Analysis of data quality issues in real-world industrial data

Thomas Hubauer[1], Steffen Lamparter[2], Mikhail Roshchin[3], Nina Solomakhina[4], and Stuart Watson[5]

[1,2,3]*Siemens AG, Munich, 81739, Germany*
fname.lname@siemens.com

[4]*Siemens AG, Munich, 81739, Germany*
nina.solomakhina.ext@siemens.com

[5]*Siemens Industrial Turbomachinery Ltd, Lincoln, LN5 7FD, United Kingdom*
watson.stuart@siemens.com

ABSTRACT

In large industries usage of advanced technological methods and modern equipment comes with the problem of storing, interpreting and analyzing huge amount of information. Handling information becomes more complicated and important at the same time. So, data quality is one of major challenges considering a rapid growth of information, fragmentation of information systems, incorrect data formatting and other issues. The aim of this paper is to describe industrial data processing and analytics on the real-world use case. The most crucial data quality issues are described, examined and classified in terms of Data Quality Dimensions. Factual industrial information supports and illustrates each encountered data deficiency. In addition, we describe methods for elimination data quality issues and data analysis techniques, which are applied after cleaning data procedure. In addition, an approach to address data quality problems in large-scale industrial datasets is proposed. This techniques and methods comprise several well-known techniques, which come from both worlds of mathematical logic and also statistics, improving data quality procedure and cleaning results.

1. INTRODUCTION

Caused by decreasing software cost and technological improvements, the amount of data produced, processed and stored by companies grows continuously. This data contains information regarding work process, equipment, staff involved and even more. Based on this data decisions are made, long-term plans are drawn up and statistics are compiled. Therefore, even small amounts of poor quality data may cause problems and costly consequences. Examples for complications caused by dirty information include wrong decisions, inadequate prognoses based on imperfect statistics, troublesome handling and analysis of data. Hence, data cleansing is one of the most important tasks in information technologies, especially in knowledge-based systems. The current work examines data analysis and cleaning using an example from the Siemens Energy Sector, in particular its subdivision Oil and Gas solutions. Part of its operational data is analyzed for possible data quality problems and a number of approaches to their solution are considered.

This paper is structured as follows. Firstly, an introduction to the topic of data quality and a description of related work is provided. The third section describes an industrial use case and particular data schemes. The forth section examines primary data characteristics describing its quality conditions. First subsection here comes with the list of generally defined data quality dimensions. Next, they are discussed in conjunction with our industrial scenario and illustrated with factual examples. In the fifth section there are proposed techniques and methods, which help to overcome difficulties of low data quality and make use of such information. In addition, data analysis techniques are described. The paper concludes with a summary of findings and statements of further requirements and needs for future development in data quality assessment and data cleaning for industrial data-related procedures.

2. RELATED WORK

Unsatisfactory data quality affects each field of action in both IT-related procedures and business-related tasks. Many companies elaborate their approaches to data quality assessment with respect to their own data purposes and types. Huge amounts of data, including names, addresses, numerical and categorical values have to be stored and manipulated. Towards to improvement of information quality assessment there is a number of research works

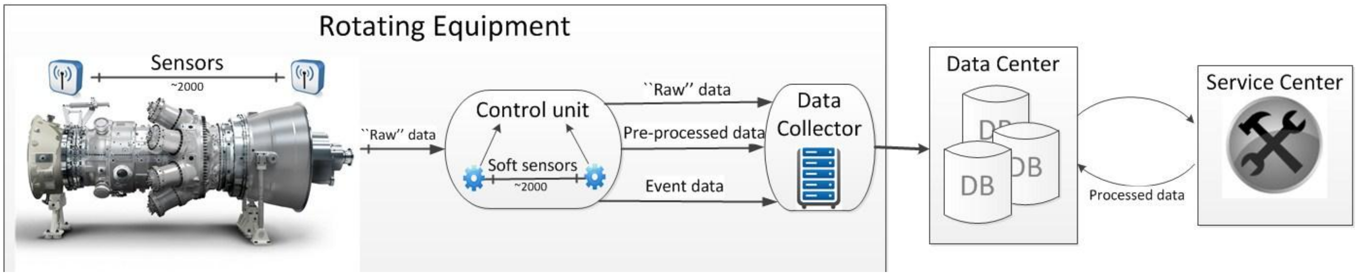


Figure 1. Appliance structure and data flow

conducted by post, insurance (Corporation & Consulting, 2011) and product trading (Pipino, Lee, & Wang, 2002) companies, criminal-record governmental system (Laudon, 1986) and many others (Wang, Strong, & Guarascio, 1996). In industry and research fields the challenge of complex data access is relevant as well: there exist a number of research works from different branches, such as industrial ecology (Weidema, B. P. & Wesnæs, 1996), healthcare industry (Safran et al., 1998; Gendron & D'Onofrio, 2001), meteorology (Foken et al., 2005), sensor networks (Wallis et al., 2007). Moreover, recently there have been launched a project "Optique" intended to improve data quality and to provide a quick end-user access to Big Data. It is conducted jointly by several European universities and two big industrial companies: Siemens AG and Statoil USA. The goals of the project are (Optique, 2012):

