Corroborating Information from Disagreeing Views

Alban Galland\textsuperscript{1} \hspace{1cm} Serge Abiteboul\textsuperscript{1}
Amélie Marian\textsuperscript{2} \hspace{1cm} Pierre Senellart\textsuperscript{3}

\textsuperscript{1} INRIA Saclay–Île-de-France \hspace{1cm} \textsuperscript{2} Rutgers University \hspace{1cm} \textsuperscript{3} Télécom ParisTech

February 4, 2010, \textit{WSDM}
What are the capital cities of European countries?

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>Poland</th>
<th>Romania</th>
<th>Hungary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Paris</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>Bob</td>
<td>?</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>Charlie</td>
<td>Paris</td>
<td>Rome</td>
<td>Katowice</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>David</td>
<td>Paris</td>
<td>Rome</td>
<td>Bratislava</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>Eve</td>
<td>Paris</td>
<td>Florence</td>
<td>Warsaw</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>Fred</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>George</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Sofia</td>
</tr>
</tbody>
</table>
## Voting

### Information: redundancy

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>Poland</th>
<th>Romania</th>
<th>Hungary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Paris</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>Bob</td>
<td>?</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>Charlie</td>
<td>Paris</td>
<td>Rome</td>
<td>Katowice</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>David</td>
<td>Paris</td>
<td>Rome</td>
<td>Bratislava</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>Eve</td>
<td>Paris</td>
<td>Florence</td>
<td>Warsaw</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>Fred</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>George</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Sofia</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frequence</th>
<th>P. 0.67</th>
<th>R. 0.80</th>
<th>W. 0.60</th>
<th>Buch. 0.50</th>
<th>Bud. 0.43</th>
<th>S. 0.57</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R. 0.33</td>
<td>F. 0.20</td>
<td>K. 0.20</td>
<td>Bud. 0.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B. 0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Corroboration A. Galland WSDM 2010

Introduction 3/26
## Evaluating Trustworthiness of Sources

**Information:** redundancy, trustworthiness of sources (= average frequency of predicted correctness)

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>Poland</th>
<th>Romania</th>
<th>Hungary</th>
<th>Trust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Paris</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.60</td>
</tr>
<tr>
<td>Bob</td>
<td>?</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.58</td>
</tr>
<tr>
<td>Charlie</td>
<td>Paris</td>
<td>Rome</td>
<td>Katowice</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.52</td>
</tr>
<tr>
<td>David</td>
<td>Paris</td>
<td>Rome</td>
<td>Bratislava</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.55</td>
</tr>
<tr>
<td>Eve</td>
<td>Paris</td>
<td>Florence</td>
<td>Warsaw</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.51</td>
</tr>
<tr>
<td>Fred</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.47</td>
</tr>
<tr>
<td>George</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Sofia</td>
<td>0.45</td>
</tr>
</tbody>
</table>

| Frequence | P. 0.70 | R. 0.82 | W. 0.61 | Buch. 0.53 | Bud. 0.46 |
| weighted  | R. 0.30 | F. 0.18 | K. 0.19 | Bud. 0.47 | S. 0.54   |
| by trust  | B 0.20 |        |        |          |          |
## Iterative Fixpoint Computation

**Information:** redundancy, trustworthiness of sources with iterative fixpoint computation

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>Poland</th>
<th>Romania</th>
<th>Hungary</th>
<th>Trust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Paris</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.65</td>
</tr>
<tr>
<td>Bob</td>
<td>?</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.63</td>
</tr>
<tr>
<td>Charlie</td>
<td>Paris</td>
<td>Rome</td>
<td>Katowice</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.57</td>
</tr>
<tr>
<td>David</td>
<td>Paris</td>
<td>Rome</td>
<td>Bratislava</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.54</td>
</tr>
<tr>
<td>Eve</td>
<td>Paris</td>
<td>Florence</td>
<td>Warsaw</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.49</td>
</tr>
<tr>
<td>Fred</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.39</td>
</tr>
<tr>
<td>George</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Sofia</td>
<td>0.37</td>
</tr>
</tbody>
</table>

| Frequence | P. 0.75 | R. 0.83 | W. 0.62 | Buch. 0.57 | Bud. 0.51 |
| weighted  | R. 0.25 | F. 0.17 | K. 0.20 | Bud. 0.43 | S. 0.49   |
| by trust  |         |        |        | B 0.19     |           |
Context and problem

- **Context:**
  - Set of sources stating facts
  - (Possible) functional dependencies between facts
  - **Fully unsupervised setting:** we do not assume any information on truth values of facts or inherent trust in sources

- **Problem:** determine which facts are true and which facts are false

- **Real world applications:** query answering, source selection, data quality assessment on the web, making good use of the wisdom of crowds
Outline

Introduction

Model

Algorithms

Experiments

Conclusion
Outline

Introduction

Model

Algorithms

Experiments

Conclusion
General Model

- Set of facts $\mathcal{F} = \{f_1...f_n\}$
  - Examples: “Paris is capital of France”, “Rome is capital of France”, “Rome is capital of Italy”
- Set of views (= sources) $\mathcal{V} = \{V_1...V_m\}$, where a view is a partial mapping from $\mathcal{F}$ to $\{T, F\}$
  - Example:
    - $\neg$ “Paris is capital of France” $\land$ “Rome is capital of France”
- **Objective**: find the most likely real world $\mathcal{W}$ given $\mathcal{V}$ where the real world is a total mapping from $\mathcal{F}$ to $\{T, F\}$
  - Example:
    - “Paris is capital of France” $\land$ $\neg$ “Rome is capital of France” $\land$ “Rome is capital of Italy” $\land$ ...
Generative Probabilistic Model

\[ V_i, f_j \]

