DQ METRICS: A NOVEL APPROACH TO QUANTIFY TIMELINESS AND ITS APPLICATION IN CRM

(Completed Paper)

Topic Category: IQ Concepts; IQ Assessment

Bernd Heinrich

Marcus Kaiser

Mathias Klier

Abstract: The importance of customer relationship management (CRM) has increased during the last decade. Thus information systems and particularly the quality of customer data are highly relevant for conducting CRM efficiently. Due to the importance of using up-to-date data, this paper analyzes how the data quality dimension timeliness can be quantified. Based on several requirements (e.g. normalization and interpretability) as well as a literature review it proposes a procedure for developing metrics for timeliness that can be for instance adapted to the specific characteristics of the customer data attribute considered. A case study demonstrates the application of our metric by presenting a business case for a CRM campaign of a major mobile services provider.

Key Words: Data Quality (DQ), Metric, Timeliness, Customer Relationship Management (CRM)

INTRODUCTION

In recent years, firms have increasingly focused on their customers, leading to an enhanced importance of CRM. In order to conduct CRM in an efficient way, information systems and especially the quality of customer data, transaction data, and contract data play a decisive role. For instance, only quality assured data enables firms to segment and address their customers. Therefore, data quality (DQ) is a key factor of success especially within CRM campaigns (cf. [15]; [23]; [26]). However, surveys reveal that most of the data within a customer data base are inconsistent and frequently outdated (cf. [4]; [8]). Hence, the following problems are imminent:

- Customers are wrongly selected or wrongly not selected for a campaign due to outdated data. E.g. a customer is addressed in a student-specific campaign although he is no longer studying.
- Customers that have been selected for a mailing campaign (whether correctly or not) can not be contacted for instance due to outdated address data leading to poor outcomes of campaigns.
- Limited capability to individualise an offer due to wrong transaction and contract data.

All of these problems decrease the probability of successful sales and may significantly reduce the success of CRM campaigns. However, the problems are frequently ignored or insufficiently addressed in business case considerations. As a result, the business cases are valuated too optimistically. Furthermore, firms seldom take DQ measures or the existing DQ level into consideration when planning a campaign. This is often due to the limited capability to quantify DQ, particularly the timeliness aspect. Therefore, procedures and metrics for DQ are needed to predict the impact of low DQ for instance on planning and economically valuating CRM campaigns. This article addresses this shortcoming by presenting a procedure for developing metrics for the DQ dimension timeliness.

Taking the design guidelines defined by Hevner et al. [17] into account, we consider the procedure as our artifact and organize the paper as follows: After briefly illustrating the relevance of the problem in the introduction, the next section defines requirements that guide the process of searching for an adequate procedure when developing a metric for the DQ dimension timeliness. On this basis, existing approaches are analyzed with regard to the requirements in the third section. This analysis reveals the contribution of our research. The fourth section designs an innovative procedure for developing a metric for timeliness based on probabilistic considerations and includes an example. In order to study the proposed procedure in depth, a case study in the fifth section illustrates how it was applied within the campaign management of a German mobile services provider (MSP). The last section summarises our findings from a management view and critically reflects on the results.

REQUIREMENTS ON DATA QUALITY METRICS

Metrics are required to quantify DQ in order to support economic management of DQ. Figure 1 depicts the closed loop of an economic management of DQ (which is e.g. of importance when planning CRM campaigns). It illustrates, that the benefits of DQ can be influenced by DQ measures, e.g. data cleansing measures, buying external address data etc. Taking measures improves the current level of DQ, which is quantified by means of metrics. A higher level of DQ leads to a corresponding economic benefit (e.g. enabling a more effective customer contact). Moreover, based on the level of DQ and considering benchmarks and thresholds, firms can decide on taking further measures or not. From an economic view, only those measures must be taken that are efficient with regard to costs and benefit (cf. [1]; [5]; [13]; [21]; [27]). E.g., given two mutually exclusive measures having equal economic benefit, it is rational to choose the one with lower costs.

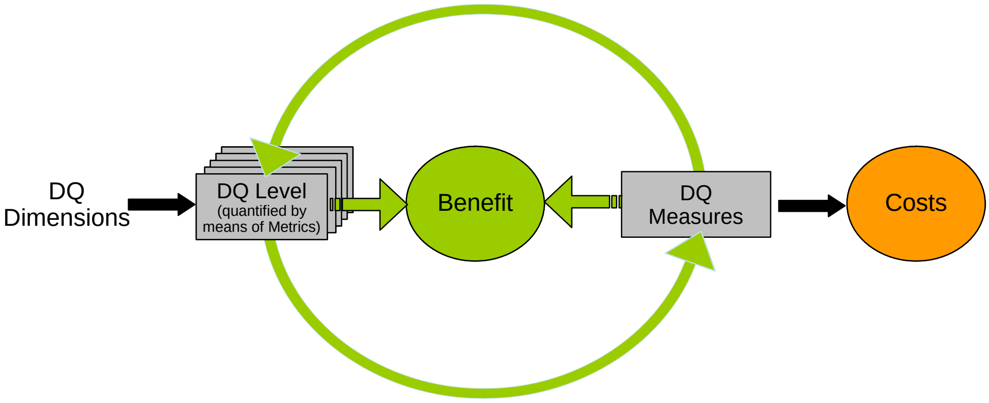

Figure 1 Data quality loop

Based on such considerations, this paper aims at quantifying the quality of datasets by designing a procedure for developing a metric for the dimension timeliness. The identification and classification of DQ dimensions has been addressed from both a scientific and a practical point of view in many publications (cf. [3]; [8]; [9]; [19]; [20]; [25]). In the following we focus on the dimension timeliness as it has been widely neglected in scientific literature (cf. next section for more detail). In addition, the main problem of CRM is usually not data being incomplete. Instead, it is more important to keep huge sets of customer data, transaction data and contract data up-to-date. Therefore it is important to allow for automatic quantification of the DQ dimension timeliness wherever possible. This leads to reduced measurement costs especially when dealing with huge sets of data.

