NameClarifier: A Visual Analytics System for Author Name Disambiguation

Qiaomu Shen, Tongshuang Wu, Haiyan Yang, Yanhong Wu, Huamin Qu and Weiwei Cui
Name ambiguity
by Wang Wei

王伟，王维，王威，王玮，汪卫，汪伟，汪威

Name ambiguity: People with different names

[ WEI WANG ]

WEI WANG  WEI WANG  WEI WANG

WEI WANG  WEI WANG  W. WANG

W. WANG  W. WANG  W. WANG
Manual Check

Library in Universities (No error allowed)

Automatic approach

Large public bibliography database

(A small number of errors are allowed)

- Purely author names.
- Publication attributes: titles, shared coauthors, venues, self-citation. Etc.
- Additional web information.

Small scale library
Major challenge 1/2:

The name ambiguity problem are **case by case**.

- Limited collaborators or wide range of collaborators
- One research interest or multiple research interests
Major challenge 2/2:

**Uncertainties** of every attribute:

- Venue cover different size of scopes (IJCV for Vision V.S. TVCG for computer graphics + visualization)
- Shared coauthors: suffer from name ambiguity themselves!
- Etc.

→ No universal model

**Get people involved**
Our solutions:

- Customize the disambiguation on a case-by-case basis
- Mining metrics + visualization
- Traditional black box solution -> white box procedure
System framework
Pre-processor

Publication database → Extract

Publication title, venue, timestamp, author list

Analyzer

Confirmed paper group

Venue + Co-authorship
Similarity + Uncertainty

Ambiguous case

Categorization

Evaluation

Visualizer

Relation View

Group View

Temporal View

Interaction
Preprocess and Data Analysis

- Confirmed authors and confirmed papers
  - Indexed authors who have been identified.
  - Each conformed author will be associated with multiple papers (confirmed papers group).

Search “Rui Wang” from dblp:

Confirmed Authors:
- Rui Wang 0002
  - Microsoft Research
- Rui Wang 0003
  - University of Massachusetts Amherst
- Rui Wang 0004
  - Zhejiang University
- Rui Wang 0005
  - Universität des Saarlandes
- Rui Wang 0006
Preprocess and Data Analysis

• Confirmed authors and confirmed papers
  • Indexed authors who have been identified by the system.
  • Each conformed author will associated with multiple papers (confirmed papers group).

• Ambiguous names and ambiguous papers:
  • The author names which have not been identified.
  • The papers with no confirmed authors are ambiguous papers.
Preprocess and Data Analysis

Input: NM

Given an author name **NM**, a collection of publications with the name NM (or approximate to NM) listed as an author will be extracted from digital library.
Preprocess and Data Analysis

- With the input name NM:

  Confirmed papers and ambiguous papers

  Grouped by confirmed authors

  Subset_NM1  Subset_NM2  ...  Subset_NMn

  Reconstruct

  Name: NM  ID: 0001

  Coauthor Set
  Venue Set
  Time series publications

  Matching

  Title: Paper1
  Coauthor List
  Venue
  Publication Time

  Paper1  Paper2  ...  Papern

  Reconstruct

  Allocation likelihood (AL)
Visual Design
Each row: an confirmed author

Saturation: Allocation likelihood (AL)

Red line indicates the position of selected ambiguous paper
Relation View

Blue: ambiguous name currently under analysis
Orange: other authors
Relation View

Relations (Venn Diagram)

Collaboration:
- coauthor
- Venue

Overall confidence

Indirectly connected coauthors similarities
Temporal View

Each rectangle indicates one confirmed paper.

Orange bars indicate the matched venue.
Light blue bars indicate the unmatched coauthors.
Dark blue bars indicate the matched coauthors.

Stack paper rectangles according to their publication years.

Red border: The year when the ambiguous paper was published.

Light blue bars indicates the unmatched coauthors.
Dark blue bars indicates the matched coauthors.
Group View

Outer ring (R1):
- Ambiguous paper group
- In each arc papers only share coauthor/venue with those in the same group

Inner ring (R2):
- Confirmed authors
- Every arc: a confirmed author

Central angle: The total number of papers in a potential group
Arc saturation: group quality
Stroke: share confirmed vs. confirmed
Group Quality

VIS 2016
(F) Nodes: papers in a selected ambiguous(paper) arc

Edges:
- Two ambiguous papers share coauthors
- Ambiguous papers share coauthors with confirmed authors

