

# Assuring Data Quality by Placing the User in the Loop

Nadia El Bekri  
Fraunhofer IOSB  
Karlsruhe, Germany  
nadia.elbekri@iosb.fraunhofer.de

Elisabeth Peinsipp-Byma  
Fraunhofer IOSB  
Karlsruhe, Germany  
elisabeth.peinsipp-byma@iosb.fraunhofer.de

**Abstract**— Advanced analytical techniques such as data mining, text mining or predictive analytics are concepts that are increasingly important in the area of discovering large data sets. Various business areas recognize that data in all formats and sizes can provide significant support for decision-making. Large amounts of data can contain explicit knowledge in form of patterns. Errors within the data can falsify extracted patterns. Data is useful if it is correct, organized and interpreted correctly. Data mining algorithms can help improve data quality. Algorithms can suggest hints on possible errors. Possible errors need a mechanism that decides whether the error is true or false. The solution this paper introduces is to integrate users in the quality assurance process for decision support systems. The user can assess whether an error is true or false.

**Keywords**—data quality; user in the loop; data mining.

## I. INTRODUCTION

The main goal of the work is to ensure the data quality for a decision support system (DSS). Decision support systems can assist users in decision-making and planning [1]. Although decision support systems should only contain approved knowledge to support the decision-making, errors can occur by various reasons. For example, by using multiple data sources, duplicates easily occur if no procedure is reviewing new data entries. Another major error source can be the manual entry of new data by multiple users. For example typing errors can occur.

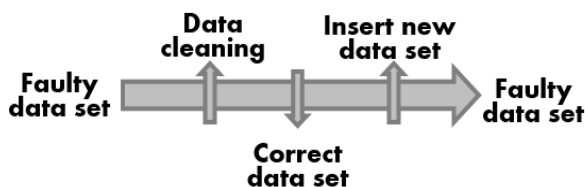


Figure 1. Traditional data-cleaning process.

Figure 1 illustrates the traditional data-cleaning process. After the data cleaning is done, a corrected data set is the result. The problem is that this correct data set is only valid until a new

data set, which contains errors, is added. There is no process to ensure that the data is constantly approved. Therefore, the approach that this paper suggests is the application of a continuous quality assurance process. The process catches errors before adding them to the data set by integrating a quality assurance process with the user in the loop.

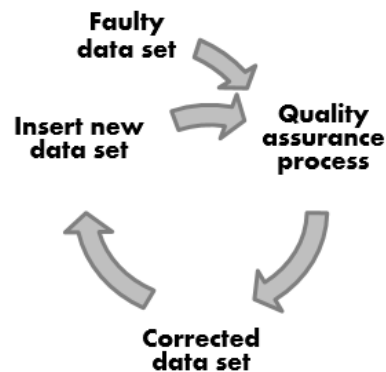


Figure 2. Adapted data cleaning process.

Figure 2 illustrates the adapted process. The quality assurance process ensures that data entries do not contain errors and thereby violate quality dimensions. The concept allows the users to correct the errors before adding them into the database. As the quality assurance procedures relies on an interactive approach for correcting the error, it is necessary to analyze data in a very short time, to ensure the interactivity. Heeter et al. [2] describe that the interactivity of a system is given when the user gets a response in less than one second. When selecting the data mining algorithms, it is important to ensure the analysis of the data is quick enough.

## II. RELATED WORK

There are different data quality dimensions. The three data quality dimensions considered for improvement are correctness, completeness and consistency. Batini et al. [3] [4] give a detailed description of each dimension. After the system identifies an error, the user gets involved. The user corrects the

error. Dallachiesa et al. [5], Yakout et al. [6], Luebbbers et al. [7] and Wu et al. [8] already introduced a concept, which involves the user in the loop. In this case, the user performs the correction of the data. The correction is done after the data has been entered into the database. The new approach is an adapted approach. The user, which wants to enter data into the database, is part of the process. This represents a major difference because existing data-cleaning methods described in Kandel et al. [9] and Luebbbers et al. [7] are used to detect errors in retrospect and thereby correct them afterwards. Holzinger et al. [10] introduce the concept of the *Doctor in the loop*. They describe a new paradigm in information driven medicine, where the doctor is part of the loop with an doctor system to support the decision-making.

