Searching for Statistical Diagrams

Michael Cafarella
University of Michigan

Joint work with
Shirley Zhe Chen and Eytan Adar

U.S. Frontiers of Engineering Symposium
2011
Statistical Diagrams

- Everywhere in serious academic, governmental, scientific documents
- Our only peek into data behind docs
- Previously rare and precious, Web gives us a flood
  - In small Web crawl, found 319K diagrams in 153K academic papers
- Google makes it easy to find docs, images; very hard to find diagrams
- Searching for diagrams part of larger semantic processing trends
Figure 7.1 Comparison of the Percent of Males Ever Married, U.S. Population and Sacramento County Sheriff’s Department Officers
Previous Work

- Searching for diagrams requires some amount of understanding.
- Lots of work in image search, most inapplicable to diagrams.
- But even understanding diagrams isn’t new.
Previous Work

- Understanding diagrams isn’t new
- Understanding a Web’s worth of diagrams is new
  - Need to search statistical diagrams in medicine, economics, biology, physics, etc
- The phone company can afford a system tailored for manhole diagrams, but we can’t
- Effective scaling with # of topics is central goal of topic-independent information extraction
**Topic-Independent IE**

- Information extraction topic since early 1990s
- Goal is to obtain structured information from unstructured raw documents
  - [Title, Price] from online bookstores
  - [Director, Film] from discussion boards
  - [Scientist, Birthday] from biographies
- Traditional solutions require topic-specific code, features, data
- Costs of TI IE do not grow with # topics
Figure 2 shows the results. As expected, Hadoop’s conventional MapReduce execution time is almost wholly insensitive to the selectivity of the program. MANIMAL execution time, in contrast, decreases as the conditional test becomes more restrictive, dropping to 63% of Hadoop in the case of $^[b]S^+$. There is nothing about the MANIMAL approach...
Our Approach

Typical Web search pipeline
- Crawl Web for documents
- Obtain and index text
- Make index queryable

Our novel components
- Diagram metadata extraction
- Custom search ranker
- Snippet generator
1. Recover good (text, x, y) from PDFs
2. Apply simple role label: title, legend, etc
3. Group texts into “model diagram” candidates, throw away unlikely ones
   - E.g., must include something on x scale
4. Relabel text using geometric relationships
   - Distance, angle to diagram’s origin?
   - Leftmost in diagram? Under a caption?
Snippet Generation

Tested five versions

1. Original-snippet
2. Small-snippet
3. Text-snippet
4. Integrated-snippet
5. Enhanced-snippet
Experiments

- Crawled Web for scientific papers
  - From ClueWeb09
  - Any URL ending in `.pdf` from `.edu` URL
  - 319K diagrams
- Fed data to prototype search engine
- Evaluated
  - Metadata extraction
  - Rank quality
  - Snippet effectiveness
- All results compared against human judgments
Human Population Billions

The figure shows the increase in human population and the usage of nitrogen fertilizer from 1900 to 2000. The graph indicates that both population and fertilizer usage have increased significantly over the past century, with population reaching billions of people. The fertilizer usage also shows a notable increase, especially in recent years. The context explains that this increase is a result of agricultural strategies and practices that have improved crop yields and food production.
## 1. Experiments - Extraction

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Text</td>
<td>All</td>
</tr>
<tr>
<td>title</td>
<td>0.256</td>
<td>0.651</td>
</tr>
<tr>
<td>Y-scale</td>
<td>0.782</td>
<td>0.796</td>
</tr>
<tr>
<td>Y-label</td>
<td>0.835</td>
<td>0.864</td>
</tr>
<tr>
<td>X-scale</td>
<td>0.903</td>
<td>0.835</td>
</tr>
<tr>
<td>X-label</td>
<td>0.241</td>
<td>0.681</td>
</tr>
<tr>
<td>legend</td>
<td>0.520</td>
<td>0.623</td>
</tr>
<tr>
<td>caption</td>
<td>0.952</td>
<td>0.887</td>
</tr>
<tr>
<td>nondiag</td>
<td>0.768</td>
<td>0.924</td>
</tr>
</tbody>
</table>
1. Experiments - Extraction

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Text</td>
<td>All</td>
</tr>
<tr>
<td>title</td>
<td>0.256</td>
<td>0.651</td>
</tr>
<tr>
<td>Y-scale</td>
<td>0.782</td>
<td><strong>0.796</strong></td>
</tr>
<tr>
<td>Y-label</td>
<td>0.835</td>
<td>0.864</td>
</tr>
<tr>
<td>X-scale</td>
<td><strong>0.903</strong></td>
<td>0.835</td>
</tr>
<tr>
<td>X-label</td>
<td>0.241</td>
<td><strong>0.681</strong></td>
</tr>
<tr>
<td>legend</td>
<td>0.520</td>
<td>0.623</td>
</tr>
<tr>
<td>caption</td>
<td>0.952</td>
<td><strong>0.887</strong></td>
</tr>
<tr>
<td>nondiag</td>
<td>0.768</td>
<td><strong>0.924</strong></td>
</tr>
</tbody>
</table>
2. Experiments - Ranking

![Bar Chart](image-url)

- naïve-rank
- reference-rank
- field-rank
- weight-rank

- Top-1 Precision
- Top-3 Precision
- Top-5 Precision
3. Experiments - Snippets

![Graph showing precision for different snippet types]

- **small-snippet**
- **original-snippet**
- **text-snippet**
- **integrated-snippet**
- **enhanced-snippet**

The graph compares the precision of different snippet types, with **enhanced-snippet** showing the highest precision.
Other Applications

Working now:
- Search by axis label
- Search by range
- Given a query diagram (or paper), find related papers

In future:
- Improved academic paper search
- Show plots that support my hypothesis
Future Work

- Spreadsheets
  - Has experiment X ever been run before?
  - WY GDP vs coal production in 2002
  - Preemptively compute good diagrams

- Deeper questions for messy data
  - HTML tables, data files, spreadsheets
  - Lots of structured data lives outside DBMS

- Structured search
Conclusions

- Metadata extraction enables 52% better search ranking
- Extraction-enhanced snippets allow users to choose 33% more accurately
- We rely on open information extraction, but extracted data not the main product
  - Can be successful even with imperfect extractors
Thanks

- Academy of Engineering
- FOE sponsors
- Google
- You!
Related Work

Suitable for Web search settings
- Huang et al, “Associating text and graphics...”, ICDAR 2005
- Huang et al, “Model-based chart image recognition”, GREC 2003

Diagram parsing
- E.g., Futrelle, “Summarization...”, 1999

Visually-impaired access
- E.g., Demir et al, “Generating textual...”, INLG 2008