Sensors & Transducers, Vol. 185, Issue 2, February 2015, pp. 140-148

Sensors & Transducers

http://www.sensorsportal.com

Data Quality Indicators Composition and Calculus: Engineering and Information Systems Approaches

[1] Leon REZNIK, [2] Sergey Edward LYSHEVSKI
[1] Department of Computer Science
[2] Department of Electrical and Microelectronic Engineering
Rochester Institute of Technology, 102 Lomb Memorial Drive, Rochester, NY 14623, USA
[1] Tel.: 1585 475 7210, fax: 1585 475 7100
[1] E-mail: lrvcs@rit.edu

Received: 14 November 2014 /Accepted: 15 January 2015 /Published: 28 February 2015

Abstract: Big Data phenomenon is a result of novel technological developments in sensor, computer and communication technologies. Nowadays more and more data are produced by nanoscale photonic, optoelectronic and electronic devices. However, their quality characteristics could be very low. The paper proposes new methods of the data management with huge data amounts that is based on associating of data quality indicators with each data entity. To achieve this goal, one needs to define the composition of the data quality indicators and to develop their integration calculus. As data quality evaluation involves multi-disciplinary research, various metrics have been investigated. The paper describes two major approaches in assigning the data quality indicators and developing their integration calculus. The information systems approach employs traditional high-level metrics like data accuracy, consistency and completeness. The engineering approach utilizes signal characteristics processed with the probability based calculus. The data quality metrics composition and calculus are discussed. The tools developed to automate the metrics selection and calculus procedures are presented. The user-friendly interface examples are provided. *Copyright © 2015 IFSA Publishing, S. L.*

Keywords: Data quality, Quality evaluation, Computer security evaluation, Sensor systems, Nanotechnology.

1. Introduction

The advances in computing, instrumentation and communication technologies over the last decade laid a strong foundation for data generation and storage on a staggering scale. For example, the Large Hadron Collider at CERN can generate 40 terabytes of data every second during experiments. Boeing 737 jet engines' sensors produce 10 terabytes of data for every 30 minutes [1]. The phenomenon of Big Data is in a large degree the result of the current and emerging sensor systems, which are creating ever-increasing amounts of data. Enabling nano-scale instruments, communication and processing equipment results in generating even larger amounts of data. We entered a new era of an exponential growth of data collected and made available for various applications. The existing technologies are not able to handle such big amounts of data. This phenomenon was called the big data. Photonics and nanotechnology enabled microsystems perform multiple generations and fusions of multiple data streams with various data quality [2-7]. The development and application of quantum-mechanical nanoscale electronic, photonic, photoelectronic communication, sensing and processing devices

significantly increase an amount of data which can be measured and stored. These organic, inorganic and hybrid nanosensors operate on a few photons, electrons and photon-electron interactions [2, 3, 5, 7]. Very low current and voltage, high noise, large electromagnetic interference, perturbations, dynamic non-uniformity and other adverse features result in heterogeneous data with high uncertainty and poor quality. The super-large-density quantum and quantum-effect electronic, optoelectronic and photonic nanodevices and waveguides are characterized by:

1) Extremely high device switching frequency and data bandwidth (~1000 THz);

2) Superior channel capacity (~10^{13} bits);

3) Low switching energy (~10^{-17} J) [8, 9].

The importance of DQ analysis, data enhancements and optimization is emphasized due to:

1) High noise-to-signal ratio (ratio of mean to standard deviation of measured signals is ~0.25 in the emerged electrons-photons interaction devices);

2) High probability of errors (p is ~0.001);

3) High distortion measure, reaching ~0.1 to 0.3;

4) Dynamic response and characteristic non-uniformity. These characteristics must be measured, processed and evaluated and provided to a data used along with the data.

New generations of information systems provide communication and networking capabilities to transfer, fuse, process and store data. Various applications require the data delivery from their origin to the point of use that might be far away. The data transfer may lead to information losses, attenuation, distortions, errors, malicious alterations, etc. Security, privacy and safety aspects of data communication and processing systems nowadays play a major role and may have a dramatic effect on the quality of data delivered.

New DQ management methods, quality evaluation and assurance (QE/QA) tools and robust algorithms are needed to ensure security, safety, robustness and effectiveness of various sensor-based engineering and technological systems. As the amount of data available multiplies every year, current information systems are not capable to process these large data arrays to make the best decision. Big data applications require better data selection of high quality inputs. The absence of DQ indicators provided along with the data hinders the recognition of potential calamities and makes data fusion and mining procedures as well as decision making prone to errors.

