Volume 3, Issue 7, July 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Using Dynamic Forms to Improve Data Quality

**Yarram.Ravi kumar**[*1], [*1]M.Tech Student , Avanthi Institute of Engineering & Technology , Makavarapalam, India

**M.Satyanarayana**[*2] [*2]Assistant.Professor, Avanthi Institute of Engineering & Technology , Makavarapalam, India

***Abstract— In modern databases data quality is crucial problem. Data- entry forms grant the primitive and supportable opportunity for identifying and modifying errors, besides, for upgrading data quality at entry time, there has been a provincial research into automatic procedures. We in this paper proposing USHER, which is an end-to-end system for designing a form, entry and data quality assurance. From the previous form submissions, USHER attains an apparent model over the form questions. At every step of the data-entry process, USHER applies this model to improve the data quality. It induces a form layout which obtains the frequent data values of a form immediately and deducts the complication of error prone questions before entry of the data values. It readjusts the form to the values those are entered dynamically by accommodating real-time interface feedback and asking queries with improbable responses and streamlines the queries by formulating them. It revisits the question responses that assume likely to have been entered wrongly by asking the question. We take the measures of these components of USHER by using two real-world datasets which demonstrates that USHER can upgrade data quality substantially at moderate cost when it is in contrast with the current approach.***

***Keywords – Data , USHERS, Dynamic Forms, feedbacks, Errors.***

## 1. Introduction

Organizations and individuals routinely make important decisions based on inaccurate data stored in supposedly authoritative databases. Data errors in some domains, such as medicine, may have particularly severe consequences. These errors can arise at a variety of points in the life cycle of data, from data entry, through storage, integration, and cleaning, all the way to analysis and decision making. These models form a principled foundation on which we develop information-theoretic algorithms for form design, dynamic form adaptation during entry, and answer verification: Since form layout and question selection is often ad hoc, USHER optimizes question ordering according to a probabilistic objective function that aims to maximize the information content of form answers as early as possible—we call this the greedy information gain principle.

1. Applied before entry, the model generates a static but entropy-optimal ordering, which focus on important questions first; during entry, it can be used to dynamically pick the next best question, based on answers so far appropriate in scenarios where question ordering can be flexible between instances.

2. Applying its probabilistic model during data entry, USHER can evaluate the conditional distribution of answers to a form question, and make it easier for likely answers to be entered—we call this the appropriate entry friction principle. For difficult-to answer questions, such as those with many extraneous choices, USHER can opportunistically reformulate them to be easier and more congruous with the available information. In this way, USHER effectively allows for a principled, controlled trade-off between data quality and form filling effort and time.

3. Finally, the stochastic model is consulted to predict which responses may be erroneous, so as to re ask those questions in order to verify their correctness we call this the contextualized error likelihood principle.

## 2. Background Work

*2.1 Data Cleaning*

In the database literature, data quality has typically been addressed under the rubric of data cleaning [1], [2]. Our work connects most directly to data cleaning via multivariate outlier detection; it is based in part on interface ideas first proposed by Hellerstein [8]. By the time such retrospective data cleaning is done, the physical source of the data is typically unavailable—thus, errors often become too difficult or time-consuming to be rectified. USHER addresses this issue by applying statistical data quality insights at the time of data entry. Thus, it can catch errors when they are made and when ground-truth values may still be available for verification.

*2.2 User Interfaces*

Past research on improving data entry is mostly focused on adapting the data-entry interface for user efficiency improvements. Several such projects have used learning techniques to automatically fill or predict a top-k set of likely values [5], [6], [7], [8], [9], [10], [11]. For example, Ali and Meek [5] predicted values for combo-boxes in web forms and measured improvements in the speed of entry, copod [11] generated type-ahead suggestions that were improved by

geographic information, and Hermens and Schlimmer [7] automatically filled leave of absence forms sing decision trees and measured predictive accuracy and time savings.

