A NOVEL DATA QUALITY METRIC FOR TIMELINESS CONSIDERING SUPPLEMENTAL DATA

Heinrich, Bernd, Department of Information Systems, University of Innsbruck, Universitaetsstrasse 15, A-6020 Innsbruck, Austria, bernd.heinrich@uibk.ac.at

Klier, Mathias, Department of Information Systems, University of Innsbruck, Universitaetsstrasse 15, A-6020 Innsbruck, Austria, mathias.klier@uibk.ac.at

Abstract

It is intensively discussed in both science and practice how data quality (DQ) can be assured and improved. The growing relevance of DQ has revealed the need for adequate metrics because quantifying DQ is essential for planning quality measures in an economic manner. This paper analyses how DQ can be quantified with respect to the DQ dimension timeliness. Based on an existing approach, we design a new metric to quantify timeliness in a well-founded manner that considers so-called supplemental data (supplemental data are additional data attributes that allow drawing conclusions about the timeliness of the data attribute considered). In addition, it is possible to take the values of the metric into account when calculating expected values, an advantage that in turn leads to improved and comprehensible decision support. We evaluate the presented metric briefly with regard to requirements for designing DQ metrics from literature. Then, we illustrate the metric's applicability as well as its practical benefit. In cooperation with a financial services provider, the metric was applied in the field of customer valuation in order to support the measurement of customer lifetime values.

Keywords: Data Quality, Data Quality Metrics, Design Research, Customer Valuation

1 INTRODUCTION

Both the benefit and the acceptance of application systems depend heavily on the quality of data processed and provided by these systems (Ballou et al. 1999, Fisher et al. 2003). Executives and employees need high-quality data in order to perform business, innovation, and decision-making processes properly (Al-Hakim 2007, Even et al. 2007). This in mind, it is not surprising that insufficient data quality (DQ) may lead to wrong decisions and correspondingly high costs. According to an international survey on DQ, 75% of all respondents have already made wrong decisions due to incorrect or outdated data. In addition, the respondents and their staff spend up to 30% of their working time on checking the quality of data provided (Harris Interactive 2006). Therefore, ensuring completeness, correctness, and timeliness of data – these properties are known as DQ dimensions (Wang et al. 1995) – still remains an important problem for many companies (Ballou et al. 1998, Jiang et al. 2007). Non-surprisingly, many scientific papers (e.g. Ballou et al. 1998, Even et al. 2007, Lee et al. 2002, Parssian et al. 2004, Pipino et al. 2002, Wang 1998) deal with the question of how DQ can be quantified. This is essential for analysing the economic effects of poor or improved DQ as well as for realising DQ measures considering cost-benefit aspects (e.g. Heinrich et al. 2007a, Pipino et al. 2002).

In the following, we propose a metric for the DQ dimension timeliness. The reason is that – according to several studies – timeliness is a serious issue in DQ management (Klein et al. 2007, Yu et al. 2007). Therefore, this dimension has already been discussed from both a scientific and a practical point of view (e.g. Al-Hakim 2007, Klein et al. 2007, Knight et al. 2005, Lee et al. 2002, Wand et al. 1996).

Referring to the guidelines for conducting design science research defined by Hevner et al. (2004), we consider the metric for timeliness as our artifact and organize the paper as follows: After discussing the relevance of the problem, section 2 briefly compiles the related work regarding timeliness and identifies the research gap. Our contribution is a novel approach to quantify timeliness. Hence, a metric is proposed in section 3 which is based on probabilistic considerations. This metric enables to quantify timeliness in a well-founded manner and to consider so-called supplemental data (supplemental data are additional data attributes that allow drawing conclusions about the timeliness of the data attribute considered). In section 4, we illustrate the application of the new approach and its practical benefit by means of an extensive real world example in the field of customer valuation at a financial services provider. Section 5 summarizes our findings and critically reflects on the results.

2 RELATED WORK

In literature, there is a wide range of definitions with respect to the DQ dimension timeliness. In some publications, timeliness is also referred to as currency or recency. Table 1 contains some selected definitions.