This paper represents an extended and enhanced version of [10] that was presented at the SecureWare 2014 conference in November 2014. In the paper we offer a novel system approach to the data management that aims at shifting a sensor system target from collecting more and more data regardless of either they are needed or could be used in a particular application to the efficient and effective data collection schemes, where data of a required quality are collected when and delivered to where they are needed. We propose to associate the DQ indicators with each data entity, and replace one-dimensional data processing and delivery with multi-dimensional data processing and delivery along with the corresponding DQ indicators. To realize this approach, we need to develop and describe the structure and content of these DQ indicators, develop the calculus of processing, and, develop interactive tools to automate this process. The current situation in DQ research is presented in Section 2. As DQ evaluation represents a multidisciplinary field, where various DQ indicators have been tried in various applications. However, we believe there are currently exist two major approaches to the DQ evaluation. Engineering approach attempts to evaluate the quality of electrical signals and works on the physical level (see [11] for more details). The approach tends to develop the calculus for the evaluation process. The DQ indicators suitable to be employed for signal quality evaluation are given in Section 3. In information system approach, which is presented in Section 4, while the DQ calculus is reported in Section 5, higher level indicators dealing with time and data based characteristics are employed. The automation tools are documented in Section 6. The conclusions are outlined in Section 7.

2. Current Environment and Achievements in DQ Evaluation

DQ represents an open multidisciplinary research problem, involving advancements in computer science, engineering and information technologies. In all those fields, it is essential to develop technologies and methods to manage, ensure and enhance quality of data. Related research in a networking field attempted to investigate how the network characteristics, standards and protocols can affect the quality of data collected and communicated through networks. In sensor networks, researchers started to investigate how to incorporate DQ characteristics into sensor-originated data [12]. Guha, *et al.* proposed a single-pass algorithm for high-quality clustering of streaming data and provided the corresponding empirical evidence [13]. Bertino, *et al.* investigated approaches to assure data trustworthiness in sensor networks based on the game theory [14] and provenance [15]. Chobsri, *et al.* examined the transport capacity of a dense wireless sensor network and the compressibility of data [16]. Dong and Yinfeng attempted to optimize the quality of collected data in relation to resource consumption [17-18].

Current developments are based on fusing multiple data sources with various quality and creating big data collections. Novel solutions and technologies, such as nano-engineering and technology are emerged in order to enable DQ assessment. Reznik and Lyshevski outlined

integration of various DQ indicators representing different schemes ranging from measurement accuracy to security and safety [19], as well as micro- and nano-engineering [20]. The aforementioned concepts are verified, demonstrated and evaluated in various engineering and science applications [21-22].

3. DQ Indicators Composition: Engineering Approach

3.1. Preliminary Definitions and Formulas

This method utilizes information theory measures and employs the probability based metrics and calculus techniques. It assumes deriving quantitative estimates and capability measures with the goal to evaluate processing performance and quality. The processing schemes are evaluated by analyzing the Shannon and quantum-mechanical performance limits in classical and quantum domains. For a random variable X_i with n outcomes $\{x_i: i=1, 2, \ldots, n-1, n\}$ and pdf $p(x_i)$, the entropy is calculated as [23]

$$H(X) = -\sum_{x\in X} p(x)\log_2 p(x) \tag{1}$$

The mutual information between the input X and output Y quantitatively defines:

1) The amount of information received on average;

2) The dependence of X and Y. The classical and quantum mutual information $I(X;Y)$ and $I(X;Y)$ depend on the classical and quantum entropies $H(X)$ and $H(X)$. These quantitative information amounts are estimated as I and I [3, 4, 11, 24-25]

$$I(X;Y)=I(Y;X)=H(X)–H(X|Y)=H(Y)–H(Y|X)= H(X)+H(Y)–H(X,Y),$$

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p_{X,Y}(x,y)\log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \text{ [bits/symb]},$$

$$I(X;Y)=H(X) – H(X|Y), \tag{2}$$

where $p_{X,Y}(x,y)$ is the joint pdf of X and Y; $p_X(x)$ and $p_Y(y)$ are the marginal probability density functions of X and Y.

Example 3.1. If X and Y are independent, $p(x,y)=p(x)p(y)$. Hence $\ln\frac{p(x,y)}{p(x)p(y)} = \ln 1 = 0$, and $I(X;Y)=0$.

The positively defined mutual information $I(X;Y)\geq 0$ determines the average amount of information received per symbol transmitted or processed.

The conditional entropy $H(X|Y=y)$ of a random variable X, that is conditional on a particular realization y of Y, defines the expected conditional information content with respect to both X and Y. We have

$$H(X|Y) = -\sum_{x\in X}\sum_{y\in Y} p_{X,Y}(x,y)\log_2 p_{X|Y}(x|y), \quad H(X|Y)\geq 0 \tag{3}$$

The conditional entropy $H(X|Y)$ corresponds to the average loss of information $L_I=H(X|Y)$. Here, $H(X)\geq H(X|Y)$.

If X and Y are independent, $H(X|Y)=H(X)$ and $I(X;Y)=0$.

