Using Web Information for Author Name Disambiguation

Denilson A. Pereira\textsuperscript{1}  
Nivio Ziviani\textsuperscript{1}  
Marcos André Gonçalves\textsuperscript{1}  
Berthier Ribeiro-Neto\textsuperscript{1,2}  
Alberto H. F. Laender\textsuperscript{1}  
Anderson A. Ferreira\textsuperscript{1}  

\textsuperscript{1}Department of Computer Science  
Federal University of Minas Gerais, Brazil  

\textsuperscript{2}Google Engineering, Belo Horizonte, Brazil  

4\textsuperscript{th} Workshop on the Future of Web: Semantic Search, Ibiza, Apr 18\textsuperscript{th}, 2009
The Problem

- Mixed Citation

“D. Pereira” may refer to “Denilson Pereira” or “David Pereira”, two different people
The Problem

- **Mixed Citation**
  “D. Pereira” may refer to “Denilson Pereira” or “David Pereira”, two different people

- **Split Citation**
  “Denilson Alves Pereira” may appear under different name abbreviations, such as “Denilson Pereira”, “D. Pereira”, or “D. A. Pereira”, or a misspelled name such as “Denilson Fereira”
The Problem

- **Mixed Citation**
  “D. Pereira” may refer to “Denilson Pereira” or “David Pereira”, two different people

- **Split Citation**
  “Denilson Alves Pereira” may appear under different name abbreviations, such as “Denilson Pereira”, “D. Pereira”, or “D. A. Pereira”, or a misspelled name such as “Denilson Fereira”

- **Goal:** given a list of citations, we want to (i) find the set of distinct authors and (ii) attribute each citation to the corresponding author
Proposed Solutions

- Supervised learning methods
  - require human labeling and training time
  - unfeasible in large-scale digital libraries

- Unsupervised clustering methods
  - use a specific clustering algorithm
  - select the most discriminative metadata for the disambiguation task
Our Proposal

- Unsupervised hierarchical clustering method
  - uses the Web as a source of additional information for author name disambiguation

- It consists of
  - gathering information from input citations and submitting queries to a Web search engine
  - parsing the answer sets looking for single author documents
  - clustering citations from each document in a bottom-up fashion
## Related Work (1/2)

<table>
<thead>
<tr>
<th>Paper</th>
<th>Mixed Citation</th>
<th>Split Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee, On, Kang, Park (IQIS, 2005)</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Kang, Na, Lee, Jung, Kim, Sung, Lee (IPM, 2009)</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Tan, Kan, Lee (JCDL, 2006)</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>On, Lee, Kang, Mitra (JCDL, 2005)</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Han, Zha, Giles (JCDL, 2005)</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Laender, Gonçalves, Cota, Ferreira, Santos, Silva (DocEng, 2008)</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Our Work</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>
## Related Work (2/2)

<table>
<thead>
<tr>
<th>Paper</th>
<th>Basic Metadata</th>
<th>Additional Metadata</th>
<th>Web Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>Han, Giles, Zha, Li, Tsioutsiouliklis (JCDL, 2004)</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Han, Zha, Giles (JCDL, 2005)</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huang, Ertekin, Giles (PKDD, 2006)</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Song, Huang, Councill, Li, Giles (JCDL, 2007)</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Tan, Kan, Lee (JCDL, 2006)</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Kang, Na, Lee, Jung, Kim, Sung, Lee (IPM, 2009)</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Our Work</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
</tbody>
</table>
Web Author Disambiguation (WAD) Method

Step 1
Web Author Disambiguation (WAD) Method

Step 1

Search Engine

Page Collector

Web

Web Document Collection

Citations

Step 2

Info Extractor

Citation Document Info
Web Author Disambiguation (WAD) Method

Step 1: Search Engine

Step 2: Web Page Collector

Step 3: Clustering Procedure

Citation Document Info

Disambiguated Authors
Step 1 - Obtaining Information

- Extract a citation string
- Submit as a query to a search engine to find pages containing publications of the authors
  - Query 1: author name + “publications” + work title
    Denilson Pereira publications Using Web Information for Author Name Disambiguation
  - Query 2: author name + quoted work title
    Denilson Pereira “Using Web Information for Author Name Disambiguation”
- Collect top 10 docs in the answer set of each query
Step 2 - Extracting Information (1/3)

- Look for each citation in the Web doc collection
  - Index documents using a local search engine
  - Query: quoted work title + first author names + publication venue name
    (allow one word error for long work titles)
Step 2 - Extracting Information (2/3)

- Identify single author documents
  - Look for an author name (must appear alone in page)
  - Every citation must contain the author name
  - Document cannot contain words such as “Abstract”, “Introduction”, “References” in a line
List of publications from the DBLP Bibliography Server - FAQ

Coauthor Index - Ask others: ACM DL/Guide - CiteSeer - CSB -
Google - MSN - Yahoo

Home Page

- Weight documents generating a rank of importance
  (IHF: Inverse Host Frequency factor)
Weight documents generating a rank of importance (IHF: Inverse Host Frequency factor)
Step 2 - Extracting Information (3/3)