- to provide a semantic end-to-end connection between users and data sources;
- to enable users to rapidly formulate intuitive queries using familiar vocabularies and conceptualizations;
- to integrate data spread across multiple distributed data sources, including streaming sources;
- to exploit massive parallelism for scalability far beyond traditional RDBMSs and thus to reduce the turnaround time for information requests to minutes rather than days.

3. Case study: Industrial Data

This paper relates to data quality at Siemens Energy Sector. Data handling and processing in energy domain is becoming a big challenge, while power generation is getting more and more important in the course of time.

Siemens Energy Services maintains thousands of power generation facilities, specifically, the major core components: gas and steam turbines, called in the latter "appliances" or "rotating equipment". Operational support is provided through a global network of more than 50 service centers. These centers are in turn linked to a common database center, which stores the information coming from the appliances in several thousands databases. Further in this chapter data organization, processing and data types used in the tables are presented (see also Figure 1).

Each appliance comprises several industrial computers, which operate based upon information from sensors and serve the functions of (i) control unit and (ii) data collector. Overall approximately 2000 sensors are used to monitor the functioning of a single appliance.

The control unit serves the following functions: receiving sensor measurements, real-time monitoring of the appliance and communication of all information to a data collector. To conduct monitoring, it processes received sensor data in several ways and generates corresponding short messages ("events"), that describe the status of a unit and its functioning. There are three levels of data processing offered by a control unit:

- no processing applied at all, data remains as it was generated by hardware sensors ("raw" data);
- soft sensors: small chunks of code, which use predefined rules (i.e., thresholds, trends) in order to generate events for condition monitoring; there are usually approx. 2000 soft sensors and mostly each soft sensor is assigned to one or several hard sensors;
- simple analysis: information preprocessing, based on hard sensors' measurements and soft sensors' calculations (e.g., Fast Fourier Transformation).

The main function of a data collector is to accumulate the information, passed by control unit and to send it regularly to the central database. Below are described different types of tables stored in databases:

- **Serial numbers, identification codes and all general characteristics of unit components.** Below is shown exemplary Table 1 with such information.

  In the addition, location of the appliance, weather conditions, history of operation and conducted maintenance, performance indices are provided. Generally, information of that kind is polytypic: it contains strings, numerical data and other types.

Table 1. Main characteristics of an appliance

ID	Engine Type	Power Output	Frequency	…
T1	TurbineType1	12.90MW(e)	50/60 Hz	…
T2	TurbineType2	19.10MW(e)	50/60 Hz	…

Table 2. Measurement data

SensorID	Timestamp	Value1	Value2	…
TMP23-1	2010/07/23 23:11:55	44	49	…

- **Measurements of sensors and monitoring devices.** Table 2 depicts the schema of such data, which is often called as “raw” data, referring to the fact, that it represents unprocessed data incoming from machinery itself. Tables of that category contain mostly numerical data and have extremely large size.
- **Pre-processed data and events.** Typically data preprocessed by control unit and soft sensors is stored in different tables. Though these tables are distinguished, they have the same structure, showed below in the Table 3. Tables of this category have a huge size as well and consist mostly of text and date/time data.
- **Processed data.** In that category databases store results of analyses, conducted previously by service centre for a particular appliance. All diagnostics results based on data from central database, store in the database as well and might be used for further diagnostics.

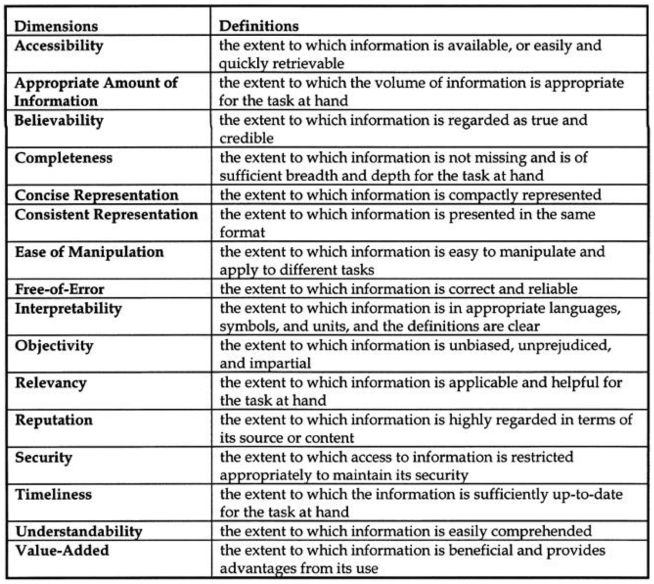
Each table has up to 20 attributes and contains various data formats, including scaled (nominal, ordinal, interval, ratio types), separated (with comma, tabulation), binary, floating-point (single, double) data types. Per a single appliance overall amount of tables exceeds 150. In sum, tabulations with sensor and event data result in 100 TB of timestamped data. Moreover, sensors continuously produce measurements at a rate between 1 and 1000 Hz and about 30 GB of a new sensor and event data are generated per day. Due to numerous causes, such as different vendors of devices or historical reasons, for a database scheme there exist more than 10 various logical schemes.