*2.3 Clinical Trials*

Data quality assurance is a prominent topic in the science of clinical trials, where the practice of double entry has been questioned and dissected, but nonetheless remains the gold standard [12], [13]. In particular, Kleinman takes a probabilistic approach toward choosing which forms to reenter based on the individual performance of data-entry staff [14]. This cross-form validation has the same goal as our approach of reducing the need for complete double entry, but does so at a much coarser level of granularity. It requires historical performance records for each data-entry worker, and does not offer dynamic reconfirmation of individual questions.

*2.4 Survey Design*

The survey design literature includes extensive work on form design techniques that can improve data quality [3], [14]. This literature advocates the use of manually specified constraints on response values. These constraints may be univariate (e.g., a maximum value for an age question) or multivariate (e.g., disallowing gender to be male and pregnant to be yes). Some constraints may also be "soft" and only serve as warnings regarding unlikely combinations (e.g., age being 60 and pregnant being yes). The manual specification of such constraints requires a domain expert, which can be prohibitive in many scenarios. By relying on prior data, USHER learns many of these same constraints without requiring their explicit specification. When these constraints are violated during entry,USHER can then flag the relevant questions, or target them for reasking.

## 3. System Design & Implementation

*3.1 Architecture*

In the Fig.1 : The USER Components and Data Flow are specified where the process flow is as ;Initially it model a form and its data then generates question ordering according to Greedy information gain , then instantiate the form in a data entry interface and immediately it provide dynamic recording , feedback and reconformation according to contextualized error likelihood.

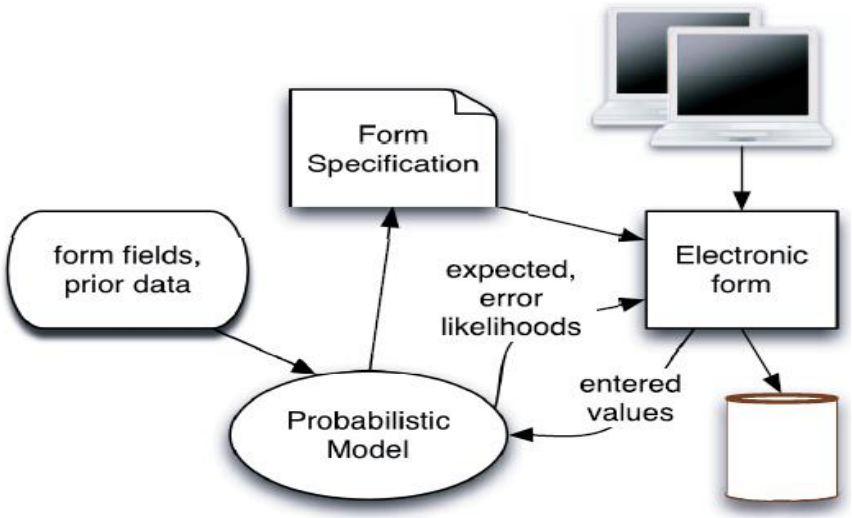



Fig.1 :USER Architecture

*3.2 Algorithm*

To address this spectrum of data quality challenges, we have developed USHER, an end-to-end system that can improve data quality and efficiency at the point of entry by learning probabilistic models from existing data, which stochastically(no chance to conform) relate the questions of a data-entry form. USHER addresses this issue by applying statistical data quality insights at the time of data entry.

**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**
  $F \leftarrow \text{argmax}_{F_i \notin \mathbf{O}} H(F_i \mid \mathbf{O})$;
  $\mathbf{O} \leftarrow (\mathbf{O}, F)$;
**end**

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

Thus, it can catch errors when they are made and when ground-truth values may still be available for verification.