- Weight documents generating a rank of importance (IHF: Inverse Host Frequency factor)
Step 3 - Clustering Citations (1/8)
Step 3 - Clustering Citations (2/8)

A\_1
\begin{array}{c}
C_1 \\
C_2 \\
\text{Doc}_1 \\
\end{array}
\begin{array}{c}
\text{IHF} \geq \text{minIHF} \\
\end{array}

A_1 \quad C_1 \\
A_1 \quad C_2 \\
A_2 \quad C_3 \\
A_3 \quad C_4 \\
A_3 \quad C_5 \\
A_4 \quad C_6
Step 3 - Clustering Citations (3/8)

A_1

\[ \text{IHF} \geq \text{minIHF} \]

C_1
C_2

Doc_1

A_1 C_1 C_2

A_1 C_1
A_1 C_2
A_2 C_3
A_3 C_4
A_3 C_5
A_4 C_6
Step 3 - Clustering Citations (4/8)

IHF ≥ minIHF

LATIN - LAboratory for Treating INformation (www.dcc.ufmg.br/latin)
Step 3 - Clustering Citations (5/8)

\[
\text{IHF} \geq \text{minIHF}
\]

LATIN - LAboratory for Treating INformation (www.dcc.ufmg.br/latin) 22
Step 3 - Clustering Citations (6/8)

IHF < minIHF
(DBLP)
Step 3 - Clustering Citations (7/8)

Doc\textsubscript{3}

A\textsubscript{3} \quad \text{IHF} < \text{minIHF}

C\textsubscript{4}
C\textsubscript{5}
C\textsubscript{6}

IHF \text{(DBLP)}

Doc\textsubscript{4}

A\textsubscript{3} \quad \text{IHF} < \text{minIHF}

C\textsubscript{4}
C\textsubscript{5}
C\textsubscript{6}
C\textsubscript{7}

LATIN - LAboratory for Treating INformation (www.dcc.ufmg.br/latin)
Step 3 - Clustering Citations (8/8)

LATIN - LAboratory for Treating INformation (www.dcc.ufmg.br/latin)
Experimental Evaluation: Test Dataset

- Han, Zha and Giles’ dataset (JCDL’05)
- 8,442 citation records, with 480 distinct authors (14 ambiguous groups)
Experimental Evaluation: Metrics

- **$K$:** geometric mean between $ACP$ and $AAP$
  - $ACP$: Average Cluster Purity (level of mixed citation)
  - $AAP$: Average Author Purity (level of split citation)
- **pairwise F1 ($pF1$):** harmonic mean of $PP$ and $PR$
  - $PP$: Pairwise Precision
  - $PR$: Pairwise Recall
- **cluster F1 ($cF1$):** harmonic mean of $CP$ and $CR$
  - $CP$: Cluster Precision
  - $CR$: Cluster Recall
Results
Query 1: unquoted author name + “publications” + unquoted work title

<table>
<thead>
<tr>
<th>Name</th>
<th>K</th>
<th>pF1</th>
<th>cF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Gupta</td>
<td>0.90</td>
<td>0.88</td>
<td>0.21</td>
</tr>
<tr>
<td>A. Kumar</td>
<td>0.86</td>
<td>0.88</td>
<td>0.20</td>
</tr>
<tr>
<td>C. Chen</td>
<td>0.71</td>
<td>0.47</td>
<td>0.18</td>
</tr>
<tr>
<td>D. Johnson</td>
<td>0.87</td>
<td>0.93</td>
<td>0.13</td>
</tr>
<tr>
<td>J. Lee</td>
<td>0.69</td>
<td>0.62</td>
<td>0.09</td>
</tr>
<tr>
<td>J. Martin</td>
<td>0.89</td>
<td>0.90</td>
<td>0.29</td>
</tr>
<tr>
<td>J. Robinson</td>
<td>0.84</td>
<td>0.86</td>
<td>0.24</td>
</tr>
<tr>
<td>J. Smith</td>
<td>0.72</td>
<td>0.62</td>
<td>0.06</td>
</tr>
<tr>
<td>K. Tanaka</td>
<td>0.87</td>
<td>0.93</td>
<td>0.04</td>
</tr>
<tr>
<td>M. Brown</td>
<td>0.76</td>
<td>0.73</td>
<td>0.16</td>
</tr>
<tr>
<td>M. Jones</td>
<td>0.78</td>
<td>0.81</td>
<td>0.03</td>
</tr>
<tr>
<td>M. Miller</td>
<td>0.79</td>
<td>0.79</td>
<td>0.07</td>
</tr>
<tr>
<td>S. Lee</td>
<td>0.76</td>
<td>0.69</td>
<td>0.20</td>
</tr>
<tr>
<td>Y. Chen</td>
<td>0.72</td>
<td>0.58</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.76</strong></td>
<td><strong>0.14</strong></td>
</tr>
</tbody>
</table>
Results
Query 2: unquoted author name + quoted work title

