Data Quality Assistance – The Use of Data Mining Algorithms to Enhance Data Quality

Nadia El Bekri and Elisabeth Peinsipp-Byma
Fraunhofer IOSB, Karlsruhe, Germany.
nadia.elbekri@iosb.fraunhofer.de

***Abstract*—Large and over the years grown databases are a persistent concern in the field of data quality. Data sets grow over time from multiple sources and various users. Data Quality is one of the key issues that needs to be considered. This paper introduces a further development of an interactive data mining assistance system for ensuring data with high quality. What exactly is data quality? Data Quality in our approach is that the data that need to fulfill special requirements. Therefore, in a first instance, data mining algorithms are used to find outliers and duplicates. In the next step, the data mining assistance system generates rules that describe the whole data set. Furthermore, a rule administration is part of the concept. Interesting rules that have been found within the data set through the application of various data mining techniques are supposed to be added at this stage. The system serves, therefore to store and review rules that can be applied to the decision support system. For generating rules, various algorithms from the field of data mining are used. These rules have to be evaluated by experts to see if they can be applied as a type of suggestion rule to the decision support system.**

***Index Terms*—KDD; Data Mining; Duplicate; Outlier; Association Rule.**

I. INTRODUCTION

Decision support systems can assist humans during various types of tasks. There are different types of decision support systems and one of them that was developed for the task of aerial reconnaissance is the recognition assistance system RecceMan® (Reconnaissance Manual). Aerial image analysts on missions use RecceMan® to classify, and thereby recognize different types of object by their certain characteristics. The characteristics are extracted visually by the aerial image analyst from aerial or satellite images. The recognition assistance offers them the possibility to interactively apply the visible characteristics into the system. As soon as they apply the characteristics into the system, the solution set of the objects is adjusted. The list arranges and shows up the objects whose characteristic matches the most at the top. A detailed explanation of the system can be found in [1]. The problem we address in this paper is the data quality of the data within the recognition assistance system. Different aerial image analysts add the data manually to the system. Obviously, this is an error-prone process. After adding new data to the data set, they are sent to a central unit that fuses the data to the core data set. This central unit actually performs the data quality assurances manually. Apparently, this is a time consuming and error-prone process. In order to support the central unit with semi-automated mechanisms, the next step is to design an interactive data mining assistance system that supports the data quality of the entire data set. The interactive data mining system is not limited to the topic of aerial image reconnaissance. As long as there exists an object-feature relation between the data, it can be transferred to any reference data set. A major factor that needs to be considered while designing an interactive data mining assistance is the human factor. The interactive data mining assistance should assist the aerial image analyst while reviewing the data set in an appropriate way. Therefore, we applied different algorithms from the field of data mining to create a simple and easy interactive data mining assistance.

Today, numerous applications are based on a reference set with an underlying database. Errors consequently can happen when a user makes an erroneous input. This can lead to very simple mistakes caused by a typographical error or user error caused by an incorrect use of the software. The increase of erroneous records in the database has a decrease in the data quality of the recognition assistance. Since a subsequent correction of errors is always associated with an increased cost, the objective is to design a quality assurance process for this application. This process is supposed to support the user while adding new data to the reference set.

A quality assurance process is a continuous process that seeks to ensure that the quality of the data set does not decrease at any time. To ensure the quality of the data set, the knowledge needs to be extracted from the existing records and is then provided to the users. The aim is to guide users through the extracted knowledge and to call his attention on possible errors in a newly created database object. The user can then decide with the help of his expert knowledge, whether the hint is right or wrong. This process allows a quality assurance of the data set and represents an improvement in the usability of the application. To determine whether a quality assurance process provides benefit or not, it must be possible to evaluate the data quality. The quality of the data set needs to be evaluated before and after the introduction of the quality assurance process.

This paper introduces the state of the art data quality mechanisms, the term of data quality and the algorithms that are applied within the interactive data quality assistance. A first description of the interactive data quality assistance can be found in [2].

II. STATE OF THE ART – DATA QUALITY MINING

Hipp et al. [3] define the term “data quality mining”. They integrate the data mining procedures in the step of pre-processing in order to increase the quality of the data set. The quality of the data set has been enhanced by replacing missing or incorrect data with the application of association rules. The association of rules is extracted from the existing and underlying data set. The application of the association of rules enables that the missing and incorrect data are checked and

replaced by correct data. Another procedure introduced in Luebbers et al. [4] is a data auditing process. This data auditing process improves the existing data cleaning approaches introduced in Batini et al. [5] under usage of the C4.5 algorithm. The C4.5 algorithm [6] uses a training data set to generate a decision tree. The correction of the incorrect entries is under the responsibility of the user and so far not automated.