Thus, in described situation there arises a number of challenges that complicate access to information and its processing. Their overcoming requires great amount of time and resources. Further in this paper these challenges and approaches to them are discussed more thoroughly.

4. Data Quality Dimensions

In order to point out and classify the defects of data, special data characteristics have been defined. Deficient condition of any one of them has an impact on effective analysis and processing of the information. They are called *Data Quality Dimensions.*

The first part of this section lists general data characteristics used to describe data of any purpose. The second part explores dimensions of industrial data and provides some explanatory factual examples.



Dimensions	Definitions
Accessibility	the extent to which information is available, or easily and quickly retrievable
Appropriate Amount of Information	the extent to which the volume of information is appropriate for the task at hand
Believability	the extent to which information is regarded as true and credible
Completeness	the extent to which information is not missing and is of sufficient breadth and depth for the task at hand
Concise Representation	the extent to which information is compactly represented
Consistent Representation	the extent to which information is presented in the same format
Ease of Manipulation	the extent to which information is easy to manipulate and apply to different tasks
Free-of-Error	the extent to which information is correct and reliable
Interpretability	the extent to which information is in appropriate languages, symbols, and units, and the definitions are clear
Objectivity	the extent to which information is unbiased, unprejudiced, and impartial
Relevancy	the extent to which information is applicable and helpful for the task at hand
Reputation	the extent to which information is highly regarded in terms of its source or content
Security	the extent to which access to information is restricted appropriately to maintain its security
Timeliness	the extent to which the information is sufficiently up-to-date for the task at hand
Understandability	the extent to which information is easily comprehended
Value-Added	the extent to which information is beneficial and provides advantages from its use

Figure 2. Data Quality Dimensions

4.1. Main characteristics of a data

Overall there are 16 typical data quality dimensions describing data features (Kahn, Strong, & Wang, 2002) as listed in Figure 2.

Typically, classification of dimensions slightly differs depending on the purpose of information and used data types. From time to time some dimensions are omitted and others are split up to several more concrete attributes. The reason is that in various fields of actions some particular characteristics are more important and more attention is paid to them. For instance, for military government information security is a major feature, whereas for postal services complete and free-of-errors address database is more of a priority. For easier prioritizing and handling data quality issues, dimensions can be clustered in three *hyperdimensions* (Karr, Sanil, & Banks, 2006):

- **Process:** characteristics related to a maintenance of data, such as Ease of Manipulation, Value-Added, Security.
- **Data:** characteristics of the information itself, such as Believability, Completeness, Free of Error, Objectivity, Relevancy.
- **User:** characteristics related to usage and interaction with users, such as Appropriate Amount of Information, Accessibility, Timeliness, Understandability.

Nevertheless, all above-listed data attributes are important for databases of any purpose and there exist different techniques and methods to estimate them and correct existing data to improve its attributes.

Table 3. Processed data

ApplianceID	Time	Class	ErrorCode	Downtime	…
XX476	2010/07/23 21:10:35	Warning	OilTemperatureHigh	00:00:05	…

To obtain acceptable data quality however often requires a lot of time and resources and at times even manual correction to ensure cleanliness of data.

From now on we focus on industrial data and its significant data quality attributes, in particular in the domain of energy solutions.

4.2. Data Quality Dimensions in industry

In this section we analyze the quality of real industrial data based on (a relevant subset of) the data quality dimensions defined previously. One of the tools used during this project for analyzing data in the Siemens database and exploring its quality is the "Diagnostics of rotating equipment" software. Its main features include:

- loading from a database sensor and event data corresponding to one particular or several appliances, components or devices during a certain time period;
- visualization of data using tables and graphs;
- analyzing sensor signals by means of statistical methods;
- identifying patterns in event data i.e., revealing regularities preceding occurrences of a particular event.

In the following we give concrete examples of Data Quality Dimensions presented in Section 4.1. In order to illustrate relevancy of data quality problems there are used thermocouples measurements monitoring functioning of a gas turbine.

