A Methodological Approach to Data Quality Management Supported by Data Mining

Udo Grimmer
DaimlerChrysler AG
Research & Technology, FT3/AD
Ulm, Germany
udo.grimmer@daimlerchrysler.com

Holger Hinrichs
Oldenburg Research and Development
Institute for Computer Science Tools
and Systems (OFFIS)
Oldenburg, Germany
holger.hinrichs@offis.de

Abstract: In this paper, we use the example of the car manufacturing domain to illustrate how data quality problems are addressed in practice today. We then propose a process model for data quality management (DQM) which meets the requirements of the current ISO 9001 standard and thus introduces a methodological, process-oriented approach to DQM. Data mining methods that are typically applied to find interesting and previously unknown patterns in large amounts of data are being used to support several phases of this process model. The main idea behind the application of data mining methods is to deem data anomalies deviations from a 'normal' quality state. The primary advantage of our approach is an increased degree of automation and enhanced thoroughness and flexibility of DQM.

1. Data Quality Challenges in Automotive Manufacturing

While product quality has always been a central focus at DaimlerChrysler, data quality has not yet received the attention it deserves. This does not mean that data quality has been completely neglected thus far, but most data quality initiatives have had solely a strong local focus. With the growing demand for the integration of distributed, heterogeneous databases into corporate warehouse applications, data deficiencies have become obvious, necessitating corporate-wide DQM to address data quality issues across system boundaries. This global data quality view implicates additional data quality perspectives and presents challenges regarding the related tools and methodologies.

As correcting data already stored in a database system is much more expensive than setting up appropriate measures to prevent substandard-quality data to be entered into the systems, precautions should be given top priority. However, as real environments are complex, there will be always a need for measuring, monitoring, and improving the quality of data after it has initially been stored. This was one of the motivations for initiating a research project which focuses on the application of data mining technologies in the context of DQM. We have introduced the term *Data Quality Mining*, which we believe to have great potentials for both future research work and a successful transfer of data mining technology into daily work processes.

In section 2, we propose a quality management system for data integration that meets the requirements of the current ISO 9001 standard. Section 3 presents a case study where a subset of these concepts has been applied to the QUIS (QUality Information System) database of the Global Services and Parts division of the Group using data mining techniques. Finally, we give an overview of related work and further research issues in section 4.

2. ISO 9001 Compliant Data Quality Management

In [16], the term 'quality' is defined as the 'degree to which a set of inherent characteristics fulfils requirements'. Quality characteristics form the backbone of *quality management (QM)*, defined as 'coordinated activities to direct and control an organization with regard to quality'. The system within which QM is performed is called *quality management system (QMS)*.

2.1. The ISO 9001 Standard

The ISO 9000 family of standards was developed by the International Organization for Standardization (ISO) to assist organizations in implementing and operating effective quality management systems. The current ISO 9000:2000 revision was published in December 2000. In this section, we give an introduction to the ISO 9001 standard, i.e. that part of the series that specifies requirements for a QMS.

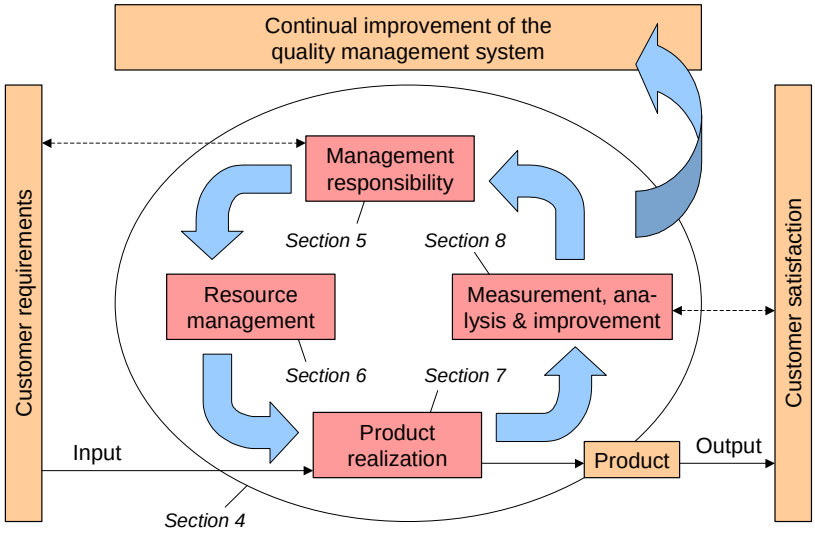

Figure 1: Model of a Process-based QMS [17]

ISO 9001 promotes the adoption of a process approach when developing, implementing, and improving a QMS to enhance customer satisfaction by meeting customer requirements [17]. Within this process, an organization has to identify various activities, then link them and assign resources to them, thus building up a system of communicating processes. Figure 1 depicts such a process-based QMS. Customers play a key role in this model since their requirements are used as input for the product creation process and customer satisfaction is continually subjected to analysis.

ISO 9001 is made up of eight sections. The first three contain general information about the scope of the standard, normative references, and terms. Sections 4 to 8 describe requirements for a QMS and its components, as indicated in Figure 1.