The presented procedures do not ensure a long-term assurance of data quality. Why? As soon as a user adds new data into the system the quality assurance process needs to be performed again on the data set in order to assure the data quality again. The repeated application of the data quality assurance process is associated with the same expense as performed on the first time.

Another procedure, used for the evaluation of smart grids, is based up on different machine learning techniques [7]. Smart Grids describe new electricity networks that monitor themselves and ensure an optimal distribution. The principal focus is to predict failures of electronic devices within the electricity network. The quality of the data is assured by offering the user support through visually presented relevant attributes. The evaluation system supports the user with the maintenance of the smart grids in which potential sources of errors are predicted. The error probability is approximated by various machine-learning procedures.

III. Data Quality

What exactly is data quality? Nowadays it is one of the most important things to consider while you are working with data because the data set increases in a short time. Data quality includes the correctness, the relevance and the reliability of the data. The data are added manually to the data set by different users. Obviously, this is a major error source during the task of adding data to the data set. Potential sources of errors that can occur during adding the new object to the data set are:

- Spelling mistakes
- Not adding potential information
- Adding wrong information
- Different representations (e.g. in spelling)

All different types of errors can have far-reaching effects on the reliability of the recognition system. Spelling mistakes or different representations in spelling can lead to duplicates in the data set. A duplicate is present when two entries point on one real-word-object. This error can have a major influence on the recognition task for aerial image analysts. Information can be either stored in the first object or in the second object. This is a huge information loss for aerial image analysts on missions, if some important characteristics of the object that they need to recognize are missing. Another error that can happen is when adding an object to the data set is an outlier error. An outlier is an observation that lies an abnormal distance from other values. In order to detect the listed type of errors algorithms from the field of data mining need to be applied. Data mining belongs to the Knowledge Discovery in Databases (KDD) framework. KDD is the process of extracting knowledge from databases. The knowledge is valid in a statistical sense, so far unknown and useful for a specific application. Data mining offers various algorithms that can be applied on the data set to detect relevant information. Thus, the aerial image analyst should be part of the interactive knowledge discovery process. The goal is to support the human intelligence combined with computational intelligence. Why is this combination so important? In order to be able to explore the data set and make assumptions on the data set, the human factor cannot be isolated from the computational. Therefore, it needs to be combined. “Quality is the degree to which a set of inherent characteristics of an object fulfils requirements” [8]. The quality is a measure which is not easily measured and is dependent on the user. The so-called Data Quality Assessment procedure deals with the evaluation of the quality of a dataset. In the work by Batini et al. [5] they summarize some Data Quality Assessment methods. The quality of a dataset is defined by various dimensions. The four basic dimensions in Batini et al. [5] are accuracy, completeness, consistency and timeliness. Although most data quality assessment methods integrate all these dimensions, the exact definitions of these dimensions differ and are based on an intuitive understanding [9]. For example, Kriebel [10] defines the accuracy as “the correctness of the output information”. Ballou and Panzer [11] interpret the term accuracy as “the recorded value is in conformity with the actual value”. Wand and Wang [9] mention that the term appears to be equivalent to the term of correctness. They point out that the “notion of data or information quality depends on the actual use of the data” [9]. The second term that describes the data quality is the completeness. In literature, this means that all needed values are included within the data set. Consistency relates to several aspects of the data. In literature, they relate to the values of the data and to the representation of the data, either internal or physical. Timelineness is described as whether the data is out of date and the availability of output on time [10] [11].

A. *Process of adding data and the structure of the data set*

In order to add new objects into the data set, RecceMan® offers different types of editors. The object editor serves therefore to add new objects. The feature editor adds various types of features into the system. The data sets of both are then fused by a data management component that creates a master database for the recognition component. The data set is structured as an object-feature model. As underlying record set for the analysis, we used two different data sets. The first idea of designing a data quality procedure is the operational data set of the recognition assistance system. Because this operational data set is confidential, the results cannot be published. Another very similar underlying data set was used as reference set. This reference set is built very similar to the structure of the operational data set of the recognition assistance system. The data set contains the data of the World Development Indicators. The primary World Bank collection of development indicators is compiled from officially recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates [12].

