Information Extraction from Wikipedia: Moving Down the Long Tail

Fei Wu, Raphael Hoffmann, Daniel S. Weld
Department of Computer Science & Engineering
University of Washington
Seattle, WA, USA
Clearfield County, Pennsylvania

<table>
<thead>
<tr>
<th>Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Founded</td>
<td>March 26, 1804</td>
</tr>
<tr>
<td>Seat</td>
<td>Clearfield</td>
</tr>
<tr>
<td>Area</td>
<td></td>
</tr>
<tr>
<td>- Total</td>
<td>2,988 km² (1,154 mi²)</td>
</tr>
<tr>
<td>- Land</td>
<td>sq mi (km²)</td>
</tr>
<tr>
<td>- Water</td>
<td>17 km² (6 mi²), 0.56%</td>
</tr>
<tr>
<td>Population</td>
<td></td>
</tr>
<tr>
<td>- (2000)</td>
<td>83,382</td>
</tr>
<tr>
<td>- Density</td>
<td>28/km²</td>
</tr>
</tbody>
</table>

Clearfield County was created on 1804 from parts of Huntingdon and Lycoming Counties but was administered as part of Centre County until 1812. Its county seat is Clearfield.

2,972 km² (1,147 mi²) of it is land and 17 km² (7 mi²) of it (0.56%) is water.

As of 2005, the population density was 28.2/km².
Preliminary Evaluation

- **Kylin Performed Well on Popular Classes:**
  - Precision: mid 70% ~ high 90%
  - Recall: low 50% ~ mid 90%

- **Long-Tailed Challenges**
  - Sparse Classes
  - Incomplete Articles
Outline

● Long-Tailed Challenges
  ● Sparse infobox classes
  ● Incomplete articles

● Moving Down the Long Tails
  ● Shrinkage
  ● Retraining
  ● Extracting from the Web

● Conclusion
Long-Tail 1: Sparse Infobox Class

- Kylin Flounders on Sparse Classes – Little Training Data

1442/1756 (82%) <100 instance; 709/1756 (40%) <10 instance

[July 2007 of Wikipedia]
Long-Tail 2: Incomplete Articles

- Desired Information Missing from Wikipedia

800,000/1,800,000 (44.2%) stub pages  [July 2007 of Wikipedia]
Search over Wikip’s Whole Spectrum

Moving Down the Long Tails
Outline

- Long-Tailed Challenges
  - Sparse infobox classes
  - Incomplete articles
- Moving Down the Long Tails
  - Shrinkage
  - Retraining
  - Extraction from the Web
- Conclusion
Shrinkage

[McCallum et al., ICML98]
Shrinkage

- KOG (Kylin Ontology Generator) [Wu & Weld, WWW08]
Shrinkage

- KOG (Kylin Ontology Generator) [Wu & Weld, WWW08]

Diagram:
- person (1201)
- performer (44)
- actor (8738)
- comedian (106)
Shrinkage

- KOG (Kylin Ontology Generator) [Wu & Weld, WWW08]
Shrinkage Experiment

• Settings:
  o Dataset: 07/16/2007 snapshot of en.wikipedia.org

Testing cases:

<table>
<thead>
<tr>
<th>Target class</th>
<th>Parent</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irish Newspaper(20)</td>
<td>Newspaper(1559)</td>
<td>–</td>
</tr>
<tr>
<td>Performer(44)</td>
<td>Person(1201)</td>
<td>Actor(8738)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comedian(106)</td>
</tr>
<tr>
<td>Baseball stadium(163)</td>
<td>Stadium(1642)</td>
<td>–</td>
</tr>
<tr>
<td>Writer(2213)</td>
<td>Person(1201)</td>
<td>Sci-fi writer(36)</td>
</tr>
</tbody>
</table>
Shrinkage Experiment

(a) Irish Newspaper

(b) Performer

(c) Baseball Stadium

(d) Writer
Outline

● Long-Tailed Challenges
  ● Sparse infobox classes
  ● Incomplete articles

● Moving Down the Long Tails
  ● Shrinkage
  ● Retraining
  ● Extraction from the Web

● Conclusion
Retraining

Complementary to Shrinkage:
Harvest extra training data from **broader** Web

**Key:**
- Identify relevant sentences given the sea of Web data?

Andrew Murray was born in Scotland in 1828 ......

<Andrew Murray, was born in, Scotland>
<Andrew Murray, was born in, 1828>
Retraining

Kylin Extraction: \( \{ t | t = \langle \text{subject, attribute, value} \rangle \} \)

TextRunner Extraction: \( \{ r | r = \langle \text{obj}_1, \text{predicate, obj}_2 \rangle \} \)

Query TextRunner for relevant sentences:

\( t = \langle \text{Ada Cambridge, location, “St Germans , Norfolk , England”} \rangle \)

\( R = \{ r | r.\text{obj}_1 = “Ada Cambridge” \text{ and } r.\text{obj}_2 \sim “St Germans , Norfolk , England” \} \)

- \( r1 = \langle \text{Ada Cambridge, was born in, England} \rangle \)
  
  Ada Cambridge was born in England in 1844 and moved to Australia with her curate husband in 1870.

- \( r2 = \langle \text{Ada Cambridge, was born in, “Norfolk , England”} \rangle \)
  
  Ada Cambridge was born in Norfolk , England , in 1844 .
Retraining Experiment

(a) Irish Newspaper
- Baseline
- Retraining
- Shrinkage
- Shrink–retrain
- Irish newspaper (20)
  - Newspaper (1559)

(b) Performer
- Baseline
- Retraining
- Shrinkage
- Shrink–retrain
- Performer (44)
  - Person (1201)
  - Actor (8738)
  - Comedian (106)

(c) Baseball Stadium
- Baseline
- Retraining
- Shrinkage
- Shrink–retrain
- Baseball stadium (163)
  - Stadium (1642)

(d) Writer
- Baseline
- Retraining
- Shrinkage
- Shrink–retrain
- Writer (2213)
  - Person (1201)
  - Sci-fi writer (36)
Outline

● Long-Tailed Challenges
  ● Sparse infobox classes
  ● Incomplete articles

● Moving Down the Long Tails
  ● Shrinkage
  ● Retraining
  ● Extraction from the Web

● Conclusion
Extraction from the Web

- Idea: apply Kylin extractors trained on Wikipedia to general Web pages
- Challenge: maintain high precision
  - General Web pages are noisy
  - Many Web pages describe multiple objects
- Key: retrieve relevant sentences

Procedure
- Generate a set of search engine queries
- Retrieve top-\(k\) pages from Google
- Weight extractions from these pages
Extraction from the Web

Recall AUC Benefit from Shrinkage + Retraining…
Extraction from the Web

AUC Benefit from Shrinkage + Retraining + Web
Summary

IE from Wikipedia: Moving Down the Long Tail

● Shrinkage
  ● Sparse infobox classes

● Retraining based on TextRunner
  ● Sparse infobox classes

● Extracting from the Web
  ● Incomplete articles
Summary

IE from Wikipedia: Moving Down the Long Tail