Most DQ metrics are designed on an ad hoc basis to solve specific, practical problems [24] and thus are often highly subjective [6]. In order to enable a scientific foundation and a design evaluation of the metrics, we state the following general requirements, which are not restricted to the dimension timeliness [16]:

First, we refine the *representation consistency* by Even and Shankaranarayanan [10] to requirements R 1 to R 3:

R 1. [*Normalization*] An adequate normalization is necessary to assure comparability of metrics (e.g., when comparing different levels of DQ over time, cf. [24]). DQ metrics are often ratios with a value between 0 (perfectly bad) and 1 (perfectly good) (cf. [10]; [24]).

R 2. [*Interval scale*] To support both the monitoring of DQ level over time and the economic evaluation of measures, we require the metrics to be interval scaled. This means, the difference between two levels of DQ must be meaningful. Thus a difference of 0.2 between the values 0.7 and 0.9 and the values 0.4 and 0.6 of the metric for correctness implies that the quantity of correct data changes to the same extent in both cases.

R 3. [*Interpretability*] Even and Shankaranarayanan demand the measurement being "easy to interpret by business users" [10]. For this reason, the values of the DQ metrics have to be comprehensible. E.g., considering a metric for timeliness, it could be interpretable as the probability of a given attribute value being up-to-date.

R 4 integrates the consistency principles *interpretation consistency* and *aggregation consistency* stated by [10].

R 4. [*Aggregation*] In case of a relational data model, the metrics shall enable a flexible application. Therefore, it must be possible to quantify DQ on the level of attribute values, tupels, relations and the whole data base in a way that the values have consistent semantic interpretation (*interpretation consistency*) on each level. In addition, the metrics must allow aggregating the quantified values on a given level to the next higher level (*aggregation consistency*). E.g., the quantification of the correctness of a relation should be computed based on the values of the correctness of those tupels being part of the relation. Moreover, the resulting values must have identical meaning as the DQ measurement on the level of tupels.

Even and Shankaranarayanan demand *impartial-contextual consistency* of the metrics ([10]). This refers to our requirement on the metrics being adaptive and thereby enabling a contextual perception of DQ in R 5.

R 5. [*Adaptivity*] For targeted quantifying of DQ the metrics need to be adaptable to the context of a particular application. If the metrics are not adapted, they should fold back to the non-adapted (impartial) measurement.

In addition, we state one more property that refers to the measurement procedure.

R 6. [*Feasibility*] To ensure practicality, the metrics should be based on input parameters that are determinable. When defining metrics, measurement methods shall be defined and if exact measurement is not possible or cost-intensive, alternative rigorous methods (e.g. statistical) shall be proposed. From an economic point of view, it is also required that the measurement procedure can be accomplished at a high level of automation.

The next section reviews the literature considering the requirements listed above.

LITERATURE REVIEW

A number of approaches to quantify DQ have been presented in academic as well as applied literature. They differ in the DQ dimensions taken into account and in the underlying measurement procedures [28]. In the following, we briefly describe the existing approaches to quantify timeliness.

Timeliness specifies to what extend the values of attributes are up-to-date (for a literature survey about existing definitions of this DQ dimension see [7]). In contrast to other dimensions (mainly correctness), quantifying timeliness does not necessarily require a real world test. Instead, the metric for timeliness shall deliver an indication, not a verified statement under certainty, whether an attribute value has changed in the real world since its acquisition and storage within the system. Therefore, Hinrichs proposed the following quotient [17]:

$$Timeliness = \frac{1}{(mean\ attribute\ update\ time) \cdot (attribute\ age) + 1}$$

This formula quantifies if the current attribute value is outdated. Related to the input factors taken into account, the quotient returns reasonable values: On the one hand, if the *mean attribute update time* is 0 (i.e. the attribute value never becomes out of date), timeliness is 1 (attribute value is up-to-date). If on the other hand *attribute age* is 0 (i.e. the attribute value is acquired at the instant of quantifying DQ) we get the same value. For higher values of *mean attribute update time* or *attribute age* the value of the metric approaches 0. That means that the (positive) indication (the attribute value is still corresponding to its real world counterpart) decreases. Hinrichs also provided formulas allowing for the aggregation of attributes thereby fulfilling R 4. Moreover, the parameter *attribute age* required to compute the value of the metric can be extracted automatically (R 6) from the meta-data.

Despite these benefits, there are some shortcomings to consider which hinder economic planning as well as prohibit evaluating the efficiency of realized DQ measures ex post:

- Typically, the value range [0; 1] is not covered, because a value of 0 is only returned if the value of *mean attribute update time* or *attribute age* respectively is ∞ (cf. R 1).
- The metrics are hardly applicable within an economic DQ management, since both absolute and relative changes can not be interpreted easily (R 3). Therefore, the resulting values of the metric are not interval scaled (R 2).

Table 1 depicts these limitations: In order to improve the value of timeliness from 0 to 0.5, the corresponding value of *mean attribute update time* multiplied with *attribute age* has to be decreased from ∞ to 1. In contrast, an improvement from 0.5 to 1 only requires a reduction from 1 to 0. Thus the interpretation of timelines improvements (e.g. by 0.5) is impeded.

Improvement of the metric	Necessary change of (*mean attribute update time*) · (*attribute age*)
0.0 → 0.5	∞ → 1.0
0.5 → 1.0	1.0 → 0.0

Table 1 Improvement of the metric and necessary change of parameters

Furthermore, by building a quotient the values returned become hardly interpretable (cf. R 3). That means that the value can not be interpreted for example as a probability that the stored attribute value still corresponds to the current state in the real world. This also implies, that the parameter *mean attribute update time* is not based on a probability distribution, i.e. all considered attributes decline according to the same pattern, which is obviously not true (see discussion below). Another limitation relates to the aggregation: It is not possible to emphasize particular attributes or relations and thereby the metric can not be adapted to the context of a particular application. Hence R 5 is not met.