Completeness, accessibility

The fullness of information i.e., the fact that data is not missing and sufficiently detailed, is the most important characteristic of a data. Nevertheless, data loss is not uncommon in industry for several reasons. These reasons include the inability to access the required data: the appliance might be located in a remote region and due to a bad (or absent) connection between the data collector in the unit and the main database, the information may be unavailable. Another reason is device faults. Depending on causes, there might be absent only one type of data tables: "raw" or event data, and in that case it is still possible to make use of available information in order to conduct an analysis. More severely is the case that no data for a particular period is available at all. Figure 3 depicts loss of sensor measurements whereas Figure 4 shows absence of event pre-processed data for a week between 20th and 26th of September.

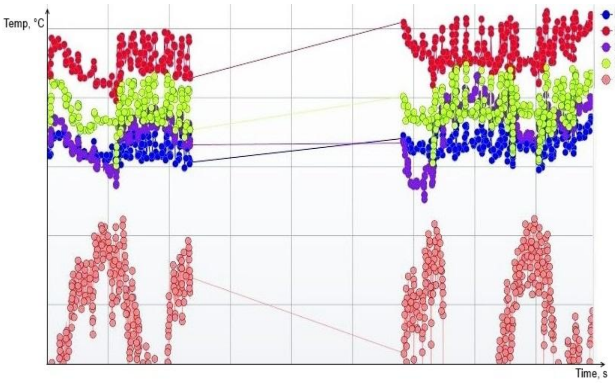
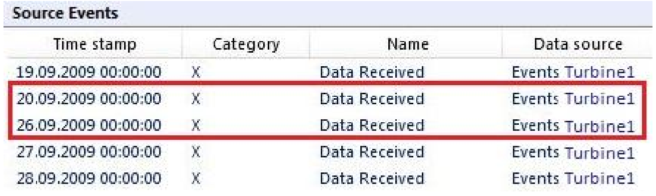


Figure 3. Signal data loss



Source Events

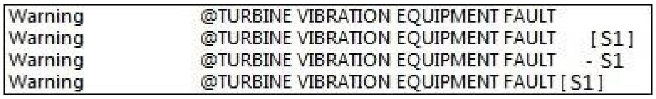
Time stamp	Category	Name	Data source
19.09.2009 00:00:00	X	Data Received	Events Turbine1
20.09.2009 00:00:00	X	Data Received	Events Turbine1
26.09.2009 00:00:00	X	Data Received	Events Turbine1
27.09.2009 00:00:00	X	Data Received	Events Turbine1
28.09.2009 00:00:00	X	Data Received	Events Turbine1

Figure 4. Event data loss

Consistent Representation

When information comes from multiple sources, it is essential to have data represented in the same format. In the current use case there exist a number of contraventions:

- various recordings of timestamps, as date and time can be written in several ways. For instance, devices of one kind write timestamps as *DD/MM/YYYY~hh:mm:ss* while another have a format *YYYY-MM-DD~hh:mm:ss* and many more of other types of devices having other date and time formats.
- data types of some information sources and monitoring devices require conversion from one format to another e.g., from *String* to *Float* or from *String* to *Integer*.
- different monitoring systems and control units indicate the same event in different ways. That happens due to diverse reasons such as various device vendors, different software versions or even location. Therefore a lack of standardization might occur and the same entities and events might be denoted differently. As shown in Figure 5, when a device S1 measurements show the failure of vibration devices, the corresponding event is denoted in several ways: different quantity of spaces between



```
Warning          @TURBINE VIBRATION EQUIPMENT FAULT
Warning          @TURBINE VIBRATION EQUIPMENT FAULT    [S1]
Warning          @TURBINE VIBRATION EQUIPMENT FAULT    - S1
Warning          @TURBINE VIBRATION EQUIPMENT FAULT [S1]
```

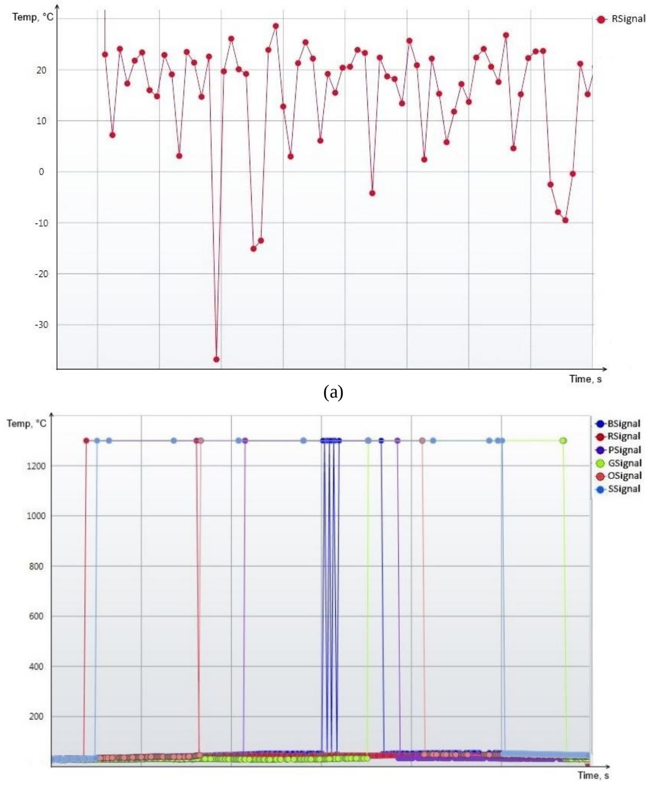
Figure 5. Different denotations of the same event

