Data Quality Metrics for Data Entry Time Quality Evaluation

Kyung Mi Lee, Keon Myung Lee

***Abstract*—** Acquisition of quality data is crucial in data analysis, machine learning and data mining. Once poor quality data have been admitted into the database, it is costly to detect and remove. One of best practices is to prevent such poor data from being admitted to the system. This paper is concerned with a data entry framework and quality metrics which evaluate newly issued data on their quality. An automatic data entry framework to evaluate incoming data and to detect suspected attribute values enables the data entry personnel to reevaluate or check suspicious data at their entry stage. The proposed framework with integrity metrics has been applied to a cohort data collection system and it has shown the promising results.

***Keywords*—**data quality, data quality metrics, outlier detection, data integrity.

I. Introduction

TECHOLOGICAL advances have enabled almost everything happened to be measured and recorded into digital repository. Accumulated data are regarded as valuable assets in most sectors in business, government, military, science, and engineering. They expect to discover new solutions or opportunities by analyzing their data, and have invested much time and money to collect data and to establish infrastructure for their analysis. Analysis results from good quality data are trustworthy, but analysis for poor data may produce misleading results. For data to be a valuable asset, it is prerequisite to maintain their quality.

These days most of data acquisition has been done through automatic devices such as sensors, scanners. There are yet many data collected with the help of human personnel. Human involvement may incur human factor-related errors such as observation mistakes, mistyping, and missing fields. Once such errors are injected into data records, it is not easy to detect and remove them as their volume gets large.

This paper is concerned with how to reduce human factor-related errors in data acquisition. One of best ideas is to detect such errors earlier on at data entry point, and to assist data entry personnel to make sure what they have read and inserted by alerting suspicious data. This paper addresses a data entry framework that monitors the input, evaluates their quality in terms of integrity to existing ones, and notifies data entry personnel of the point of attentions. It also presents the integrity metrics which are used to evaluate input data quality.

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea(NRF) (grant no.: 2015R1D1A1A01061062)

Keon Myung Lee is with Dept. of Computer Science, Chungbuk National University, Cheongju, Chungbuk 28644, Korea (corresponding author to provide e-mail: kmlee@cbnu.ac.kr).

Kyung Mi Lee is a PhD course student at Dept. of Computer Science, Chungbuk National University, Cheongju, Chungbuk 28644, Korea.

II. Related Works

There have been many works to deal with data quality. Some managerial efforts have been exerted to develop and enforce the data acquisition guidelines for quality assurance. The environmental protection agency of the United States established a data quality assessment process that guides project managers and planners to check whether the required data quality has been attained. [1] Various approaches have been proposed to measure the data quality. Pipino et al. [2] suggested 16 subjective and objective data quality dimensions including appropriate amount of data, free-of-error, relevancy, and so on. Heinrinch et al. [3] addressed six desiderata for data quality measures such as normalization, interval scale, interpretability, and so on.

Outliers might be data containing errors, and hence outlier detection techniques can be used to help data entry personnel at data entry stage. Kriegel et al. [4] categorize outlier detection methodologies into statistical approaches, depth-based approaches, deviation-based approaches, distance-based approaches, density-based approaches, and high-dimensional approaches.

The statistical approaches model a statistical distribution for the given data set and decide data points with low probability as outliers. When a multivariate Gaussian is used to model the distribution, it assumes that the Mahalanobis distance of a data point to the distribution follows the a χ^2-distribution, and applies the χ^2-test [5] to determine potential outliers.

The depth-based approaches build nested convex hulls [6] for the data set and suggest the data points located at the border of the outmost convex hulls as potential outliers. It is hard to compute convex hulls in higher dimensional spaces, and hence it is usually not applicable to such data set.

The deviation-based approaches find outliers by searching for data points of which removal minimizes the variance of the data set.[7] The smoothing factor is computed for each subset of data set, which indicates how much the variance of the remaining data set is reduced when the designated subset is removed. The maximal subset with the largest smoothing factor is selected as potential outliers.