Reference	Term and Definition
Ballou et al. (1985), Ballou et al. (1998)	Timeliness: “the recorded value is not out of date […]. A stored value, or any data item, that has become outdated is in error in that it differs from the current (correct) value.”
Wang et al. (1996)	Timeliness: “The extent to which the age of the data is appropriate for the task at hand.”
Redman (1996)	Currency: “degree to which a datum in question is up-to-date. A datum value is up-to-date if it is correct in spite of possible discrepancies caused by time-related changes to the correct value.”
Hinrichs (2002)	Timeliness: “Property that the attributes or tuples respectively of a data product correspond to the current state of the discourse world, i.e. they are not out-dated” (own translation)
Price et al. (2005)	Timeliness: “The currency (age) of data is appropriate to their use”
Batini et al. (2006)	Timeliness: “Timeliness expresses how current data are for the task at hand.”
Heinrich et al. (2007a), Heinrich et al. (2007b)	Timeliness: ”Timeliness can be interpreted as the probability that an attribute value is still up-to-date”

Table 1. Selected definitions of the DQ dimension timeliness

The main issue of most definitions is that timeliness expresses whether an attribute value stored in a database is still up-to-date. This means that an attribute value, which was correct when it was stored, still corresponds to the current value of its real world counterpart at the (later) instant when DQ is quantified. In other words, the attribute value has not become outdated (due to its temporal decline). This is also reflected in the authors' approaches to quantify timeliness. In contrast to the DQ dimension correctness, quantifying timeliness does not necessarily require a real world test. Instead, a metric for timeliness should provide an estimation, not a verified statement under certainty (which is necessary for correctness), on whether an attribute value is still up-to-date. Heinrich et al. (2007a, b) refer to this fact explicitly. They interpret timeliness as the probability that an attribute value is still up-to-date. For huge data sets and when the shelf life of attribute values is not explicitly known, it seems to be quite reasonable to quantify timeliness by means of such an estimation. This is because comparing attribute values to their real world counterparts (real world test) is often by far too time- and cost-intensive and not practical at all.

In this context the following questions arise: (1) How can well-founded estimations for the timeliness of attribute values be derived? (2) Avoiding real world tests, what kind of data can alternatively be used for quantifying timeliness? Some authors mentioned above consider so-called *attribute metadata*. Such metadata are the instant t_0, when the attribute value's corresponding real world counterpart was created (e.g. for an attribute value "student" of a data attribute "professional status": the instant of the student's enrolment, at which the data value "student" became valid), and the attribute value's shelf life T (e.g. for the attribute value "student": the duration of study that represents how long this value is valid). Depending on whether these metadata are known or not, we have to quantify the timeliness of an attribute value under certainty or uncertainty. According to the definition of timeliness given above, we have to quantify whether an attribute value still corresponds to the current value of its real world counterpart at the instant t_1 of quantifying DQ. In other words, we have to analyse whether $(t_1\text{-}t_0)\leq T$ holds. If the instant t_0 when the attribute value's corresponding real world counterpart was created and the attribute value's shelf life T are known (e.g. based on a temporal rule), it is possible to determine whether the attribute value is still up-to-date under certainty. In this case the value of the metric equals zero (minimum value) if $(t_1\text{-}t_0)>T$ holds (attribute value is outdated). Otherwise, we can definitely state that the attribute value considered still corresponds to the current value of its real world counterpart ($(t_1\text{-}t_0)\leq T$) and the value of the metric equals one (maximum value). In contrast to this scenario of temporal rules, we will focus in the following on quantifying timeliness under uncertainty. This case is much more interesting (in the scenario under certainty you have to check only if $(t_1\text{-}t_0)\leq T$ holds) and often more realistic because the shelf life T of an attribute value is usually unknown. Thus, the question arises of how to accomplish well-founded estimations for the timeliness of data when the shelf life of attribute values is not known.