2.2. Data Quality Management

In the following section, we sketch a QMS for the process of data integration from heterogeneous sources as an exemplary data processing activity that is especially important in data warehouse applications like customer relationship management or supply chain management. As Figure 2 illustrates, the data integration process can be viewed as a kind of production process. Following this analogy, we can adapt the well-established QM concepts known from the manufacturing/service domain to the context of data integration, hereafter called *data quality management (DQM)*.

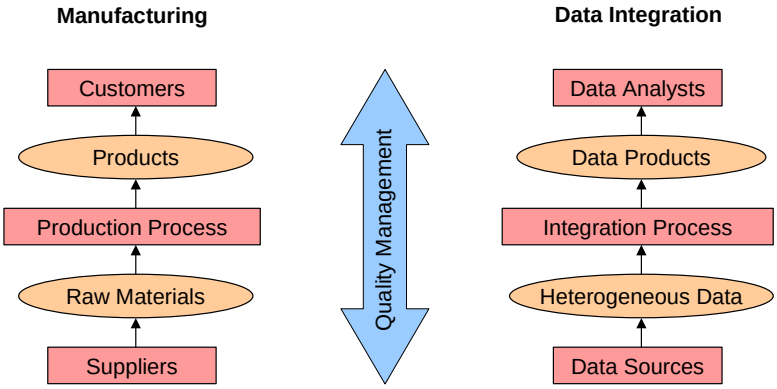

Figure 2: Analogy Between Manufacturing and Data Integration

An essential aspect of DQM is data quality measurement. As DeMarco so rightly states: 'You cannot control what you cannot measure' [3]. We need metrics to be able to calculate the degree of conformance with given requirements. In the manufacturing domain, we have to measure characteristics like lengths, weights, speeds, etc. For databases, on the other hand, we need to measure characteristics like consistency, timeliness, and completeness of data products (see also section 3.2). Yet, metrics for data quality characteristics are still a matter of research [1], [13], [22], [26].

2.3. A Data Quality Management System for Data Integration

In this section, we present a process model for data integration which defines the exact integration steps to be executed. Based on ISO 9001, this process model is enriched with DQM steps to ensure that customer requirements are fulfilled. Integration steps plus DQM steps along with organizational DQM activities form a QMS for data integration called *data quality management system (DQMS)*.

The DQMS should be viewed as an integral part of an organization. It is closely coupled to the organization's management, its human and technical resources, and – of course – its (data) suppliers and (data) customers. Customers specify quality-related requirements and provide feedback concerning their satisfaction with the data products supplied.

In this paper, due to the space constraints, we concentrate on ISO 9001 sections 7 (product realization) and 8 (measurement, analysis, and improvement). For details on the remaining ISO 9001 sections that concern organizational aspects such as documentation, resource management, etc. and their influence on DQM, see [9]. Furthermore, we will only describe the operational

and their influence on DQM, see [9]. Furthermore, we will only describe the operational phases of data integration organized as a 10-step process model, which forms the core of our DQMS (see Figure 3).

