Data in the first mile

Usher: Improving Data Quality with Dynamic Forms

Chen et al.
Problems

• Getting quality data from developing nations
• Sub Saharan Africa
• Tanzania
Problems

- Collecting data with
  - Limited Resources
  - Power, Bandwidth, Education
  - Limited expertise
  - Attrition
Problems

- Innovation pile up
- Naïve mobile form interfaces
Problems

- Old QA Metrics in use
- Popular gold standard - double entry
Goal

- Get Quality Data!
- Partner with an org. and solve their problems!
Projects

- Shreddr
- Usher
Transcribe the following.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Marital status</td>
<td>Single</td>
</tr>
<tr>
<td>Street</td>
<td></td>
</tr>
<tr>
<td>City</td>
<td></td>
</tr>
</tbody>
</table>

Mark the values that are NOT 'Michael'.

A   B

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael</td>
<td>MICHAEL</td>
<td>Michael</td>
<td>Michael</td>
<td>Phillip</td>
<td>Michael</td>
<td>Michael</td>
<td>Michael</td>
</tr>
<tr>
<td>Michael</td>
<td>Mike</td>
<td>Michael</td>
<td>Mike</td>
<td>Micheld</td>
<td>Michael</td>
<td>Steve</td>
<td></td>
</tr>
<tr>
<td>Michael</td>
<td>Michael</td>
<td>Wesley</td>
<td>Michael</td>
<td>Michael</td>
<td>Michael</td>
<td>Michael</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Usher Goals

• Automatically improve data quality
• Done during data collection
• Improve form efficiency
• Applicable for arbitrary data
• On Cloud
curbstoning

• Get surveyee to answer few questions
• Fill up rest of the form
Usher

- Learns and applies a probabilistic model over form questions
- Model specific to form and data set
Datasets

- Patient Dataset xxxx 15q
  - transcribed from paper patient-registration forms at an HIV/AIDS program at Tanzania
- Survey Dataset 1986 9q
  - Phone survey of political opinion in SF, entered into an electronic form
Usher - Model

Bayesian Network for Patient data set
Usher - Model

Bayesian Network for Patient data set

- Edge -> dependency b/w 2 random variables
- No path -> probabilistically independent
Question Layout

<table>
<thead>
<tr>
<th>ReferredFrom</th>
</tr>
</thead>
<tbody>
<tr>
<td>DistrictCode</td>
</tr>
<tr>
<td>DateConfirmedHIVPositive</td>
</tr>
<tr>
<td>DateOfBirth</td>
</tr>
<tr>
<td>MaritalStatus</td>
</tr>
<tr>
<td>PriorExposure</td>
</tr>
<tr>
<td>DateFirstPositiveHIVTest</td>
</tr>
<tr>
<td>Sex</td>
</tr>
<tr>
<td>RegionCode</td>
</tr>
</tbody>
</table>

- Question layout generated by Usher
Example

- Question Reformulation
  - Is Region code xxxxx? (yes/no)
Usher: Key concepts

• Greedy Information gain
  • Entropy-optimal question ordering
• Appropriate Entry Friction
• Contextualized error likelihood
  • Dynamic feedback
Implementation

- BANJO
- JavaBayes
- Infer.NET
Learning the model

• Build a Bayesian network over form questions

• Estimate Parameters of the resulting network
Learning the model

• Naïve Approach
  • assume complete dependence of each question on every other question
Learning the model

• Usher Approach
• Restrict Feature Space
• Use prior form submissions and choose best structure using Bayesian Dirichlet Equivalence with simulated annealing
Learning the model

\[ P(F_i = f_i \mid \{F_j = f_j : F_j \in \mathcal{P}(F_i)\}) = \frac{N(F_i = f_i, \{F_j = f_j : F_j \in \mathcal{P}(F_i)\})}{N(\{F_j = f_j : F_j \in \mathcal{P}(F_i)\})} \]

- Conditional Probability
Learning the model

\[ P(F_i = f_i \mid \{ F_j = f_j : F_j \in \mathcal{P}(F_i) \}) = (1 - \alpha) \frac{N(F_i = f_i, \{ F_j = f_j : F_j \in \mathcal{P}(F_i) \})}{N(\{ F_j = f_j : F_j \in \mathcal{P}(F_i) \})} + \frac{\alpha}{m}, \]

- After Smoothing
- Jelinek-Mercer
Question Ordering

- Greedy information gain
- Information Entropy
Question Reordering

- Total information per form is fixed
- Maximize conditional information entropy

\[ H(F_i | G) = - \sum_{g=(f_1,\ldots,f_n)} \sum_{f_i} P(G = g, F_i = f_i) \log P(F_i = f_i | G = g), \]
Question Reordering

**Input:** Model $\mathcal{G}$ with questions $\mathbf{F} = \{F_1, \ldots, F_n\}$

**Output:** Ordering of questions $\mathbf{O} = (O_1, \ldots, O_n)$

$\mathbf{O} \leftarrow \emptyset$

while $|\mathbf{O}| < n$ do

\[
F \leftarrow \arg\max_{F_i \notin \mathbf{O}} H(F_i \mid \mathbf{O});
\]

$\mathbf{O} \leftarrow (\mathbf{O}, F)$;

end

**Algorithm 1:** Static ordering algorithm for form layout.

- Offline Static Algorithm
Question Reordering

- Dynamic reordering
  - E.g. If gender == male skip isPregnant
  - Use previous responses
Error Model

- $F_i$: Correct Value
- $D_i$: Question Response
- $R_i$: Binary hidden variable specifying error
- $\theta_i$: Probability distribution of mistakes
Error Model

\[ P(R_i \mid D = d) \]

- Contextualized Error Likelihood
Question Reasking

• When to reask?
  • When errors occur
    • Pros
      • caught as errors occur
    • Cons
      • context ignored
Question Reasking

- Batch
- Recency vs Ease, Accuracy
- Usher
- Error Probability, Budget
Question
Reformulation

1. How did you come to the clinic?
   - Ambulance
   - Bicycle
   - Bus (Daladala)
   - Car taxi (Special hire)
   - Foot
   - Motorcycle taxi (Bodaboda)
   - Private vehicle
   - Other

2. How did you come to the clinic?
   - Foot
   - <another answer>
Question Reformulation

- Static
  - During Form Layout
- Dynamic
  - During Form Filling
- Post-Entry
  - Applied in Conjunction with reasking
Evaluations

• 80pc Training/ 20pc Test
• Predicting missing responses
• Identify Erroneous responses accurately
• Question Reformulation
Evaluation

Survey Dataset

Patient Dataset

Dynamic Reordering
Static Ordering
Original Ordering
Random
Evaluation
Mockups

(a) Select the referring organization
- People living with HIV/AIDS group (31%)
- Sexually transmitted infections clinic (21%)
- Home based care programme (09%)
- In patient department of hospital (01%)

(b) Select the district code
- Dodoma Rural
- Dodoma Urban

(c) Choose the patient's gender
- Male (40%)
- Female (59%)
Summary

• Benefits
  • Potential to reduce error

• Detractions
  • Training data
  • A wrong model learned might be worse than no model learned