# Data Quality Enhancement with Novel Search Technique to Avoid Repetition of Records

**V. Govindasamy, V. Akila, K.S.Raajesh, Muralidhar Moka, B.Augustin Raj**

***Abstract—Data quality is the assessment of data's fitness to serve its purpose in a given context. Characteristics of data quality include: Accuracy, Completeness, Update status, Relevance, Reliability, Appropriate presentation, Accessibility. Data quality is the major problem experienced by many data entry operators. Our project reduces the possible errors more effectively by incorporating a novel search technique which will avoid repetition of data. During a survey, our system initially will create forms dynamically and the required questions can be entered. Then, the questions can be automatically re-ordered by setting necessary constraints to the questions. The default entry values can be entered for any question where the data needs to be constant. While entering data during the process of survey, the system will automatically re-ask the data-entry operators to enter the appropriate data. Then the search technique will search for the previous data and show whether the particular data is already in database or not.***

***Index Terms— Data quality, Novel search technique, Re-asking, Default entry.***

## I. INTRODUCTION

This system can be used to design data-entry forms which improve data quality and controls data-driven insights. Before entering data the form elements are unordered. After re-ordering, the forms elements will be reformulated in such a way that the accurate responses will be promoted by using greedy information gain principle. During entry, we can dynamically select the form needed which makes use of re-asking, default entry, and real-time interface feedback for providing appropriate data entry. After entry, we identify erroneous data by default entry technique and it re-ask those questions. Our system will show the benefits of data quality for the components like question ordering, default entry and re-asking.

## II. PROBLEM DEFINITION

The existing System contains many double entered data. This system reduces the possible errors more effectively by incorporating a novel search technique which will avoid repetition of data also reduces uncertainty using double data entry technique. The existing system has the major disadvantage of having duplicate data which occupies more space. This motivates to implement novel search technique. No default entry for effective adaptations of user interface. No re-asking before entry.

Cross-validation questions were not incorporated to provide accuracy. It will suggest the operator, the possible answers for the respective field.

## III. LITERATURE SURVEY

In the database literature, data quality has typically been addressed under the rubric of data cleaning [3]. Our work connects most directly to data cleaning via multivariate outlier detection [1]. 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. The system 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.

Past research on improving data entry is mostly focused on adapting the data-entry interface for user efficiency improvements. Predicted values for combo-boxes in web forms and measured improvements in the speed of entry [2], [7], generated type-ahead suggestions that were improved by geographic information automatically filled leave of absence forms using decision trees and measured predictive accuracy and time savings[4].

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 [9],[10]. In particular, Kleinman takes a probabilistic approach toward choosing which forms to reenter based on the individual performance of data-entry staff [8]. The survey design literature includes extensive work on form design techniques that can improve data quality [5],[6].

## IV. IMPLEMENTATION

The proposed system for novel search technique has been implemented and tested using 10 datasets from different fields. The datasets are taken from University of California Irvine data repository (UCI) and Table.1 describes the 10 datasets that we have used. The selected datasets are varied in number of Instances and attributes. There occurs a problem that many of the data's were repeated twice and more. So to reduce the double data entry and to save space in memory, we introduced this new novel search technique for the purpose of enhancing data quality. Further the testing has been done in some areas (i.e., census, survey about social websites) with the existing one and made a better result. The dataset we described here is as follows:

Manuscript received on April, 2013.

**Mr. V. Govindasamy**, Information Technology, Pondicherry Engineering College, Puducherry, India.

**Mrs. V. Akila,** Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India.

**K. S. Raajesh**, Information Technology, Pondicherry Engineering College, Puducherry, India.

**MuralidharMoka**, Information Technology, Pondicherry Engineering College, Puducherry, India.

**B. Augustin Raj**, Information Technology, Pondicherry Engineering College, Puducherry, India.

| DATASET | INSTANCES |
|---|---|
| Dermatology | 366 |
| Hepatitis | 155 |
| Lung Cancer | 32 |
| Iris | 150 |
| Diabetes | 768 |
| Heart-Stalog | 270 |
| Waveform | 5000 |
| Audiology | 226 |
| Splice | 3190 |
| Sonar | 208 |

**Fig 1. Dataset for survey**

Our proposed system has a web application where the User Interface loads the dynamic forms which we need on that particular time and prompt to enter the form name with unique ID. Then by selecting the appropriate form name, it is possible to create questions whatever may be. The questions can be ordered according to the specified constraints, data type and prompt. The server will automatically take these values and saves it. After creating questions, our model will enter into the default entry where the constant values will be given (common for all forms). Then the static form values will be saved in database. Once after the completion of default entry, the re-ordering phase will occur where the form elements will be arranged using priority (constraint and prompt). During entry, the form elements ordered will be displayed and ask the questions in order to save the time by only asking the required questions. This avoids uncertainty in data during entry.