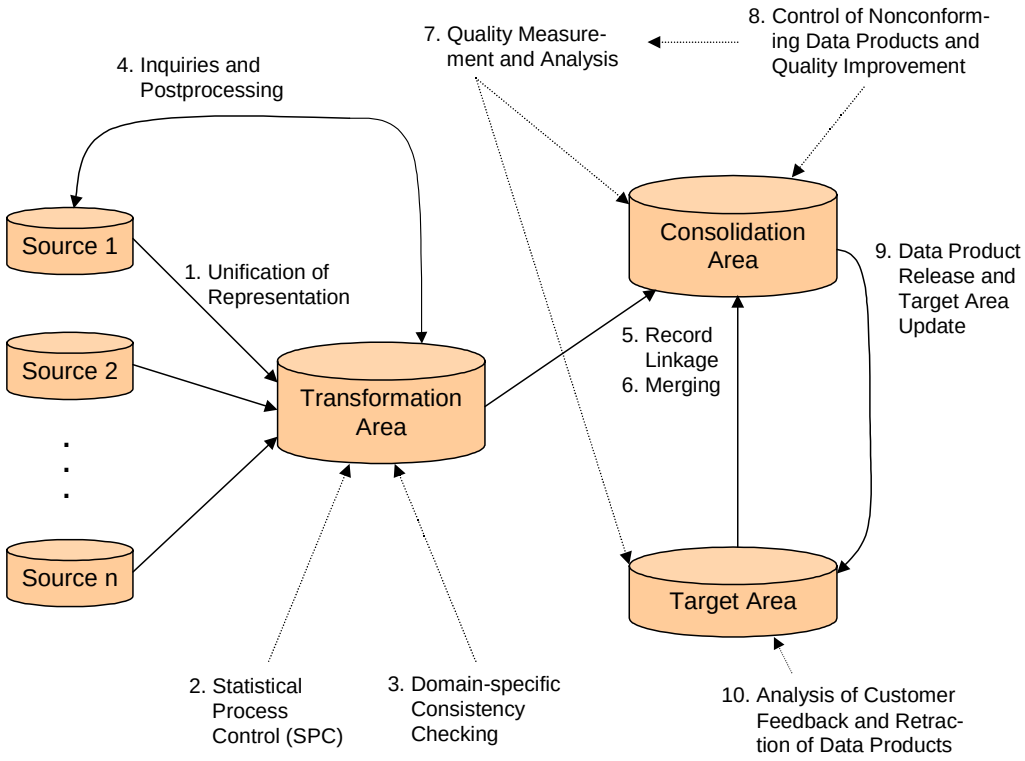

Figure 3: Operational Steps of Data Integration

Step 1: Unification of Representation

In this step, source data are moved into a temporary storage called the *transformation area*. The transformation area is assumed to have a global schema that is covered (in terms of content) by the source schemata. The main task of this step is to map the heterogeneous source data structures to the global data structures of the transformation area. The following mapping tasks are especially important:

- Generating unique keys which refer to source system identifiers.
- Unifying character strings syntactically (with regard to umlauts, white spaces, etc.).
- Unifying character strings semantically (in case of synonyms).
- Decomposing complex attribute values into atomic ones (e. g. addresses).
- Aggregating atomic attribute values into complex ones (e. g. date values).
- Converting codings (e. g. gender values m/f to 1/2).
- Converting scales (e. g. inches to cm).

Step 2: Statistical Process Control

After unification, *statistical process control (SPC)* is performed on the transformation area data as per classical SPC theory [27]. The idea is to compute attribute-specific statistical figures (mean, variance, etc.) and log them over time. The newly computed figures are then compared to previously logged key figures. This allows data deficiencies (e. g. transfer errors) to be detected at a very early stage and appropriate actions such as initiating a new data transfer to be taken.

Step 3: Domain-specific Consistency Checking

In this step, the transformation area records are checked with regard to consistency using domain-specific knowledge. The latter should be represented in such a way that it can be processed automatically. Different types of representation are possible:

- *Rules*, e. g. 'IF RepairDate < ManufacturingDate THEN Error (Severity 0.9)'.
- *Lookup tables*, e. g. engine types.
- *Regular expressions*, e. g. for special equipment codes.
- Arbitrary domain-specific functions.

In addition, business rules could also be discovered at runtime by means of data mining methods (see section 3) and then be applied to transformation area data. If an inconsistency is detected, an appropriate action has to be executed, e. g. by generating an error message or warning.

Step 4: Inquiries and Postprocessing

If appropriate domain knowledge is available (or data mining methods are being applied), the major proportion of inconsistencies can be *detected* automatically. However, very few inconsistencies can be *corrected* automatically. Consequently, if an inconsistency is not tolerable, an inquiry has to be sent to the data source affected (generated automatically from an error message, if possible). Corrected records then have to be integrated into the transformation area appropriately.

Step 5: Record Linkage

The goal of this step is to detect duplicate records, i. e. records that describe the same real-world entity, both within the transformation area and between the transformation area and the so-called *target area* where the consolidated data are to be stored in the end.

Because of the heterogeneity and potential internal redundancy of data sources, records have to be linked by means of non-key attributes like name, city, gender, delivery date, for example. Several methods suitable for automating this task have been proposed in literature. The most prominent ones are (i) Probabilistic Record Linkage [18], (ii) Duplicate Elimination Sorted-Neighborhood Method [14], and (iii) Neighborhood Hash Joins [8]. All these methods result in a set of record pairs which potentially belong together. These pairs, more precisely their transitive closures, now have to be analyzed with respect to whether or not the linkage is correct. In marginal cases, an interactive review is inevitable.

Step 6: Merging

Records that describe the same real-world object must now be merged to a single record in order to avoid unintentional redundancy. By applying certain criteria (information content, attribute-specific priorities on data sources, timeliness, etc.), the best pieces of information have to be extracted from the records in question and written into a target record. In our process model, this target record is *not* written to the target area as one could expect, but into another temporary storage, the so-called *consolidation area*, instead. The target area is not updated until the very last step of the process model, thus ensuring that only data which have passed all the conformance tests (some of which will follow in the subsequent steps) are written to the target area and thus made available for analysis tasks.

The records that merged in a consolidation area record are then deleted from the transformation area. The remaining transformation area records are moved to the consolidation area without any modification. The consolidation area now serves as the starting point for the following DQM activities.

Step 7: Quality Measurement and Analysis

In this step, a check must be done to ensure that the data in the target area meet the specified customer requirements (in compliance with ISO section 8) even after they have been updated with the current consolidation area data. To do this, the actual quality of data must be measured, using appropriate metrics and measuring software according to ISO section 7. These measurements[1] must span both the (present) target area data and the consolidation area data. (Note that the target area has *not* been updated yet!)

In a subsequent analysis phase, the results of the quality measurements have to be compared to a priori-specified quality requirements. If data do not meet a given requirement, appropriate action has to be taken (see step 8). Conflicts resulting from contradicting requirements (e. g. high timeliness vs. high consistency) have to be resolved, e. g. by data replication and different treatment of the replicas.