The concept of quality assurance with the user in the loop contains the process of the indirect error identification that uses a classifier. A classifier belongs to the group of the supervised data mining algorithms. This means that classifiers assign with the help from already classified data, new and unknown objects. In general, there are two basic approaches: the instance-based and the model-based classification. The instance-based approach uses existing classified records to assign the class of a new object. An example of the instance-based approach is a classification using the k-nearest neighbor algorithm. By the use of a model-based approach a classifier is trained and a calculated model determines how new unknown objects need to be classified. Support vector machines, neural networks and decision trees are examples of algorithms in which a model is calculated.

In the following, the approach for the classification based on a decision tree is explained more detailed because it is used in the quality assurance process. C4.5 is an algorithm used to generate a decision tree. A decision tree is used as an underlying index structure for the classification. The assignment to a class occurs in the leaves of the tree. The nodes that are on the way up to the leaf contain decisions based on an attribute. In order to be able to generate a decision tree, the attributes are necessary to separate the classes. For this purpose, the entropy of an attribute is used. The entropy of an attribute always refers to the present training set and is zero if only one expression for this attribute exists. Thus, an exact prediction of the attribute is possible. For purposes of illustration, the information content is measured as a rate of disorder within the attribute. The more disorderly the attribute is, the higher its entropy is. More disorderly means that no prediction about the expression of the attribute can be made.

### III. CONCEPT OF THE QUALITY ASSURANCE WITH THE USER IN THE LOOP

Figure 3 illustrates the quality assurance process with the user in the loop. The first step in the process is to analyze the data set. Data mining algorithms can help analyze the complete data set, find patterns and derive interesting rules. El Bekri et al. [11] describe the use of association rules to extract patterns from the data set and use them to improve the data quality while entering data. After this step, the results are used to

build a model. In case of data entry, the model approves the data and may give error notes to the user. Then the user can modify the data record on his own and thereby can verify whether the suggested error note was correct.

After the data analysis, the results are available. The main goal while the data analysis is to identify certain patterns and the detection of relationships between the data. Patterns can appear frequently or rare in a database. It depends on the relevance for the specific field. There are various types of patterns: subgraphs, associations, trends, periodic patterns, sequential rules etc. Data Mining offers different techniques to find patterns within the data:

- Association analysis
- Classification
- Sequence analysis
- Clustering
- Forecasting

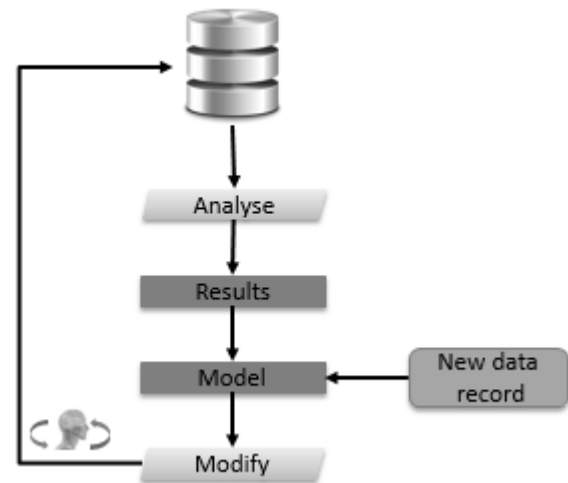


Figure 3. Quality assurance process.

Association analysis discovers patterns where one issue relates to another issue. Classification can result in a reorganization of the data. Clustering divides the data set into groups and thereby can find previously hidden information. The sequence analysis searches for patterns where one issue guides to another issue. Forecasting seeks for patterns that can guide to future predictions.

After the data analysis results are available, the user can evaluate them. The user can approve or disapprove the derived patterns and thereby evaluate the results. This reflects the concept of the user in the loop. In addition, a rating of how supportive the found results were can be done. Those results can return into the analyzing step, to improve the results and the extracted patterns over time. By the approval of discovered patterns, a quality model is built to apply them as rules. The rules are inserted into the system as a type of notification for

the user, who inserts new data entries. The notification informs the user that might a violation against some rules occurred. If the user accepts the hint on the possible error, the data set and thereby the error is corrected. The following section illustrates the idea more detailed using a data-mining algorithm.

