Ranking-Based Name Matching for Author Disambiguation in Bibliographic Data

Jialu Liu — Department of Computer Science, UIUC, jliu64@illinois.edu

Kin Hou Lei — Department of Computer Science, UIUC, klei2@illinois.edu

Jeffery Yufei Liu — Department of Statistics, UIUC, jliu64@illinois.edu

Chi Wang — Department of Computer Science, UIUC, chiwang1@illinois.edu

Jiawei Han — Department of Computer Science, UIUC, hanj@illinois.edu

Background

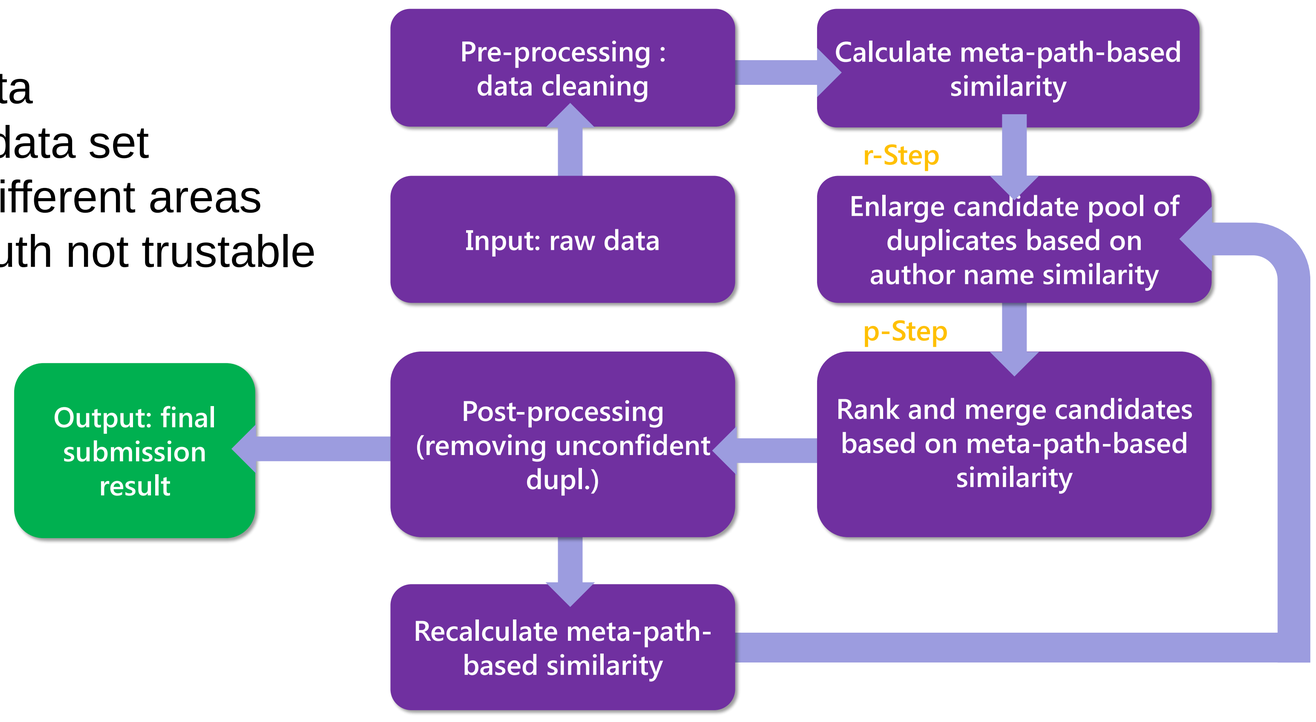
Team name: SmallData

Achievement: 2nd @ 2nd Track

Performance: **99.157 (F1 score)**

Challenge

- No training data
- Noises in the data set
- Names from different areas
- Test ground truth not trustable



Pre-process: Data Cleaning

- Noisy First or Last Names
  - Eytan H. Modiano and Eytan Modianoy
- Mistakenly Separated or Merged Name Units
  - Sazaly Abu Bakar and Sazaly AbuBakar
- Build statistics of name units
  - Count[“Modianoy”] << Count[“Modiano”]

The r-Step: Improving Recall

- Improving the recall of the algorithm means that given an **author ID (input)**, one should find as many **potential duplicates (output)** as possible.
- What do we need to consider? **Name!**

String-based Consideration

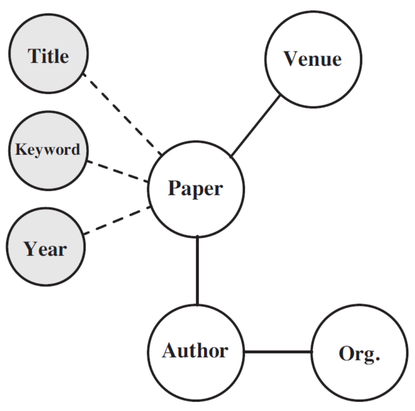
- **Levenshtein Edit Distance**
  - Spelling or OCR error
- **Soundex Distance**
  - “Michael”,“Mickel” and “Michal”
- **Overlapping Name Units**
  - Name reordering brought by parser
  - Wing Hong Onyx Wai and Onyx Wai Wing Hong
- **Asian Names and Western Names**
  - Andrew Chi-Chih Yao and Michael I. Jordan

