<td>F1-score</td>
<td>Recall</td>
</tr>
<tr>
<td>0</td>
<td>0.856</td>
<td>0.954</td>
<td>0.892</td>
<td>0.856</td>
</tr>
<tr>
<td>2</td>
<td>0.857</td>
<td>0.954</td>
<td>0.893</td>
<td>0.867</td>
</tr>
<tr>
<td>5</td>
<td>0.855</td>
<td>0.954</td>
<td>0.891</td>
<td>0.873</td>
</tr>
<tr>
<td>10</td>
<td>0.863</td>
<td>0.956</td>
<td>0.897</td>
<td>0.885</td>
</tr>
<tr>
<td>20</td>
<td>0.889</td>
<td>0.963</td>
<td>0.917</td>
<td>0.905</td>
</tr>
<tr>
<td>30</td>
<td>0.903</td>
<td>0.964</td>
<td>0.927</td>
<td>0.915</td>
</tr>
</tbody>
</table>

How F1-score varies with number of queries
Thank you!