Model Selection Strategies for Author Disambiguation

TIR’11

Roman Kern, Mario Zechner, Michael Granitzer
Outline

- Problem Statement
- Disambiguation Workflow
- Clustering & Model Selection
- Data Set & Experiments
- Conclusion
Problem Statement

Definition

Given a set of scientific publications/citations, our aim is to **identify distinct authors and their respective publications** within the set.
Problem Statement

Examples

Where does ambiguity come from?

☐ Two distinct authors share the same name

☐ A single author is referred to by different orthographic variations of her name

☐ An author has changed her name due to marriage or other causes
Application Scenarios

- Citation and Impact Analysis

- Creating author profiles in social research networks like Mendeley

- Recommendation engine for research papers

- Facetted Search
A Disambiguation Framework

Overview

Many systems presented in the literature for author disambiguation share the same workflow:

1. Extract author name occurrences from publications/citations
2. Block author names and the publications they occur in
3. Disambiguate authors within each block
A Disambiguation Framework

1. Author Name Extraction

Extraction of author names

- **Rule-based** extraction based on known publication layouts
- Machine learning techniques for **sequence tagging** (HMMs, CRFs, SVMs)

**Achievable Performance** (not part of this paper)
- 0.8-0.9 Precision
- 0.5-0.8 Recall

⇒ The result of this stage is a set of author names for each publication/citation
A Disambiguation Framework

2. Blocking

- Blocking is the process of grouping sufficiently similar author names and the publications/citations associated with them.

- Blocking is performed for performance and tractability reasons.

- Similarity measures for author names:
  - Phonetic hashing via Soundex or Metaphone
  - String hashing methods

- Focus on Recall – a block must contain all possible unique authors and their publications.
A Disambiguation Framework

3. Disambiguation

- Disambiguation is most often achieved via **clustering**
  - For every block
  - Consider all pairs of author names represented as strings and their occurrence in publications
  - Cluster all pairs to groups with unique authors, so that
    - all pairs in a group represent the same unique author
    - All publications of one authors are contained in one group

- Core decisions to apply clustering
  - **Features** to represent pairs
  - **Similarity Measures** between pairs
  - **Model Selection** Method, i.e. guessing the number of authors in one block
  - **Clustering algorithm**
Clustering Properties
Features & Similarity Measure

- Noun, adjective and adverbs of publications plain text and information obtained from search engines
- Keywords extracted from a publications plain text (TextRank)
- Tokenized title text
- Tokenized author names

- Cosine as similarity measure
Clustering Properties
Model Selection – I

- Guess the number of unique authors in one block
  - Large variance, correct number can range from 1 to 100 (and more)
  - Hard problem, often neglected by related work

- Standard methods exists (e.g. density based, stability based)
  - Preliminary test showed very low accuracy
  
  Development of a task specific model selection strategies (main contribution)
Clustering Properties
Model Selection - II

Approach/Assumption

- Clustering groups textually similar publications of a block
- Use different feature kind for model selection: Co-authorship
- The more co-authors overlap in a cluster and the less they are spread between cluster, the better

Use conditional probabilities as measures therefore

\[
\bar{P}_{ac} = \frac{1}{|pairs|} \sum_{\forall pairs} 1 - P(\text{author} | \text{cluster})
\]

\[
\bar{P}_{ca} = \frac{1}{|pairs|} \sum_{\forall pairs} 1 - P(\text{cluster} | \text{author})
\]

\[
F(C) = \frac{\bar{P}_{ac} + \bar{P}_{ca}}{2}
\]
Clustering Properties
Model Selection - III

- Similar formulation using point-wise conditional entropy

\[ H_{\text{pointwise}}(Y|X) \overset{\text{def}}{=} - \sum_C p(x, y) \log p(y|x) \]

\[ C = \{ c | c \in \text{Clusters}_\text{Co-Author} \} \]

