**Personal Name Resolution of Web People Search**

**Motivation, task**
- People-related search tasks
  - E.g., building profiles, creating biographies, finding experts, etc.
  - 5-10% of web searches contain person names (Schein et al., SIGIR 2002)
- Task of personal name resolution
  - Given a set of documents, all of which refer to a particular person name
  - Identify which documents are associated with each single individual (referent)
  - Generally approached as a clustering problem

**The person clustering hypothesis**
- Cluster hypothesis
  (Jardine and van Rijjsbergen, 2001)
  - Similar documents tend to be relevant to the same request
- Re-stated in the context of personal name resolution: “person clustering hypothesis”
  - Similar documents tend to represent the same person (referent)

**In this paper...**
- Examine to which extent the person clustering hypothesis holds under the most general conditions
  - Only feature: distribution of terms in documents
  - Two forms of clustering, identifying relationships between documents
    - Term level
    - Latent space

**Outline**
- **Clustering approaches**
  - Assumptions
  - Single Pass Clustering
  - Probabilistic Latent Semantic Analysis
- Evaluation platform
- Experiments and results
- Conclusions
### Assumptions
1. One document is associated with one referent
2. The distribution of documents assigned to referents follows a power law
3. Every document refers to a distinct person sense, unless there is evidence to the contrary
4. The number of person senses is not known a priori (but is limited by the number of documents available)
5. Documents are unstructured (no guarantees about the format or structure within documents)

### Single Pass Clustering (SPC)
- **Mimic user behavior**
- **For each document**
  - If a cluster representing that person already exists, then assign document to that cluster
  - Otherwise assign it to a new cluster
- **Capitalize on the fact that most popular (dominant) senses of the person name are highly ranked**
- **Very efficient, can be computed online**

### SPC (2)
- **Document is assigned to the most similar cluster as long as**
  1. similarity is higher than a threshold
     \[ \text{SIM}(D, C) > \gamma \]
  2. maximum number of clusters has not been reached
    - if reached, assign document to the last cluster ("left overs")

### SPC (3)
**Measuring document and cluster similarity**
- **(SPC-NB) Naive Bayes**
  \[ \text{sim}(D, C) = O(D, C) \]
  \[ O(D, C) = \frac{p(D|\theta_C)}{p(D|\theta_{\bar{C}})} = \frac{\prod_{t \in D} p(t|\theta_C)^{n(t,D)}}{\prod_{t \in D} p(t|\theta_{\bar{C}})^{n(t,D)}} \]
- **(SPC-COS) Cosine using TF.IDF weighting**
  \[ \text{sim}(D, C) = \cos(\vec{t}(D), \vec{t}(C)) = \frac{\vec{t}(D) \cdot \vec{t}(C)}{\|\vec{t}(D)\| \cdot \|\vec{t}(C)\|} \]

### Probabilistic Latent Semantic Analysis (PLSA)
- **Decomposition of the term-document matrix into a lower dimensional latent space**
  \[ p(t, d) = p(d) \sum_z p(t|z)p(z|d) \]
- **Obtained using the EM algorithm**
- **Each latent topic z represents one of the different senses of the person name**

### PLSA (2)
- **A document d is assigned to one of the person-topics z, if**
  1. \( p(z|d) \) is the maximum argument
  2. odds of the document given z is greater than a threshold: \( O(z, d) > \gamma \)
  \[ O(z, d) = \frac{p(z|d)}{p(z'|d)} = \frac{p(z|d)}{\sum_{z', z' \neq z} p(z'|d)} \]
PLSA (3)

- Automatically finding the number of person senses (i.e., |z|)
  1. set $z=2$, compute the log-likelihood of the decomposition
  2. increment $z$ and compute the log-likelihood again
     - if log-likelihood increased (>0.001), then repeat (2)
     - else goto (3)
  3. STOP

Outline

- Clustering approaches
- Evaluation platform
  - Data set
  - Performance measures
  - Document representation
  - Experiments and results
  - Conclusions

Data set

- WePS 2007 platform (Web People Search track at the Semantic Evaluation Workshop 2007)
- Web pages obtained from the top (up to) 100 results for a person name query to a web search engine
- Each page from the result list is stored
  - URL, title, position in the ranking, snippet

Data set (2)

- Annotators manually classified each web page
- Original task statement allows a document to be assigned to multiple clusters
- Some documents were discarded (e.g. out-of-date)
- Training (49 names) and test (30 names) sets
- Names from 4 different sources
  - US Census, Wikipedia, ECDL06, ACL06

Data set - sources

<table>
<thead>
<tr>
<th>Data set / source</th>
<th>#names</th>
<th>avg(docs)</th>
<th>discarded</th>
<th>refersnts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>49</td>
<td>71.02</td>
<td>20.00</td>
<td>16.74</td>
</tr>
<tr>
<td>US Census</td>
<td>32</td>
<td>47.20</td>
<td>18.00</td>
<td>5.90</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>7</td>
<td>99.00</td>
<td>8.29</td>
<td>23.14</td>
</tr>
<tr>
<td>ECDL06</td>
<td>10</td>
<td>99.20</td>
<td>30.30</td>
<td>15.30</td>
</tr>
<tr>
<td>Test set</td>
<td>30</td>
<td>98.93</td>
<td>15.07</td>
<td>45.93</td>
</tr>
<tr>
<td>US Census</td>
<td>10</td>
<td>99.10</td>
<td>14.90</td>
<td>50.30</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>10</td>
<td>99.30</td>
<td>17.50</td>
<td>56.50</td>
</tr>
<tr>
<td>ACL06</td>
<td>10</td>
<td>98.40</td>
<td>12.80</td>
<td>31.00</td>
</tr>
</tbody>
</table>

