Disambiguating Web Appearances of People in a Social Network

Ron Bekkerman
Andrew McCallum

University of Massachusetts at Amherst
<table>
<thead>
<tr>
<th>Title</th>
<th>Presenters</th>
<th>Info</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP04: MDA Standards for Ontology Development</td>
<td>Dragan Gašević, Dragan Djurić, and Vladan Devetić</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TA05: Introduction to RDF Query with SPARQL</td>
<td>Dave Beckett, Steve Harris, Eric Prudhommaux and Andy Seabone</td>
<td></td>
<td>101A</td>
</tr>
<tr>
<td>TP05: Web-based Interactive Collaboration using Semantic Web Technology - Introduction of the Annota and its Comparison to the Blog</td>
<td>Nobuhisa Shiraishi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TA06: Networked Arts - Methods and Tools to create and maintain Virtual Museums</td>
<td>Alfredo Ronchi</td>
<td>INFO</td>
<td>INFO</td>
</tr>
<tr>
<td>TP06: Matching Words and Pictures - Problems, Applications and Progress</td>
<td>Latifur Khan</td>
<td></td>
<td>301A</td>
</tr>
<tr>
<td>TA07: Web Content Mining</td>
<td>Bing Liu</td>
<td></td>
<td>301A</td>
</tr>
<tr>
<td>TP07: Location-based Services in Mobile Information Systems - Architectures, Description, and Systems</td>
<td>Ling Liu</td>
<td></td>
<td>101B</td>
</tr>
</tbody>
</table>

**Full Day Tutorials**

Please refer to Description of Full Day Tutorials for the details.
Associate Professor
Department of Computer Science
University of Illinois at Chicago
PhD, 1989, University of Edinburgh
Interests: Data Mining, Machine Learning

Full Day Tutorials
Please refer Description of Full Day Tutorials for the details.
Which one to choose?
Ph.D Student

Bing Liu

Ph.D Student
Medical Imaging and Computing Group
National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences

Tel: +85 10 6265 9278
Fax: +85 10 6266 1993
Email: bliu@nlpr.ia.ac.cn

Personal Homepage:
http://nlpr-web.ia.ac.cn/english/mic/BingLiu/index.htm
Looking for a person on the Web

- Web appearance disambiguation
  - Which pages refer to the particular person
- Web page content filtering
  - Which section is relevant to the person
- Decision making
  - E.g., whether there are *no* relevant pages
Looking for a person on the Web

- Web appearance disambiguation
  - Which pages refer to the particular person
- Web page content filtering
  - Which section is relevant to the person
- Decision making
  - E.g., whether there are no relevant pages
Objectives

- High precision
  - Retrieve information only about the desired person
  - Precision error can be a disaster
- High recall
  - Retrieve as much relevant info as possible
  - High recall is the major goal
Localization

- **Coreference**: unite different mentions of the same entity
- **Disambiguation**: distinguish between identical mentions of different entities
Literature on name disambiguation

- Bagga & Baldwin, ACL-98
- Mann & Yarowsky, CoNLL-03
- Fleischman & Hovy, ACL-04
- Pedersen et al., CICLing-05
- Han et al., JCDL-05

Increasing interest!
Requirements

- Asymmetry
  - Accept pages of the desired person
  - Reject pages of his/her namesakes
- Unsupervised approach
  - No training set can model the Web
- Single-link preferred over average-link

How to represent the desired person?
Name itself tells nothing

- Given just a name, the task is ill-defined
- Additional information required!
  - Keywords? (e.g. “professor”, “student”)
    - Nelken et.al., WWW 2003
    - But how to obtain them?!
The idea

- Consider a list of names!
  - Of people in one social network
    - Yanhong Zhai + Bing Liu (I)
    - Bing Liu (II)

- Not burdensome to obtain such list
  - Any name appears in context of other names
Approaches

- Pages of acquaintances are interconnected
- Link Structure Model (LS)
  - Build a core of interconnected pages
    - Of *different* people!
  - Add proximate pages to the core
Approaches

- Pages of acquaintances are interconnected
- Link Structure Model (LS)
  - Build a core of interconnected pages
    - Of different people!
  - Add proximate pages to the core
- Distributional Clustering Model (DC)
  - Simultaneously cluster pages and their words
    - Double clustering is usually more accurate
  - Pick cluster with most interconnected pages
Link Structure model (LS)

- Nodes are pages, edges are hyperlinks
- This picture shows pages of 3 people
Technical details of LS

- Pages are *interconnected* if they share a hyperlink
  - URL’s domain & first dir:
    - [http://www.cs.umass.edu/~ronb/enron_dataset.html](http://www.cs.umass.edu/~ronb/enron_dataset.html)
  - And the domain is not too common

Use Google’s link: `link:` operator
Technical details of LS

- Pages are *interconnected* if they share a hyperlink
  - URL’s domain & first dir:
    - http://www.cs.umass.edu/~ronb/enron_dataset.html
  - And the domain is not too common
- *Cosine similarity* between page & the core
  - With novel `google_tfidf` weighting:
    \[
    \text{google}_\text{tfidf}(w) = \frac{tf(w)}{\log(\text{google}_\text{df}(w))}
    \]
Distributional Clustering model (DC)