The joint entropy is the entropy $H(X,Y)$ is

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y), \quad H(X,Y) = -\sum_{x,y} p_{X,Y}(x,y)\log_2 p_{X,Y}(x,y) \tag{4}$$

Analog and digital deterministic computing of real-valued physical variables guarantee exceptional performance. Quantum communication and processing on a few photons and electrons result in significant uncertainties, distortions and errors [8]. An inherent quantum determinism ensures quantum-deterministic communication and processing on the *utilizable* initial (I) and final (F) *state* transductions $S=[S_I,S_F]^T$, $S_I:\boldsymbol{\nu}_I\rightarrow S_F:\boldsymbol{\nu}_F$ performed on real-valued, directly *detectable*, *measurable* and *processable* physical variables $\boldsymbol{\nu}$ [5, 23-24]. The measured X and Y result in $p(x)$, $H(X)$, $H(X;Y)$, $I(X;Y)$ and other measures and estimates. This concept is substantiated by *natural* systems, as well as by commercialized quantum-effect optoelectronic and photonic devices [6, 8]. A quantum-effect processing primitive P_j exhibits transductions $S_j(\boldsymbol{\nu})$ on *detectable*, *measurable* and *processable* variables $\boldsymbol{\nu}_j$ yielding *distinguishable* and *computable* transforms $T_j(S,\boldsymbol{\nu})$. These quantum transductions $S_j(\boldsymbol{\nu})$ result in processing tasks [5, 23-24].

3.2. Communication Channel Capacity

For conventional and quantum-deterministic communication, the channel capacity of a stationary memoryless channel with finite input and output alphabets is

$$C = \max_{p_X(\cdot)} I(X;Y) \text{ and } C = \max_{p_X(\cdot)} I(X;Y) \text{ [bits/symbol]} \tag{5}$$

If the transition probabilities vary, for nonstationary memoryless channels

$$C = \max_{p_{X_1}(\cdot),\ldots,p_{X_n}(\cdot)} \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i) = \frac{1}{n}\sum_{i=1}^{n} C_i.$$

3.3. Data Fusion and Data Exchange Rate

These are the DQ related characteristics, which demonstrate the achievable performance of the

processing platforms that could be reached despite errors, distortions, non-uniformity, inconsistencies, sensitivity and uncertainties. Using the probability of a bit error p_b, the maximum data fusion rate is

$$r_{\max}(p_b) = \frac{C}{1 - H_2(p_b)}, \quad H_2(p_b) = -[p_b \log_2 p_b + (1-p_b)\log_2(1-p_b)], \tag{6}$$

where H_2 is the binary entropy function.

3.4. Distortion Measure

Consider a sequence $X_1, \ldots, X_n$ with $p(x)$ and a finite alphabet A, $x \in A$. Using the reproduced alphabet A_r with symbols $x_r \in A_r$, the finite distortion measure d: $A \times A_r \to R$ is

$$d_{\max} = \max_{x \in A, x_r \in A_r} d(x, x_r) < \infty \tag{7}$$

The distortion depends on the sequences, encoding and decoding functions, etc. The rate distortion function for a source X with $d(x,x_r)$ can be defined as

$$r_I(D) = \min I(X;X_r), \; r_I(D) = \min I(X;X_r), \tag{8}$$

$I(X;X_r) = H(X) - H(X \mid X_r) \geq H(p) - H(D)$, $r_I(D) \geq H(p) - H(D)$.

The quantity $\min I(X;X_r)$ is found with respect to all condition distributions $p(x_r \mid x)$ for which the joint distribution $p(x,x_r) = p(x)p(x_r \mid x)$ satisfies the imposed distortion constraints.

3.5. Data Processing Capability

The data processing capability is estimated as

$$D = BI(X;Y) r^{-1}(p_b) r_I^{-1}(D), \; D = RI(X;Y) r^{-1}(p_b) r_I^{-1}(D), \tag{9}$$

where B is the bandwidth; R is the quantum transduction rate in the *microscopic* processing system.

3.6. Data Processing Complexity

The entropies H and H define the data set complexity. A finite length of the string $x \in \{0, 1\}$ is denoted as $l(x)$. The Kolmogorov descriptive complexity

$$K_U(x) = \min_{p U(p) = x} l(p), \tag{10}$$

provides the minimal description length l of a string x with respect to a universal processor U within a processing realization p. Here, U is the computable function of arguments x and p.

Binary strings are the words in the alphabet $A = \{0,1\}$. For any computable function U: $A \to A$. The complexity of $x \in A$ is defined with respect to U. For any processor P

$$K_U(x) \geq c K_P(x), \; \forall x, \; c > 0, \tag{11}$$

where c is the constant which depends on U and P. Using $K_U(\cdot)$, we define the mutual complexity as

$$I_K(X;Y) = K_U(Y) - H(Y|X, K_U(X)) \tag{12}$$

The data processing complexity estimates are given as

$$L = H(X) K_U(x) I_K(X;Y), \; L = H(X) K_U(x) I_K(X;Y) \tag{13}$$

3.7. Data Quality

A Markov information source is a pair (M, f) of stationary Markov chain M and function f of reachable states s_k, $f(s_k)$: $S \to A$. The transductions $S_j(\mathbf{v})$ are on *detectable*, *measurable* and *processable* variables $\mathbf{v}_j$. The mapping $f(s_k)$ maps states S into the Markov chain to entities in the alphabet A. To estimate the data quality of information sources, we use $I \geq 0$ or $I \geq 0$. The sequence of length n has a complexity $\sim O(n)$. The probability of an input p is $\sim 2^{-l(p)}$. The universal probability of a binary string x is

$$P_U(x) = \sum_{p U(p) = x} 2^{-l(p)} = \Pr(U(p) = x), \quad P_U(x) \approx 2^{-K(x)} \tag{14}$$