The second approach by Ballou et al. defines the metric as follows (the notation was slightly adapted, cf. [2]):

$$Timeliness = \{\max[(1 - \frac{currency}{shelf\ life}), 0]\}^{s}$$

The *currency* of an attribute value is – in contrast to [17] – computed as follows: The time between quantifying timeliness and acquiring the attribute value is added to the age of the attribute value in the instant

of acquiring it. This corresponds to the age of the attribute value at the instant of quantifying DQ. *Shelf life* is according to Ballou et al. an indicator for the volatility of an attribute value. Thus, a relatively high *shelf life* results in a high timeliness and vice versa. The exponent *s* – which has to be assigned by experts – impacts to what extent a change of the quotient (*currency/shelf life*) affects the value of the metric. Thereby the computation can be adapted according to the attribute considered and to the particular application to certain extend (R 5).
Moreover, the values of the metric are normalized to [0; 1] (R 1). However, due to the impact of the exponent *s* the values become hard to interpret (cf. R 3) and are not interval scaled (R 2). Table 2 illustrates the effect for $s = 2$ corresponding to our approach shown in table 1. Again, an improvement of the metric by 0.5 is hard to interpret:

Improvement of the metric	Necessary change of (*currency/shelf life*)
0.0 → 0.5	1.00 → 0.29
0.5 → 1.0	0.29 → 0.0

Table 2 Improvement of the metric and necessary change of parameters

It seems that it is the aim of Ballou et al. to derive mathematical relations, so they do not deal with the topic of aggregating the values of the metric to higher levels (R 4). Moreover, they do not focus on designing a metric whose values would be interpretable within an economic DQ management (cf. R 3) and thus easily understandable by operating departments, e.g. a marketing division conducting a CRM campaign. Indeed, the values are interpretable as the probability that the attribute value in the information system still corresponds to its real world counterpart, if $s = 1$, which means assuming a uniform distribution. However, a uniform distribution assumes a fixed maximum lifetime and a constant (absolute) decline rate with regard to the initial value for the random variable considered. This means within the context of quantifying DQ: There is a maximum lifetime that can not be exceeded for each attribute considered. This does not hold for a lot of important attributes (e.g. 'surname' or 'date of birth') as they possess neither a fixed maximum shelf life nor a constant (absolute) decline rate. For $s \neq 1$, the value of the metric can not be regarded as a probability relying upon common distribution functions. Therefore, it is obvious that such a metric can not be adapted to contexts where interpretation of the values of the metric as a probability is required (R 5). Furthermore, the measurement of the parameter *currency* of an attribute value can mostly not be accomplished at a high level of automation (R 6).

The approach presented in Heinrich et al. [16] suggests a metric using probabilistic theory to improve the interpretability of the metrics results and to enable automated analysis. In this context, timeliness can be interpreted as the *probability* of an attribute value still being up-to-date. They assume shelf life of the underlying attribute values to be exponentially distributed. The exponential distribution is a typical distribution for lifetime. However this assumption does not hold for all attributes (this fact will be discussed later). The metric described in [16] takes two variables into account: *age*(*w, A*) and *decline*(*A*). *age*(*w, A*) denotes the age of the attribute *A*'s value, which is derived by means of two factors: the instant of quantifying DQ and the instant of data acquisition. The decline rate *decline*(*A*) of attribute *A*'s values can be determined statistically. The metric on the layer of an attribute value is therefore noted as:

$$Timeliness = \exp(-decline(A) \cdot age(w, A))$$

Thus following the determination of the decline rate, the *timeliness* of each attribute value can be quantified automatically (R 6) by means of the meta data. In addition, the value for *timeliness* as defined above denotes the probability that the attribute value is still valid. This interpretability (R 3) is an advantage over the approaches mentioned earlier. In addition, cases where *decline*(*A*) = 0 (e.g. for attributes as 'date of birth' or 'place of birth', that never change) are taken into account correctly by

$$Timeliness = \exp(-decline(A) \cdot age(w, A)) = \exp(0 \cdot age(w, A)) = \exp(0) = 1.$$

The same holds for cases where *age*(*w*, *A*) = 0 (the attribute value is acquired at the instant of quantifying DQ) by calculating

$$Timeliness = \exp(-decline(A) \cdot age(w, A)) = \exp(-decline(A) \cdot 0) = \exp(0) = 1.$$

Thereby the metric fulfils the requirements normalization (R 1) and interval scale (R 2). Moreover, [16] provide formulas allowing the aggregation of values to higher levels (R 4). Their metric is also adaptable to different applications as it allows incorporating weights in order to emphasize particular attributes and relations. Regarding this aspect of adaptivity, R 5 is (partly) fulfilled.

Summarizing, the metric meets all of the above stated requirements, if the attribute considered is exponentially distributed with the parameter *decline*(*A*). However, it can be criticized that the exponential distribution is memoryless in the following way:

$$P(X \geq x + t | X \geq x) = P(X \geq t).$$

I.e. if an exponentially distributed random variable *X* exceeds the value *x*, then exceeding *x* by at least *t* is as probable as an exponentially distributed random variable exceeding the value *t*.

In the context of quantifying DQ this means: The probability that a particular attribute value is outdated is equally high for each period of time considered. I.e. this probability is – with regard to a particular instant or a particular period of time – independent of the current age of the attribute value. If two attribute values *a* and *b* are up-to-date at the instant of quantifying DQ and *a* is older than *b*, then both values become out-of-date within the subsequent period of time with identical probability.

Thus the assumption of the validity being exponentially distributed can not hold for all attributes (e.g. shelf life of the attribute value *student* as professional status within a customer data base). Therefore, the metric in [16] is not applicable within contexts, where the attribute values are not exponentially distributed. Hence, R 5 is only partly met.

Due to the shortcomings of all of the above described approaches, we present a way of developing adequate metrics in the following.

A Data Quality Metric for Timeliness

The existing metrics to quantify timeliness either do not explicitly take into account the distribution of the shelf-life [18] or assume a particular distribution ([2]; [16]). Therefore, they are not applicable for a number of important attributes. In order to overcome this problem, we propose a procedure to develop metrics for timeliness based on the distribution of the shelf-life of the attribute to be valuated. From that point of view the procedure supports developing adequate metrics that meet all six requirements, particularly including adaptivity (R 5), which is an advantage compared to existing approaches.