To improve the estimations for the timeliness of data and to enhance existing approaches, we follow the idea to use additionally other data attribute values w_i (i=1, …, n) to draw conclusions about the timeliness of an attribute value ω considered (e.g. about its unknown shelf life T). In the following, the values w_i, which are used to improve the estimation, are called *supplemental data*. A short example illustrates their importance: Figure 1 (based on data from Federal Statistical Office of Germany 2007, Heublein et al. 2003, Heublein et al. 2008) shows that the duration of study (including dropouts) – i.e. the shelf life T of the attribute value "student" – and the type of university (university vs. university of applied sciences) are statistically contingent. Consider a number of customers stored in a database whose professional status "student" was for example stored 5.5 years ago at the beginning of their studies of economics and social sciences (age t of the attribute values: $(t_1\text{–}t_0)$=5.5 years=11 semesters): Then it is expected that in average about 90% of the customers enrolled at a *university of applied sciences* have already finished their studies (see Figure 1). This means that the attribute value "student" is up-to-date for only 10% of them. In contrast, it is expected that only 66% of the customers enrolled at a *university* have already finished their studies (see Figure 1) – i.e. the attribute value "student" is still up-to-date for about 34% of them in average.

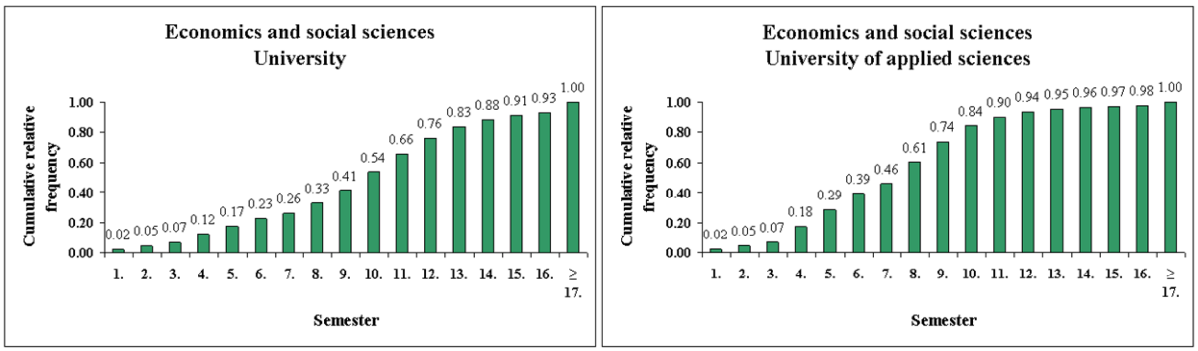

Figure 1. Cumulative frequency distribution of duration of study (incl. study dropout).

Consequently, supplemental data – like the type of university – are relevant for quantifying the timeliness of data. Thus, it seems to be necessary to consider not only *attribute metadata*, but also *supplemental data*. Below, we discuss how existing approaches deal with this kind of data.

We found the approaches by Hinrichs (2002), Ballou et al. (1998), Even et al. (2007), and Heinrich et al. (2007a) as these are – to the best of our knowledge – the only approaches which (1) design metrics for timeliness, (2) are formally noted, and (3) are based for the most part on a Quality of Conformance definition. Regarding (3), literature distinguishes two different concepts and definitions of quality: Quality of Design and Quality of Conformance (Heinrich et al. 2003, Juran 1998, Teboul 1991). Quality of Design denotes the degree of correspondence between the users' requirements and the specification of the information system. In contrast, Quality of Conformance represents the degree of correspondence between the specification and the existing realization in the information systems (e.g. correspondence between a data schemata and a set of stored data values). In the following, we focus on metrics for quantifying Quality of Conformance as they can be applied in many different situations and are more reusable (because they are more independent from particular users' requirements in a specific business context). Heinrich et al. (2009) have already analysed the approaches mentioned above and compared them with respect to six requirements (due to space restrictions we can not describe this here): *Normalisation, Interval scale, Interpretability*, *Aggregation*, *Adaptivity*, and *Feasibility*. This analysis revealed weaknesses particularly with respect to *Adaptivity*, which requires a metric to be adaptable to the context of a particular application in order to enable a goal-oriented quantification of DQ (for a detailed analysis see (Heinrich et al. 2009)).