- Ambiguity in the test data is much higher than in the training data
- To measure performance as reliably as possible, we use all names

Distribution of documents to person senses

- Size of the clusters follows a power law
  - Exponent of approx. 1.31
  - Confirms our assumption (2) about the data
Performance measures

• Standard clustering measures
  • Purity — “precision”  
    • Rewards methods that introduce less noise in each cluster
  • Inverse purity — “recall”  
    • Rewards methods that gather more elements of each class into a corresponding single cluster
• F-measure (weighted average of purity and inv. purity)
  • $F_{0.5}$ harmonic mean
  • $F_{0.8}$ user’s point of view (more importance to inv. purity)
  • $F_{0.8}$ machine’s point of view (more importance to purity)

Document representation

• Separate index for each person
• Document is represented using
  • Title and snippet from the search engine’s output
  • Body text extracted from HTML
  • Segments of the page, separated by block-level HTML tags, that contain 10 or more words

Outline

• Clustering approaches
• Evaluation platform
• Experiments and results
  • SPC, PLSA
  • Comparing methods
  • Group-level analysis
  • Comparison to other approaches
• Conclusions

Research questions

• What factors affect performance?
  • Similarity threshold
  • Limiting the number of clusters
• How stable is performance?
• What is the best number of clusters to use? Can we determine this automatically?

SPC

Similarity threshold

<table>
<thead>
<tr>
<th>SPC-NB</th>
<th>SPC-COS</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="SPC-NB Graph" /></td>
<td><img src="image2" alt="SPC-COS Graph" /></td>
</tr>
</tbody>
</table>

• Performance is stable w.r.t. the threshold
• Best performance is obtained with low threshold

Limiting the number of clusters

<table>
<thead>
<tr>
<th>SPC-NB</th>
<th>SPC-COS</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3" alt="SPC-NB Graph" /></td>
<td><img src="image4" alt="SPC-COS Graph" /></td>
</tr>
</tbody>
</table>

• Enforcing a limit on the number of clusters hurts (independent of the similarity threshold)
Findings

- **SPC**
  - Good estimate of person senses
  - High purity scores

- **PLSA**
  - Underestimates the number of person senses
    - Identifies the prominent person senses, but fails when only limited examples (1-2 docs) of the other referents are available
  - Very high inverse purity
    - Referents are usually not dispersed among clusters

Comparing methods

<table>
<thead>
<tr>
<th>Method</th>
<th>All names</th>
<th>pur.</th>
<th>invp.</th>
<th>F0.5</th>
<th>F0.2</th>
<th>F0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPC-NB</td>
<td>0.828</td>
<td>0.562</td>
<td>0.623</td>
<td>0.579</td>
<td>0.705</td>
<td></td>
</tr>
<tr>
<td>SPC-COS</td>
<td>0.808</td>
<td>0.641</td>
<td>0.681</td>
<td>0.651</td>
<td>0.736</td>
<td></td>
</tr>
<tr>
<td>PLSA</td>
<td>0.517</td>
<td>0.782</td>
<td>0.543</td>
<td>0.622</td>
<td>0.515</td>
<td></td>
</tr>
</tbody>
</table>

Comparison to other approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>Test set</th>
<th>pur.</th>
<th>invp.</th>
<th>F0.5</th>
<th>F0.2</th>
<th>F0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Naive baselines”</td>
<td></td>
<td>1.000</td>
<td>0.470</td>
<td>0.610</td>
<td>0.520</td>
<td></td>
</tr>
<tr>
<td>ONE-IN-ONE</td>
<td></td>
<td>0.290</td>
<td>1.000</td>
<td>0.400</td>
<td>0.580</td>
<td></td>
</tr>
</tbody>
</table>

This paper

<table>
<thead>
<tr>
<th>Method</th>
<th>Test set</th>
<th>pur.</th>
<th>invp.</th>
<th>F0.5</th>
<th>F0.2</th>
<th>F0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPC-NB</td>
<td>0.884</td>
<td>0.688</td>
<td>0.747</td>
<td>0.707</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPC-COS</td>
<td>0.850</td>
<td>0.777</td>
<td>0.791</td>
<td>0.780</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLSA</td>
<td>0.370</td>
<td>0.885</td>
<td>0.442</td>
<td>0.581</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SemEval 2007 Top 3

<table>
<thead>
<tr>
<th>Method</th>
<th>Test set</th>
<th>pur.</th>
<th>invp.</th>
<th>F0.5</th>
<th>F0.2</th>
<th>F0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>CU_COMSTEM</td>
<td></td>
<td>0.720</td>
<td>0.880</td>
<td>0.780</td>
<td>0.830</td>
<td></td>
</tr>
<tr>
<td>IRST-BP</td>
<td></td>
<td>0.750</td>
<td>0.800</td>
<td>0.750</td>
<td>0.770</td>
<td></td>
</tr>
<tr>
<td>PSNUS</td>
<td></td>
<td>0.730</td>
<td>0.820</td>
<td>0.750</td>
<td>0.780</td>
<td></td>
</tr>
</tbody>
</table>
Wrap up

• Task of person name resolution in web search
• Two approaches
  • SPC (term based)
  • PLSA (semantic based)
• SPC outperforms PLSA and delivers excellent performance
• The “person clustering hypothesis” holds to a large extent

Future work

• Combine advantages of both methods
• Richer feature set (e.g., named entities)
• Pre-processing documents (removing irrelevant content)

Questions?

Krisztian Balog
kbalog@science.uva.nl
http://www.science.uva.nl/~kbalog