Measurement results concerning the effectiveness of processes have to be recorded and analyzed (ISO section 8), leading to process improving activities if necessary (see below).

Step 8: Control of Nonconforming Data Products and Quality Improvement

In this last step before the target area update, data products which do not conform to the given requirements must be treated appropriately in accordance with ISO section 8. The following options may be taken:

- Sort out and re-request data.
- Restrict the use of data to specific scenarios, e. g. by flagging.
- Eliminate detected nonconformities and then continue with step 7.

[1] Including consistency checks as in step 3 (and, if required, postprocessing as in step 4), since a merging of records may introduce new combinations of attribute values and thus new inconsistencies.

While all these activities tackle only the symptoms of a problem, further (cause-oriented) measures may be taken to increase the system's ability to fulfill quality requirements in the future according to ISO section 8:

- Improve the integration process, especially by tuning process parameters such as attribute mappings, SPC parameters, consistency rules, record linkage parameters, merging criteria, for example.
- Improve quality planning and quality control processes, e. g. by finding better means to capture user requirements, by optimizing measurement methods, or by improving feedback methods.

Step 9: Data Product Release and Target Area Update

Depending on the analysis results of step 7, the approved proportion of data is now released, i. e. the consolidation area records affected are flagged as 'passed'. The passed proportion of data is then moved from the consolidation area to the target area, replacing obsolete target area data if necessary. With this step, the newly integrated data are made available for customer use.

Step 10: Analysis of Customer Feedback and Retraction of Data Products

The organization concerned has to record customer feedback and evaluate it as a measure of the DQMS performance (ISO section 8). If a deficiency of a released data product is detected during current use (i. e. by a customer), and this deficiency significantly impairs the usability of the data product, the organization has to retract the product from the target area and 'repair' it if possible (see step 8) before re-releasing it. If necessary, cause-oriented measures should be taken into account (see step 8).

3. Case Study: DQM for QUIS

In the following, we describe our experiences with the implementation of selected steps from the DQMS described in section 2.3 to the QUIS (QUality Information System) database that is running on a 20-processor HP-UX machine with 16 GB main memory and 720 GB disk space. QUIS is a central database in the Global Services and Parts division of DaimlerChrysler. It contains technical and commercial data of passenger cars and trucks from the warranty and goodwill periods. It is used for different tasks such as product quality monitoring, early error detection and analysis, or reporting. An application which is targeted at deviation detection of warranty and goodwill costs is presented in detail in [12]. Current data quality problems like inconsistent or incomplete data are generally related to the complex system environment, which consists of various operational source systems with partially non-conforming data models and sophisticated data transfer processes. In this case study, data mining methods have been used for data quality assessment.

3.1. Data Quality Mining

According to Fayyad et al., the typical task of data mining is the investigation of large amounts of data to discover '*valid, novel, potentially useful, and ultimately understandable patterns*' [7].

Although the discovery itself can be automated, the subsequent interpretation of these patterns (rules, decision trees, etc.) always requires human interaction. Let's take a look at the following rule, which was found by a rule learning algorithm during the analyses of the QUIS *axles* table:

"IF Model = 210 AND Plant = 050 AND PDate <= 1999/09/22
THEN PID = B (86514 cases, 100.00%)"

This rule states that in all known (86514) cases, the value for the attribute *PID* is 'B', if all three conditions from the IF part hold. Hence, we could apply this rule for consistency checks for any new data, for example. If we find a record where the conditions from the IF part hold, but the value of the attribute *PID* does not equal 'B', we need to check whether this is a new, valid finding or whether one or more of the values for *Model*, *Plant*, or *PDate* are incorrect.

As this example leads us to conclude, there are only marginal differences between the application of a data mining algorithm for discovering new patterns, and the application of the same algorithm for discovering data anomalies. What makes the difference is merely the way patterns are applied and results interpreted. The term *data quality mining* is used to indicate this differentiation. In [10] we define data quality mining as the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of data quality mining is to detect, quantify, explain and correct data quality deficiencies in very large databases.

Naturally, there are some data mining algorithms that are more appropriate for data quality mining tasks than others. In particular, algorithms from the fields of deviation detection and dependency analysis bear the largest potentials for data quality mining.

3.2. Selection of Relevant Data Quality Aspects

In the initial phase of the QUIS data quality project, we conducted a series of workshops with QUIS data customers (business end users, knowledge engineers, management, and database administrators) to identify the key data quality characteristics. Starting from a list of potential data quality aspects as found in [29], we applied the Quality Function Deployment Matrix approach in the context of DQM as published in [25]. The goal was to map the subjective data quality requirements to objective, quantifiable criteria. As a prerequisite for the application of data mining methods, we had to focus on aspects that could directly be derived from the QUIS data. In accordance with the above constraints, the following five data quality aspects have been identified:

- *Completeness* (with respect to both records and attributes): for records, the percentage of data objects stored in QUIS compared to the number of real entities; for attributes, the percentage of missing and/or blank values.
- *Timeliness*: the period between the date any data have been entered in the operational source systems and the date they become available in QUIS.
- *Consistency*: the number of anomalous records compared to the whole number of records. Consistency rules do not need to be defined manually, as they are discovered from historic data by data mining methods.
- *Accuracy*: the degree of conformance of the data objects in QUIS with the real-world objects.

