Improving Data Quality With Dynamic Forms

Kuang Chen, Harr Chen, Neil Conway, Heather Dolan, Joseph M. Hellerstein, and Tapan S. Parikh

***Abstract*—Organizations in developing regions want to efficiently collect digital data, but standard data gathering practices from the developed world are often inappropriate. Traditional techniques for form design and data quality are expensive and labour-intensive. We propose a new data-driven approach to form design, execution (filling) and quality assurance. We demonstrate USHER, an end-to-end system that automatically generates data entry forms that enforce and maintain data quality constraints during execution. The system features a probabilistic engine that drives form-user interactions to encourage correct answers.**

I. INTRODUCTION

Governments, companies, and individuals routinely make important decisions based on inaccurate data stored in supposedly authoritative databases. In healthcare, a simple error may have fatal consequences. While data quality can be addressed at every stage of the data life-cycle, from creation to archival, we believe that entry-time is the first and best opportunity to improve the quality of manually-entered data. There is much prior work on improving the quality of data that already resides in a database [1]. However, relatively little attention has been paid to improved techniques for data entry.

Survey design [2] has long informed the design of data entry forms, applying principles for data encodings, constraints, and validation rules. For electronic forms, quality assurance during entry has centered on the ubiquitous and costly practice of double-entry [3]. Current standards have failed to take advantage of new technology: pervasive cellular networking and low-cost mobile devices allows even remote users to interact with data entry systems that could potentially provide rich feedback.

For organizations with limited resources, existing standards are neither practical nor attainable. In such settings, designing data collection instruments is too often an ad hoc practice, consisting of mapping desired information elements to a set of entry widgets (text fields, combo boxes, etc.), guided only by the designer's intuition. According to recent work on data collection in resource-poor settings, lack of expertise and difficulty of remote data collection are the chief obstacles to high data quality [4]. In our previous fieldwork with a well-funded HIV/AIDS treatment program in East Africa, we found that little thought was given to form design, and a haphazard double-entry program bottlenecked the data entry process to a substantial degree; in fact, the program's health clinic operations were limited to using paper forms to ensure timely information access. Only after a labor-intensive delay did the medical researchers enjoy the benefits of digital data for research and analysis.

We have built a system called USHER that maximizes data quality at entry-time using statistical data modeling, dynamic interfaces, and collaborative insight. Guided by prior data, USHER learns probabilistic relationships in the data to train a model, which is then applied to automatically generate forms with the appropriate constraints. USHER then provides real-time feedback during the data entry process to dynamically guide (or usher) the user toward better data quality.

Based on a list of form questions and a sufficient set of answers, USHER optimizes the form's question-ordering and layout, mimicking survey design principles. During form entry, USHER provides dynamic data-quality feedback to the user. When the user enters a value, USHER automatically decorates the interface with hints and warnings if the answer is deemed "risky." Decoration choices are probabilistically guided, and include auto-complete, correctness-thermometers, warning/error flags, and other scented widgets [5]. USHER also invites the user to write and view comments about form questions or data instances for and by other users. Finally, USHER mimics double-entry by choosing to re-ask questions with responses likely to be erroneous, based on the probabilistic model.

II. DEMONSTRATION

Our demonstration will show USHER's ability to approximate expert form design and double-entry based only on prior data, both on a PC and a mobile device. Using a real dataset from a rural health organization, users will be able to 1) automatically extract training data from a Microsoft Access database; 2) refine the automatically designed form; and 3) execute the forms with and without smart decorations and quality assurance.

K. Chen, N. Conway and J. M. Hellerstein are with the Computer Science Division, University of California at Berkeley (email: {kuangc, nrc, hellerstein}@cs.berkeley.edu).

H. Chen is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (email: harr@csail.mit.edu)

H. Dolan and T. S. Parikh are with the School of Information, University of California at Berkeley (email: {dolan, parikh}@ischool.berkeley.edu)

Manuscript accepted Feb 10, 2009.

REFERENCES

[1] J. M. Hellerstein, "Quantitative data cleaning for large databases," United Nations Economic Commission for Europe, 2008.

[2] R. M. Graves, F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau, *Survey Methodology*. Wiley-Interscience, 2004.

[3] S. Day, P. Fayers, and D. Harvey, "Double data entry: what value, what price?" *Controlled Clinical Trials*, 1998.

[4] J. V. D. Broeck, M. Mackay, N. Mpontshane, A. K. K. Luabeya, M. Chhagan, and M. L. Bennish, "Maintaining data integrity in a rural clinical trial," *Controlled Clinical Trials*, 2007.

[5] W. Willett, "Scented widgets: Improving navigation cues with embedded visualizations," *IEEE TVCG*, 2007.