## 4. Results

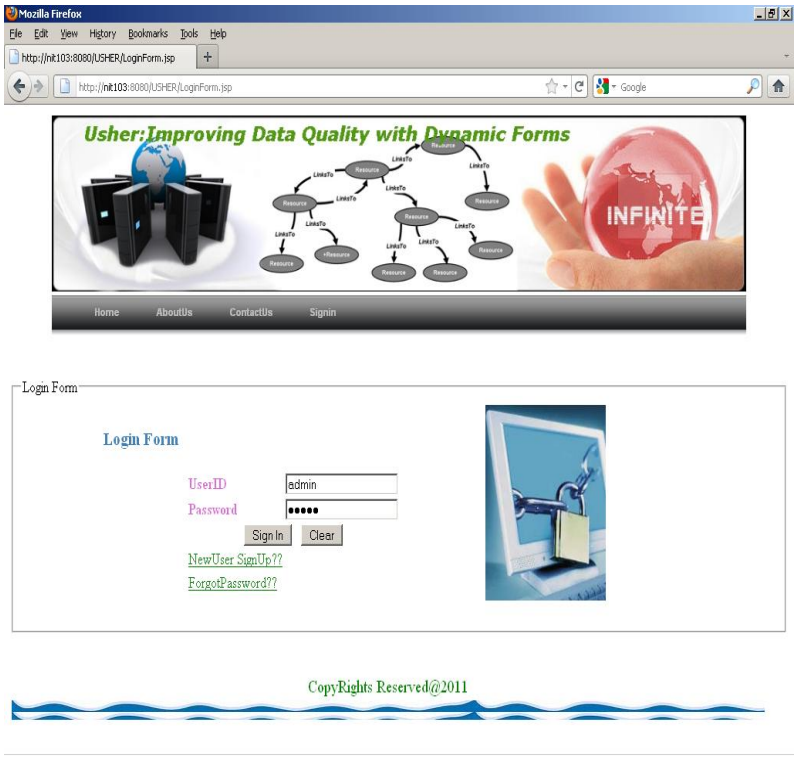
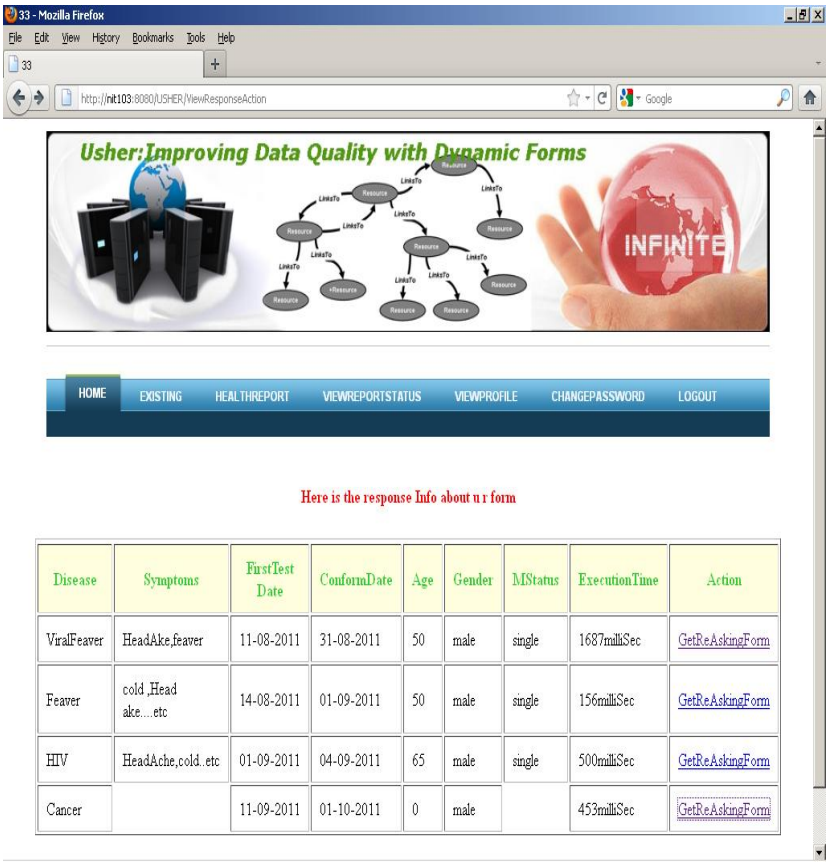