IV. Interactive Quality and Knowledge Mining

The KDD process extracts knowledge from the existing dataset and serves as an abstract framework. This section describes how the adoption of the KDD process results in a quality assurance process. The quality assurance process serves as abstract framework too. The results of the KDD process are used to detect erroneous records. Therefore, the

complete KDD process needs to be implemented in the first step. The KDD process consists of five steps: the selection, the pre-processing, the transformation, data mining and the interpretation.

The three steps, which are the selection, pre-processing and transformation must be designed and implemented separately for a specific data set. These steps are very dependent on the data structure. For these steps, it is important to be aware of the specific data structure in order to edit the database. Every data set in the database needs to be prepared before applying the data mining algorithms on it. The following subsections introduce the data mining algorithms that were implemented in the interactive data mining assistance.

A. Association Rule Mining

Association Rule Mining describes the approach of finding rules, which represent an if-then linkage between two sets. The sets contain concrete characteristic values from the database. For example, an association rule can be depicted as follows:

{diaper, baby cream } → {baby food}

The if-then linkage is described by the → arrow symbol. This example implies that if a customer bought diapers and baby cream, he also bought baby food. For every association rule, certain performance indicators are displayed to quantify the quality of the rule. This performance indicators are the Support supp (X) and the Confidence conf (X). The Support supp (X) indicates the frequency of the occurrence of this association rule in the whole data set. The Confidence conf (X) specifies the probability in which an association rule matches. That means that the confidence is the probability of that when the set x was found, the set y has been found also. In the example, this means how often diapers, baby cream and baby food were bought together divided by the amount of how often only diapers and baby cream were bought together. The Confidence is represented as conf (X→Y) = supp (X u Y)/ supp(X). The quality assurance process uses the Apriori algorithm with different parameters that can improve the results. The user can specify the minimum support and the minimum confidence. The minimum support indicates which amount of objects with a specific attribute range that at least need to be linked in order to be considered as association rule. The minimum confidence indicates which confidence that at least needs to be accomplished in order to be displayed as association rule.

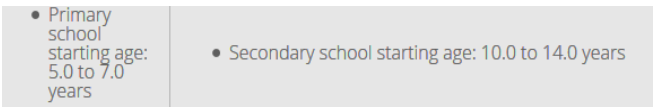

Figure 1: Association rule

A very strong association rule that was found through the interactive assistance system within the WDI data set is displayed in figure 1. Pupils that start the primary school at the age of 5 to 7 years are starting secondary school between the ages of 10 to 14 years.

B. Duplicate Mining

Duplicate Mining describes the process to crawl the given data set after duplicates and very similar records are identified. A duplicate refers to a multiple stored objects that point on the same real world object. In literature, it is called record linkage or record matching. As mentioned in the introduction, duplicates can happen very easily through misspelling. There are various types of duplicates:

- True positives
- False positives
- False negatives
- True negatives

In order to find duplicates in the data set, the first step is to reduce the search space within the data. Without partitioning the data set into certain groups of records, the comparisons within the whole data set would take an enormous computing effort. This is because every data object needs to be compared to every single other data object. For example, if we have 5000 objects that would make 12497500 comparisons. This method is not an appropriate approach to be considered as the optimal solution while searching for duplicates in the data set. In literature, there are various methods to reduce the search space:

- Sorted Neighborhood
- Blocking

The sorted neighborhood method is a part of the windowing methods. The method is divided into three phases. In the first phase, a sorting key is allocated to every record of the data set. After this step, all of the records are sorted according to the sorting key. The last phase includes sliding a window of a fixed size above this sorted list of records. Blocking methods divide the record set into disjoint partitions. The method only compares all pairs within the block [13]. We decided to choose the Sorted Neighborhood method for our purposes. After choosing a search method, we have to decide which comparison method we want to choose. There are different comparison methods that can be chosen for this task:

- Hamming distance
- Edit distance

The hamming distance is used to denote the difference between two strings. More specifically, the distance is the number of the bit positions that differs in-between the strings. The Edit distance measures the dissimilarity of two strings by counting the number of operations needed in order to transform one string into the other string. In the literature, there are various definitions of the edit distance, and the most common is the Levenshtein distance. The operations allowed within this method are insert, delete or substitute. We applied the search algorithm Sorted Neighborhood in combination with the Levenshtein distance for the comparisons of the strings.