Node colors: publication years

2000  Publication Year  2016
Case study
Case Studies

Case1: Wei Chen

Sort by Max Group Relation Allocation Likelihood

The most cases can be easily addressed directly by **Relation View**

# Total paper: 1170
- 573 ambiguous
- 597 confirmed papers for 25 confirmed authors
Case Studies

Case 1: Wei Chen

Sort by Max Group Relation Allocation Likelihood

# Total paper: 1170
- 573 ambiguous
- 597 confirmed papers in 25 confirmed authors

Click to see the temporal view
Case Studies

Case 1: Wei Chen

In some cases, the allocation likelihood is different from the visual pattern.
Case Studies

# Total paper: 560
- 179 ambiguous + 381 recognized papers
- 15 recognized authors

Case2: Rui Wang

Sort by Max Group Relation Allocation Likelihood

The most tricky one: It cannot be easily distinguished through comparison link and temporal view
Case Studies

Case 2: Rui Wang

Sort by Max Group Relation Allocation Likelihood

The most tricky one: It cannot be easily distinguished through comparison link and temporal view.
Case Studies

Case 2: Rui Wang

Papers closely connected to both these two confirmed authors
Some nodes with the black strokes are *loosely connected* with those Rui Wang 0003’s papers.
Nearly all the nodes with the black strokes are tightly connected with those Rui Wang 0004’s confirmed papers.

Case Studies

Case2: Rui Wang

Release the papers of the Rui Wang 0004

We tend to think all the ambiguous papers belong to Rui Wang 0004
Case Studies

Case 2: Rui Wang

Start exploration from the farthest one from 0003

Expand the Rui Wang 0003
Case Studies

Case2: Rui Wang

Think back to the most tricky one:

More evidence are provided to make relations distinguishable.
Case Studies

Case 2: Rui Wang

Start from the largest ambiguous arc. Select this part and form a new confirmed author.
Case Studies

Case2: Rui Wang

Start from the largest ambiguous arc. Notice that there is one connect with a confirmed authors.
Case Studies

Case 2: Rui Wang

Start from the largest ambiguous arc. Notice that there is one connect with a confirmed authors.
Conclusion

- NameClarifier, an interactive visual system for name disambiguation;
- Turn the traditional black-box solution into a white-box procedure;
- The system provides guidance instead of classification results for ambiguous cases.
Future work

- Extension to more attributes;
- Visual alarming for the improper operation;
Thank you!

Q&A

NameClarifier: A Visual Analytics System for Author Name Disambiguation

Qiaomu Shen, Tongshuang Wu, Haiyan Yang, Yanhong Wu, Huamin Qu and Weiwei Cui
Back Up

- Automatic Evaluation
  - Allocation Likelihood
    - Co-author Matching
    - Venue Match
  - Confidence Measurements
    - Co-author Confidence
    - Venue Confidence
• Automatic Evaluation
  • Allocation Likelihood
    • Co-author Matching
    • Venue Match
  • Confidence Measurements
    • Co-author Confidence
    • Venue Confidence

Confirmed paper $p_i$ and Ambiguous paper $p_A$

$$cm_{i,p} = \frac{|C(p_i) \cap C(p_A)|}{|C(p_i) \cap C(p_A)|}$$
Back Up

- Automatic Evaluation
  - Allocation Likelihood
    - Co-author Matching
    - Venue Match
  - Confidence Measurements
    - Co-author Confidence
    - Venue Confidence

Confirmed paper $p_i$ and Ambiguous paper $p_A$

$$vm_{i,p} = sgn(vs(v_A, v_i) - s)$$

where

$$vs(v_A, v_i) = \left| \frac{|A(v_A) \cap A(v_i)|}{|A(p_i) \cap C(p_A)|} \right|$$
• **Automatic Evaluation**
  • **Allocation Likelihood**
    • Co-author Matching
    • Venue Match
  • **Confidence Measurements**
    • Co-author Confidence
    • Venue Confidence

Confirmed paper $p_i$ and Ambiguous paper $p_A$

$$AL(A, G) = \frac{1}{n} \sum_{i=1}^{n} (\alpha_c \cdot cm_i + \alpha_\beta \cdot vm_i)$$
Back Up

- Automatic Evaluation
  - Allocation Likelihood
    - Co-author Matching
  - Venue Match
- Confidence Measurements
  - Co-author Confidence
  - Venue Confidence

Confirmed paper $p_i$ and Ambiguous paper $p_A$

$$cc(c) = 1_{DC}(c) \cdot (cf(c) + gq(c))$$
Back Up

- Automatic Evaluation
  - Allocation Likelihood
    - Co-author Matching
  - Venue Match
- Confidence Measurements
  - Co-author Confidence
  - Venue Confidence

Confirmed paper $p_i$ and Ambiguous paper $p_A$

$$vc(p_i, p_A) = 1_{vr}(v_i) \cdot (ad(v_i) + vs(v_i, v_A))$$