The distance-based approaches assume that normal data points have dense neighbors and outliers are far apart from their neighbors. The nested loop method first computes all pair-wise distances and chooses the data points which are far away from the others. [8] Distance-based approaches cannot handle successfully the data set with different density distributions.

The density-based approaches use the information of density around a point and the density around its local neighbors. They assume that the density around a normal data is similar to the density around its neighbors, and outliers have considerably different density compared to its neighbors. [9]

To reduce or prevent human factor-oriented mistakes in data entry, most application systems are equipped with some input validation module. Such validation module has some prespecified logic about what criteria the input should meet, such as range of legal values, relationship among some attributes, uniqueness of values, allowance of missing values, and so on. Well-designed validation modules can make great enhancement of data quality. Extraction of validation logic demands intensive work of domain experts, and it is rarely possible to extract enough validation logic and moreover it is practically impossible for human expert to have sufficient knowledge for changing environment.

III. The Proposed Data Entry Assisting System for Data Quality Enhancement

To help data entry personnel avoid human factored mistakes in data entry time, we propose a data entry system framework. The framework adaptively detects suspicious values in input data by evaluating its integrity with reference to the existing data set. We focus on two types of integrity and propose some metrics for them.

A. The Proposed Data Entry Framework

The proposed data entry framework consists of data entry module, data integrity check module, data integrity profile, and data repository.

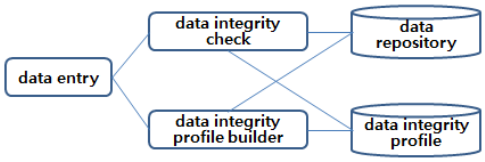

Fig. 1 The proposed data entry framework

The data entry module provides the data entry interface to receive the input from the user, i.e., data entry personnel, and hands it over to the data integrity check module to ask whether the input looks valid. If the data integrity check module replies it with some identified suspicious parts in the input, the data entry module displays the received information to the user. Upon the attention messages, the user comes to reexamine their input and consult the domain expert or personnel in charge of the original data, if necessary. After reconfirming the validity of data, the user keeps the original values or corrects the erroneous parts.

The data integrity check module compares the new input with the existing data set to determine whether the new one is compatible with existing one. For efficient integrity evaluation, existing data sets for each table are compiled into some statistics by the data integrity profile builder and are maintained in the data integrity profile. For each data entry, the data integrity check model evaluates its integrity using the information from the data integrity profile, and if the data passes the integrity test or proper correction has made upon it, input data is processed to contribute to the corresponding data integrity profile by the data integrity profile builder.

B. Data Integrity Metrics

Although there are various criteria or dimensions for data quality, this study assumes that data quality can be claimed by their integrity. This assumption is made because the other quality dimensions can usually not be improved once a database schema is designed. The integrity is one of the main concerns once the data architecture like database schema and data acquisition method is fixed. Data integrity refers to maintain and assure the accuracy and consistency of data. To measure integrity we need some metric for accuracy and consistency.

In data entry application, the integrity needs to be checked on the fly and its evaluation results should be used simultaneously. It is hard to have true value for input on the fly, and hence it is not suitable to define accuracy measures in this setting. The consistency in database community refers to the property that data transaction changes the data only in allowed way. We use the integrity in data entry system to claim that the new input goes well with existing data set. As data have been collected, we are not sure which input is valid and trustworthy. In this study, we assume that input data which are not similar to existing data set are suspected, ask the user to reexamine the input before committing her/his input. This approach does always not detect erroneous data, but is very helpful to reduce the human factored errors. The false alerts do not much harm to the user although too frequent false alarm annoys the user.

When the integrity-based method is used that evaluates how input goes well with existing data set, it is adaptive to changing environment. If the integrity profile can be more adjusted to recent inputs, the profile also keeps track of the recent patterns inherent in the data set. It allows the data entry assisting system for quality enhancement to be applicable in dynamically changing environment.

