Creating Probabilistic Databases from Information Extraction Models

Rahul Gupta, Sunita Sarawagi

Presented by Guozhang Wang

DB Lunch, April 13rd, 2009

Several slides are from the authors
Outline

• Problem background and challenges
• Proposed Solutions
  ◦ Segmentation-per-row model
  ◦ One-row model
  ◦ Multi-row model
• Experiments and conclusion
Extracting and Managing Structured Web Data

- **Information Extraction (using CRF, etc):**
  - Text Segmentation (McCallum, UMASS)
  - Table Extraction (Cafarella, UW)
  - Preference Collection (Wortman, UPenn)

- **Uncertainty Management:**
  - RDBMS
  - Prob. RDBMS
Challenges in Presenting Data

- Segmentation-per-row model
- Storage efficiency v.s. query accuracy
  - Top-1 v.s. all segmentation for each string

52-A Goregaon West Mumbai 400 062

<table>
<thead>
<tr>
<th>House_no</th>
<th>Area</th>
<th>City</th>
<th>Pincode</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>Goregaon West</td>
<td>Mumbai</td>
<td>400 062</td>
<td>0.1</td>
</tr>
<tr>
<td>52-A</td>
<td>Goregaon</td>
<td>West Mumbai</td>
<td>400 062</td>
<td>0.2</td>
</tr>
<tr>
<td>52-A</td>
<td>Goregaon West</td>
<td>Mumbai</td>
<td>400 062</td>
<td>0.5</td>
</tr>
<tr>
<td>52</td>
<td>Goregaon</td>
<td>West Mumbai</td>
<td>400 062</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Confidence = Probability of Correctness

![Graph showing the relationship between confidence and probability of top segmentation. The x-axis represents the probability of top segmentation, ranging from 0 to 0.9, and the y-axis represents the fraction correct, ranging from 0 to 1. The graph shows an upward trend with increasing confidence.]
Trade-off Between Accuracy and Efficiency

- **Query Accuracy**

![Graph showing trade-off between number of columns in projection query and square error. The graph compares 'Only best extraction' with 'All extractions with probabilities.' The red line represents 'Only best extraction,' and the green line represents 'All extractions with probabilities.' As the number of columns increases, both lines show an increase in square error, with the 'Only best extraction' line generally higher than the 'All extractions with probabilities' line.](image-url)
Trade-off Between Accuracy and Efficiency II

- Storage Efficiency

Number of segmentations required to cover 0.9 probability
Goal of This Paper

- Design data models to achieve good trade-offs between storage efficiency and query accuracy
  - To achieve query accuracy
    - Approximate the extracted segmentation distribution as similar as possible
    - Similarity metric: KL-Divergence

\[
KL(P||Q) = \sum_s P(s) \log \left( \frac{P(s)}{Q(s)} \right)
\]
Outline

- Problem background and challenges
- Proposed Solutions
  - Segmentation-per-row model
  - One-row model
  - Multi-row model
- Experiments and conclusion
Proposed Data Models

- Segmentation-per-row model (Exact)
- One-row model (Column Independence)
- Multi-row model (Mixture of the two)
### Segmentation-per-row Model

<table>
<thead>
<tr>
<th>HNO</th>
<th>AREA</th>
<th>CITY</th>
<th>PINCODE</th>
<th>PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>Bandra West</td>
<td>Bombay</td>
<td>400 062</td>
<td>0.1</td>
</tr>
<tr>
<td>52-A</td>
<td>Bandra</td>
<td>West Bombay</td>
<td>400 062</td>
<td>0.2</td>
</tr>
<tr>
<td>52-A</td>
<td>Bandra West</td>
<td>Bombay</td>
<td>400 062</td>
<td>0.5</td>
</tr>
<tr>
<td>52</td>
<td>Bandra</td>
<td>West Bombay</td>
<td>400 062</td>
<td>0.2</td>
</tr>
</tbody>
</table>

- Exact but impractical. We can have too many segmentations!
One-row Model

Each column has an independent multinomial distribution \( Q_y(t,u) \)

- E.g. \( P(52-A, \text{Bandra West}, \text{Bombay}, 400 062) = 0.7 \times 0.6 \times 0.6 \times 1.0 = 0.252 \)

- Simple model, but computed confidences are approximated (even wrong)
Populating One-row Model

\[
\begin{align*}
\text{Min } KL(P\|Q) &= \text{Min } KL(P\| \prod_y Q_y) \\
&= \text{Min } \sum_y KL(P_y\|Q_y)
\end{align*}
\]

- Has a **closed form** solution \( Q_y(t,u) = P(t,u,y) \) where \( P(t,u,y) \) is marginal dist’n.
- Marginal \( P(t,u,y) \) can be computed using **forward-backward** message passing algorithm:
Forward-Backward Algorithm

\[ P(t,u,y) = c \beta_u(y) \sum_{y'} \alpha_{t-1}(y') \text{Score}(t,u,y,y') \]
### Multi-row Model

- **Rows with same ID are mutually exclusive with row probability \( \pi_k \)**
- **Columns in same row are independent**
  - E.g. \[ P(52-A, \text{Bandra West}, \text{Bombay}, 400 062) = 0.833 \times 1.0 \times 1.0 \times 1.0 \times 0.6 + 0.5 \times 0.0 \times 0.0 \times 1.0 \times 0.4 = 0.50 \]

<table>
<thead>
<tr>
<th>HNO</th>
<th>AREA</th>
<th>CITY</th>
<th>PINCODE</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>52 (0.167)</td>
<td>Bandra West (1.0)</td>
<td>Bombay (1.0)</td>
<td>400 062 (1.0)</td>
<td>0.6</td>
</tr>
<tr>
<td>52-A (0.833)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52 (0.5)</td>
<td>Bandra (1.0)</td>
<td>West Bombay (1.0)</td>
<td>400 062 (1.0)</td>
<td>0.4</td>
</tr>
<tr>
<td>52-A (0.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Populating Multi-row Model (fix $k$)

$$\text{Min } KL(P||Q) = \text{Max } \sum_s KL(P_s|| \sum_k \pi_k Q^k_s)$$

- We cannot obtain the optimal parameter values in closed form because of the summation within the log.
- However, we can reduce this to a well-known mixture model parameter estimation problem, and solve it using EM algorithm.
Enumeration-based EM Approach

- Initially guess the parameter values $\pi_k$ and $Q^k_y(t,u)$
- E Step: soft assign each segmentation $s_d$ to segmentation $k$
- M Step: update the parameters with ML values using the above soft assignment

Note the E step need to enumerate all segmentations $s_d$
Enumeration-less Approach

- **Observation:**
  - We need to enumerate segmentations at E step since we use soft assignment.

- **Idea:**
  - Use **hard assignment** instead, so that each $s_d$ belongs to exactly one component.
    - We use a decision tree to make the hard assignment (use information gain to split node)
    - Then we can have a closed form solution to the optimization problem
    - **Merge mechanism** to remove the disjointness limit
Outline

- Problem background and challenges
- Proposed Solutions
  - Segmentation-per-row model
  - One-row model
  - Multi-row model
- Experiments and conclusion
Experiment I

- Comparing multi-row with SPR
Experiment II

- Comparing multi-row with one-row
Lessons Learned?

- Column Independence might not be suitable in some cases (8% v.s. 25%)
- Multi-row model has a good illustration of the correlations between columns

(but) How to implement this probabilistic model?
  - One single row in Multi-row model will take more space

Are accuracy and space efficiency equally important in this application scenario?
Questions?