<table>
<thead>
<tr>
<th>Name</th>
<th>K</th>
<th>pF1</th>
<th>cF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Gupta</td>
<td>0.82</td>
<td>0.78</td>
<td>0.13</td>
</tr>
<tr>
<td>A. Kumar</td>
<td>0.82</td>
<td>0.85</td>
<td>0.16</td>
</tr>
<tr>
<td>C. Chen</td>
<td>0.74</td>
<td>0.56</td>
<td>0.16</td>
</tr>
<tr>
<td>D. Johnson</td>
<td>0.81</td>
<td>0.80</td>
<td>0.19</td>
</tr>
<tr>
<td>J. Lee</td>
<td>0.71</td>
<td>0.63</td>
<td>0.18</td>
</tr>
<tr>
<td>J. Martin</td>
<td>0.88</td>
<td>0.89</td>
<td>0.29</td>
</tr>
<tr>
<td>J. Robinson</td>
<td>0.94</td>
<td>0.96</td>
<td>0.53</td>
</tr>
<tr>
<td>J. Smith</td>
<td>0.79</td>
<td>0.77</td>
<td>0.10</td>
</tr>
<tr>
<td>K. Tanaka</td>
<td>0.82</td>
<td>0.86</td>
<td>0.14</td>
</tr>
<tr>
<td>M. Brown</td>
<td>0.82</td>
<td>0.82</td>
<td>0.13</td>
</tr>
<tr>
<td>M. Jones</td>
<td>0.81</td>
<td>0.84</td>
<td>0.08</td>
</tr>
<tr>
<td>M. Miller</td>
<td>0.68</td>
<td>0.61</td>
<td>0.13</td>
</tr>
<tr>
<td>S. Lee</td>
<td>0.78</td>
<td>0.81</td>
<td>0.23</td>
</tr>
<tr>
<td>Y. Chen</td>
<td>0.75</td>
<td>0.65</td>
<td>0.13</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>0.80</td>
<td>0.77</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Hierarchical Agglomerative Clustering (HAC)
Baseline 1

- Web based
- Tan, Kan and Lee (JCDL, 2006)
  - unsupervised clustering method
  - each citation is represented by a feature vector
  - features: the relevant URLs obtained by querying a search engine
  - feature weight: IHF of the URLs
  - the number of correct clusters is previously known
K-way spectral clustering (KWAY)
Baseline 2

- Han, Zha and Giles (JCDL, 2005)
  - unsupervised clustering method
  - each citation is represented by a feature vector
  - features: terms in authors, work and publication venue titles
  - feature weight: tf-idf
  - the number of correct clusters is previously known
Support Vector Machine (SVM)
Baseline 3

- Han, Giles, Zha and Li (JCDL, 2004)
  - supervised learning method
  - each citation is represented by a feature vector
  - features: terms in authors, work and publication venue titles
  - feature weight: tf-idf
  - 10 running, 50% of the data for training and 50% for test
Comparison with Baselines

- Unsupervised clustering methods

<table>
<thead>
<tr>
<th>Method</th>
<th>K</th>
<th>pF1</th>
<th>cF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAD</td>
<td>0.80</td>
<td>0.76</td>
<td>0.14</td>
</tr>
<tr>
<td>HAC</td>
<td>0.63</td>
<td>0.46</td>
<td>0.05</td>
</tr>
<tr>
<td>KWAY</td>
<td>0.50</td>
<td>0.36</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Comparison with Baselines (2/2)

- Supervised learning based method

<table>
<thead>
<tr>
<th>Method</th>
<th>K</th>
<th>pF1</th>
<th>cF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAD</td>
<td>0.82</td>
<td>0.80</td>
<td>0.29</td>
</tr>
<tr>
<td>SVM</td>
<td>0.77</td>
<td>0.66</td>
<td>0.20</td>
</tr>
</tbody>
</table>

statistically tied
Combining additional information

- Use the WAD method in a first phase of a hierarchical clustering process
- Apply other evidences in a second phase to continue fusing clusters
- We experimented applying coauthor information to fuse clusters

<table>
<thead>
<tr>
<th>Method</th>
<th>K</th>
<th>pF1</th>
<th>cF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAD</td>
<td>0.82</td>
<td>0.81</td>
<td>0.15</td>
</tr>
<tr>
<td>WAD + coauthor</td>
<td>0.86</td>
<td>0.84</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Failure Cases

- Some citations not found on the Web due to
  - misspelling errors in more than one word in work titles
  - existence of incorrect citations in our test dataset
- Some citations found only in pages of digital libraries
  - WAD needs more than one evidence
  - digital libraries contain errors
- WAD failed on its heuristics to identify
  - citations in documents
  - single author documents
Conclusions

- Novel method to disambiguate author names
  - uses information extracted from the Web
  - takes advantage of the use of the Web by scientific researchers and of sophisticated matching procedures of Web search engines
- Results indicate gains of up to 65.2% in the quality of disambiguation
- WAD method need not adjust complex parameters, and just requires a single threshold
- It can be used in conjunction with other disambiguation strategies
Future Work

- Extract information such as e-mail, address, affiliation, and author’s full names from single author documents
- Create author name authority files
Reference


- Nivio Ziviani (nivio@dcc.ufmg.br)