The proposed system evaluates the horizontal integrity and vertical integrity for categorical and numerical attributes. The horizontal integrity is the notion to examine the association among attribute values appearing in the same data record. The vertical integrity is the notion to evaluate the historical tendency shown in data set which are collected sequentially over the time.

The integrity measures depend on the value types and the integrity dimensions. The value types under consideration are either categorical or numerical. The ordinal value type is considered as categorical. The numerical type indicates that the values come from a continuous domain. They integrity dimensions are either horizontal or vertical.

When horizontal integrity is evaluated, first we need to

determine the associated attribute sets in each table, where a table stores the data records which consist of attribute values. Associated attribute sets are identified by applying an association rule mining algorithm to attribute values of data records. Each data record is treated as a basket data in which an item is a pair of attribute name and its value. Any association rule mining algorithms can be applied including Apriori [13], and FPGrowth [13]. Those algorithms find the frequent itemsets, but we are interested in the associated attributes sets. Hence we cumulate the number of the occurrences of associated attribute values identified by the employed association rule mining algorithm. The attributes sets are regarded as associated attributes sets when their relative cumulated count is not less than the pre-specified threshold. Because the association rule mining algorithms cannot be directly applied to the numerical attributes, discretization is conducted to such attributes. The discretization can be done into equal-length intervals, equal-frequency intervals, or cluster-based intervals. The number of intervals and type of intervals are empirically determined using the existing data set at the development stage of the proposed data quality evaluation system. That needs some engineering work because that depends on the characteristics of the data.

For each associated numerical attribute set, a clustering algorithm like *k*-means algorithm [13] is applied to identify clusters C = {C_1, C_2, …, C_K}. The information about identified clusters are stored in the data integrity profile module for later integrity check. When a data record is given as input, the distances of the input to the clusters are computed. When spherical clusters are assumed, the distance δ_i of data d_i on attribute set AS_j to a cluster C_k is the distance to 2σ boundary of the cluster where σ is the standard deviation. Its outlierness $o_k(\delta_{i,} AS_j)$ to cluster C_k is computed using a sigmoid function as follows:

$$o_k(\delta_{i,} AS_j) = \exp(2\delta_i)/(1+ \exp(2\delta_i)) \quad (1)$$

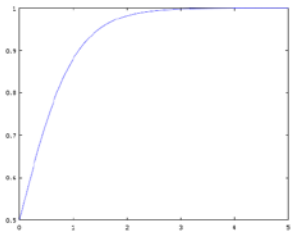

Fig. 2 A sigmoid function used in numerical attribute outlierness computation

The outlierness $o(d_i)$ of data d_i on attribute set AS_j is defined as the minimum value of the outlierness values for the clusters:

$$o(d_{i,} AS_j) = \min\{o_k(\delta_{I,} AS_j) \mid C_k \in C\} \quad (2)$$

When there are multiple associated sets AS = {AS_1, AS_2, …, AS_m}, the overall outlierness of data d_i on the whole attribute sets is the maximum of outlierness values for associated sets.

$$o(d_i) = \max\{o(d_{i,} AS_j) \mid AS_j \in AS\} \quad (3)$$

At the entry of new numerical data d_i, the outlierness is first computed; if its value is greater than the prespecified threshold, the attribute values associated with the highest outlierness is presented to the user for reexamination of data.

For categorical attributes, the local outlierness $oc(a_{i,} AS_j)$ of a value combination a_i of data d for an associated attribute set AS_j is defined as follows:

$$oc(a_{i,} AS_j) = -\log_2 p_i/(p_i N \log_2 N) \quad (4)$$

where p_i is the relative frequency of a_i in the data set and N is the number of data in the data set. The measure of Eq.(4) gives higher value as long as the relative frequency p_i gets lower. The measure gives a bounded value in the interval [0,1). The categorical outlierness $oc(d_i, AS)$ of data d_i over the associated attribute set AS is defined as the maximum of the local outlierness over AS.

$$oc(d_i, AS) = \max\{oc(a_{k,} AS_j) \mid a_k \in AS_k, AS_j \in AS\} \quad (5)$$

The outlierness index is bounded in the interval [0,1).