Procedure for developing an adequate metric for timeliness

As already mentioned, each attribute can differ with regard to the distribution of the shelf life of its values (as defined by [2], cf. above). Therefore we present a procedure for developing timeliness metrics which – on the one hand – follows the approach presented in [16] and provides a value that can be interpreted as a probability. On the other hand, the procedure shall assure that the metric can be adapted to the specific characteristics of the shelf life of the attribute considered. This will enable us to eliminate limiting assumptions about the shelf life and the timeliness (in a majority of cases). Figure 2 illustrates the steps of the procedure, which will be described in the following.

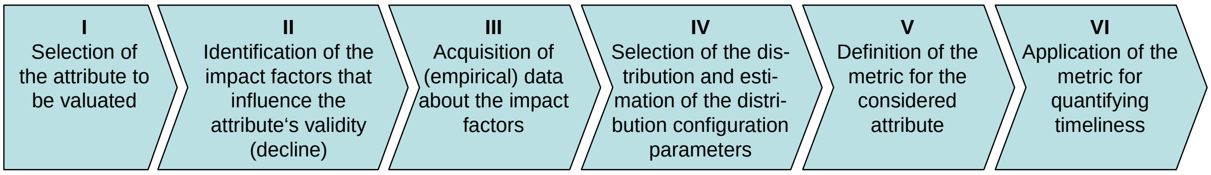
I
Selection of the attribute to be valuated
II
Identification of the impact factors that influence the attribute‘s validity (decline)
III
Acquisition of (empirical) data about the impact factors
IV
Selection of the distribution and estimation of the distribution configuration parameters
V
Definition of the metric for the considered attribute
VI
Application of the metric for quantifying timeliness

Figure 2 Procedure for developing metrics for timeliness

Step I chooses the attribute to be valuated. Hereby it has to be examined from an economic point of view whether the development of a metric for a particular attribute is necessary with respect to the given purpose (note that the development of a metric can be very costly, whereas the step of quantifying DQ itself can often be automated). E.g., it is not necessary to develop metrics for all attributes within a CRM campaign. Instead, one should focus on relevant attributes, for example those to be used as a selection criterion to identify customers for a particular target group (segmentation). Only if the given purpose justifies quantifying DQ, step II should be conducted.

Before acquiring empirical data, factors influencing the shelf life or the decline rate of the attribute values (i.e., what does the decline rate of the validity of an attribute value depend on?) have to be determined in step II. Where a number of these factors exist, steps III to V typically have to be carried out for each factor. Before choosing an adequate distribution, empirical data about the decline rate of the shelf life has to be acquired. Sources for such data might be external statistics (e.g. Federal Statistical Offices or scientific studies) or internal statistics of the firm as well as experience values and experts’ estimations. In step IV an adequate probability distribution has to be determined based on the acquired data of step III. Thereby one has to consider the properties of the different distributions in order to represent the decline rate of the shelf life correctly in approximation. Table 3 states important properties of selected cumulative distribution functions:

Cumulative distribution function	**Properties**	**Example**
Uniform distribution: A random variable being equally distributed over [a; b] X~U(a; b) has the following cumulative distribution function: $F(x)=\begin{cases}0 & \text{for } x \le a\\ \frac{x-a}{b-a} & \text{for } a < x < b\\ 1 & \text{for } x \ge b\end{cases}$	– Constant, absolute decline rate within the given period of time – Fixed maximum period of validity and lifetime – Continuous distribution	Analyses on the validity of customers’ debit or eurocheque cards (unknown date of issue, fixed expiry date of all distributed cards)
Geometric distribution: A geometric distributed random variable X_n with parameter $q = 1 - p$ (with q as probability for a failure) has the following cumulative distribution function: $F(n) = 1-(1-p)^n$	– Constant, absolute probability of decline within each period – Memoryless discrete distribution	Analyses about the validity of a contract (e.g. labor agreement) with the option to terminate at quarter-end
Exponential distribution: A exponentially distributed random variable X (defined on IR^+) with rate parameter λ is characterized by the following cumulative distribution function: $F(x)=\begin{cases}1\text{-}\exp(\text{-}\lambda x) & \text{for } x \ge 0\\ 0 & \text{for } x < 0\end{cases}$	– Memoryless – The conditional probability that the attribute value considered becomes out of date in the next period of time is independent of the current age of the attribute – Constant, relative decline rates	Analyses about timeliness of address data (e.g. house moving)

Weibull distribution: A Weibull distributed random variable X (defined on IR^{+}) with shape parameter $k > 0$ and scale parameter $\lambda > 0$ has the following cumulative distribution function: $F(x)=\begin{cases}1-\exp(-\lambda x^{k}) & \text{for } x\geq 0 \\ 0 & x<0\end{cases}$	– Not memoryless – Applicable for increasing or decreasing, constant (relative) decline rates	Analyses about duration of study and professional status student (cf. below)
Gamma distribution A gamma distributed random variable X (defined on IR^{+}) with shape parameter $k > 0$ and scale parameter $\theta > 0$ has the following cumulative distribution function: $F(x)=\begin{cases}\dfrac{\gamma\left(k,\dfrac{x}{\theta}\right)}{\gamma(k)} & \text{for } x\geq 0 \\ 0 & x<0\end{cases}$	– Not memoryless – Applicable for the description of changing, relative decline rates of the shelf life of one attribute	Analyses about the lifespan of end devices (e.g. within marketing campaigns for accessories)

Table 3 Important properties of selected cumulative distribution functions

Since the distributions usually can be adapted via distribution configuration parameters (cf. the shape and scale parameters mentioned above). These parameters have to be determined by means of common estimation procedures and empirical data. In cases where several factors have impact on the decline rate, it is not sufficient to conduct steps III-V for each factor. Moreover, the distribution functions have to be combined. This is done in step V, in which the metric is defined based on the combined distribution (this ensures that the requirements R 1, R 2 and R 3 are fulfilled, since the result is a probability). To allow a flexible application and enable a measurement of DQ at the layer of tupels, relations, and the whole data base (cf. R 4) the developed metric on the layer of attribute values can be aggregated to the next higher layers as shown in [16]. Since particular attributes and relations can be emphasized when aggregating the values to higher levels and the procedure is designed in order to adapt the metric to the characteristics of an attribute value, R 5 is met. Step VI allows quantification of timeliness by means of the metric (cf. [16]). Therefore the age of each attribute value has to be calculated automatically from the instant when DQ is quantified and the instant of data acquisition (cf. R 6). Afterwards, the value of the metric for timeliness is determined using the combined distribution function of step V. Finally, the results can be applied within an economic management of DQ. The following section illustrates the procedure in a real-world scenario and shows that the resulting metrics are applicable.

