Corroborate and learn facts from the web

Shubin Zhao and Jonathan Betz

Presentation by Yang Yu

Problem Definition

Known Facts

Entity_Name	Angelina Jolie
Date of Birth	June 4, 1975

More Facts

Entity_Name	Angelina Jolie
Date of Birth	June 4, 1975
Academy Awards	?
Place of birth	?

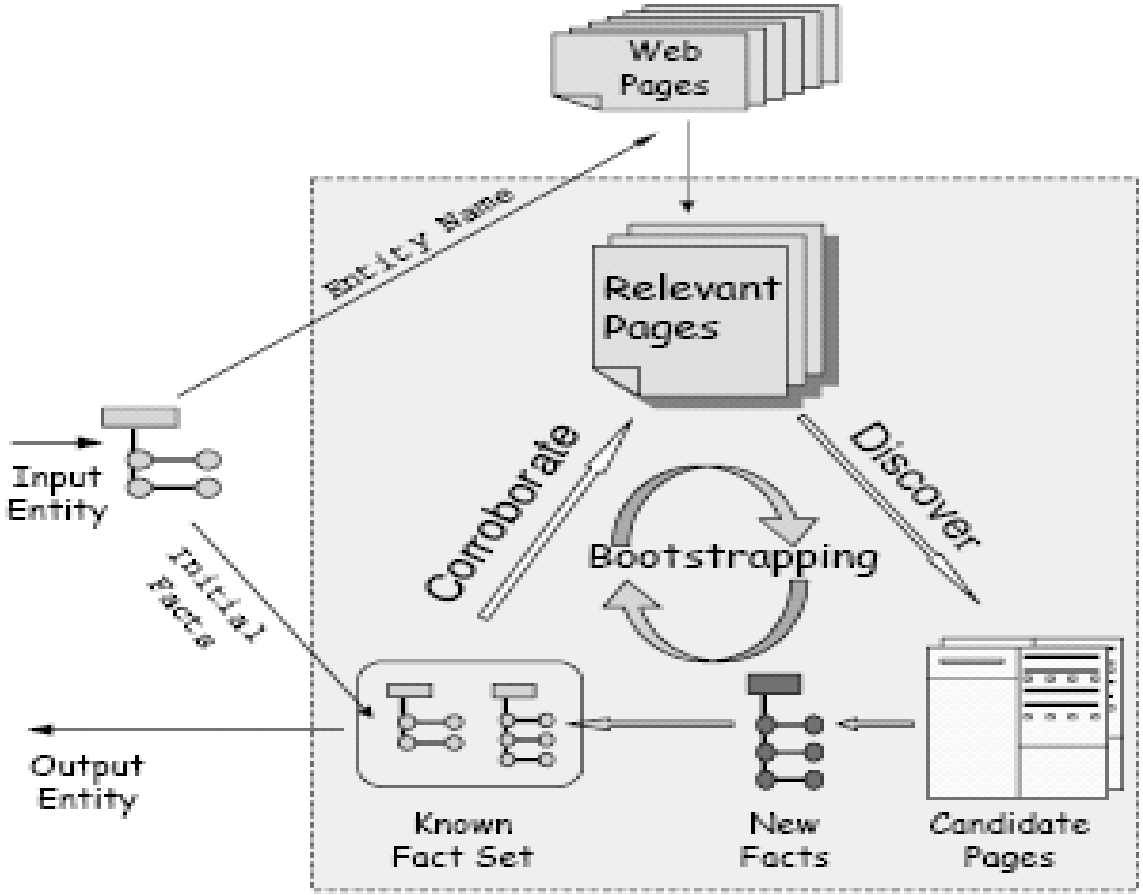
Problem Definition

More Facts

Entity_Name	Angelina Jolie
Date of Birth	June 4, 1975
Academy Awards	?
Place of birth	?

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

GRAZER System Overview



Retrieve Relevant Pages

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

```
Mapper:
  Input:  (Null-key, Crawled-page)
  Output: (Entity-name, Page) or Nothing
Reducer:
  Input:  (Entity-name, Page)
  Output: (Entity-name, Page-list)
```

Corroborate Known Facts

- Avoid wrong corroboration on common facts

$$P(v|A) = freq(v|A) / \sum_{v_i} (freq(v_i|A)). \qquad p = \prod_i p(v_i|A_i).$$

```
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

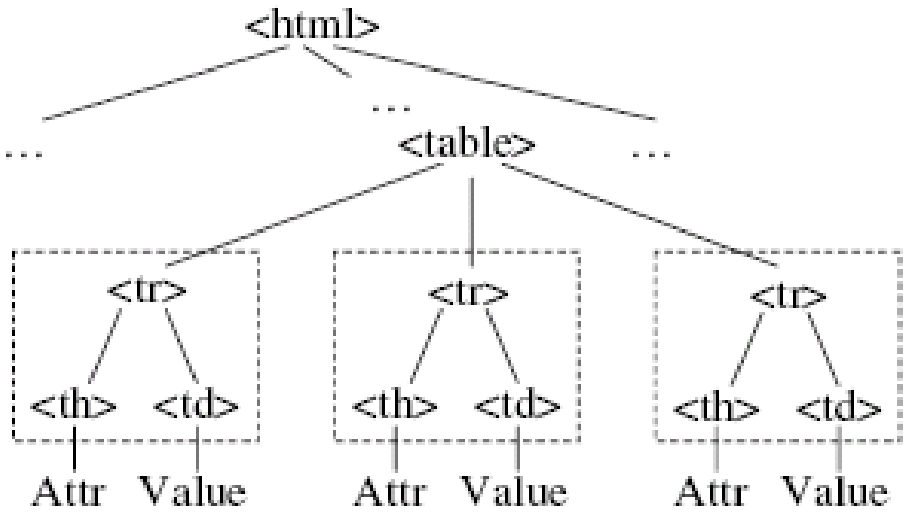
```
Mapper:
 Input: (key=entity-name,
        value1=entity, value2=relevant-pages-set)
 Output:(entity-name, new-entity)
```

Extract New Facts

```
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
        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
    end for
  end for
  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*

```
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);
        Ē = ExtractFacts(E, P);
        if there is new facts in Ē then
          E = Ē;
          terminate = false;
        end if
      end if
    end for
  end for
```

Experiments

- *Experiments on Country Facts*

Table 2: The Learning Results on Country Entities

Category	Count	Precision
Corroborated Seed Facts	230	–
New sources for Seed Facts	28920	99.9%
Corroborated New Facts*	10656	98.40%
Uncorroborated New Facts*	106337	92.61%

* 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

- *Experiments on Wikipedia Facts*

Table 5: Stats of the Learning Results Per Round on Wikipedia Seeds

Category	Round 1 (in millions)	Round 2 (in millions)	Round 3 (in millions)
Corroborated Seed Facts	1.393	1.393	1.393
New sources for Seed Facts	5.150	5.150	5.150
Corroborated New Facts*	0.618	0.815	0.862
Sources of Corroborated New Facts	4.138	5.941	6.298
Uncorroborated New Facts	4.176	5.152	5.956
Corroborated Entities*	0.290	0.290	0.290

* Corroborated new facts refer to the extracted facts being corroborated later. They must have more than one sources.
* Corroborated entities refer to entities with at least one fact corroborated.

Experiments

- *Experiments on Wikipedia Facts*

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

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

*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 !