Information Extraction for Structured Web Queries

Michael J. Cafarella, Christopher Re, Dan Suciu, Oren Etzioni, Michele Banko

UW Affiliates Day
October 30, 2007
Structured Queries, Unstructured Data

“Show me some people, what they invented, and the years they died”

\[ q(?a, ?b, ?c) :\text{ invented}(?a, ?b), \text{ died-in}(?a, \text{<year> ?c}) \]

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kepler</td>
<td>log books</td>
<td>1630</td>
<td>.7902</td>
</tr>
<tr>
<td>Heisenberg</td>
<td>matrix mechanics</td>
<td>1976</td>
<td>.7897</td>
</tr>
<tr>
<td>Galileo</td>
<td>telescope</td>
<td>1642</td>
<td>.7395</td>
</tr>
<tr>
<td>Newton</td>
<td>calculus</td>
<td>1727</td>
<td>.7366</td>
</tr>
</tbody>
</table>
In 1877, Edison invented the phonograph. Although he...
In 1877, Edison invented the phonograph. Although he ...

1. Run extractors
2. Populate data model
3. Queries
Each concept has an IE mechanism

<table>
<thead>
<tr>
<th>Example</th>
<th>Description</th>
<th>IE technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>invented(Edison, phonograph)</td>
<td>Arity-2 fact</td>
<td>TextRunner</td>
</tr>
<tr>
<td>&lt;scientist&gt; Einstein</td>
<td>Type (hyponymy)</td>
<td>KnowItAll</td>
</tr>
<tr>
<td>has-invented = invented</td>
<td>Synonymy</td>
<td>DIRT</td>
</tr>
<tr>
<td>invented ⊆ discovered</td>
<td>ID (troponymy)</td>
<td>?</td>
</tr>
<tr>
<td>FD: has-capital(x, y) → has-capital(y)</td>
<td>FD (rule)</td>
<td>?</td>
</tr>
</tbody>
</table>
no one could have... 

In 1877, Edison invented the phonograph. Although he... 

RDBMS 

Query middleware 

invented(Edison ?e, ?i) 

1. Run extractors 
2. Populate data model 
3. Queries
It was big news when Edison invented the light bulb. He visited cities such as Boston and New York. We all know that Edison invented the light bulb. In 1877 Edison created the light bulb.
In 1877, Edison invented the phonograph. Although he...
Our prototype

- Web crawl: 90M pages
- Facts: 338M tuples, 102M objects
- Types: 6.6M instances
- Synonyms: 17k pairs
- Most queries in ~30 seconds
- Built on DB2 with custom middleware; we want to try a compressed C-store
Related Work

- Query Systems:
  - CI Mple (CIDR07), AVATAR (DEBul06)
  - Liu, Dong, Halevy (WebDB06)
  - Gubanov and Bernstein (WebDB06)

- Extraction: Sarawagi (VLDB06 and others), Etzioni (WWW04), …

- Probabilistic DBs: MYSTIQ, Trio, …

- Deep web, reference reconciliation, …