### A. DYNAMIC FORM CREATION:

Each time, when we want to enter data for a particular requirement, we need to create forms according to our requirements and instant form creation cannot be possible at times. Hence, dynamic form creation is to generate forms dynamically by assigning a form name according to our requirements. Once the name for the form is given, the form will be automatically generated and we are allowed to add form elements as per our requirements. We are also allowed to create numerous dynamic forms and once we realize that we no longer in need of a particular form, it can be deleted from the database permanently.

***Goals Of Bayesian Network:***

A Bayesian Network(BN) is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a Directed Acyclic Graph (DAG). For example, a Bayesian Network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

-To improve the quality of data,

-To avoid the duplication of the data,

-To avoid the need of entering the known data again

### B. CREATING FORM ELEMENTS:

Form elements can be added into the form that has been generated dynamically. This system uses a text box to give our inputs as form elements and once the input is given, the form element will be added to the form by means of labels. The inputs that are given in the text box will be converted into labels and be inserted into the form which has been already generated dynamically. These labels will be given as the questions to the users to enter only the appropriate data and once the data is entered, it will be stored in the database. The user may tend to give irrelevant data to the questions that has been asked in the form but that can be avoided by means of "Default Entry".

### C. ORDERING THE FORM ELEMENTS:

The major problem we often face while entering the data is to avoid the unnecessary data. Sometimes we may be required some data but it is not so important and when we are searching for the data, there is every possibility that this "not so important data" will consume our valuable time. Hence, reordering plays a major role to avoid this "not so important data" while we are searching for the data. In reordering, the data that has been entered can be reordered by using the constraints "necessary, needed and may be." Reordering will be done according to the priority of the constraints such as the constraints: "necessary" will take "first or highest priority", "needed" will take "second priority", and "may be" will take" third or lowest priority." By doing so we can search the data which is only necessary to us and we can also reduce the search time by avoiding the unnecessary data. We have used "Static Ordering Algorithm" for reordering the data.

Step 1: Create a New Norm F with a required name and Form ID.
Step 2: Input Form Question Q= {$q_1$, $q_2$, $q_3$….} by selecting appropriate form F.
Step 3: While giving question names, give question criteria, such as necessary or not, data type, prompt needed or not etc.
Step 4: Re-ordering the Form by selecting form name F, this is done using the criteria given in Step 3.
Step 5: Default value for by choosing a particular form F.
Step 6 : Insert Form data to database.
Step 7 : Check for default value for form, if not matching, re-ask the question 'q'.
Step 8 : Update data for the updated q for the form F.

### D. DEFAULT ENTRY:

The major problem we often face while entering the data is to avoid unnecessary data entry because while filling up a particular data form, the user can fill whatever he wish and the entered data may not be relevant to the requirements. Hence, Default entry is to avoid entering unnecessary data by the user. Default entry is nothing but assigning "a single restriction or multiple restrictions" to make the data entry to be particular. Default entry supports both "character as well as numeric data" and also number of default entries can be more than one (i.e., generic).

### E. RE-ASKING:

If the data is irrelevant to the particular question, the system will generate an error message indicating that the data cannot be accepted. Then the user needs to give the appropriate data to the respective question. Re-asking is being done with the help of default entry. If the user did not give relevant data, then the data cannot be inserted into the database.

### F. SEARCHING:

Once the above processes are completed, one of the answers in the form will be considered for searching which is to avoid the repetition of data in database. The search item will be inserted in text box and it looks for same thing matched with the items in database. If the particular item is found, then it will not add the details into the database otherwise it will add the item with corresponding details. The search term is not only for particular data, but also it compares with other details of that particular data. The algorithm we used for novel searching technique is as follows:

1. Get the ID.
2. Select the record from the database using the given ID.
3. Display the record if it is available in the database. Otherwise get the correct ID.