#### A. Implicit error identification using a classifier

This section describes the development of a new approach for identification of errors based on a classifier and placement of the user in the loop. Classifiers belong to the group of supervised data mining algorithms; therefore, a training set is necessary. By using the training data set, classifiers need to be trained and optimized for their task. The generation of a representative training data set for error identification is a challenge. A training data set is created, by applying data mining algorithms that derive rules for error identification. Therefore, the cluster-rule based algorithm described by El Bekri et al. [11] is used. There, clustering and association analysis is combined to generate association rules that are used for the error identification and correction.

The algorithm is executed on the data set and classifies each data record as correct or incorrect. This results in a classified data set, which is used to train the classifier. However, one big challenge is still that the correctness of the training data is not guaranteed. The algorithm for error identification does not provide a perfect result. The classified data set can contain data records that are marked as errors but actually error-free records. The model of the classifier then transfers these errors to newly classified records. Furthermore, no new errors can be found by this procedure, but instead only errors, which are also found by the previous method. This is because the model was trained with the errors, which also were recognized by the previous procedure. The new model only contains the knowledge that the algorithm for the classification of the training set contained.

Another possibility is to create artificial errors. In this approach, a correct data set is manipulated by inserting errors. The challenge in this case, is the generation of errors which are similar to the errors occurring in the real application. Luebbers et al. [7] describe a classifier used in a system to improve data quality. There the C4.5 classifier was used to identify errors. The training data set is generated from artificial errors. For the generation of artificial errors the user needs to define special rules in the system. Based on these rules, the system then derives artificial errors. The generation of artificial errors requires enormous domain knowledge and a definition of the rules by the user. In summary, it can be concluded that previous approaches such as Luebbers et al. [7], Yakout et al. [6] and Kandel et al. [9] attempt to correct the error by training a model directly. Directly means in this context that a model is trained with help of a data record containing already classified records. The model, generated by this approach, makes a decision whether the single data record is correct or incorrect because of the attributes.

In our approach, the classifier is not trained directly to identify the error, but to predict individual attributes of the data

records. Figure 4 illustrates the approach of the indirect error identification.

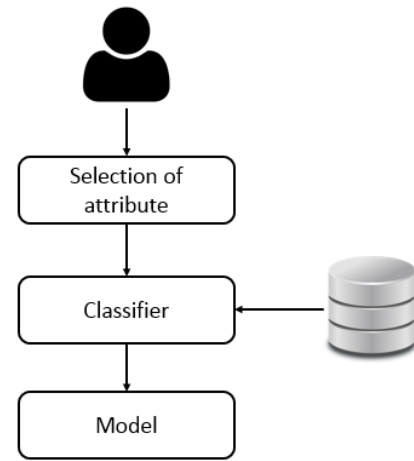


Figure 4. Process of indirect error identification.

A possible error is indirectly identified by making a prediction of the attribute "X" and the actual occurrence of the feature "Y". Based on such a deviation a possible error can be predicted. This approach is independent of the specific classifier that is used. As classifier, any classification algorithm can be used that suits best for the application. The advantage of this approach is that the previous training data set can be used. No new artificial data set needs to be generated by the user or an already classified data set needs to exist. Thereby, the question of how representative the artificial data set really is does not arise. Furthermore, the task of a user is limited to providing meaningful information on relevant attributes on which the classifier needs to be trained. With this approach the indirect error identification is simplified. Indirect error identification can be achieved by using several classifiers that are trained on different attributes. If a classifier predicts an error, this record will be marked as incorrect. The following section illustrates the process of indirect error identification with a specific example.