Referring to *Adaptivity*, supplemental data are valuable to improve the quantification of timeliness in many fields of application. Just consider the example mentioned above (see Figure 1). The existing metrics by Hinrichs (2002), Ballou et al. (1998), Even et al. (2007), and Heinrich et al. (2007a) use *metadata* like instant of creation t_0 and shelf life T. However, they do not use *supplemental data*. That is, values of data attributes – as in our example *type of university* and *course (field of study)* –, which are relevant for quantifying timeliness and improving estimations, cannot be considered at all so far. Concluding, further research is needed to design a metric for timeliness that supports supplemental data in a methodically well-founded way.

3 DESIGNING A NOVEL METRIC FOR TIMELINESS

In the following, we take the metric for timeliness defined by Heinrich et al. (2007a) as starting point. We do so because the metric is based on probabilistic theory and its value can be interpreted as the probability that the considered attribute value ω is still up-to-date. Assuming a finite shelf life for an attribute value, the value of the metric and consequently the probability decrease when the age t ($=t_1-t_0$) increases (and vice versa). In contrast, an infinite shelf life would lead to a situation that an attribute value acquired once is always up-to-date and never changes, which refers to a quantification under certainty and is therefore quite simple.

We generalize the metric proposed by Heinrich et al. (2007a) as follows: The metric quantifies the timeliness of an attribute value ω, which is characterized by the corresponding real world counterpart's instant of creation t_0. Together with the instant t_1 of quantifying DQ (with $t_1 \geq t_0$), it is possible to determine the age t of the attribute value ω: $t = t_1 - t_0$. The limited shelf life $T \in R^+$ is unknown and therefore defined as a continuous random variable. Consequently, timeliness is defined as the probability that the shelf life T is greater than or equal to the age t of the attribute value ω. Given the probability distribution function $F^{\omega}(t) := P^{\omega}(T \leq t)$ of the shelf life T, we define the metric for timeliness as follows (The distribution function can be determined in several ways. This is discussed below):

$$Q^{\omega}_{Time.}(t) := P^{\omega}(T \geq t) = 1 - P^{\omega}(T < t) = 1 - F^{\omega}(t) \tag{1}$$

In the particular case of an exponential distribution, which is a typical distribution for lifetime and has already proven to be useful in quality management (especially for address data etc.), Heinrich et al. (2007a) define the metric as shown in (2). Assuming that the attribute value ω is correct at the instant t_0' of its acquisition, we may use this instant t_0' to calculate the age t instead of the corresponding real world counterpart's instant t_0 of creation. This is because the exponential distribution is memoryless in the following way: $P(X \geq x + t \,|\, X \geq x) = P(X \geq t)$; i.e. the conditional probability that the attribute value becomes outdated in the next period of time is independent of its current age.

$$Q^{\omega}_{Time.}(t) := \exp(-decline(A) \cdot t) \tag{2}$$

The parameter $decline(A)$ is the decline rate indicating how many of the attribute's values become outdated on average within one period of time. For example, a value of $decline(A) = 0.2$ has to be interpreted like this: on average 20% of the attribute A's values lose their validity within one period of time. Obviously, the definitions (1) and (2) do not support the use of supplemental data w_i ($i = 1, \ldots, n$). Therefore, values of data attributes like *type of university* and *course* cannot be considered at all when quantifying timeliness of the attribute value "student", for example.