- *Validity*: the percentage of valid (=known) attribute values for different data fields compared to the number of values stored in certain QUIS reference tables.

Regarding appropriate data mining techniques, we have chosen the following for an initial application and evaluation (the data quality aspects addressed are listed below):

- *Descriptive statistics*: completeness (w. r. t. attributes), validity.
- *Statistical outlier detection*: completeness (w. r. t. records), timeliness, accuracy.
- *Decision rules* and, initially, *association rules*: consistency.

3.3. Data Quality Mining to Support Operational Steps of Data Integration

Following the methodology presented above, we partially implemented some of the steps for QUIS. Initially, we had to investigate the historic data stored in the QUIS tables (this would have been unnecessary if the DQMS proposed in section 2.3 had been applied from the very beginning). For this investigation we developed a prototype that will be described in the section entitled Descriptive Statistics for Relational Database Content Summarization.

Now we discuss the relationship of the individual operational steps 1 through 10 from the data integration process model to the QUIS application, including sample applications of data mining methods.

Step 1: Unification of Representation was met by operational procedures, i.e. records from the source systems are uniquely identified by key attributes such as the vehicle or part identification number.

Step 2: Statistical Process Control was of special importance, since data collecting and transferring processes are susceptible to interference. The corresponding actions are set out in the section on Statistical Approach to Deviation Detection.

Step 3: Domain-specific Consistency Checking and *Step 4: Inquiries and Postprocessing* were already operationalized for a number of business rules. Here, the problem was that new or as yet unknown inconsistencies are not covered by the existing rules because changes in the operational systems are sometimes not communicated to the QUIS administrators quickly enough. The application of a rule generating method to identify potential data anomalies (including new, but still unknown patterns) is described in Decision Rules to Discover Potential Data Anomalies.

Step 5: Record Linkage and *Step 6: Merging* are currently irrelevant for QUIS since the procedures in place are sufficient.

Step 7: Quality Measurement and Analysis, as described in section 2.3, can only be applied to the target area (QUIS) as currently there is no physical consolidation area installed. Finding general (in the sense of 'can be agreed by all data customers') measures to automatically decide on whether the data quality of the system after the update is still satisfactory or not remains a great challenge for the operationalization of this step.

Different approaches are taken for *Step 8: Control of Nonconforming Data Products*, depending on the error type and the source system. For some of the source systems, incorrect data is rejected and re-loaded after the errors have been corrected in the source system. Alternatively, erroneous data is corrected and loaded immediately, with a note to that effect being passed to the source system's administrator. However, not all errors are caused by substandard-quality source

system data. Also, data might be corrupted or lost during transfer processes. In these cases, root causes have to be analyzed, and the particular process has to be fixed and re-executed. As such procedures were already in place, they were therefore not included within the scope of the current project.

Step 9: Data Product Release and Target Area Update are operationalized for QUIS and not directly affected by our research activities.

Finally, *Step 10: Analysis of Customer Feedback and Retraction of Data Products*, is viewed as a mid-term activity. We are, of course, interested in monitoring how our results influence QUIS data quality. We have already started collecting certain data quality measures regarding completeness that will allow for comparisons over the next few months. This part, however, still needs further refinement and completion.

Following this overview, we now discuss some sample applications of data mining methods which directly concern steps 2 to 4 of the process model described in section 2.3.

Descriptive Statistics for Relational Database Content Summarization

One of the key tasks for any data analysis, and for data mining projects in particular, is the so-called explanatory phase. According to the CRISP-DM process model (Cross Industry Standard Process for Data Mining) [2], this phase consists of two subphases: data understanding and data preparation. Experiences from many sources have shown that up to 80 percent of all efforts expended in data mining projects is related to these phases, and that the quality of the results of the analyses carried out depends more on thorough preparation during these phases than on the optimization of any parameter taken from data mining methods in the subsequent modeling phase. Main efforts during the explanatory phase are related to the collection and selection of the right – i.e. the most relevant – data and to statistical, and visual analysis in order to achieve a clear understanding of data contents and representations and data transformations in preparation for the subsequent modeling phase.

For this reason, we took great care to obtain a sound, univariate description of the QUIS data in a first step before applying further, more sophisticated methods. We developed a prototype that automatically generates detailed documentation of the entire QUIS database at attribute level either in Postscript, PDF, or HTML format,. Compared to the functionality of state-of-the-art data mining tool suites, we believe this prototype to have enhanced scalability and the capability to provide a more sophisticated output as it contains a description of each field, together with content and index references.

Figure 4 shows the output of the prototype for the attribute *AXLE_TYPE* from the QUIS table AXLE. Additional textual descriptions extracted from modeling tools, or simple text files can be included for each attribute, too. This is a key prerequisite if database contents need to be discussed with business customers, who normally do not know the mapping between their specific business terms and database attribute names.