Define the data quality measures as

$$D_q = I(X;Y) P_U(x), \; D_q = I(X;Y) P_U(x) \tag{15}$$

The mutual information does not depend entirely on the input – response mapping. The $I(X;Y)$ is predefined by the input probabilities. Hence, some measures and estimated may not be fully characterized and evaluated.

3.8. Data Reliance

Using the real-valued deterministic $\phi_d(\cdot)$ and probabilistic $\phi_p(\cdot)$ characteristics, given as a known set of functions in the defined function space, the data reliance measure is defined by using an operator L as

$$L: F(D_c \phi_d(\cdot) \phi_p[p_i(x), p_j(x,y)]) \to D_r, \quad L: F(D_c \phi_d(\cdot) \phi_p[p_i(x), p_j(x,y)]) \to D_r \tag{16}$$

The deterministic parametric characteristics $\phi_d(\cdot)$, such as accuracy, linearity, noise, error, signal-to-noise ratio and others, are available. In addition, the

pdfs of faults, failures, defects, characteristic variations, sensitivity, noise, errors and other quantities may be known and characterized by $\phi_b(\cdot)$. For example, the normal $N_i(\mu_i,\sigma_i^2)$ and extreme value $V_i(\mu_i,\sigma_i,k_i)$ pdfs are found. The n-dimensional deterministic and statistic analyses can be performed using factor, principal component, classification and other models. The data reliance D_r degrades with the decrease of device size which leads to the parameter variations, increase of noise, etc.

Example 3.2. For the nanoscaled optoelectronic devices, we use the descriptive pdfs of parameter variations $N_v(\mu_v,\Sigma_v)$, noise $N_n(\mu_n,\Sigma_n)$ and reliability $V_r(\mu_r,\sigma_r,k_r)$. The analysis can be accomplished. The measured signal is $X=S+N$, where S and N are the not-perturbed signal and noise. Thus, $p_{S,N}(s,n)=p_S(s)p_N(n)$ with $p_{N|S}(n|s)=p_N(n)$, $p_{X|S}(x|s)=p_N(n-s)$. For $N\sim L(\alpha)$, $S\sim L(\beta)$, the parameters α and β are estimated,

$$p_X(x)=\frac{\alpha\beta}{2(\alpha^2-\beta^2)}\left(\alpha e^{-\beta|x|}-\beta e^{-\alpha|x|}\right)$$

$$p_{S|X}(s|x)=\tfrac{1}{2}\left(\alpha^2-\beta^2\right)\frac{e^{-\alpha|x-s|-\beta|x|}}{\alpha e^{-\beta|x|}-\beta e^{-\alpha|x|}}.$$

4. DQ Metrics Composition in Information Systems

Data may have various quality aspects, which can be measured. These aspects are also known as data quality dimensions, or metrics. Traditional dimensions are as follows, some of them are described in [26-27]:

- **Completeness:** Data are complete if they have no missing values. It describes the amount, at which every expected characteristic or trait is described and provided.
- **Timeliness**: Timeliness describes the attribute that data are available at the exact instance of its request. If a user requests for data and is required to wait a certain amount of time, it is known as a data lag. This delay affects the timeliness and is not desirable.
- **Validity**: It determines the degree, at which the data conforms to a desired standard or rules.
- **Consistency**: Data are consistent if they are free from any contradiction. If the data conforms to a standard or a rule, it should continue to do so if reproduced in a different setting.
- **Integrity**: Integrity measures how valid, complete and consistent the data are. Data's integrity is determined by a measure of the whole set of other data quality aspects / dimensions.
- **Accuracy**: Accuracy relates to the correctness of data and measurement uncertainty. Data with low uncertainty are correct.
- **Relevance**: It is a measure of the usefulness of the data to a particular application.
- **Reliability**: The quality of data becomes irrelevant if the data are not obtained from a reliable source. Reliability is a measure of the extent, to which one is willing to trust the data.
- **Accessibility**: It measures the timeliness of data.
- **Value added**: It is measured as the rate of usefulness of the data.

The methodologies of evaluating the DQ aspects listed above have been developed over the decades. They well represent the quality of the data at the point of their origin at the data source. However, nowadays most of the data are used far away from the point of their origin. In fact, the structured data are typically collected by distributed sensor networks and systems, then transmitted over the computer and communication networks, processed and stored by information systems, and, then, used. All those communication, processing and storage tasks affect the quality of data at the point of use, changing their DQ in comparison to one at the point of origin. The DQ evaluation should integrate accuracy and reliability of the data source with the security of the computer and communication systems. The high quality of the data at the point of their origin does not guarantee even an acceptable DQ at the point of use if the communication network security is low and the malicious alternation or loss of data has a high probability.