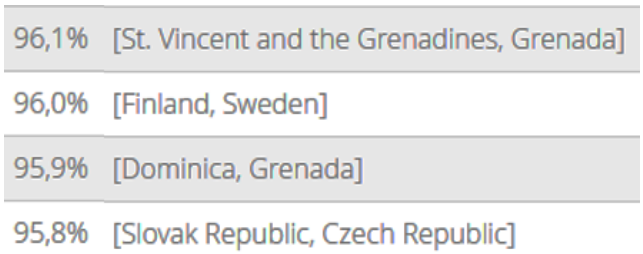

Figure 2: Similar and duplicate entries

Figure 2 presents the detection of very similar data entries. In this case, the characteristics of the countries Finland and Sweden have 96 per cent. After applying these methods on the record set to find possible duplicates, a decision model is needed. Not every duplicate found via the application of the methods automatically needs to be a real duplicate. At the beginning of this section, we mentioned the different types of

duplicates. In the literature, there are several models for the decision model possible [4]:

- Probabilistic model
- Clustering Model
- Hybrid model

Currently, we are not applying any decision model. In the first stage, an expert evaluates the found duplicates whether they are real world duplicates or not. After the evaluation by the expert, we implement a decision model to compare the results of the expert with the results of the decision model.

C. Outlier Mining

A very popular and often used process in the field of data mining is the outlier detection. There are various algorithms to detect outliers. The quality assurance process implements the following outlier algorithms:

- Boxplot
- Local Outlier Factor
- Local Outlier probability
- K-nearest neighbour

A boxplot is a graphical display to characterize the behavior of the data at the beginning, in the middle as well as at the end of the distribution. The box plot uses the median and the lower and upper quartiles. These are described as the 25th and 75th percentiles. If the lower quartile is Q1 and the upper quartile is Q3, then the difference between them is called the interquartile range. Boxplots can easily visualize outliers. Points beyond the inner fences and the outer fence are regarded as outliers.

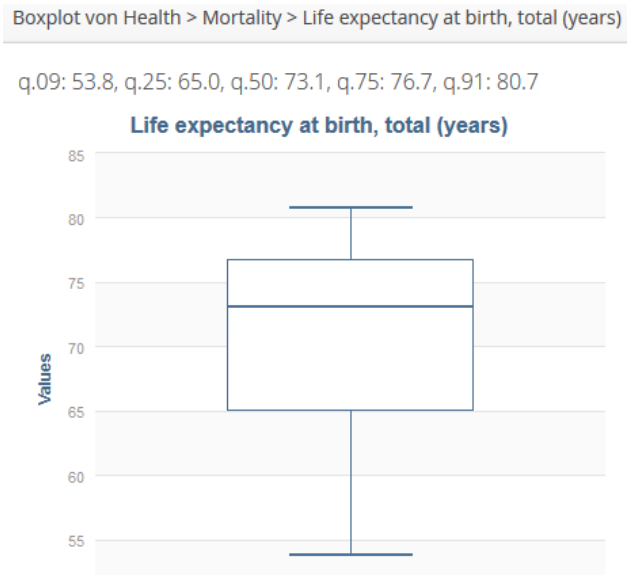

Figure 3: Boxplot diagram from data assistance system

Figure 3 visualizes a boxplot from the interactive data assistance system. In this case, we see the boxplot from the parameter "life expectancy at birth". This parameter describes the life expectancy in years for every country. Figure 1 also displays the minimum, the lower quartile, the median, the upper quartile and the maximum values. Values that not within the minimum or the maximum are considered as outliers.

Element	Life expectancy at birth, total (years)
Sierra Leone	44,839
Botswana	46,44

Figure 4: Outlier countries

Figure 4 displays countries that are considered as outliers within the data set. In this case, the countries, namely Sierra Leone and Botswana, have a very low life expectancy in years compared with the other countries.

The local outlier factor (LOF) is another algorithm for outlier detection. This algorithm works based on density. The density of a data point is compared with the density of its comparison. The local outlier probability is an extension of LOF, which additionally normalizes (normal distribution) the output of LOF for the density estimation. It approximates the probability that a particular data point is an outlier. The KNN is based upon the graph of k nearest neighbors. Every vector in the data set forms one node and every node has pointers to its k nearest neighbors. As examples, Figure 3 and Figure 4 show the results of applying two of the presented algorithms (LOF and k-nearest neighbor) on the data set of the WDI.