event message and sensor ID, sensor ID in parentheses etc, which badly affects analysis and statistics.

Free of Errors, Believability, Accuracy

In order to rely on results of analysis, data should be correct, precise and relevant. The possible causes of occurrence of erroneous and inaccurate data are very diverse: (i) one or several devices of the appliance faulted and gave inaccurate or wrong measurements; (ii) control unit failure occurred and there was an error during data preprocessing; (iii) there are three data transfer segments - from sensors to control unit, from control unit to data collector and between the appliance and data warehouse, for each frequencies of data transfer and speeds of data flow differ. It might happen that poor connection distorted information on one of these segments. Below are listed a few examples of discrediting data or insufficient data accuracy.

- Time Synchronization - timestamps of events and measurements incoming from several different devices might slightly differ due to such reasons as (1) time settings of a particular devices; (2) frequency and duration of data transfers between components, control unit and data collector.
- Range of values. Figure 6a shows an example where thermocouple sensor measure values are out of domain, namely minus temperatures. Additionally, occasionally outliers occur – spikes or sudden changes of value within the domain. They should be treated properly during the analysis. Figure 6b depicts an example of outliers - all sensors show alternately range maximum and minimum.
- Oscillations and noise. Figure 7a shows heavy oscillations of all signals. Figure 7b depicts the case, when signal measurements contain too much noisy data.
- Vast difference in measures. If there are several sensing elements, which duplicate each other, and they measure completely different values, then it is problematically to rely on these measurements. On a Figure 7c RSignal measurements differ from all other measurements for more than 100 degrees. In the case shown on Figure 7d, duplicating sensors measure the similar values, but as soon as temperature drops or rises, sensors measurements change with the different amplitude, as it is marked inside of black rectangles.



(a)

(b)

Figure 6. (a) Values out of range: minus temperatures. (b) Outliers: measurements of range minimum, maximum.

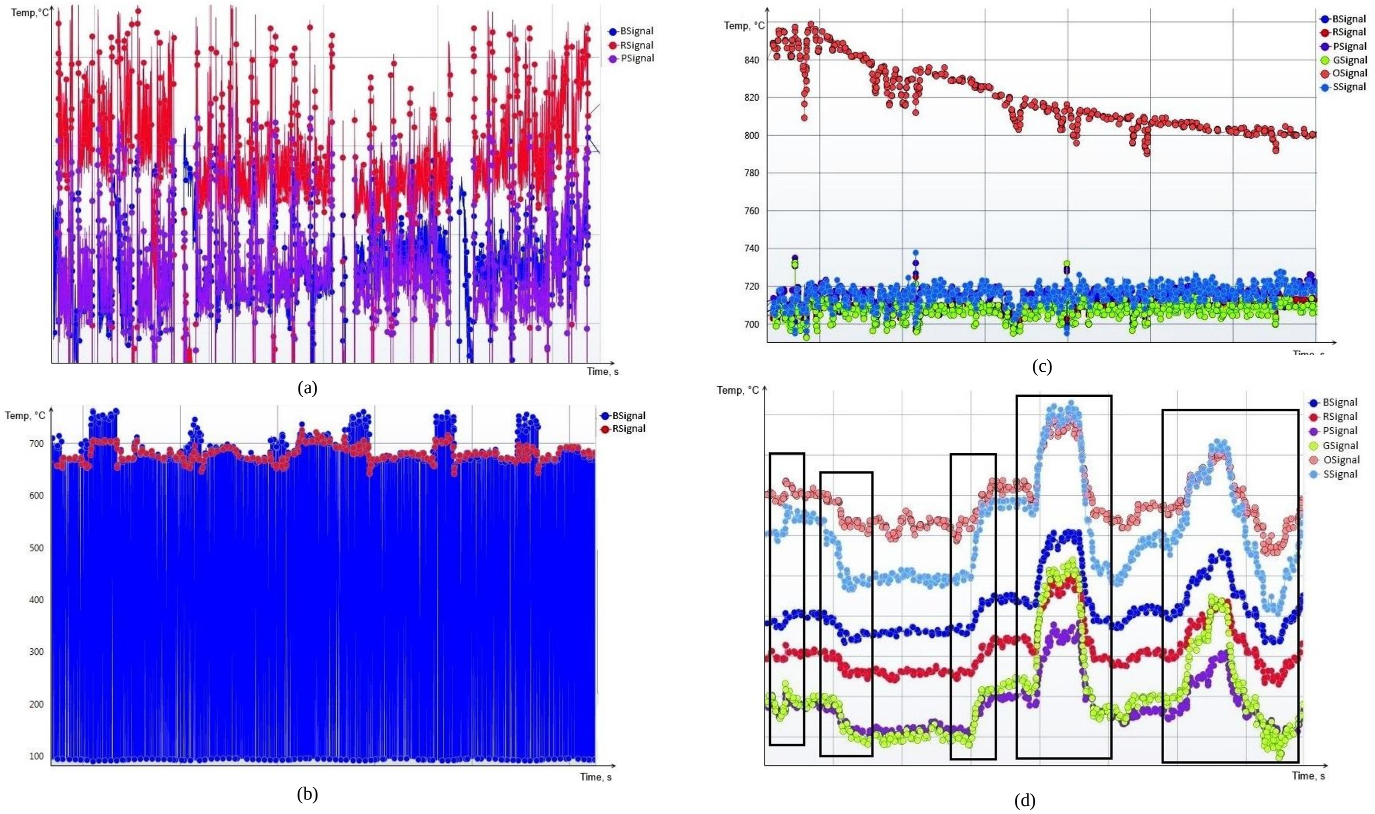
- Signal alternation. On the Figure 8 is shown the case, when two signal at some moment alternated each other and swapped their measurements, as it is also marked with the black rectangles.