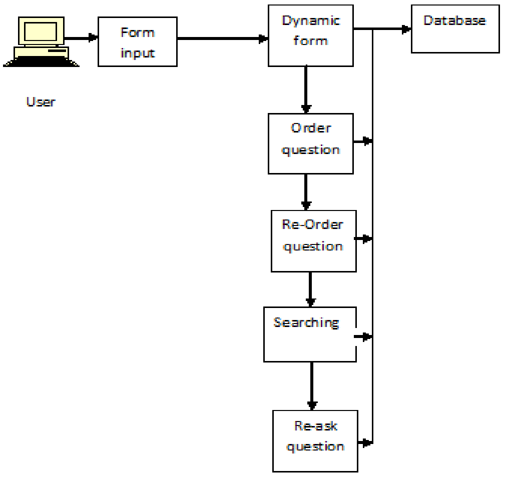

**Fig.2 System for data quality enhancement using novel search technique.**

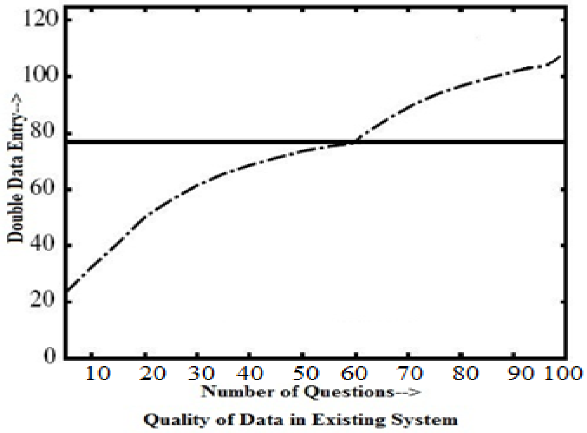

**Fig.3 Existing System**

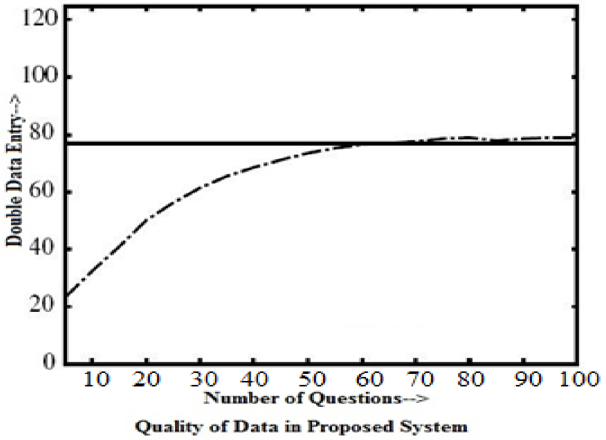

**Fig.4 Proposed System**

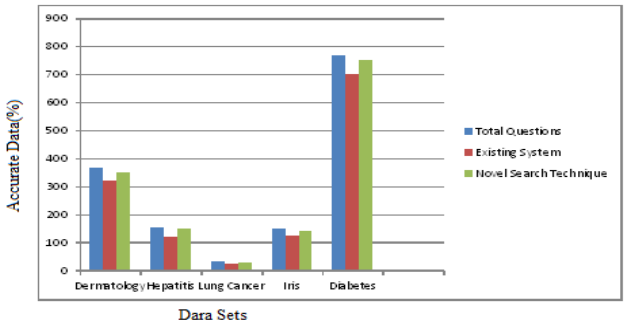

**Fig.5 Overall Comparison**

## V. CONCLUSION

A novel search technique and default entry technique has been introduced to improve the quality of data than the existing system does. Default entry has been introduced to avoid the repetition of data. Probabilistic model: A statistical model is a collection of probability distribution functions or probability density functions (collectively referred to as *distributions* for brevity). The Probabilistic model has more space complexity than our system. So here we reduce the time complexity also to some extent by simply giving the details to the questions which remains constant for that particular survey.

## FUTURE WORK

The work can be extended by integrating search technique with re-asking module instead of entering the data twice. The default entry can be avoided by introducing machine learning technique. The response time for the question re-asking can be reduced. The machine learning technique should be capable of rectifying and correcting the errors even while entering the data.

## REFERENCES

1. J.M. Hellerstein, “Quantitative Data Cleaning for Large Databases,” United Nations Economic Commission for Europe (UNECE), 2008.
2. A. Ali and C. Meek, “Predictive Models of Form Filling,” Technical Report MSR-TR-2009-1, Microsoft Research, Jan. 2009.
3. Dasu and T. Johnson, Exploratory Data Mining and Data Cleaning. Wiley, 2003.
4. 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.
5. R.M. Groves, F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau, Survey Methodology. Wiley-Interscience, 2004.
6. K.L. Norman,, “Online Survey Design Guide,” http://lap.umd. edu/survey_design, 2011.
7. 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.
8. K. Kleinman, “Adaptive Double Data Entry: A Probabilistic Tool for Choosing Which Forms to Reenter,” Controlled Clinical Trials, vol. 22, no. 1, pp. 2-12, 2001.
9. 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.
10. D.W. King and R. Lashley, “A Quantifiable Alternative to Double Data Entry,” Controlled Clinical Trials, vol. 21, no. 2, pp. 94-102, 2000.
11. Kuang Chen, Student Member, IEEE, Harr Chen, Neil Conway, Joseph M. Hellerstein, Member, IEEE Computer Society, and Tapan S. Parikh, ”Improving Data Quality with Dynamic Forms” IEEE transactions on knowledge and data engineering, vol. 23, no. 8, august 2011.