We describe the DQ evaluation structure as a multilevel hierarchical system. In this approach, we combine diverse evaluation systems, even if they vary in their design and implementation. The hierarchical system should be able to produce a partial evaluation of different aspects that will be helpful in flagging the areas that need urgent improvement. In our initial design we will classify metrics into five groups (see Fig. 1):

1. Accuracy evaluation;
2. Measurement and reliability evaluation;
3. Security evaluation;
4. Application functionality evaluation;
5. Environmental impact.

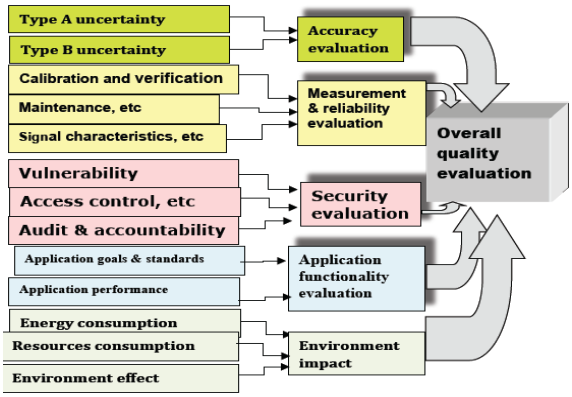

Fig. 1. Integral quality evaluation composition.

While the first three groups include rather generic metrics, groups #4 and #5 are devoted to metrics, which are specific to a particular application. Our

metrics evaluation is based on existing approaches and standards, such as [28] for measurement accuracy and [29] for system security. Table 1 gives a sample of generic metrics representing all first three groups, while Table 2 lists the metrics, which are considered specific to a particular sensor and an application.

Table 1. Samples of Generic Metrics.

Generic Attribute Name	DQ indicator/group (Fig.1)	Description
Time-since-Manufacturing	Maintenance/reliability	The measure of the age of the device
Time-since-Service	Maintenance/reliability	The measure of the days since last service was performed in accord with the servicing schedule
Time-since-Calibration	Calibration/reliability	The measure of the days since last calibration was performed in accord with the calibration schedule
Temperature Range	Application/performance	The measure of temperature range within which the device will provide optimum performance
Physical Tampering Incidences	Physical security/security	The number of reported incidents that allowed unauthorized physical contact with the device
System Breaches	Access control/security	The measure of the number of unauthorized accesses into the system, denial of service attacks, improper usage, suspicious investigations, incidences of malicious code
System Security	Security/security	Measures presence of intrusion detection systems, firewalls, anti-viruses
Data Integrity	Vulnerabilities/securities	Number of operating system vulnerabilities that were detected
Environmental Influences	Environment/environment	Number of incidences reported that would subject the device to mechanical, acoustical and triboelectric effects
Atmospheric Influences	Environment/environment	Number of incidences reported that would subject the device to magnetic, capacitive and radio frequencies
Response Time	Signals/reliability	Time between the change of the state and time taken to record the change

Table 2. Samples of Specific DQ Metrics (examples of electric power and water meters).

Device Name	Application specific Quality indicator	Description
Electric / Power Meters	Foucalt Disk	Check to verify the material of the foucalt disk
	Friction Compensation	Difference in the measure of initial friction at the time of application of the compensation and the current friction in the device
	Exposure to Vibrations	Measure of the number of incidences reported which would have caused the device to be subjected to external vibrations
Water Meters	Mounting Position	The measure of the number of days since regulatory check was performed to observe the mounting position of the device
	Environmental Factors	Number of incidences reported which may have affected the mounting position of the device
	Particle Collection	Measure of the amount of particle deposition

5. DQ Metrics Calculus

In DQ calculus implementation we investigate a wide number of options of calculating integral indicators from separate metrics ranging from simple weighted sums to sophisticated logical functions and systems. Those metrics and their calculation procedures will compose the DQ calculus. To simplify the calculus, we organize it as a hierarchical system calculating first the group indicators and then combining them into the system total. We follow the user-centric approach by offering an application user a choice of various options and their adjustment. We plan to introduce a function choice automatic adjustment, verification and optimization.

To realize a wide variety of logical functions, the expert system technology is employed as the main implementation technique. The automated tool set includes the hierarchical rule-based systems deriving values for separate metrics, then combining them into groups and finally producing an overall evaluation. This way, the tool operation follows up the metrics structure and composition (see Fig. 1). This system needs to be complemented by the tools and databases

assisting automation of all stages in the data collection, communication, processing and storage for all information available for data quality evaluation. The developed tools facilitate automated collection, communication and processing of the relevant data. Based on the data collected, they not only evaluate the overall data quality but also determine whether or not the data collection practice in place is acceptable and cite areas that are in need of improvement.