- Bekkerman, El-Yaniv & McCallum, 2005
- Submitted to ICML
Hybrid model (LS+DC)

- DC starts with small but clean clusters
  - One of which, $C_{DC}$, is most interconnected
- Overlap LS’s core with $C_{DC}$
  - Obtain larger but still clean core
- Add proximate pages to the new core
  - Just as in LS model
LS+DC system overview

Google API

Remove markup
Extract hyperlinks

Match hyperlinks

Cluster of relevant pages

Choose core

Overlap cores

Add proximate pages

Clusters

Names (Queries)

Pages

Other pages

Core cluster

Core cluster

DC

LS
## Dataset

<table>
<thead>
<tr>
<th>Personal name</th>
<th>Position</th>
<th>Pages</th>
<th>Namesakes</th>
<th>Relevant pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam Cheyer</td>
<td>SRI Manag</td>
<td>97</td>
<td>2</td>
<td>96</td>
</tr>
<tr>
<td>William Cohen</td>
<td>CMU Prof</td>
<td>88</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Steve Hardt</td>
<td>SRI Eng</td>
<td>81</td>
<td>6</td>
<td>64</td>
</tr>
<tr>
<td>David Israel</td>
<td>SRI Manag</td>
<td>92</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>Leslie Pack Kaelbling</td>
<td>MIT Prof</td>
<td>89</td>
<td>2</td>
<td>88</td>
</tr>
<tr>
<td>Bill Mark</td>
<td>SRI Manag</td>
<td>94</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>Andrew McCallum</td>
<td>UMass Prof</td>
<td>94</td>
<td>16</td>
<td>54</td>
</tr>
<tr>
<td>Tom Mitchell</td>
<td>CMU Prof</td>
<td>92</td>
<td>37</td>
<td>15</td>
</tr>
<tr>
<td>David Mulford</td>
<td>Stanf Undergrad</td>
<td>94</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Andrew Ng</td>
<td>Stanf Prof</td>
<td>87</td>
<td>29</td>
<td>32</td>
</tr>
<tr>
<td>Fernando Pereira</td>
<td>UPenn Prof</td>
<td>88</td>
<td>19</td>
<td>32</td>
</tr>
<tr>
<td>Lynn Voss</td>
<td>SRI Eng</td>
<td>89</td>
<td>26</td>
<td>1</td>
</tr>
</tbody>
</table>

**OVERALL:** 1085 187 420

- 12 names out of Melinda Gervasio’s social network
Results

- 20% higher than the baseline
- Hybrid method increases recall
  - As we could predict

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agglom. clust.</td>
<td>61.7</td>
<td>53.3</td>
<td>57.2</td>
</tr>
<tr>
<td>LS</td>
<td>84.2</td>
<td>71.8</td>
<td>77.5</td>
</tr>
<tr>
<td>DC</td>
<td>87.3±1.7</td>
<td>71.3±2.5</td>
<td>78.4±0.9</td>
</tr>
<tr>
<td>LS+DC Hybrid</td>
<td>86.9</td>
<td>74.5</td>
<td>80.3</td>
</tr>
</tbody>
</table>
### Results of LS+DC by person

<table>
<thead>
<tr>
<th>Personal name</th>
<th>Found correct</th>
<th>Not found</th>
<th>Found wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam Cheyer</td>
<td>62</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>William Cohen</td>
<td>6</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Steve Hardt</td>
<td>16</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>David Israel</td>
<td>19</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Leslie Pack Kaelbling</td>
<td>84</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Bill Mark</td>
<td>6</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Andrew McCallum</td>
<td>54</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Tom Mitchell</td>
<td>14</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>David Mulford</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Andrew Ng</td>
<td>30</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Fernando Pereira</td>
<td>21</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>Lynn Voss</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>OVERALL:</strong></td>
<td><strong>313</strong></td>
<td><strong>107</strong></td>
<td><strong>47</strong></td>
</tr>
</tbody>
</table>
Distinguishing “doubles”
- Intermediate iteration of DC algorithm
- 98% precision, 45% recall
- Doubles well discriminated

Homepage finding
- 9 of 10 homepages are found
- Except for Steve Hardt’s
- Mulford and Voss have no homepages
Ongoing research

Heuristic search in the Web graph

Two people are in one social network
  - If there’s a path between their pages

89.6% precision
  - With up to 4-length paths only
Conclusion

- First attempt to tackle the problem
  - Of finding people’s web appearances
- Many applications
  - Web, email, social network analysis…
- Proposed methods can be also used:
  - For acronym disambiguation
  - For word sense disambiguation
Conclusion

First attempt to tackle the problem
  Of finding people’s web appearances

Many applications
  Web, email, social network analysis…

Proposed methods can be also used:
  For acronym disambiguation
  For word sense disambiguation

Thank you!