- \( \varphi(V_i)\varphi(f_j) \): probability that \( V_i \) “forgets” \( f_j \)
- \( \epsilon(V_i)\epsilon(f_j) \): probability that \( V_i \) “makes an error” on \( f_j \)
- Number of parameters: \( n + 2(n + m) \)
- Size of data: \( \bar{\varphi}nm \) with \( \bar{\varphi} \) the average forget rate
Obvious Approach

- **Method:** use this generative model to find the most likely parameters given the data
  - Inverse the generative model to compute the probability of a set of parameters given the data

- **Not practically applicable:**
  - Non-linearity of the model and boolean parameter $\mathcal{W}(f_j)$
    $\Rightarrow$ equations for inverting the generative model very complex
  - Large number of parameters ($n$ and $m$ can both be quite large)
    $\Rightarrow$ Any exponential technique unpractical

$\Rightarrow$ Heuristic fix-point algorithms
Baselines

Counting (does not look at negative statements, popularity)

\[
\begin{cases} 
T & \text{if } \frac{|\{V_i : V_i(f_j) = T\}|}{\max_f |\{V_i : V_i(f) = T\}|} \geq \eta \\ 
F & \text{otherwise} 
\end{cases}
\]

Voting (adapted only with negative statements)

\[
\begin{cases} 
T & \text{if } \frac{|\{V_i : V_i(f_j) = T\}|}{|\{V_i : V_i(f) = T \lor V_i(f_j) = F\}|} \geq 0.5 \\ 
F & \text{otherwise} 
\end{cases}
\]

TruthFinder [YHY07]: heuristic fix-point method from the literature
3-Estimates

- Iterative estimation of 3 kind of parameters:
  - truth value of facts
  - error rate or trustworthiness of sources
  - hardness of facts
- Tricky normalization to ensure stability
So far, the models and algorithms are about positive and negative statements, without correlation between facts.

How to deal with functional dependencies (e.g., capital cities)?

**pre-filtering:** When a view states a value, all other values governed by this FD are considered stated false.

*If I say that Paris is the capital of France, then I say that neither Rome nor Lyon nor . . . is the capital of France.*

**post-filtering:** Choose the best answer for a given FD.
Outline

Introduction

Model

Algorithms

Experiments

Conclusion
Datasets

- Synthetic dataset: large scale and highly customizable
- Real-world datasets:
  - General-knowledge quiz
  - Biology 6th-grade test
  - Search-engines results
  - Hubdub
http://www.hubdub.com/

- 357 questions, 1 to 20 answers, 473 participants
### Hubdub (2/2)

<table>
<thead>
<tr>
<th></th>
<th>Number of errors (no post-filtering)</th>
<th>Number of errors (with post-filtering)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>278</td>
<td>292</td>
</tr>
<tr>
<td>Counting</td>
<td>340</td>
<td>327</td>
</tr>
<tr>
<td>TruthFinder</td>
<td>458</td>
<td>274</td>
</tr>
<tr>
<td>3-Estimates</td>
<td>272</td>
<td>270</td>
</tr>
</tbody>
</table>
General-Knowledge Quiz (1/2)

1. Where is the city of Ushuaia located?
   - Don’t know
   - In Italy
   - In Greece
   - In Argentina
   - In the Ivory Coast
   - In Sweden
   - In Malaysia

2. What is the last word of all three parts of Dante’s *Divine Comedy* (Hell — Purgatory — Paradise)?
   - Don’t know
   - “Stars” (“Stelle”)
   - “God” (“Dio”)
   - “Hope” (“Speranza”)
   - “Beatrice”

3. Who discovered the planet Uranus?
   - Don’t know
   - Sir William Herschel (in 1781)
   - Urbain Le Verrier (in 1846)
   - Clyde Tombaugh (in 1930)
   - Percival Lowell (in 1894)

http://www.madore.org/~david/quizz/quizz1.html

- 17 questions, 4 to 14 answers, 601 participants
## General-Knowledge Quiz (2/2)

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of errors (no post-filtering)</th>
<th>Number of errors (with post-filtering)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Counting</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>TruthFinder</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3-Estimates</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>
Outline

Introduction

Model

Algorithms

Experiments

Conclusion
In brief

- We believe truth discovery is an important problem, we do not claim we have solved it completely
- Collection of fix-point methods (see paper), one of them (3-Estimates) performing remarkably and regularly well
- Cool real-world applications!

All code and datasets available from http://datacorrob.gforge.inria.fr/
Thanks.

Foundations of Web data management
Perspectives

- Exploiting dependencies between sources [DBES09]
- Numerical values (1.77\text{m} and 1.78\text{m} cannot be seen as two completely contradictory statements for a height)
- No clear functional dependencies, but a limited number of values for a given object (e.g., phone numbers)
- Pre-existing trust, e.g., in a social network
- Clustering of facts, each source being trustworthy for a given field

Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Truth discovery with multiple conflicting information providers on the Web. In Proc. KDD, San Jose, California, USA, August 2007.