In our automated procedures, the DQ score is computed by applying either linear, exponential or stepwise linear reduction series to the maximum score of an attribute. In case an attribute defines a range for ideal working, the linear series is substituted by a trapezoidal drop linear series and exponential is replaced by a bell drop series.

When considering both accuracy and security DQ metrics, assessing whether fusion enhances DQ is not obvious as one has to tradeoff between accuracy, security and other goals. While adding up a more secure data transmission channel improves both security and accuracy indicators, using a more accurate data stream will definitely improve data accuracy but could be detrimental to certain security indicators (see [30] for further discussion). If resources are limited, as in the case of sensor networks, one might consider trying to improve accuracy of the most secure data source versus more or less even distribution of security resources in order to achieve the same security levels on all data channels. The concrete recommendations will depend on the application.

6. Generic Tool Design

The proposed design of the tool divides the procedure for automated data collection into three main stages. First stage involves mainly a device configuration. Since the tool is generic, it provides certain flexibility in configuring a large variety of diverse devices. These devices could be electric meters, power meters, water meters and marine sensors. The second stage computes data quality indicators of the configured device. The final stage performs the detailed analysis of the computed data quality indicators. It highlights low data quality and help flag erroneous data. Also, it provides recommendations on improving low data quality and helps ensure that the data being utilized are fit for the purpose it is intended to be used. Fig. 2 presents the architecture of the tool. Currently, the first and the second stages are implemented.

The generic tool allows for a configuration of a large variety of devices. Each automated data collection device has DQ factors, which are common to other similar devices. These factors are referred by the tool as generic attributes. Other attributes, which are unique to a particular device are called dynamic attributes. These attributes are assigned the maximum score based on the significance of the contribution they would add to the data quality. The greater the significance, the greater is the score.

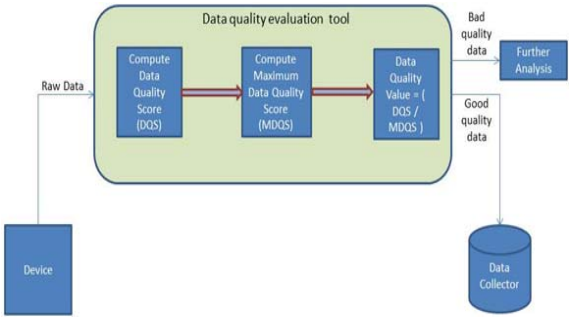

Fig. 2. Data quality evaluation procedure.

The configuration step mainly involves recognizing the generic and application-specific attributes, as well as assigning the max possible score to each of them. Generic attributes are common to most devices, for example, timeliness and quality of common device servicing such as calibration. Application-specific attributes are unique to a device, for example, exposure to vibration, shock and radiation. This is important for a particular application because certain devices, like electric meters, produce misleading results when exposed to the external adversary affects. If, for some reason, a generic attribute does not apply to a particular device, the max score of zero would be applied in order to eliminate the attribute from the analysis. Table 1 describes the generic attributes being considered by the tool. Fig. 3 illustrates configuring some of the generic attributes for an electric meter. Table 2 describes some application specific attributes, which are device and application specific. Fig. 4 illustrates configuring an application-specific attribute for an electric meter, provided as an example.

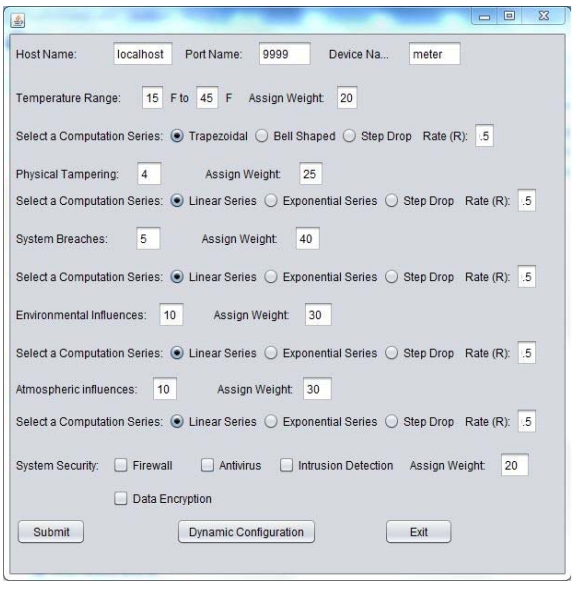

Fig. 3. Generic attribute configuration.

