Usher: Improving Data Quality with Dynamic Forms

Kuang Chen
kuangc@cs.berkeley.edu

Joint work with Harr Chen, Neil Conway, Joseph M. Hellerstein, Tapan S. Parikh

March 2, 2010
Agenda

• Reform the form!
• A new approach to data collection
• Usher
Paper data entry

National AIDS Control Programme
CTC2 Database

Patient Registration

MS Access data entry forms for Tanzanian HIV/AIDS monitoring
Mobile data entry

Community health workers in Ruhiira, Uganda
Direct user entry

SurveyMonkey

http://www.flickr.com/photos/bkrealist/3695750517
Why we should care?

We depend on its accuracy.
The form needs reform

• Quality data is critically important
• DBMS research has focused on data cleaning (*post-hoc*)
• **What about collection time?**
  – Still an opportunity to correct an answer, if we can detect the error
• Let’s take a **data-driven** approach!
Beyond query interfaces

Entry interfaces!
Best practices from survey methodology

• Form design principles
  – question ordering, grouping, encoding
  – constraints and cross-validations

• Data quality assurance
  – Double, triple entry

• Form design experts and resources required!
  – What about the low resource organizations?
Data collection in low resource orgs

• Lack of required expertise
  – Form design
  – Domain knowledge

• Lack of resources
  – Ad hoc form design
  – Double entry is costly and slow

“the one way street of public health”
Data collection is evolving

- Mobile data collection
  - Displaces traditional paper-transcription + double entry
  - Data quality 10x worse
    [Patnaik et al, ICTD 2009]

- Millions of web surveys designed by amateurs
  - No double entry here either

Time for a more principled and automatic approach?
Agenda

• Reform the form
• A new approach to data collection
• Usher
<table>
<thead>
<tr>
<th>Challenge</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of expertise</td>
<td>Automatic form generation</td>
</tr>
<tr>
<td>No intelligence in the interface</td>
<td>Forms with dynamic feedback</td>
</tr>
<tr>
<td>Quality assurance (double-entry) too</td>
<td>Smart question confirmation</td>
</tr>
<tr>
<td>costly</td>
<td></td>
</tr>
</tbody>
</table>
A data-driven approach

• Learn probabilistic model over arbitrary form
  – Input: form variables & previous form data

• Apply model in each stage of data entry
  1. Before: automatic form layout (ordering)
  2. During: adapt the form to entered values
  3. After: re-confirm likely-wrong responses
Question ordering

• Some questions are more important than others
  – Measure by expected information or entropy
  – How to win at 20 questions: maximize information gain
Dynamic ordering and widgets

• For direct entry: *dynamically pick the next best question*

• In entry interfaces: apply *friction* in proportion to likelihood
Smart re-asking

- During or immediately after entry: perform **online outlier detection**
Agenda

• Reform the form
• A new approach to data collection
• Usher

or
System components and data flow

Form Specification
- form fields, prior data
- expected, error likelihoods
- entered values

Electronic form

Probabilistic Model

Data entry clerks in Ruhiira, Uganda
Data

• Datasets
  – Survey of race and politics (survey)
  – Patient registration data from an HIV/AIDS program in Tanzania (patient)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Survey</th>
<th>Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td># Records</td>
<td>1113</td>
<td>1605</td>
</tr>
<tr>
<td># Questions</td>
<td>15</td>
<td>9</td>
</tr>
</tbody>
</table>

• Data preparation
  – Remove free-text fields
  – Discretize continuous values
  – Absent data as separate value
Building the probabilistic model 1/2

Goal:
• Given a subset of answers, predict values for unanswered questions

Step 1
• Learn Bayesian network from relationships between questions via structural learning
Building the probabilistic model 2/2

Step 2

• Calculate a conditional probability table (CPT) for each question, given its related questions (parameter estimation)
  – CPT values: proportion of each possible answer value, given each possible assignment of parent values
  – Smooth estimates
Before entry: a better static form

• Induce optimal ordering for questions that reduce uncertainty as early as possible

• Technique: greedy algorithm calculating expected conditional entropies

• Benefits
  – More reliably predict remaining fields
  – Allows for better dynamic feedback during entry
Question ordering evaluation

Hypothesis
• Usher orderings asks the most uncertain questions first, improves predictive ability

Results
• Usher orderings yield greater predictive power than the original and random orderings
During entry: a dynamic form

- Dynamically pick the next best question

- Enables dynamic feedback through the entry interface...
During entry: adaptive entry interfaces

- Use probabilities to *assess* an entered value
- Use expectations to *nudge* users toward expected values
- Use probabilities to *guide (bias)* users towards expected results

**Automating constraints**

*friction!"
After entry: smart reconfirmation

- Apply model to identify potential errors
- Graphical model captures probabilistic view of error
  - Entered value ($D$) is either true ($F$) or noise ($\Theta$)
  - Inference algorithm predicts probability of error ($R$)
- Re-ask questions with high $R$ and flag discrepancies for reconciliation

Error model:

- $F_i$: true unknown value
- $D_i$: noisy entered value
- $\theta_i$: error distribution
- $R_i$: binary variable indicating whether there is an error
- $Z$: connects correlated fields
Re-assembling evaluation

Hypothesis
• Usher re-asking can identify erroneous responses accurately

Results
• Usher answer confirmation can correct erroneous responses with less user effort
Summary

• Automated probabilistic approaches can promote high data quality through intelligent forms
• Usher addresses each step of data entry
• Data collection is pertinent DBMS research
DB research in developing regions

• Data is everywhere, but we are not
  – Many non-profit orgs, doing interesting work, are more than willing to share data
  – Incidental help for disadvantaged communities

• Information and Communication Technology for Development (ICTD) as DBMS research
  – It’s possible, fun and personally rewarding!
  – ICTD & SIGCAS, SIGDEV?
Questions?

kuangc@cs.berkeley.edu
Backup slides
Future Directions

• Form optimization service
• Mobile implementation
• Shreddr
FAQ

• Data collection vs survey research
• Electronic vs paper
• What if there is no/insufficient data
• Generalizability of models