Fig.2: Admin Login Page




| Disease | Symptoms | FirstTest Date | ConfornDate | Age | Gender | MStatus | ExecutionTime | Action |
|---|---|---|---|---|---|---|---|---|
| ViralFeaver | HeadAke,feaver | 11-08-2011 | 31-08-2011 | 50 | male | single | 1687milliSec | GetReAskingForm |
| Feaver | cold ,Head ake....etc | 14-08-2011 | 01-09-2011 | 50 | male | single | 156milliSec | GetReAskingForm |
| HIV | HeadAche,cold..etc | 01-09-2011 | 04-09-2011 | 65 | male | single | 500milliSec | GetReAskingForm |
| Cancer | | 11-09-2011 | 01-10-2011 | 0 | male | | 453milliSec | GetReAskingForm |

Fig.3: Report Status of the User

## 5. Conclusion

This paper, has presented a probabilistic method which can be used to design intelligent data-entry forms that promote high data quality. In the data entry pipeline USHER influences data-driven awareness to automate multiple steps. We discover a prescribed form of fields which simulates very quick information capture, driven by greedy information gain principle, and reformulate questions statically to promote more accurate responses before entry and also we dynamically accept the form depending upon the entered values by facilitating re-asking, formulation and real time interface feedback

with the spirit of providing suitable entry friction. The data quality benefits of each of the components are simulated by the empirical evaluations and we automatically locate the erroneous inputs which are guided by contextualized error likelihoods

**References**

[1] T. Dasu and T. Johnson, Exploratory Data Mining and Data Cleaning. Wiley, 2003.
[2] C. Batini and M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques. Springer, 2006.
[3] R.M. Groves, F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau, Survey Methodology. Wiley-Interscience, 2004.
[8] J.M. Hellerstein, "Quantitative Data Cleaning for Large Databases," United Nations Economic Commission for Europe (UNECE), 2008.
[9] A. Ali and C. Meek, "Predictive Models of Form Filling," Technical Report MSR-TR-2009-1, Microsoft Research, Jan. 2009.
[10] L.A. Hermens and J.C. Schlimmer, "A Machine-Learning Apprentice for the Completion of Repetitive Forms," IEEE Expert: Intelligent Systems and Their Applications, vol. 9, no. 1, pp. 28-33, Feb. 1994.
[11] D. Lee and C. Tsatsoulis, "Intelligent Data Entry Assistant for XML Using Ensemble Learning," Proc. ACM 10th Int'l Conf.Intelligent User Interfaces (IUI), 2005.
[12] J.C. Schlimmer and P.C. Wells, "Quantitative Results Comparing Three Intelligent Interfaces for Information Capture," J. Artificial Intelligence Research, vol. 5, pp. 329-349, 1996.
[13] S.S.J.R. Warren, A. Davidovic, and P. Bolton, "Mediface: Anticipative Data Entry Interface for General Practitioners," Proc.Australasian Conf. Computer Human Interaction (OzCHI), 1998.
[14] J. Warren and P. Bolton, "Intelligent Split Menus for Data Entry: A Simulation Study in General Practice Medicine," Proc. Am. Medical Informatics Assoc. (AMIA) Ann. Symp., 1999.
[15] Y. Yu, J.A. Stamberger, A. Manoharan, and A. Paepcke, "Ecopod: A Mobile Tool for Community Based Biodiversity Collection Building," Proc. Sixth ACM/IEEE CS Joint Conf. Digital Libraries (JCDL), 2006.
[16] S. Day, P. Fayers, and D. Harvey, "Double Data Entry: What Value, What Price?" Controlled Clinical Trials, vol. 19, no. 1, pp. 15- 24, 1998.
[17] D.W. King and R. Lashley, "A Quantifiable Alternative to Double Data Entry," Controlled Clinical Trials, vol. 21, no. 2, pp. 94-102, 2000.