Corroborate and learn facts from the web

Shubin Zhao and Jonathan Betz

Presentation by Yang Yu
Problem Definition

## Known Facts

<table>
<thead>
<tr>
<th>Entity_Name</th>
<th>Angelina Jolie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date of Birth</td>
<td>June 4, 1975</td>
</tr>
</tbody>
</table>

## More Facts

<table>
<thead>
<tr>
<th>Entity_Name</th>
<th>Angelina Jolie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date of Birth</td>
<td>June 4, 1975</td>
</tr>
<tr>
<td>Academy Awards</td>
<td>?</td>
</tr>
<tr>
<td>Place of birth</td>
<td>?</td>
</tr>
</tbody>
</table>
### Problem Definition

#### More Facts

<table>
<thead>
<tr>
<th>Entity_Name</th>
<th>Angelina Jolie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date of Birth</td>
<td>June 4, 1975</td>
</tr>
<tr>
<td>Academy Awards</td>
<td>?</td>
</tr>
<tr>
<td>Place of birth</td>
<td>?</td>
</tr>
</tbody>
</table>

```html
<tr>
  <td> Date of Birth </td>
  <td> June 4, 1975 </td>
</tr>
<tr>
  <td> Academy Awards </td>
  <td> …… </td>
</tr>
```
Retrieve Relevant Pages

- The page contains the entity names
- Matching anchor text of a page with entity names
- Address ambiguity
- MAPREDUCE
Corroborate Known Facts

- Avoid wrong corroboration on common facts

\[ P(v|A) = \frac{freq(v|A)}{\sum_{v_i}(freq(v_i|A))}. \]
\[ p = \prod_{i} p(v_i|A_i). \]

procedure CorroborateFacts(Entity E, Page P)
  for each fact F in entity E, do
    Search the value F.val in P.
    for each match M_v of F.val, do
      Match attribute name F.attr before and after M_v;
      if there is an attribute match M_a, then
        Cache (F, M_a, M_v) into MentionList;
      end if
    end for
  Compute random prob p of all M_v in MentionList;
  if p is below a threshold then
    for each (F, M_a, M_v) in MentionList do
      Annotate M_a and M_v in page P as a mention of
      fact F.
      Add the url of P to the source list of F.
    end for
  end if
end for
Corroborate Known Facts

- Corroboration Strategies
  - Lexicographical sorting of tokens
  - Using synonyms of attribute names
  - Not counting stopwords
  - Matching of attribute name is optional
- MAPREDUCE

Mapper:
Input: (key=entity-name, 
   value1=entity, value2=relevant-pages-set) 
Output:(entity-name, new-entity)
procedure DiscoverPatterns(HtmlNode Node)
    for (int i=0; i<size(Node.Children); i++) do
        string child_tags = GetHtmlTags(Node.Children[i]);
        NodeData[i] = child_tags;
    end for
    for (i=0; i<size(NodeData); i++) do
        for (j=i+1; j<size(NodeData); j++) do
            if IsSimilar(NodeData[i], NodeData[j]) then
                matched = true;
                for (k=1; k<j-i; k++) do
                    if not IsSimilar(NodeData[i+k], NodeData[j+k])
                        then
                            matched = false;
                            break;
                        end if
                end for
            end if
        end for
    end if
    if (matched) then
        /* Find a pattern of j - i nodes */.
        1) Repeat the previous loop from NodeData[j]
           to find the maximum span of the pattern;
        2) Save the pattern and its span into PatternList;
        end if
    end if
    if not PatternList.empty() then
        1) Save the pattern $P_{max}$ with max span in PatternList;
        2) Mark nodes uncovered by $P_{max}$;
    end if
    for each node $N$ not covered by $P_{max}$ do
        DiscoverPatterns($N$)
    end for
Extract New Facts

An example
Extract New Facts

- **Bootstrapping**

```plaintext
procedure Bootstrap(E, set[P])
    terminate = false;
    for (round=1; terminate is false; round++) do
        terminate = true;
        for each page P in set[P], do
            CorroborateFacts(E, P);
            if fact examples found in P then
                DiscoverPatterns(P);
                \( \bar{E} = \text{ExtractFacts}(E, P) \);
                if there is new facts in \( \bar{E} \) then
                    \( E = \bar{E} \);
                    terminate = false;
                end if
            end if
        end for
    end for
```