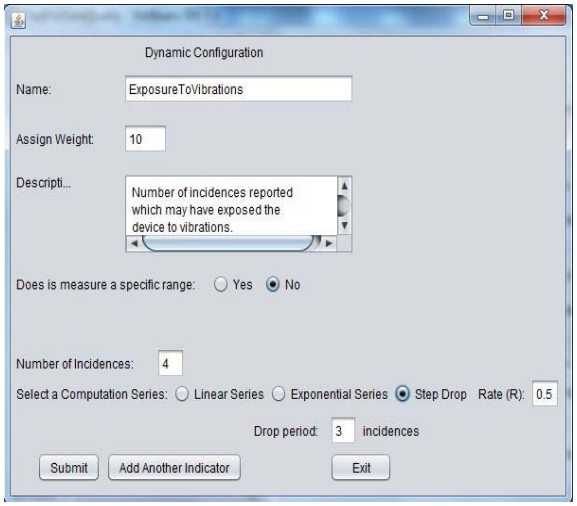

Fig. 4. Application specific attribute configuration.

The second stage involves data quality computation. The configured generic and application specific attributes help compute the individual quality scores. Each attribute is considered a quality indicator, whose significance will be dependent on its max score. These quality indicators produce a quality score using a chosen logic procedure. For example, we can consider a generic attribute called time-since-calibration. Some devices need to get calibrated every year. If a device has not been calibrated for an entire year or a couple of years, the quality factor for that indicator will go down. If the device has never been calibrated since its installation it can affect the quality score even more. The tool allows a user to define the procedure for calculating the application-specific quality indicators.

7. Conclusions

The paper introduces a novel approach to data management in data collection and processing systems, which might incorporate SCADA, sensor networks and other systems with nanoscale devices. According to this methodology, we associate each data entity with the corresponding DQ indicator. This indicator integrates various data characteristics ranging from accuracy to security, privacy and safety, etc. It considers various samples of DQ metrics representing communication and computing security as well as data accuracy and other characteristics. Incorporating security and privacy measures into the DQ calculus is especially important in the current development as it allows shifting the DQ assessment from the point of data origin to the point of data use.

In order to achieve this goal, one needs to develop the DQ indicators structure and composition as well as their integration calculus that would allow calculation of the integral quality indicators. Over the last decade, a variety of indicators and methods have been investigated. The paper examines two major approaches: an engineering and information systems. The engineering approach employs signal level characteristics and develops calculus based on their probability measures. The information system approach utilizes higher level metrics that could be calculated over some time and samples, such as the data accuracy, data consistency, etc.

A unified framework for assessing DQ is critical for enhancing data usage in a wide spectrum of applications because this creates new opportunities for optimizing data structures, data processing and fusion based on the new DQ information use. By providing to an end user or an application the DQ indicators which characterize system and network security, data trustworthiness and confidence, etc. Correspondingly, an end user is in a much better position to decide whether and how to use data in various applications. A user will get an opportunity to understand and compare various data files, streams and sources based on the associated DQ with integral quality characteristics reflecting various aspects of system functionality and to redesign data flows schemes. This development will transform one-dimensional data processing into multi-dimensional data optimization procedures for application-specific data applications. We describe and demonstrate an application of the DQ metrics definition and calculation tools, which enable integration of various metrics to calculate an integral indicator.

References

[1]. S. Rogers, Big Data is Scaling BI and Analytics, *Information Management*, Vol. 21, September 2011, pp. 14-18.

[2]. P. W. Coteus, J. U. Knickerbocker, C. H. Lam, Y. A. Vlasov, Technologies for exascale systems, *IBM Journal of Research and Developments*, Vol. 55, Issue 5, 2011, pp. 14.1-14.12.

[3]. Handbook on Nano and Molecular Electronics, Ed. S. E. Lyshevski, *CRC Press*, Boca Raton, FL, 2007, pp. 6-1-6.102.

[4]. B. G. Lee, *et. al*., Monolithic silicon integration of scaled photonic switch fabrics, CMOS logic, and device driver circuits, *Journal of Lightwave Technology*, Vol. 32, Issue 4, 2014, pp. 743-751.

[5]. S. E. Lyshevski, Molecular Electronics, Circuits and Processing Platforms, *CRC Press*, Boca Raton, FL, 2007.

[6]. Micro-Electromechanical Systems (MEMS), International Technology Roadmap for Semiconductors, 2011 and 2013 Editions, available at http://www.itrs.net, accessed on August 1, 2014.

[7]. A. Yariv, Quantum Electronics, *John Wiley and Sons*, New York, 1988.

[8]. Emerging Research Devices, International Technology Roadmap for Semiconductors, 2011 and 2013 Editions, available at http://www.itrs.net, accessed on August 1, 2014.

[9]. J. Warnock, Circuit and PD challenges at the 14^{nm} technology node, in *Proceedings of the ACM Int. Symposium on Physical Design*, 2013, pp. 66-67.

[10]. L. Reznik, S. Lyshevski, Data Quality and Security Evaluation Tool for Nanoscale Sensors, in

Proceedings of the 8th International Conference on Emerging Security Information, Systems and Technologies (SECURWARE'14), Lisbon, 16-21 November 2014, pp. 118-122.