Illustration of the procedure

In the following we want to exemplary illustrate the procedure and develop a particular metric for the DQ dimension timeliness. As continuing example we use the attribute *professional status* within a customer data base (step I). We chose this attribute as it is also used in our practical example in the next section, in which we consider a CRM campaign for a major MSP: Targeting students within a CRM campaign. Frequently customers are included in the target-group for such campaigns although they have finished or abandoned their studies and are thus no longer eligible for student-discounts. The implications of wrongly selecting customers for campaigns are twofold: decreased customer satisfaction and low success-rates campaigns leading to inefficient usage of resources. To reduce such deficiencies a metric for the attribute *professional status* with the value *student* (selection criterion within the campaign) is presented.

The attribute value *student* can loose its validity due to two impact factors (cf. step II): Either a study is completed or aborted. Hence, the metric consists of two different distributions, one for each factor. For the problem at hand neither a sampling survey nor any other form of internal data collection is necessary. Instead we can determine the distribution by means of external data: Many universities as well as the Federal Statistical Offices provide statistics about the duration of studies (step III, cf. [12]; [14]). For illustrational purposes we use data from the University of Vienna/Austria, since they provide a relative frequency distribution of the duration of study, which aggregates the data of several programs of study of different faculties. Otherwise it is also possible to use data of German Universities, for example provided by the Federal Statistical Office of Germany.

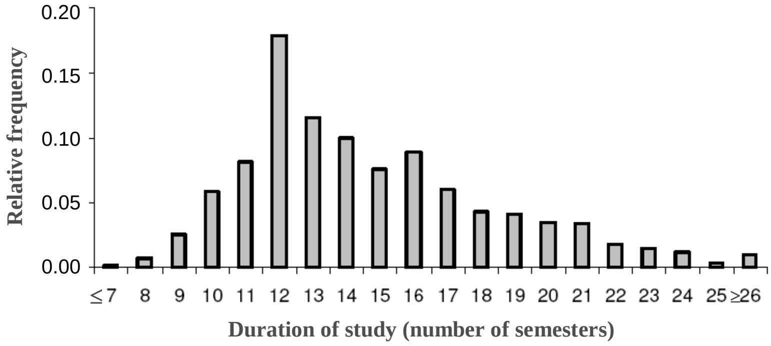

Figure 3 Relative frequency distribution of duration of study

Figure 3 shows the relative frequency distribution of the study-duration at the University of Vienna (in this figure for all students graduating in 2000). Considering the first impact factor (successful completion of degree), the distribution of the duration of study can be determined. Analyzing the historical data reveals that the percentage of students completing their degree is not constant over time. The assumption of it being constant would imply the following: The probability that a student already studying for eight semesters completes his degree within the next two semesters is equal to the probability that a student already studying for twelve semesters completes his degree within the next two semesters. This obviously does not hold as initially the relative frequency steeply increases and decreases after the 12th semester (cf. figure 3). Thereby a constant percentage of students completing their degree as well as memorylessness – an important property of the exponential distribution – can not be assumed. Hence, the approaches by Hinrichs [18], Ballou et al. [2] and Heinrich et al. [16] are not suitable within this context.

Therefore we need a distribution of the shelf life that is not memoryless and that can consider increasing or decreasing decline rates. A continuous distribution holding these properties is the Weibull distribution (step IV).

The Weibull distribution *wei*(k, λ) is based on two parameters, *shape* (k) and *scale* (λ). A number of alternatives exist to determine these parameters for the problem at hand. Marks presents an adequate method to determine the Weibull distribution parameters based on symmetric percentiles P_L and P_U (lower and upper percentile, L and U denoting the value of the distribution at the percentile P_L and P_U respectively) [22]. Percentiles are the values of a variable below which a certain percent of observations fall. An example for symmetric percentiles is the 10th percentile (P_{10}) and the 90th percentile (P_{90}). I.e. 90% of all values lie below the 90th percentile. The simple estimation procedure is based on the following equations for k and λ (with ln as natural logarithm function):

$$k = ln[ln(U)/ln(L)] \cdot (P_U / P_L) \text{ and } \lambda = \frac{P_U}{(-\ln(L))^{\frac{1}{k}}}$$

Applying a Monte Carlo simulation, Marks illustrates that the best estimation is achieved when using the 10th and 90th percentile. We can utilize this method, but have to adapt it slightly: The method by Marks implicitly assumes that the Weibull distributed values start at the origin. However, in our example the graduates complete their degrees between the 7th and 26th semester. That is why the calculated parameters have to be slightly adjusted by a left shift. Doing so we get the following parameter values: k = 0.00002 and λ = 4. The value of the coefficient R^2 of determination is 0.91, expressing an adequate approximation of the empirical distribution by means of the parameterized Weibull distribution. The cumulative distribution function can be approximated as follows:

$$P_{Gradute}(x) = 1 - exp(-0.00002 \cdot x^4) \quad for \quad x \geq 0$$

$P_{Graduate}(x)$ denotes the cumulative probability that a student has completed his degree after x semesters (step V). Furthermore, we have to analyze the distribution of dropouts as the second impact factor on the validity of the attribute value *student.*

Figure 4 illustrates the corresponding data for the University of Vienna (step III). It shows the percentage of all dropouts that aborted their studies within a particular semester (again aggregated for all programs of study): E.g. 18% of all dropouts discarded their studies within the first semester. It holds for this distribution that the dropouts' percentage remains approximately constant in relation to the students still active (in contrast to the number of absolute dropouts, which is obviously decreasing). Therefore the properties of a constant relative decline rate and memorylessness can be stated and we can apply the exponential distribution.