The list of Figure 5 and Figure 6 start to differ at the data point of Angola. For the algorithm of the local outlier probability, the different next outlier starts at Nigeria with a life expectancy of 51,289. In the case of the algorithm of the k nearest neighbor, the next outlier is San Marino with the highest life expectancy of 83,129 years.

Element	Life expectancy at birth, total (years)	LOF
Sierra Leone	44,839	2,405
Botswana	46,44	2,284
Lesotho	47,483	2,208
Central African Republic	48,099	2,165
Swaziland	48,346	2,147
Mozambique	49,137	2,093
Côte d'Ivoire	49,675	2,057
Chad	49,77	2,05
Angola	50,654	1,992
Nigeria	51,289	1,953
Equatorial Guinea	51,533	1,938
Burundi	52,624	1,873
South Sudan	53,465	1,824
Malawi	53,466	1,824
Guinea-Bissau	53,558	1,818

Figure 5: Outlier output from LOF algorithm

Element	Life expectancy at birth, total (years)	k-Nearest Neighbour
Sierra Leone	44,839	0,018
Botswana	46,44	0,016
Lesotho	47,483	0,015
Central African Republic	48,099	0,014
Swaziland	48,346	0,014
Mozambique	49,137	0,014
Côte d'Ivoire	49,675	0,013
Chad	49,77	0,013
Angola	50,654	0,012
San Marino	83,159	0,012
Hong Kong SAR, China	82,978	0,012
Japan	82,843	0,012
Italy	82,337	0,011
Iceland	81,898	0,011
Equatorial Guinea	51,533	0,011

Figure 6: Outlier output from k-nearest neighbour algorithm

V. CONCLUSION

The presented algorithms of the field of data mining are efficient methods to analyze the data set. The interactive data quality assistance system is the first step in a whole data quality assurance process. The system detects duplicates, outliers and delivers association rules that are potentially valid for the whole data set. The next step will be to apply more data mining algorithms on the data set to be able to derive more rules that describe the data set. In order to be able to perform pilot studies with experts and to improve the found rules, we need also to compare different algorithms and the results against each other.

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] El Bekri N., Angele S., Ruckhäberle M., Peinsipp-Byma E., Haelke B. 2015. RecceMan: An Interactive Recognition Assistance For Image-Based Reconnaissance: Synergistic Effects Of Human Perception And Computational Methods For Object Recognition, Identification, And Infrastructure Analysis. *SPIE Proceedings.*

[2] El Bekri N., Peinsipp-Byma E. 2015. An Approach for Min(d) the Quality of Data. *The 2015 International Conference on Data Mining (DMIN)*. 62-64.

[3] Hipp J., Günther U.,Grimmer U. 2001. Data Quality Mining – Making a Virute of Necessity. *Data Mining and Knowledge Discovery (DMKD).*

[4] Luebbers, D., Grimmer U., Jarke M. 2003. Systematic Development Of Data Mining Based Data Quality Tools. *Proceedings Of The 29th International Conference On Very Large Databases*. (29): 548-559.

[5] Batini C., Cappiello C., Francalanci C., Maurino A. 2009. Methodologies For Data Quality Assessment And Improvement. *ACM Computing Surveys (CSUR).* (41): 16 -19

[6] Michalski S., Carbonell, G., Mitchell, M. 2013. Machine learning: An Artificial Intelligence Approach. *Springer Science & Business Media.*

[7] Wu, L., Kaiser, G., Rudin, C., Anderson, R. 2011. 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*. 28-32

[8] ISO. 2015. Quality management systems - Fundamentals and vocabulary / *International Organization for Standardization*. (9000:2015).

[9] Wand Y., Wang Y. 1996. *Communications of the ACM.* (39): 86-95.

[10] Kriebel, C.H. 1979. Evaluating the Quality Of Information Systems. *Design and Implementation of Computer Based Information Systems*.

[11] Ballou, D.P., and Pazer, H.L. 1985. Modeling Data And Process Quality In Multi-Input, Multi-Output Information Systems. *Manage. Sci*. 31. (12) 150-162.

[12] From: http://data.worldbank.org/data-catalog/world-development-indicators. 2015

[13] Draisbach U., Naumann F. 2011. A Generalization of Blocking and Windowing Algorithms for Duplicate Detection. *International Conference on Data and Knowledge Engineering*.