[11]. Lyshevski S. E., Reznik L., Smith T. C., Beisenbi M. A., Jarasovna J. Y., Mukataev N. S., Omarov A. N., Estimates and measures of data communication and processing in nanoscaled classical and quantum physical systems, in *Proceedings of the IEEE 14th International Conference on Nanotechnology (IEEE-NANO)*, Toronto, 18-21 August 2014, pp. 1044–1047.

[12]. M. Klein, W. Lehner, Representing Data Quality in Sensor Data Streaming Environments, *Data and Information Quality,* Vol. 1, 2009, pp. 1-28.

[13]. S. Guha, A. Meyerson, N. Mishra, R. Motwani, L. O'Callaghan, Clustering Data Streams: Theory and Practice, *IEEE Trans. on Knowl. and Data Eng.,* Vol. 15, 2003, pp. 515-528.

[14]. H. S. Lim, K. M. Ghinita, E. Bertino, A Game-Theoretic Approach for High-Assurance of Data Trustworthiness in Sensor Networks, in *Proceedings of the IEEE 28th International Conference on Data Engineering (ICDE'12)*, Washington, DC, USA, 2012.

[15]. C. Dai, H. S. Lim, E. Bertino, Provenance-based Trustworthiness Assessment in Sensor Networks, in *Proceedings of the 7th Workshop on Data Management for Sensor Networks (DMSN)*, *in conjunction with VLDB, DMSN 2010,* Singapore, 2010.

[16]. S. Chobsri, W. Sumalai, W. Usaha, Quality assurance for data acquisition in error prone WSNs, in *Proceedings of the 1st International Conference on Ubiquitous and Future Networks (ICUFN'09)*, 2009, pp. 28-33.

[17]. W. Dong, H. Ahmadi, T. Abdelzaher, H. Chenji, R. Stoleru, C. C. Aggarwal, Optimizing quality-of-information in cost-sensitive sensor data fusion, in *Proceedings of the International Conference on Distributed Computing in Sensor Systems (DCOSS'11)*, Piscataway, NJ, USA, 27-29 June 2011, pp. 1-8.

[18]. W. Yinfeng, W. Cho-Li, C. Jian-Nong, A. Chan, Optimizing Data Acquisition by Sensor-channel Co-allocation in Wireless Sensor Networks, in *Proceedings of the International Conference on High Performance Computing (HiPC'10)*, 19-22 December 2010, Piscataway, NJ, USA, 2010, pp. 1-10.

[19]. L. Reznik, Integral Instrumentation Data Quality Evaluation: the Way to Enhance Safety, Security, and Environment Impact, in *Proceedings of the IEEE International Instrumentation and Measurement Technology Conference*, Graz, Austria, 13-16 May 2012, pp. 2138 - 2143.

[20]. S. E. Lyshevski, L. Reznik, Processing of extremely-large-data and high-performance computing, in *Proceedings of the International Conference on High Performance Computing*, Kyiv, Ukraine, 2012, pp. 41-44.

[21]. G. P. Timms, P. A. J. de Souza, L. Reznik, D. V. Smith, Automated Data Quality Assessment of Marine Sensors, *Sensors,* Vol. 11, 2011, pp. 9589-9602.

[22]. G. P. Timms, P. A. de Souza, L. Reznik, Automated assessment of data quality in marine sensor networks, in *Proceedings of the IEEE International Conference* OCEANS'10, Sydney, 2010, pp. 1-5.

[23]. S. E. Lyshevski, Molecular and Biomolecular Processing: Solutions, Directions and Prospects, Handbook of Nanoscience, Engineering and Technology (Eds. W. Goddard, D. Brenner, S. E. Lyshevski and G. Iafrate), *CRC Press*, Boca Raton, FL, 2012, pp. 125-177.

[24]. S. E. Lyshevski, L. Reznik, Information-theoretic estimates of communication and processing in nanoscale and quantum optoelectronic systems, in *Proceedings of the IEEE Conf. Electronics and Nanotechnologies*, 2013, pp. 33-37.

[25]. S. E. Lyshevski, High-performance computing and quantum processing, in *Proceedings of the IEEE Conf. High-Performance Computing*, Kiev, Ukraine, 2012, pp. 33-40.

[26]. F. G. Alizamini, M. M. Pedram, M. Alishahi, K. Badie, Data quality improvement using fuzzy association rules, in *Proceedings of the International Conference on Electronics and Information Engineering (ICEIE)*, 2010, pp. V1-468-V1-472.

[27]. L. Sebastian-Coleman, Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework, *Morgan-Kaufmann Publishers Inc.,* Waltham, MA, 2013.

[28]. ANSI/NCSL, US Guide to the Expression of Uncertainty in Measurement, ed., Z540-2-1997.

[29]. National Institute of Standards and Technology, Performance Measurement Guide for Information Security, ed. Geithersburg, MD, July 2008.

[30]. L. Reznik, E. Bertino, Data quality evaluation: integrating security and accuracy, in *Proceedings of the ACM SIGSAC Conference on Computer Communications Security*, Berlin, Germany, 2013, Poster.