When a data contains both categorical and numerical attributes, the data set is stratified according to the categorical values. The numerical outlierness measures are computed against each stratum.

The vertical integrity is evaluated for a data set which maintains a sequence of data for the same entities over the time. For such data set, the changes in attribute values have some tendency which means that abrupt changes are unusual and might be an indication of outlierness. The vertical integrity is evaluated for an attribute over the historical values.

The vertical integrity of numerical attributes is evaluated by self-similarity, self-standard deviation change minimization, and relative standard change minimization. Self-similarity $o_{ss}(v)$ of new attribute value v to the values set of A is a measure for how similar new input is to existing ones, which is defined as follows:

$$o_{ss}(v) = \exp([v\text{-}m]/s)/(1+\exp([v\text{-}m]/s)) \quad (6)$$

Here, v denotes attribute value, m denotes the mean, and s denotes the time.

The self-standard deviation change minimization O_{SSD} measures the difference of standard deviation before and after the insertion of new data.

$$o_{ssd}(v) = \exp([std'\text{-}std]/s)/(1+\exp([std'\text{-}std]/s)) \quad (7)$$

The relative standard deviation change minimization measure $o_{RSD}(v)$ examines the difference between the standard deviation std_i' after insertion of new data v and the largest standard deviation in other numerical attributes, as follows:

$$o_{RSD}(v) =\exp([std_i'-m_i]/s-mst)/(1+\exp([std_i'-m_i]/s\text{-}mst)) \quad (8)$$

where $mst = \max\{(std_j - m_j)/m_j \mid j \neq i\}$.

The three measures reflect some aspects of outlierness in numerical historical data set. The overall outlierness $o(v)$ of numerical value v is defined as the minimum value of all outlierness values.

$$o(v) = \min\{o_{ss}(v), o_{ssd}(v), o_{RSD}(v)\} \quad (9)$$

The outlierness $oc(w)$ of categorical value w over its historical data is evaluated with occurrences of categorical values. The co-occurrence information of categorical values is a decision criterion in outlierness detection. Co-occurrences between pair of categorical values are evaluated for the cases of categorical values are paired regardless of their location in the sequence, and for the cases of adjacent pair elements are accepted. The relative frequencies for each pair of data elements are counted

for possible pairs of neighboring and sequential categorical values. The relative frequency $F_{NB}(a, b)$ for neighboring categorical value pair (a, b) is the number of occurrences that a and b happen as adjacent event for the same entity in a column of target table. $F_{SQ}(a, b)$ for pair (a, b) is the number of occurrences that a and b happen sequentially (but not necessarily adjacent) for the same entity in a column of target table. Table contains the data for different entities, hence such co-occurrence information can also be extracted from other entities. $F_{NB\text{-}ALL}(a, b)$ denotes the relative frequency of neighboring categorical value pair (a, b) over all entities. $F_{SQ\text{-}ALL}(a, b)$ denotes the relative frequency for pair (a, b) that happen sequentially over all entities. The outlierness $oc(w)$ is defined as the minimum of the above four relative frequencies as follows:

$$oc(w) = \min_c\{F_{NB}(a,w), F_{SQ}(c,w), F_{NB\text{-}ALL}(a,w), F_{SQ\text{-}ALL}(c,w)\} \quad (10)$$

Here a is the preceding categorical value to the new input value w in the same entities, and c indicates any preceding value happened over the attribute under consideration.

For each data entry, the relevant outlierness metrics are applied and the attention messages are displayed when some metric gives higher than its threshold. The threshold values depend on application domains which should be determined during engineering works to deploy the proposed data entry framework.