Ease of Manipulation, Data Schemes

Data schemes and structures are highly heterogeneous, depended upon which technique was used to create it, which unit it belongs to, from where it comes historically. Moreover, not all foreign keys between databases are present. If information on the same entity is distributed among several sources, for instance, if information concerning a particular malfunction of an appliance should be extracted from tables “Incident Summary”, “Daily Event Log”, “Burner tip temperature” and others, the problem of missing foreign keys do not allow for easy merging of data.

Timeliness, Appropriate Amount of Information

For a thorough analysis it is critical to have all data available and updated. Though for each diagnostics case the considered time period always differs: it might be sufficient to consider only the last hour in order to identify a cause of an event, but in other cases one needs to analyze the last several years, for example to detect a deterioration of a particular component.



(a)

(c)

(b)

(d)

Figure 7. (a) Oscillatory signals. (b) Noisy data - BSignal measurements look like a white noise. (c) RSignal shows divergent values. (d) SSignal and GSignal have rise/drop amplitudes differing from other duplicating sensors.

Thus, usually data typically does not expire and become irrelevant in several years but on the other hand, has to be stored for decades.

In conclusion, for successful information analysis it is crucial to determine how reliable data is and to bring it to the representation convenient for required purposes. In the following section we discuss methods and techniques developed to achieve this goal.

5. Data Analysis

In this chapter there are examined techniques which help to get use of low quality data. Firstly, there are described data cleaning methods and in addition, a proposal to improve them is made. Data analysis techniques are described in the second part of the chapter.

5.1. Quality assessment and cleaning

There are several directions in data cleaning and existing techniques aimed at particular problems (Rahm & Do, 2000): duplicate identification and elimination, data transformations, schema matching, data mining approaches and others. Moreover there are also unified techniques. The main scientific approaches include statistical, machine learning and knowledge-based approaches. But in general any data cleaning technique should satisfy several requirements (Rahm & Do, 2000):

- should detect and remove all major errors and inconsistencies both in individual data sources and when integrating multiple sources.
- should be supported by tools to limit manual inspection and programming effort and be extensible to easily cover additional sources.
- should not be performed in isolation but together with schema-related data transformations based on comprehensive metadata.

Statistical methods are used to: (i) visualize the data; (ii) summarize and describe existing data by means of univariate and multivariate analysis; (iii) offer hypotheses and decisions with the aid of statistical tests; (iv) interpret data employing sampling techniques.

One of the most widely used statistical tools for data quality assessment is called quality indicator (Bergdahl et al., 2007). It is a measure of how well provided information meets criteria and requirements for an output quality. Also there exist a number of statistical/probabilistic techniques and its modifications (Winkler, 1999), 1-1 matching methods and bridging file technique (Winkler, 2004).

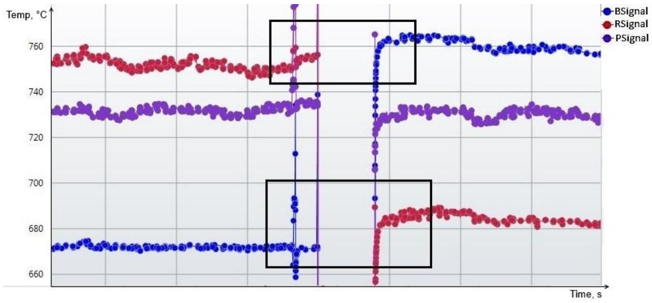


Figure 8. BSignal and RSignal traded places.