2.1.3 *AXLE.AXLE_TYPE*

Data type: VARCHAR(1)
NOT NULL
Number of different values: 3 (0.146987%)
Average occurences of a value: 680.333
Standard deviation: 1017.36
Values:
-- 1850 'S' (90.6418%)
-- 190 'D' (9.30916%)
-- 1 '0' (0.0489956%)

Chart:

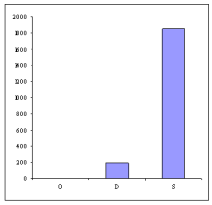

Figure 4: Database Content Documentation

General scenarios the prototype can beneficially contribute to include (i) data and system understanding (e. g. source system analysis), (ii) system design (e.g. entity relationship modeling), or (iii) process optimization.

Statistical Approach to Deviation Detection

Within this part of the project, we developed a statistical prototype for outlier detection. Outlier detection is a burning issue for most analytical applications, and a great deal of work has already been devoted to it in the past. For large real-world applications, it is impossible to provide comprehensive background knowledge specifying what exactly is to be considered an outlier. Data mining methods can be useful here, as they are able to process huge amounts of data autonomously and derive the corresponding knowledge. [4] describes such an approach for the analysis of a very large amount of time series data.

We use the same approach – i.e. deriving models or normative parameters in the simplest case – from historic data. The data used for learning must be guaranteed to be of sufficient quality regarding the parameters to be derived, e. g. mean and variance. Once models have been derived and stored in the model tables for later reuse, corresponding values for new data are computed each time new data are input. Then, a test on deviation is applied. If statistically significant deviations are discovered, suspicious records are flagged and warning messages generated. Figure 5 provides an overview of the steps involved in the model generation and model application processes.

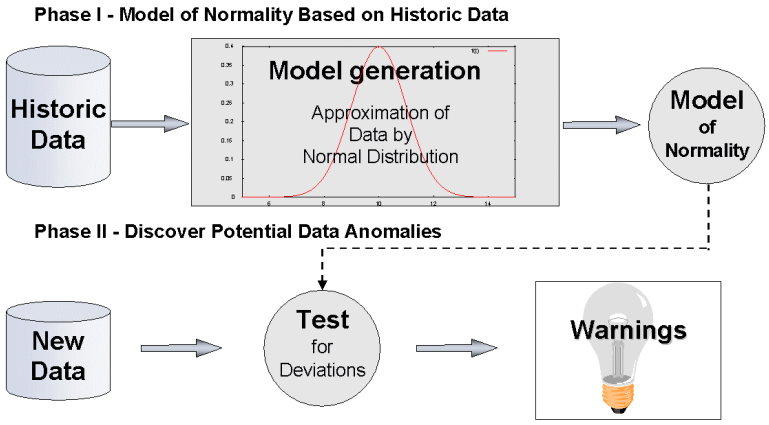

Figure 5: DataQualityMiner-Prototype for Deviation Detection

Example findings from this prototype are, for example, irregular (that is below or above configurable action limits) numbers of warranty claims on a daily or weekly basis for single repair shops as shown in Figure 6.

Claims from garage 199, Period: 01.01.2001-30.06.2001

Weekends (Saturday + Sunday)
13./14.01.2001 claims: 73
expected: 31.23 test statistic: 0.0052496

Weekdays (Monday through Friday)
16.-20.04.2001 claims: 692
expected: 1097.02 test statistic: 0.0045372

Figure 6: DataQualityMiner-Prototype: Sample Results

The prototype offers great flexibility as it covers different queries that can be specified by plain SQL statements. Different parameters such as the model type or an aggregation level regarding a second dimension such as time can be specified as well. Models are updated either as per a fixed schedule or ad-hoc on request. The application of this prototype contributes to (i) automated, early error detection and (ii) permanent process monitoring.

Decision Rules to Discover Potential Data Anomalies

To discover consistency problems in the database, we applied the commercial tool GritBot [23] to the individual QUIS database tables as well as to several joined tables (views). Unfortunately, there is no technical description available on the methods GritBot uses. What we have, however, been able to guess from the results and the works published by Quinlan is that GritBot performs several rule generating runs, each time considering another attribute from a subset of n (default n=5) attributes as target (dependent) attribute. Next, each record that violates a certain rule is assigned a significance value indicating the probability that the anomalous value might occur by chance rather than by error. Numeric attributes are discretized automatically, and groupings of values for nominal attributes are generated where appropriate. There are only few parameters that the user can change, but this seems to be somewhat of an advantage, as we were able to achieve good results using only the default settings.

```
GritBot [Release 1.02]  Mon Jul  3 13:30:52 2000      Options:  Application `AXLE1'
Read 5219164 cases (14 attributes) from AXLE1.data
1995 possible anomalies identified

case 1324925:  [0.000]
        AXLE_MDAT = 1919/02/28  (55911 cases, mean 1996/03/31, 100.00% >= 1995/09/01)
           AXLE_TYPE in {311, 919, 305, 306, 061, 018, 460, 313, 470, 054,  ..., 075} [480]
           AXLE_PDAT > 1995/09/11 and <= 1997/10/29 [1995/11/15]
case 3773780:  [0.000]
        AXLE_EPOS = 2  (118966 cases, 100.00% `1')
           AXLE_TYPE = 510
case 2746691:  [0.000]
        AXLE_MODEL = 739  (32459 cases, 100.00% `730')
           AXLE_TYPE = 018
           AXLE_EPOS = 1
           AXLE_ASSID = D
....
```