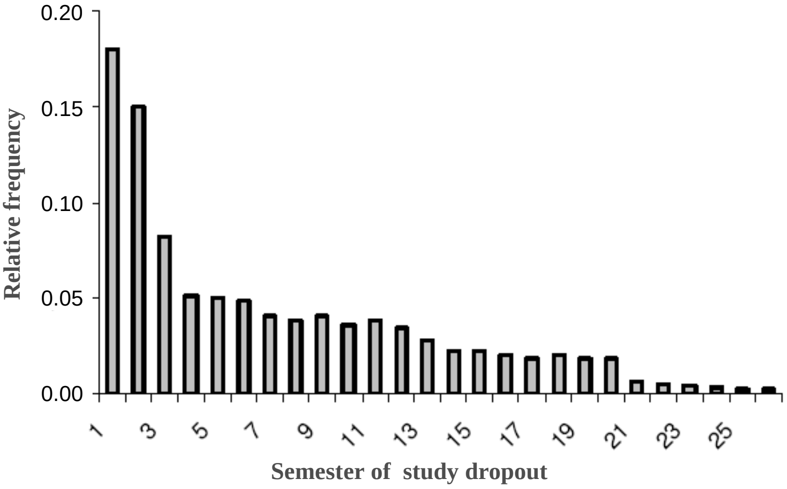

Figure 4 Relative frequency distribution of study dropout

The estimation of the parameters (step IV) for the exponential distribution can be conducted by applying the expected value: It corresponds to the reciprocal of the decline rate. The arithmetic mean of the empirical data serves as unbiased estimator for the expected value $E(x)$. In our example the arithmetic mean is about 5.5 semesters. Thereby, the distribution parameter λ of the exponential distribution is calculated as follows:

$$E(x) = \frac{1}{\lambda} \Leftrightarrow \lambda = \frac{1}{E(x)} = \frac{1}{5.5} = 0.18$$

Again, we get an adequate approximation of the empirical distribution by means of the parameterized exponential distribution. This is expressed by a value of the coefficient R^2 of determination of 0.88. Thereby, $P_{Dropout}(x)$ denotes the cumulative probability that a student has aborted his study (step IV).

$$P_{Dropout}(x) = 1 - \exp(-0.18 \cdot x) \text{ for } x \geq 0$$

In order to integrate the two distributions determined above we have to estimate the percentage of graduates and dropouts. Using historical data, the percentage of graduates can be estimated at 64%. Therefore one has to define the probability $P_{Student}(x)$ that a student is still studying, as follows (step IV):

$$P_{Student}(x) = 1 - 0.64 \cdot (P_{Gradute}(x)) - 0.36 \cdot (P_{Dropout}(x)) \text{ for } x \geq 0$$

We can use this distribution to calculate the probability that a customer with the attribute value *student* (within a data base) is still studying. Based on probability theory, the values of the metrics are normalized (R 1), interval scaled (R 2) and interpretable (R 3). Moreover, the aggregation formulas defined by [16] can be applied (R 4), which also allow to emphasize particular attributes or relations. This and adapting the metric to the shelf life of the attribute value make the metric meet R 5. As already mentioned, the timeliness of a particular customer's professional status can be calculated automatically, after the decline rate is determined and by using the above formula (R 6). After theoretically discussing the procedure with respect to the requirements, the application of the procedure within the mobile services sector is described in the next section.

CASE STUDY: APPLICATION OF THE METRIC FOR TIMELINESS

This following section illustrates the economic effects of quantifying DQ within CRM campaigns by means of a case study where metrics for timeliness were developed according to the procedure described above to quantify DQ (R 6). We exemplify the procedure by means of a particular attribute and its characteristics, but the results are reusable: If the attribute shall be used for other tasks outside CRM campaigns (e.g. in order to design new products and mobile tariffs), the metric does not have to be developed again. As the development of a metric can be costly, this is an obvious advantage.

In our case study, a MSP wants to address customers with higher sales who have the professional status *student*: They shall be offered a new premium product called *Student AAA.* For reasons of confidentiality, all client-specific figures and data had to be changed and made anonymous. Nevertheless, the procedure and the basic results remain the same.

The MSP previously acted as follows: The top customers (according to their sales values) fulfilling a given criterion (e.g. attribute value *clerk* or *student*) were selected from the customer data base. After that, the new offer was sent to them. Ex post success rates of such campaigns averaged to approximately 9%, i.e. ca. 9,000 out of 100,000 contacted customers accepted the offer.

Applying this procedure to the new campaign *Student AAA* would require selecting all customers with the attribute value *student* who (according to requirements from the marketing department) belong to the top 30% customers with regard to sales. Thereby, about 46,800 customers (out of about 156,000 customers with the attribute value *student)* would be addressed. These customers show average sales of 1,340 € p. a. Assuming the former success rate of about 9%, nearly 4,200 customers will accept the offer (estimation). If a customer accepts the offer, the MSP can increase its return on sales by 5%. Thus 4,200 customers with average sales of 1,340 € accepting the offer would imply a forecasted additional profit of approx. 281,400 € which sounds like a quite profitable business case for the campaign.

Yet before starting the campaign and addressing these 46,800 customers, its forecasted profit had to be verified and improved by means of the DQ metric developed above. Especially the success rate should be increased by raising the percentage of addressed customers who are still studying in reality (as only they are actually eligible to accept the offer and need to provide a certificate of matriculation).

Therefore the probability that a customer with the attribute value *student* is still studying was calculated. This had to be done automatically, since a manual "check" of each of the approx. 46,800 customers (top 30% sales) would have been far too time-consuming and cost-intensive.

The domain [0; 1] of the probability that a customer with the attribute value *student* is still actually a student (value of the metric) was divided into ten equal intervals. The selected 46,800 customers were assigned to these intervals according to their individual probability. Figure 5 shows the results (e.g. the number of customers within the interval]0.2; 0.3] is 3,480).