The p-Step: Improving Precision

- Improving the precision of the algorithm means that once finding **potential duplicates (input)** from r-step, we need to infer the **real author entity (output)** shared by one or more author IDs.
- What do we need to consider? **Network!**

Meta-path in networks

A meta-path P is a path defined on the graph of a network schema. For example, in this competition data set, the co-author relation can be described using the length-2 meta-path APA (author-paper-author)



Measure Matrix for Nodes Similarity

- A measure matrix is for keeping similarities for any pair of nodes based on a meta-path.
- For example, the measure matrix for Author-Paper-Venue is: $\overline{M_{A,V}} = \text{Normalize}(\text{M}_{\text{A,P}} \times M_{P,V})$
  L2 Normalization is applied to make such that the self-maximum property can be achieved.
- Measure matrix for APVPA: $\overline{M_{A,A}} = \overline{M_{A,V}} \times \overline{M_{A,V}}^{T}$

Multiple Measure Matrices

- We are interested in similarity score between authors
- Such score can be obtained via multiple measure matrices with different meta-paths.
- To support measure matrices defined on different meta-paths, we adopt the linear combination strategy:

$$Sim(a_i, a_j) = \sum W_{path} Sim_{path}(a_i, a_j)$$

- The selected meta-paths are APA, AOA, APAPA, APV PA, APKPA, APTPA and APY PA. The weights for them are decreasing progressively.

Ranking-based Merging

- Assume we have three authors and their similarity scores in the listed tables
- To infer the real entity behind each ID
  - Sort the similarity scores
  - Start merging from top ranked ID
    - (2), (3) are in conflict, skip
    - (1), (2) merge -> (1, 2)
    - (1), (3) are in conflict because (2) and (3)
    - return (1, 2) and (3)
  - Once two IDs have multiple publications and low meta-path-based similarity score, reject their merging request.
- Expand author names corresponding to the IDs once we are confident about two IDs to be the duplicate.
  - For example, as authors 1 and 2 are highly possible to be the same person and the name of author 2 has better quality than that of author 1, we can replace the name of author 1 to be Michael J. Lewis.
  - Suppose the full name of author 1 or 2 to be Michael James Lewis and we have a new author with name James Lewis.
  - If we do not adopt this name expanding mechanism, obviously author 1 and this new author are in conflict.

Table 1: Rank of author ID pairs.

Author ID Pair	Similarity	Rank
(1, 2)	0.6325	2
(1, 3)	0	3
(2, 3)	0.7071	1

Table 2: Lists of matched author IDs.

Name	Author ID	Matched IDs
Michael Lewis	1	2, 3
Michael J. Lewis	2	1
Michael P. Lewis	3	1

Table 3: Test cases for name matching.

Author Name A	Author Name B	Expected Result
Jiawei Han	*Jia Han*	In Conflict
Xiang Li	*Xiang Lin*	In Conflict
Gordon D. Moskowitz	*Gordon Blaine Moskowitz*	In Conflict
H. Murray-Rust	*D. M. Murray-Rustt*	In Conflict
Deliang L. Wang	*Liang Wang*	In Conflict
S. J. Thomas Schwarz	*Thomas J. E. Schwarz*	In Conflict
Takeshi Mori	*Taketoshi Mori*	In Conflict
Tadashi Suzuki	*Takashi Suzuki*	In Conflict
Hong-Hu Zhu	*H. H. Zhu*	Compatible
Ralph Mac Nally	*RalphMac Nally*	Compatible
V. Scott Gordont	*V. Scott Gordon*	Compatible
Jeff W. Hughes	*Jeffrey W. Hughes*	Compatible
William Hughes	*Bill Hughes*	Compatible
William Hughes	*B. Hughes*	Compatible
Valli Kumari Vatsavayi	*V. Valli Kumari*	Compatible
Mercedes Femandez-Redondo	*Mercedes Fernandez Redondo*	Compatible
Aliaa Abdel-Haleim Abdel-Razik Youssif	*Aliaa A. A. Youssif*	Compatible

Table 4: Module descriptions and corresponding performance gains.

Performance	Gain	New Module(s)	Days
95.376	-	Same Author Name Benchmark	-
95.786	0.410	+ Meta-path: Coauthor	19
96.623	0.837	+ Name Initials, Omitted Middle Name	21
97.427	0.804	+ Meta-path: Covenue	27
97.770	0.343	+ Nicknames + Asian names handling	33
98.729	0.959	+ Accepting name-compatible author pair even with zero meta-path-based similarity	36
99.020	0.291	+ Name reordering + noisy last/first name pre-processing	37
99.036	0.016	+ Rough post-processing	42
99.075	0.039	+ SoundEx distance	45
99.130	0.055	+ Name units breaking/merging pre-process, + name expansion	49
99.157	0.027	+ Iterative framework + more aggressive post-processing	54

Post-processing

- “**Unconfident**” duplicate author IDs should be removed even though their names are compatible and their meta-path-based similarity scores are acceptable.
- We define “**unconfident**” to have two factors
  - the difference between name strings in terms of unmatched name units to be large
  - the meta-path-based similarity score to be not large.
  - Wing Hong Onyx Wai and W. Hong