In the current use case statistical approach is widely used for detecting faults in sensor readings. For large-scale databases that enlarge day by day with new portions of sensor measurements it is highly essential to use fast and robust techniques detecting changes in signal behavior. The main approach is time-series analysis, cross- and autocorrelation, spectrum and Fourier analyses in particular. In addition, there are simple indicators and distribution tests are exploited in order to detect quickly changes in statistical parameters of sensor readings.

There exist a number of effective machine-learning algorithms. The most widely used are artificial neural networks, clustering algorithms, support vector machines, similarity learning. For a faulty sensor readings detection machine learning approach is successfully used for analysis of several sensor signals at once in order to establish confidence level for each device and thus to identify malfunctioning sensors straight away.

Another application of machine learning algorithms is duplicate elimination. For this task usually clustering and neural networks are exploited. One more technique is sorted neighbourhood method and its modifications (Bertolazzi, De Santis, & Scannapieco, 2003; Yan et al., 2007). All these methods are used in large-scale databases as well (Hernandez & Stolfo, 1995) and in the current use-case can be exploited to get rid of duplicates in event data.

For a knowledge-based approach the application domain can be represented (Batini & Scannapieca, 2006):

- procedurally in form of program code, or implicitly as patterns of activation in a neural network;
- as an explicit and declarative representation, in terms of a knowledge base, consisting of logical formulas or rules expressed in a representation language.

Typically the most general approach to perform data transformations are extensions of standard query language SQL (Rahm & Do, 2000), which allows flexible transformation step definitions, their easy reuse and supports query processing tasks.

Table 4. Measurement data for an exemplary sensor

Sensor ID	Timestamp	Value
TMPS1	2010/08/28 13:21:55	597.2
TMPS1	2010/08/28 13:22:00	598.5
TMPS1	2010/08/28 13:22:05	599.6
TMPS1	2010/08/28 13:22:10	600.3

$$\text{TempSensor} \sqsubseteq \forall \text{hasValue.xsd:int}[\geq 0, \leq 600]$$

Figure 9. Temperature sensor measurements range represented in a model

Additionally, there are several systems developed which improve the quality of data by means of rules extracted from domain knowledge and domain-independent transformations (Batini & Scannapieca, 2006), e.g. the Intelliclean system (Lup, Lee, & Wang, 2001) aimed at efficient duplicate elimination, the Atlas technique (Tejada, Knoblock, & Minton, 2001) which allows to obtain new rules through a learning process and Clue-Based method for record matching (Buechi et al., 2003).

As a proposition for a further work, we propose to combine existing techniques in order to increase productivity and effectiveness of the data cleaning process. The dataset introduced here can serve as a test. As a motivating example, consider measurements of a temperature sensor presented in a Table 4 both in a semantic model and as a statistical value.

In the model-based representation of a sensor data, such as indicated on Figure 9, after processing a measurements presented in a Table 4 the system would detect an outlier at a time 13:22:10.

On the contrary, analyzing data with statistical methods, there would be a trend detected. Thus, having available results both by model-based reasoning and statistical techniques would prevent a false alarm.

Likewise, it is useful to combine multivariate statistical analysis and machine learning algorithms such as clustering and neural networks for establishing a quality of several sensors measurements.

Therefore, that joint approach would help to improve the following weak points in managing low-quality data:

- efficient detection of data deficiency, such as (i) false positive errors and (ii) false negative errors;
- detecting correlations between particular sequences of events and their consequences and between measurements using numerous solutions, such as pattern-matching algorithms, independency tests and others; and

- model-driven correction of a model in case of changes in system structure.

5.2. Analysis and diagnostics

In this subsection there are shortly explained, how the cleaned data is studied and processed further in the current industry case. The main use is continuous diagnosing of the condition of the appliance in order to predict and prevent future faults of the machinery and to react instantly as anomalies or faults in operating are detected. Two main approaches for that are: (i) data-driven and (ii) knowledge-based techniques. Data-driven approaches includes pattern recognition, neural networks, numerical approaches; knowledge-based techniques include case descriptions, faults and correct behavior modeling. The following factors determine the choice of the appropriate diagnosing method in a particular case (ISO 13379-1, 2009):

- application and initial design of the equipment;
- availability of data to be analyzed and its complexity; and
- required qualifications of a resulting computations and models.

Brief summary of above-mentioned diagnosing techniques, presented in (ISO 13379-1, 2009):

Data-driven approach methods classify different functioning states of an appliance: normal, fault one, fault two etc. In order to achieve this, firstly the model is trained with the historical data from each condition and after that launched with the new data, which has to be classified.