Figure 7 : Sample GritBot Results

An example from the analysis of the QUIS table *axles* is depicted in Figure 7: for case 1324925, for example, the value for the field *AXLE_MDAT* was identified to be anomalous, whereas for case 2746691 the value of the field *AXLE_MODEL* seems to be incorrect. However, for the latter case there is no evidence that it is the value of the attribute *AXLE_MODEL*, which is wrong. This might also be a correct value, and one or more of the values for *AXLE_TYPE*, *AXLE_EPOS*, or *AXLE_ASSID* might be incorrect.

GritBot proved to be highly scalable in terms of the number of records per table, and the results generated were greatly promising. Without requiring specification of any business knowledge in advance, business rules known by QUIS data customers were proved, and, to top it off, new potential data anomalies were discovered. For our purposes, GritBot's major drawback is the missing database interface: we were continually forced to manually extract two flat files – one file containing the data and a corresponding names file describing the types and possible values for each attribute. As GritBot generated long output listings (some 270,000 lines for all single tables), additional postprocessing steps were implemented to aggregate analytical results.

The application of association rule methods as proposed in [21] seems to be another promising approach. There are, however, two main challenges related to the application of association rules: (i) the special data representation required by most association rule algorithms, and (ii) processing the large number of rules generated. We have already addressed the former issue and are now able to apply association rule algorithms directly on a relational database [11]. Further work is needed to take up the latter challenge. We are about to implement a system which automatically computes quality scores for each record depending on the percentage of its conformance with the rule set generated by the algorithms [10]. This, we hope, will make human interpretation of the rule set to a large extent dispensable.

Both approaches are generally applicable for (i) unsupervised consistency checks and (ii) detection of random (non-frequent) errors.

4. Related Work and Future Challenges

In 1992, DeLone and McLean [5] set up a model of information system success that included the quality of data as a key success factor. In the following years, two major research projects emerged, viz. MIT's *Total Data Quality Management (TDQM)* program [28], active since 1992, and the ESPRIT project *Foundations of Data Warehouse Quality (DWQ)* [19], which ran from 1996 to 1999. Apart from TDQM and DWQ, several minor research activities have been launched during the last two to three years, reflecting the rising awareness of the importance of DQM. Among these are CARAVEL [8], IntelliClean [20], HiQiQ [22], and Potter's Wheel A-B-C [24]. [21], [4], and [15] use data mining methods to detect errors automatically.

All in all, although there are some projects which deal with data quality aspects, several critical research issues remain. The overall challenge will be to promote corporate-wide data quality awareness and thus establish DQM as a primary success factor in organizations. DQM should make it evident to all the people in an organization that data quality is not just a local issue, but a global challenge of strategic importance. In the long run, we need to strive for certification of data and data processing systems, as the methodological approach proposed in section 2 suggests. To reach this goal, a number of research issues need to be investigated further:

- Commonly accepted metrics are necessary to enable organizations to assess the quality of their data. For data quality characteristics such as consistency, completeness, and absence of redundancy, metrics have already been defined (see [13]). However, 'soft' characteristics like relevance and understandability are very difficult to handle, because they are exposed to subjective influences and require extensive domain and context knowledge. Hence, methods that allow partial automation of measurement by appropriately integrating interactive components must be developed to accommodate these characteristics.
- A methodology for introducing DQM into an organization is needed. Costs related to substandard-quality data and appropriate data quality assurance steps need to be analyzed in more detail (see [6] for an initial discussion). If such costs can be identified and proved using real figures, this would greatly facilitate argumentation for data quality project funding.
- A comprehensive, flexible, and standardized metadata management, plus scalable, domain-independent, and automatable software tools for data quality measurement and improvement have to be developed. To ensure efficiency even with very large data volumes, DQM operators should be implemented as close to the database management system as possible. Furthermore, tools should offer user-oriented means to maintain data quality figures and monitor them over time.
- Automated correction of inconsistencies also requires a great deal of further research. Data mining-based approaches seem to be especially promising in this field (cf. section 3). Future research work should include the integration of domain-specific knowledge and the optimization of efficiency, precision, and recall of such methods.

Although we foresee the need for extensive research work, we believe that our combination of an ISO 9001-compliant process model and automated, data mining-based quality measures will yield a promising foundation for corporate-wide, scalable data quality management.

5. References

[1] Ballou, D. P., Tayi, G. K.: Enhancing Data Quality in Data Warehouse Environments, *Communications of the ACM*, **42** (1): 73-78, 1999.

[2] CRISP-DM Consortium: *CRISP-DM 1.0 – Step-by-Step Data Mining Guide*, http://www.crisp-dm.org, 2000.