IV. Experiments

In medical and healthcare sectors, many data are yet collected and inserted into digital repository by data entry personnel due to the nature of their domain requiring interactions with patients and care-needing persons. Especially cohort studies require keeping track of a group of people over several years and collecting data by partly laboratory tests and partly personal interviews. The personal interviews with paper sheets having many questions are prone to contain mistakes and missing fields. The experiment system has been implemented for a cohort study data collection system in Korea. In the system, the people of study are required to fill the interview sheet by themselves with the guidelines according to a specific schedule. The data in the interview sheets are read and put into the database system by data entry personnel like nurses. The original system was equipped with some data validation module which checks the conformity of input data to the pre-defined validity rules. The validity rules are fixed and cannot detect the unknown abnormal patterns. The proposed data integrity metrics evaluate the data quality based on the own statistics of collected data set in the perspective of how well new data go well with existing data.

The experimental system has been tested as the front-end subsystem to the working cohort data collection system. The statistical information was first extracted for the integrity metrics such as the centroids and standard deviations of clusters, and relative frequencies of categorical values. Then the metric scores were computed as leave-one-out method, i.e., one data was evaluated with respect to all the other data. The thresholds for each attribute were set with reference to the maximum value of the evaluated metric scores. For the artificially generated abnormal data, the proposed metrics could successfully detect their abnormality. From the pilot studies to real data input, we observed that the front-end data integrity test module discovered interesting cases such as rare disease code for different gender like breast cancer for male.

V. Conclusions

It is very important to collect good quality data for extracting meaningful analysis results. The proposed data entry framework is to reduce human factored mistakes in data entry. The data integrity metrics have been developed for categorical and numerical attributes with respect to horizontal and vertical integrity. From the experiment study to a cohort data collection system, the proposed data framework and data integrity metrics have shown promising results. It is expected that the proposed method could be applied in the data entry service in many data collection systems.

References

[1] US Environmental Protection Agency, "Data quality assessment: statistical methods for practitioners," *US Environmental Protection Agency*, Washington, DC, EPA/240/B-06/003, 2006.

[2] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, pp. 211-218, 2002.

[3] B. Heinrich, M. Kaiser, and M. Klier, "How to measure data quality? A metric-based approach," in *Proceedings of the 28th International Conference of Information Systems*, Montreal, Canada, 2007, pp. 1-15.

[4] H. Kriegel, P. Kroger, and A. Zimek, "Outlier Detection Techniques," *Tutorial at the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2009.

[5] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[6] F. Preparata and M. Shamos, *Computational Geometry: an Introduction*. Springer Verlag, 1988.

[7] A. Arning, R. Agrawal, and P. Raghavan, "A linear method for deviation detection in large databases," *Proc. Int. Conf. on Knowledge Discovery and Data Mining*, Portland, OR, 1996.

[8] E. M. Knorr and R. T. Ng, "Finding intensional knowledge of distance-based outliers," *Proc. of the 25th Int. Conf. on Very Large Data Bases*, pp.211-222, 1999.

[9] M. M. Breunig, H. -P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," *Proc. ACM SIGMOD Int. Conf. on Management of Data*, Dallas, TX, 2000.

[10] V. J. Hodge, "A survey of outlier detection methodlogies," *Artificial. Intelligence Review*, Vol.22, No.2, pp. 85-126. 2004.

[11] D. Hawkins, *Identification of Outliers*, Chapman and Hall, 1980.

[12] A. D. Chapman, *Principles of Data Quality*, Report for the Global Biodiversity Information Facility, Copenhagen, 2005

[13] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, Elseviwer, 2012.

Kyung Mi Lee studied in Kyushu Institute of Technology, Japan. She is a PhD candidate at Dept. of Computer Science, Chungbuk National University. Her research interest includes fuzzy theory, optimization, and data analysis.

Keon Myung Lee received BS, MS, and PhD from KAIST, Korea. He is a professor at Dept. of Computer Science, Chungbuk National University. His research interest includes machine learning, data mining, big data, and artificial intelligence applications.