FireCite: Lightweight real-time reference string extraction from web pages

Ching Hoi Andy Hong  Jesse Prabawa Gozali  Min-Yen Kan
School of Computing
National University of Singapore
Outline

- Introduction
- Reference String Recognition
- Reference String Parsing
- Firefox Extension
- Conclusion
Introduction: The Problem

• Recognition and parsing of references found on the Internet

• Criteria:
  – Accurate
  – Fast

<table>
<thead>
<tr>
<th>Journal Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR NL ML</td>
</tr>
<tr>
<td>IR ML</td>
</tr>
<tr>
<td>NL ML</td>
</tr>
<tr>
<td>ML</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Fuchun Peng, Dale Schuurmans and Shaojun Wang; <em>Augmenting Voka Pages</em></td>
</tr>
</tbody>
</table>
Introduction: Related Work

- Zotero
- Mendeley
- Google Scholar
- CiteSeerx

Publications are listed in reverse chronological order. Presentations slides associated with the publications are supplied when available. You can also review most recent slides and readings from talks, as well as earlier presentations from my PhD years.
Outline

- Introduction
- Reference String Recognition
- Reference String Parsing
- Firefox Extension
- Conclusion
Reference String Recognition: Definition

- Are there reference strings?
- Where are the reference strings?

### Journal Articles


Reference String Recognition: Methodology

1. Web page exclusion
   - Keyword + URL Matching

**Journal Articles**


- Shaojun Wang, Dale Schuurmans, Fuchun Peng and Yunxin Zhao; *Learning Mixture Models with the Regularized Latent Maximum Entropy Principle*,
Reference String Recognition: Methodology

2. Splitting

- Split web page text into segments
- GOAL: Each segment to contain at most 1 reference string, and each reference string to exist in only 1 segment.

### Journal Articles

<table>
<thead>
<tr>
<th>IR</th>
<th>NL</th>
<th>ML</th>
</tr>
</thead>
</table>
| ![Journal Articles](image.png)


Reference String Recognition: Methodology

3. Selection
   - Token length and word length of segment

4. Verification
   - Reject segments that do not have a title or authors
Reference String Recognition: Evaluation

- Test set: 40 staff homepages from 4 universities
- Reference strings found: 364/379 (96.0%)
- False positives: 269 (42.5%)

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>FireCite (All 40 pages)</td>
<td>0.575</td>
<td>0.960</td>
<td>0.719</td>
</tr>
<tr>
<td>FireCite (Only 20 pages with reference strings)</td>
<td>0.655</td>
<td>0.960</td>
<td>0.779</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Related Work
- Reference String Recognition
  - Reference String Parsing
  - Firefox Extension
- Conclusion
Reference String Parsing: Definition

- **Purpose**
  - Store reference according to metadata fields
  - Assist reference string recognition

- Only identify **authors**, **title**, **date**

Reference String Parsing: Methodology

- **Tokenising**


  - **Advantages**
    - Reduce number of computations
    - Allow information-richer learning features
Reference String Parsing: Methodology

- **Labelling**
  - J48 decision tree classifier
  - CORA corpus (500 reference strings) as training corpus

- **Repairs**
  - “Title” and “Authors” fields are contiguous
# Reference String Parsing: Evaluation

## 6 Faculty Staff Publication Pages

<table>
<thead>
<tr>
<th>Page (No. of ref. strings)</th>
<th>Token-level F-measure</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Title</td>
<td>Authors</td>
<td>Date</td>
<td>All Tokens</td>
</tr>
<tr>
<td>A (72)</td>
<td>0.902</td>
<td>0.893</td>
<td>0.988</td>
<td>0.708</td>
</tr>
<tr>
<td>B (52)</td>
<td>0.953</td>
<td>0.957</td>
<td>0.990</td>
<td>0.960</td>
</tr>
<tr>
<td>C (29)</td>
<td>0.684</td>
<td>0.304</td>
<td>0.774</td>
<td>0.651</td>
</tr>
<tr>
<td>D (68)</td>
<td>0.753</td>
<td>0.968</td>
<td>0.889</td>
<td>0.917</td>
</tr>
<tr>
<td>E (8)</td>
<td>0.692</td>
<td>0.875</td>
<td>1.000</td>
<td>0.889</td>
</tr>
<tr>
<td>F (45)</td>
<td>0.847</td>
<td>1.000</td>
<td>0.989</td>
<td>0.966</td>
</tr>
<tr>
<td><strong>Overall (274)</strong></td>
<td><strong>0.836</strong></td>
<td><strong>0.916</strong></td>
<td><strong>0.948</strong></td>
<td><strong>0.878</strong></td>
</tr>
</tbody>
</table>
## Reference String Parsing: Evaluation

**FLUX-CiM Computer Science Dataset (300 citations)**

<table>
<thead>
<tr>
<th>System Name</th>
<th>Field-level F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Title</td>
</tr>
<tr>
<td>FireCite</td>
<td>0.92</td>
</tr>
<tr>
<td>ParsCite</td>
<td>0.96</td>
</tr>
<tr>
<td>FLUX-CiM</td>
<td>0.93</td>
</tr>
</tbody>
</table>
## Reference String Parsing: Evaluation

<table>
<thead>
<tr>
<th>Parser</th>
<th>Classifier Type</th>
<th>Size of classifier model (KB)</th>
<th>Size of dictionary (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FireCite</td>
<td>Decision Tree</td>
<td>12.6</td>
<td>0.0</td>
</tr>
<tr>
<td>FLUX-CiM</td>
<td>Knowledge-Based</td>
<td>&gt;786 (estimated)</td>
<td>0.0</td>
</tr>
<tr>
<td>ParsCit</td>
<td>Conditional Random Fields</td>
<td>7339</td>
<td>1722</td>
</tr>
</tbody>
</table>
## Reference String Parsing: Evaluation

<table>
<thead>
<tr>
<th>Page Type</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages with reference strings</td>
<td>90</td>
<td>544</td>
<td>192</td>
</tr>
<tr>
<td>Pages without reference strings</td>
<td>6</td>
<td>222</td>
<td>74</td>
</tr>
<tr>
<td>All pages</td>
<td>6</td>
<td>544</td>
<td>133</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Reference String Recognition
- Reference String Parsing
- **Firefox Extension**
- Conclusion
Extension: Demo

(a) 

Publications

- Wei Lu, Hwee Tou Ng, Wee Sun Lee. A Generative Model for Parsi.
  [pdf]

  In Proceedings of the 23rd AAAI.

- Hanna Kurniawati, David Hsu, et al. SARSOP: Efficient point-base
  spaces.
Conclusion

- Results
  - Fast and lightweight reference string parser
  - Reference string recogniser with good recall
  - Basic, expandable Firefox extension
Conclusion

- Future work
  - Reference String Recognition
    - More rules to improve precision
  - Reference String Parser
    - Use web page reference strings as training data
    - Recognise implicit/common metadata
  - Firefox Extension
    - Add more features to the extension

Questions?