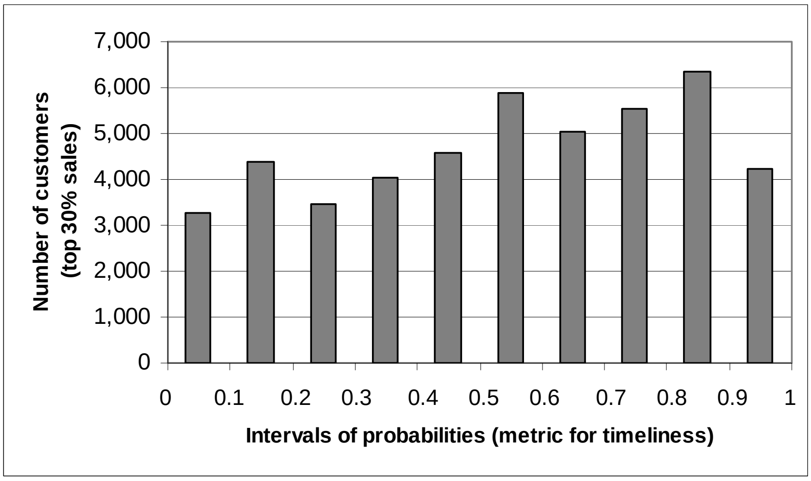

Figure 5 Numbers of customers depending on the metric for timeliness

The figure illustrates: The probability that the selected customers are still studying in reality is less than 0.5 for approx. 20,000 (about 42%) customers. I.e. it is very likely that these customers are not eligible to accept the offer. To overcome this problem, the probability (value of the metric) was automatically calculated for all 156,000 customers. Based on this, the expected profit for each customer (based on his/her sales p.a. and the additional return on sales of 5%) was determined in case the offer is accepted. This expected profit was now used to identify the top 30% customers. The results were obviously different compared to selecting only by sales, since DQ and thereby the probability whether a customer is still studying were taken into account. I.e., some of the customers with high sales were very likely not to have student-status anymore. That is why they were not selected according to the expected profit. In the end, only approx. 18,100 customers were selected according to both criteria (sales and expected profit), i.e. more than 28,700 customers that would have been addressed based on the previous procedure of the MSP were not addressed anymore when the metric for timeliness was taken into consideration. Instead, 28,700 other customers were identified for the campaign based on the criterion expected profit.

As a precaution, the marketing department decided to address all approx. 75,500 customers that were selected according to one or both of the criteria to verify whether including DQ provides better results. The analysis conducted after the campaign revealed the following results, which are depicted in Figure 6:

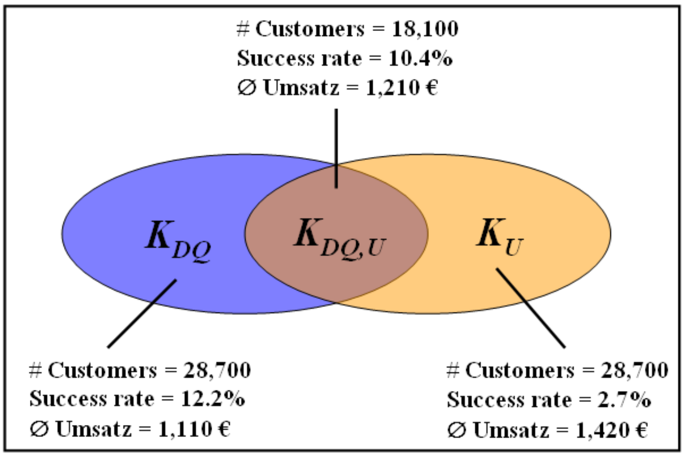

Figure 6 Differences between the two selection criteria

The ex post analysis was done separately for the 18,100 customers (= $K_{DQ,U}$) that were selected according to both criteria and for the respective 28,700 customers that were chosen either only according to sales (= K_U) or with respect to the expected profit calculated by means of the DQ metric (= K_{DQ}). The customers $K_{DQ,U}$ had a success rate of 10.4% and average sales of 1,210 €. In contrast, the success rate of the customers K_U was only 2.7%, whereas their average sales were quite high at 1,420 € as these customers are the ones with the highest sales.

The overall success rate of the 46,800 customers that were initially selected by sales was only 5.7% and therefore below the expected 9%. This can be explained as follows: Indeed, K_U selected the customers with the highest sales. However, many of these customers were not studying anymore and therefore could not accept the offer. From an economic point of view it is highly questionable to select these customers when taking into account the customer-contact cost. Considering the 28,700 customers in K_{DQ}, it can be stated that average sales were lower (1,110 €), but the success rate of 12.2% exceeded expectations. The success rate of the 46,800 customers ($K_{DQ,U}$ and K_{DQ}) that were selected by the DQ metric was 11.5%. In combination with the average sales (1,150 €) of these 46,800 customers the MSP made an additional profit of approx. 309,200 €. In contrast, by addressing only the customers with the highest sales, the additional profit would have been far lower.

The example illustrates the applicability of the metric timeliness and its impact in order to improve CRM campaigns. Moreover, we demonstrated how applying the DQ metric helps to substantiate the calculation of business cases. This is especially relevant as using only sales as selection criterion would have resulted in choosing many customers who are very likely not to study anymore. I.e. many of the customers addressed would not have been eligible to accept the offer thus leading to an unprofitable campaign.

However, we base our findings on a single case study and the application of the procedure would have to be repeated (especially in other areas) to stabilize or undermine our findings. Moreover, we need to emphasize that the selection of the customers is based on the assumption, that the attribute value *student* is already stored in the data base. However, there are customers within the data base who have become students (in the meantime), but whose attribute value is different (e.g. *pupil, apprentice*) and may thus be interesting for the target group of the campaign as well. Since they are not considered yet, metrics for timeliness have to be developed for other attribute values as for instance *pupil* that indicate the probability for a transition into the professional status *student* (cf. e.g. data from the Federal Statistical Office of Germany).