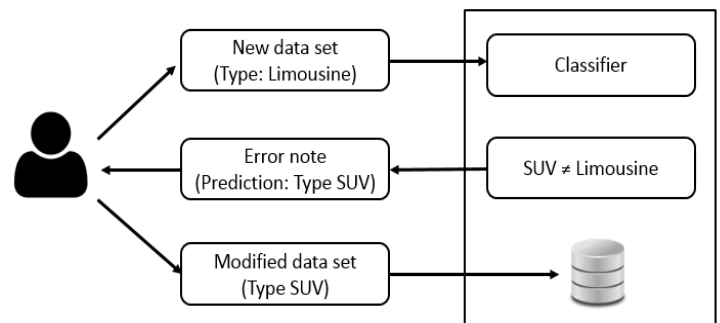


Figure 5. Example of the indirect error identification use.

Figure 5 illustrates an example of the interaction using the indirect error identification. The figure illustrates a user who

has added for example a new data record for a vehicle. As type of the vehicle, the user erroneously indicated a "limousine" instead of a "SUV". The trained model will make the prediction "SUV" for the attribute type. Therefore, an error notification automatically will be shown to the user. This gives the user the possibility to correct the error directly. As a result, the incorrect data record is not added to the data set.

#### IV. FUTURE WORK

The proposed concept describes the idea by placing a user in the loop of the quality assurance process. The concept supports user doing the manual entry. Most of the time users contain the domain-specific knowledge of the application. Although this is the fact, the process of the data entry can lead to many errors within the data set caused by the user, e.g., through typing errors. Therefore, the concept of the indirect error classification builds a model through previous derived rules out of the data set. The benefit of this approach is that no classified data set needs to be generated by users in order to train the model. The future work will be to improve the whole quality assurance algorithm through other error identification algorithms and to improve the results. Furthermore, a concept on how exactly the evaluation of the user should be integrated and weighted in the analysis step will be developed.

#### ACKNOWLEDGEMENT

The underlying project to this article is funded by the WTD 81 of the German Federal Ministry of Defense. The authors are responsible for the content of this article.

#### REFERENCES

- [1] Heeter, C.: "Interactivity in the context of designed experiences", In: *Journal of Interactive Advertising* 1, No. 1, p. 3–14, (2000).
- [2] Power D. J.: "Decision Support Systems: Concepts and Resources for managers", Greenwood Publishing Group (2002).
- [3] Batini, Carlo: "Data and Information Quality: Dimensions, Principles and Techniques", Springer International Publishing, (2016).
- [4] Batini, Carlo; Cappiello, Cinzia; Francalanci, Chiara and Maurino, Andrea: "Methodologies for data quality assessment and improvement", *ACM computing surveys (CSUR)*, Vol. 41(3): p. 16, (2009).
- [5] Dallachiesa, M; Ebaid, A.; Eldawy, A; Elmagarmid, A.; Ilyas, F.; Ouzzani, M.; Tang, N.: "NADEEF: a commodity data cleaning system", *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data ACM*, p. 541–552, (2013).
- [6] Yakout, M; Elmagarmid, A.; Neville, J.; Ouzzani, M.; Ilyas, Ihab F.: "Guided data repair", *Proceedings of the VLDB Endowment* 4, Nr. 5, p. 279–289, (2011).
- [7] Luebbers, D., Grimmer U., Jarke M., "Systematic development of data mining based data quality tools", *Proceedings of the 29th international conference on very large databases*. Vol. (29), p. 548-559, (2003).
- [8] Wu, L., Kaiser, G., Rudin, C., Anderson, R., "Data quality assurance and performance measurement of data mining for preventive maintenance of power grid", *Proceedings of the first International Workshop on Data Mining for Service and Maintenance ACM*, p. 28-32, (2011).
- [9] Kandel, Sean; Parikh, Ravi; Paepcke, Andreas; Hellerstein, Joseph M., Heer, J.: "Profiler: Integrated statistical analysis and visualization for data quality Assessment", *Proceedings of the International Working Conference on Advanced Visual Interfaces ACM*, p. 547–554, (2012).
- [10] Girardi D, Küng J, Kleiser R, et al.: *Interactive knowledge discovery with the doctor-in-the-loop: a practical example of cerebral aneurysms research*. *Brain Informatics*, (2016).
- [11] El Bekri N., Peinsipp-Byma E.: "Cluster Rule Based Algorithm for Detecting Incorrect Data Records", *UKSim-AMSS 18th International Conference on Mathematical Modelling and Computer Simulation*, (2016).