[3] DeMarco, T.: *Controlling Software Projects*, Yourdon Press, New York, 1982.

[4] Dasu, T., Johnson, T., Koutsofios, E.: Hunting Data Glitches in Massive Time Series Data, *Proc. of the 2000 Conference on Information Quality*, 2000.

[5] DeLone, W. H., McLean, E. R.: Information Systems Success: The Quest for the Dependent Variable, *Inf. Systems Research*, **3** (1): 60-95, 1992.

[6] English, L.P., Improving Data Warehouse and Business Information Quality, New York: John Wiley & Sons, 1999

[7] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview, in: Fayyad U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.), *Advances in Knowledge Discovery and Data Mining*, AAAI-Press, pp. 1-37, 1996.

[8] Galhardas, H., Florescu, D., Shasha, D., Simon, E.: Declaratively Cleaning your Data using AJAX, *Journ. Bases de Données Avancées*, Oct. 2000.

[9] Hinrichs, H., Aden, T.: An ISO 9001:2000 Compliant Quality Management System for Data Integration in Data Warehouse Systems, in: Theodoratos, D., Hammer, J., Jeusfeld, M., Staudt, M. (eds.): *Proc. Intl. Workshop on Design & Management of Data Warehouses (DMDW), Interlaken, Switzerland*, 2001.

[10] Hipp, J., Güntzer, U., Grimmer, U.: Data Quality Mining – Making a Virtue of Necessity, *Proceedings of the ACM SIGMOD/PODS 2001 Conference (Workshop DMKD'01)*, 2001.

[11] Hipp, J., Güntzer, U., Grimmer, U.: Integrating Association Rule Mining Algorithms with Relational Database Systems, *International Conference on Enterprise Information Systems* (forthcoming), 2001.

[12] Hotz, E., Grimmer, U., Heuser, W., Nakhaeizadeh, R., Wieczorek, M.: REVI-MINER, a KDD-Environment for Deviation Detection and Analysis of Warranty and Goodwill Cost Statements in Automotive Industry, *Proceedings of the ACM SIGKDD* (forthcoming), 2001.

[13] Hinrichs, H.: Datenqualitaetsmanagement in Data Warehouse-Umgebungen, *Datenbanksysteme in Buero, Technik und Wissenschaft, 9. GI-Fachtagung BTW 2001, Oldenburg*, pp. 187-206, Springer, Berlin, 2001 (in German).

[14] Hernandez, M. A., Stolfo, S. J.: The Merge/Purge Problem for Large Databases, *Proc. of the 1995 ACM SIGMOD Conference*, 1995.

[15] Hinrichs, H., Wilkens, T.: Metadata-Based Data Auditing, *Data Mining II (Proc. of the 2nd Intl. Conf. on Data Mining, Cambridge, UK)*, pp. 141-150, WIT Press, Southampton, 2000.

[16] International Organization for Standardization: *ISO 9000:2000: Quality Management Systems – Fundamentals and Vocabulary*, Beuth, Berlin, 2000.

[17] International Organization for Standardization: *ISO 9000:2000: Quality Management Systems – Requirements*, Beuth, Berlin, 2000.

[18] Jaro, M. A.: Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, **84**: 414-420, 1989.

[19] Jarke, M., Jeusfeld, M. A., Quix, C., Vassiliadis, P.: Architecture and Quality in Data Warehouses, *Proc. of the 10th Intl. Conf. CAiSE*98, Pisa, Italy*, pp. 93-113, Springer, Berlin, 1998.

[20] Lee, M. L., Ling, T. W., Low W. L.: IntelliClean – A Knowledge-Based Intelligent Data Cleaner, *Proc. of the 6th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, Boston, MA*, 2000.

[21] Maletic, J. I., Marcus, A.: Data Cleansing – Beyond Integrity Analysis, *Proc. Conf. on Information Quality IQ2000, MIT, Boston, MA*, pp. 200-209, 2000.

[22] Naumann, F., Leser, U., Freytag, J. C.: Quality-Driven Integration of Heterogeneous Information Sources, *Proc. of the 1999 Intl. Conf. on Very Large Databases (VLDB '99), Edinburgh, UK*, 1999.

[23] Quinlan, R.: *GritBot – An informal tutorial*, http://www.rulequest.com, 2000.

[24] Raman, V., Chou, A., Hellerstein, J. M.: Scalable Spreadsheets for Interactive Data Analysis, *Proc. of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), Philadelphia*, 1999.

[25] Redman, T. C.: Data Quality for the Information Age. Boston/London: Artech House, 1996.

[26] Shanks, G.: Semiotic Approach to Understanding Representation in Information Systems,

Proc. of the IS Foundations Workshop, ICS Macquarie University, Sydney, 1999.

[27] Shewhart, W. A.: *Economic Control of Quality of Manufactured Product*, D. Van Nostrand, New York, 1931.

[28] Wang, R. Y.: A Product Perspective on Total Data Quality Management, *Communications of the ACM*, 41 (2): 58-65, 1998.

[29] Wang, R. Y., Strong, D. M.: Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems (JMIS)*, 12(4), 5-34, 1996.