RESULTS AND LIMITATIONS

The paper analyzed how the DQ dimension timeliness can be quantified in a goal-oriented and economic manner. The aim was to design a procedure for developing metrics for timeliness. The metric presented may enable an objective and partly automated measurement. In cooperation with a major MSP, the procedure was applied in the context of CRM campaigns. In contrast to existing approaches, the metrics resulting from the procedure are designed according to important requirements like interpretability and feasibility. Moreover, the procedure allows developing metrics that are not based on (limiting) assumptions as for instance an exponential distribution of the validity. These metrics enable quantifying DQ and represent thereby the basis for economic analyses. The effect of DQ measures can be analyzed by comparing the realized DQ level (ex post) with the planned level (ex ante).

Despite these improvements in relation to existing techniques and metrics some limitations of our current approach provide room for further research. One important prerequisite when developing metrics is a suitable data pool. Frequently external data, for instance from Federal Statistical Offices or data provided by estimations and forecasts of experts can help to populate data pools. Conducting samples or other forms of internal data analysis is by far more time-consuming and cost-intensive. Therefore it has to be examined whether the development of a metric for an attribute is necessary and reasonable with respect to the given goal (e.g. considering the deterministic decline of the attribute values like railway schedules expiring after a certain date). However, it has to be considered that a developed metric can frequently be reused or adapted for a several fields of application. The authors currently develop a model-based approach for the economic planning of DQ measures. For implementing such a model, adequate DQ metrics and measurement procedures are necessary. The approaches presented in this paper provide a basis for those purposes. Nevertheless, further metrics for other DQ dimensions should be developed and thus further research in this area is encouraged.

BIBLIOGRAPHY

[1] Ballou, D. P., and Pazer, H. L. "Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff." *Information Systems Research* 6 (1), 1995, pp. 51-72

[2] Ballou, D. P., Wang, R. Y., Pazer, H., and Tayi, G. K. "Modeling information manufacturing systems to determine information product quality." *Management Science* 44 (4), 1998, pp. 462-484

[3] Batini, C., and Scannapieco, M. *Data Quality: Concepts, Methods and Techniques.* Springer, Berlin, 2006.

[4] Betts, M. "Data quality: The cornerstone of CRM." *Computerworld*, 2002-02-18.

[5] Campanella, J. *Principles of quality cost*. ASQ Quality Press, Milwaukee, 1999.

[6] Cappiello, C., Francalanci, Ch., and Pernici, B. "Data quality assessment from the user's perspective." *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*, Paris, 2004, pp. 68-73.

[7] Cappiello, C., Francalanci, Ch., and Pernici, B. "Time-Related Factors of Data Quality in Multichannel Information Systems." *Journal of Management Information Systems* 20 (3), 2004, pp. 71-91

[8] English, L. *Improving Data Warehouse and Business Information Quality*. Wiley, New York, 1999.

[9] Eppler, M. J. *Managing Information Quality*. Springer. Berlin, 2003.

[10] Even, A., and Shankaranarayanan, G. "Utility-Driven Assessment of Data Quality." *The DATA BASE for Advances in Information Systems* 38 (2), 2007, pp. 75-93

[11] Even, A., and Shankaranarayanan, G. "Value-Driven Data Quality Assessment." *Proceedings of the 10th International Conference on Information Quality,* Cambridge, 2005.

[12] Federal Statistical Office of Germany *Education in Germany - Statistics of Higher Education*. Wiesbaden, 2006.

[13] Feigenbaum, A. V. *Total quality control*. McGraw-Hill, New York, 1991.

[14] Hackl, P., and Sedlacek, G. "Analyse der Studiendauer am Beispiel der Wirtschaftsuniversität Wien." *Dutter, R. (edi.): Festschrift 50 Jahre Österreichische Statistische Gesellschaft*, Wien, 2001, pp. 41-59

[15] Heinrich, B., and Helfert, H. "Analyzing Data Quality Investments in CRM – a model based approach." *Proceedings of the 8th International Conference on Information Quality,* Cambridge, 2003.

[16] Heinrich, B., Kaiser, M., and Klier, M. F. "How to measure Data Quality - a metric based approach" *In appraisal for: International Conference on Information Systems, Montréal, 2007.*

[17] Hevner, A. R., March, S. T., Park, J., and Ram, S. "Design Science in Information Systems Research." *MIS Quarterly* 28 (1), pp. 75-105

[18] Hinrichs, H. *Datenqualitätsmanagement in Data Warehouse-Systemen.* doctoral thesis, Oldenburg, 2002.

[19] Jarke, M., and Vassiliou, Y. “Foundations of Data Warehouse Quality – A Review of the DWQ Project.” *Proceedings of the 2nd International Conference on Information Quality*, Cambridge, 1997.

[20] Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. “AIMQ: a methodology for information quality assessment.” *Information & Management* 40, 2002, pp. 133-146

[21] Machowski, F., and Dale, B. G. “Quality costing: An examination of knowledge, attitudes, and perceptions.” *Quality Management Journal* 3 (5), 1998, pp. 84-95

[22] Marks, N. B. “Estimation of Weibull Parameters from Common Percentiles.” *Journal of Applied Statistics* 32 (1), 2005 1, pp. 17-24

[23] Nelson, S. *What's happening to CRM in 2002*. Gartner Group, January, 2002.

[24] Pipino, L. L., Lee, Y. W., and Wang, R. Y. “Data Quality Assessment.” *Communications of the ACM* 45(4), 2002, pp. 211-218

[25] Redman, T. C. *Data Quality for the Information Age*. Arctech House, Norwood, 1996.

[26] SAS Institute *European firms suffer from loss of profitability and low customer satisfaction caused by poor data quality*. Survey of the SAS Institute, 2003.

[27] Shank, J. M., and Govindarajan, V. “Measuring the cost of quality: A strategic cost management perspective.” *Journal of Cost Management* 2 (8), 1994, pp. 5-17

[28] Wang, R. Y., Storey, V. C., and Firth, C. P. “A Framework for analysis of data quality research.” *IEEE Transaction on Knowledge and Data Engineering* 7 (4), 1995, pp. 623-640