The great advantage of data-driven approach is that it does not have need for a thorough knowledge of the system to be diagnosed. The other strong advantage is absence of constraints on the data type of independent variables. As a disadvantage it is worth to mention, that it might be computationally difficult to train a model, as it requires comparatively large amount of prescribed fault and non-fault states to construct a model. In addition, modelling by this approach does not result with an explanatory diagnosis.

The list of the most common data-driven techniques:

- Statistical data analysis, case-based reasoning;
- Neural networks;
- Classification trees;
- Random forests;
- Logistic regression;
- Support vector machines.

Knowledge-based approaches are used to represent knowledge using various knowledge representation techniques and reason over it to infer new knowledge. Their biggest advantage is the possibility of thorough diagnostics. There are two fundamental knowledge-based methods used by engineers:

- Fault/symptom diagnostic approach;
- Causal tree diagnostic approach.

In special situations several approaches may be combined for better results, but still both approaches are not disjoint, i.e. there are methods which might be referred to both types. However, each approach has its advantages as well as drawbacks and an engineer chooses appropriate diagnostic technique based on the type of an appliance, complexity of modeling, availability of necessary data and other factors.

6. Conclusion

For current industry use-case, data is employed to conduct calculations necessary for emergency diagnostics, prognosis of efficiency and further analysis. However, due to imperfect, incomplete or defective information and data schemes these tasks have become rather difficult to realize: wrong, missing or incorrectly formatted data may result in erroneous computations and false decisions, which can be quite disastrous for an industry processes, especially for large-scale industries.

The current paper studies data quality and different approaches to its assessment. We summarized and illustrated the most common defectiveness of a large-scaled industrial database by the example of Siemens Energy Domain and its equipment measurements. We also reviewed existing techniques that are used to overcome errors in data and proposed an approach to address data quality problems. And as shown in the examples, there is no doubt that data requires continuous control and quality improvement, although the design of a convenient technological solution to that challenge is far from trivial.

References

ISO 13379-1, I. D. (2009). *Condition monitoring and diagnostics of machines data interpretation and diagnostics techniques part 1: General guidelines.* ISO, Geneva, Switzerland.

Batini, C., & Scannapieca, M. (2006). *Data quality: concepts, methodologies and techniques.* Springer.

Bergdahl, M., Ehling, M., Elvers, E., Földesi, E., Körner, T., Kron, A., and others (2007). *Handbook on data quality assessment methods and tools*, 9–10.

Buechi, M., Borthwick, A., Winkel, A., & Goldberg, A. (2003) *ClueMaker: A Language for Approximate Record Matching*. IQ, 207-223.

Corporation, A., & Consulting, W. M. (2011). Data quality in the insurance market.

Foken, T., Göockede, M., Mauder, M., Mahrt, L., Amiro, B., & Munger, W. (2005) *Post-field data quality* *control*. Handbook of micrometeorology, Springer, 181-208.

Gendron, M. S., & D’Onofrio, M. J. (2001). Data quality in healthcare industry. *Data Quality*, 7(1), 23–31.

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4), 184–192.

Laudon, K. C. (1986). Data quality and due process in large interorganizational record systems. *Communications of the ACM*, 29(1), 4–11.

Lup Low, W., Li Lee, M., & Wang Ling, T. (2001). A knowledge-based approach for duplicate elimination in data cleaning. *Information Systems*, 26(8), 585–606.

Optique. (2012). *Optique: project description.* Retrieved November, 2012, from CVS: "http://www.optique-project.eu/ about-optique/about-optique/".

Pipino, L. L., Lee, Y.W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211–218.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4), 3–13.

Safran, D. G., Kosinski, M., Tarlov, A. R., Rogers, W. H., Taira, D. A., Lieberman, N., & Ware, J. E. (1998). The primary care assessment survey: tests of data quality and measurement performance. *Medical care*, 36(5), 728–739.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, 40(5), 103–110.

Tejada, S., Knoblock, C. A., & Minton, S. (2001). Learning object identification rules for information integration. *Information Systems*, 26(8), 607–633.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86–95.

Wang, R. Y., Strong, D. M., & Guarascio, L. M. (1996). Beyond accuracy: What data quality means to data consumers. *J. of Management Information Syste*ms, 12(4), 5–33.

Winkler, W. E. (1999) *The state of record linkage and current research problems.* Statistical Research Division, US Census Bureau.

Winkler, W. E. (2004). Methods for evaluating and creating data quality. *Information Systems, Elsevier, 29*, 531-550.

Yan, S., Lee, D., Kan, M.-Y., & Giles, L. C. (2007) Adaptive sorted neighborhood methods for efficient record linkage. *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 185-194