Experiments

- Experiments on Country Facts

Table 2: The Learning Results on Country Entities

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corroborated Seed Facts</td>
<td>230</td>
<td>–</td>
</tr>
<tr>
<td>New sources for Seed Facts</td>
<td>28920</td>
<td>99.9%</td>
</tr>
<tr>
<td>Corroborated New Facts*</td>
<td>10656</td>
<td>98.40%</td>
</tr>
<tr>
<td>Uncorroborated New Facts*</td>
<td>106337</td>
<td>92.61%</td>
</tr>
</tbody>
</table>

* Corroborated new facts refer to the extracted facts being corroborated later. Each fact has 14.0 sources in average.
* Uncorroborated new facts only have one source.
Experiments on Wikipedia Facts

<table>
<thead>
<tr>
<th>Category</th>
<th>Round 1 (in millions)</th>
<th>Round 2 (in millions)</th>
<th>Round 3 (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corroborated Seed Facts</td>
<td>1.393</td>
<td>1.393</td>
<td>1.393</td>
</tr>
<tr>
<td>New sources for Seed Facts</td>
<td>5.150</td>
<td>5.150</td>
<td>5.150</td>
</tr>
<tr>
<td>Corroborated New Facts*</td>
<td>0.618</td>
<td>0.815</td>
<td>0.862</td>
</tr>
<tr>
<td>Sources of Corroborated New Facts</td>
<td>4.138</td>
<td>5.941</td>
<td>6.298</td>
</tr>
<tr>
<td>Uncorroborated New Facts</td>
<td>4.176</td>
<td>5.152</td>
<td>5.956</td>
</tr>
<tr>
<td>Corroborated Entities*</td>
<td>0.290</td>
<td>0.290</td>
<td>0.290</td>
</tr>
</tbody>
</table>

* Corroborated new facts refer to the extracted facts being corroborated later. They must have more than one source.
* Corroborated entities refer to entities with at least one fact corroborated.
### Experiments on Wikipedia Facts

#### Table 6: Stats of the Learning Results Per Type on Wikipedia Seeds

<table>
<thead>
<tr>
<th>Type</th>
<th>Seed Facts (in kilos)</th>
<th>*Corroborated Facts (in kilos)</th>
<th>New Facts (in kilos)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>3,347.0</td>
<td>450.6</td>
<td>640.2</td>
</tr>
<tr>
<td>Geo-location</td>
<td>1,510.4</td>
<td>253.6</td>
<td>181.1</td>
</tr>
<tr>
<td>Organization</td>
<td>278.0</td>
<td>55.8</td>
<td>56.1</td>
</tr>
<tr>
<td>Film</td>
<td>264.0</td>
<td>413.6</td>
<td>2,867.3</td>
</tr>
<tr>
<td>Event</td>
<td>180.8</td>
<td>35.1</td>
<td>129.0</td>
</tr>
<tr>
<td>Animal</td>
<td>157.3</td>
<td>31.0</td>
<td>19.3</td>
</tr>
<tr>
<td>Character</td>
<td>151.0</td>
<td>47.5</td>
<td>88.6</td>
</tr>
<tr>
<td>Building</td>
<td>133.2</td>
<td>18.6</td>
<td>17.1</td>
</tr>
<tr>
<td>Book</td>
<td>114.8</td>
<td>106.3</td>
<td>609.3</td>
</tr>
<tr>
<td>Music</td>
<td>93.7</td>
<td>26.7</td>
<td>57.3</td>
</tr>
</tbody>
</table>

*Corroborated Facts include corroborated seed facts and corroborated new facts.
Conclusions

- Difference with related work
  - Wrappers are generated dynamically
  - Using the content examples to locate and to label the extracted data
  - Bootstrapping focused on structured text in HTML
Questions and Comments

Thank You!