- Example
Experiments
Setup

☐ Two data set

- **Giles** provides a nice dataset of citations with 12 ambiguous author names ([http://bit.ly/aBV8qP](http://bit.ly/aBV8qP))
- **Mendeley** provided us with a much larger dataset retrieved from user profiles (<3 Mendeley)
- For every publication/citation we also gathered web search results for additional data from Google, Bing, ACM (until we got blocked), e.g. plain text

☐ Workflow Setup

- Identification (Step 1) was not necessary
- Blocking – used Ground-Truth to create forename subsets: Lee, Martin, Gupta, Kumar, Chen, Johnson

➤ Error Analysis focuses solely on clustering properties
Experiments
Results Clustering Algorithms

Fix number of clusters to the true number of unique authors
HAC with average linking seems is best clustering approach
Experiments

Results Features

Again, assume number of unique authors known

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Keyword</th>
<th>Stem</th>
<th>Normalize</th>
<th>Purity</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.92</td>
<td>0.9</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.91</td>
<td>0.89</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.88</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table I

Best 5 results using HAC Clustering on the Giles-Martin subset. Sorted by F1. 16 distinct authors, 112 publications.

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Keyword</th>
<th>Stem</th>
<th>Normalize</th>
<th>Purity</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.72</td>
<td>0.50</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.70</td>
<td>0.48</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.70</td>
<td>0.46</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.61</td>
<td>0.42</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.66</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Table II

Best 5 results on HAC Clustering on the Giles Lee subset. Sorted by F1. 100 distinct authors, 1419 publications.

Results on the Martin subset are encouraging. Reason: full-text features contain less noise
Experiments

Results Model Selection

What is the difference if we have to guess the author number?

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$K_{\text{real}}$</th>
<th>$F_{\text{1best}}$</th>
<th>$K_{\text{guess}}$</th>
<th>$F_{\text{1guess}}$</th>
<th>$F_{\text{1real}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mendely-lee</td>
<td>49</td>
<td>61%</td>
<td>44</td>
<td>28%</td>
<td>27%</td>
</tr>
<tr>
<td>Giles-martin</td>
<td>16</td>
<td>90%</td>
<td>16</td>
<td>84%</td>
<td>84%</td>
</tr>
<tr>
<td>Giles-gupta</td>
<td>26</td>
<td>65%</td>
<td>14</td>
<td>43%</td>
<td>65%</td>
</tr>
<tr>
<td>Giles-kumar</td>
<td>14</td>
<td>70%</td>
<td>14</td>
<td>44%</td>
<td>44%</td>
</tr>
<tr>
<td>Giles-chen</td>
<td>61</td>
<td>46%</td>
<td>12</td>
<td>10%</td>
<td>37%</td>
</tr>
<tr>
<td>Giles-johnson</td>
<td>15</td>
<td>78%</td>
<td>11</td>
<td>60%</td>
<td>75%</td>
</tr>
<tr>
<td>Giles-lee</td>
<td>100</td>
<td>50%</td>
<td>21</td>
<td>5%</td>
<td>38%</td>
</tr>
</tbody>
</table>

Table IV
MODEL SELECTION RESULTS USING POINT-WISE CONDITIONAL ENTROPY ON KEYWORDS ONLY

Performance varies, but gives good results when clustering comes close to the real groups (i.e. Martin Subset)

Underestimate correct number of clusters
Conclusion

- HAC as empirical best algorithm for disambiguation
- New Model Selection Strategies work good given good clustering results
- Automatic Author Disambiguation still unsolved for practical scenarios
- Identification and Blocking as additional error sources

Future Work

- reduce the effect of blocking errors and model selection through outlier detection
- Improved feature selection and cleaning
Thanks for your attention

and also thanks also to our supporters

http://mendeley.com

FP 7 TEAM Project: http://team-project.tugraz.at/

You can follow our research on http://team-project.tugraz.at/blog or via Twitter @mgrani or via Mendley‘s public group „Publications (TEAM IAPP)“