To solve this problem, we developed the following idea: We redefine the metric for timeliness in term (1) to represent the *conditional* probability that the considered attribute value ω is still up-to-date. Using the supplemental data w_i as conditions $W_1 = w_1, \ldots, W_n = w_n$ when calculating the probability is a well-founded way to consider them. The values of the variables W_i (i.e. w_i) are known (they are stored in the database) and thus need not be modelled as random variables. However, they usually are subject to temporal decline as well. Hence, it is advantageous to model them – without loss of generality - as random variables, too. Given the distribution function $F^{\omega}(t|w_1, \ldots, w_n) := P^{\omega}(T \leq t | W_1 = w_1, \ldots, W_n = w_n)$ with the supplemental data w_i ($i = 1, \ldots, n$) we define the new metric for timeliness $Q^{\omega}_{Time.}(t, w_1, \ldots, w_n)$ as follows:

$$\begin{aligned} Q^{\omega}_{Time.}(t, w_1, \ldots, w_n) &:= P^{\omega}(T \geq t \,|\, W_1 = w_1, \ldots, W_n = w_n) = 1 - P^{\omega}(T < t \,|\, W_1 = w_1, \ldots, W_n = w_n) \\ &= 1 - F^{\omega}(t \,|\, w_1, \ldots, w_n) = 1 - \int_0^t f^{\omega}(\theta \,|\, w_1, \ldots, w_n) d\theta \end{aligned} \tag{3}$$

The conditional probability (=value of the metric) is calculated based on the complementary probability $P^{\omega}(T < t | W_1 = w_1, \ldots, W_n = w_n)$ – which represents the probability that the attribute value is outdated at the instant t_1 of quantifying DQ ($T < t = t_1 - t_0$) – and the distribution function $F^{\omega}(t|w_1, \ldots, w_n)$. Thereby, the conditional distribution function is defined as the integral over the conditional probability density function $f^{\omega}(\theta|w_1, \ldots, w_n)$. This function, in turn, is determined by the quotient of the combined probability density functions $f^{\omega}(\theta, w_1, \ldots, w_n)$ and $f^{\omega}(w_1, \ldots, w_n)$. As the complementary probability represents whether the attribute value ω is outdated before the age t is reached, the definite integral is calculated for the interval $[0; t]$. This in mind, we can calculate the probability in our example that the stored attribute value "student" is still up-to-date for a certain customer considering supplemental data

(see next section). Before, we briefly evaluate whether the novel metric meets the requirements defined by Heinrich et al. (2007b).

The definition as a *conditional* probability ensures that the values of the novel metric are *normalized* to [0; 1]. Moreover, the metric is equal to one for attribute values with age t=0 (i.e. t_1=t_0): $Q^{\omega}_{Time.}(0, w_1,..., w_n) = 1$. This is reasonable because the attribute value ω is correct at the corresponding real world counterpart's instant t_0 of creation (see definition of timeliness). Moreover, the values of the metric are limited to zero – due to their limited shelf life T, the following equation holds: $\lim_{t\to\infty} Q^{\omega}_{Time.}(t, w_1,...w_n) = 1 - \lim_{t\to\infty} F^{\omega}(t \mid w_1,..., w_n) = 0$. Thereby, a value of zero means, that the attribute value ω is certainly outdated. Based on probability theory, the values of the metric are *interval scaled* and *interpretable* (as a probability). Moreover, the *aggregation* formulas defined by Heinrich et al. (2007b) can be applied as well. As mentioned earlier, the timeliness of an attribute value (e.g. professional status of a particular customer) can be calculated automatically to a large extent by using the formula above as well as SQL DML statements. This ensures that the metric meets *Feasibility*. The weighting factors in the aggregation formulas and designing the metric according to the shelf life of the attribute values support *Adaptivity*. However, the *Adaptivity* of the novel metric could be further improved – related to the metric proposed by Heinrich et al. (2007a) – by integrating supplemental data. In the next section, we illustrate this advantage by an example of a financial services provider.

4 PRACTICAL EVALUATION OF THE METRIC

In this section, we evaluate our metric for timeliness by means of a real use situation. Thereby, we analyse its applicability and practical benefit. We applied the metric at a German financial services provider (FSP) and especially focus on the question of how it can support the process of customer valuation. In the context of customer valuation, the customer lifetime value (CLV) is a widely accepted approach to valuate customers and is defined as the present value of all existing and future cash flows generated with a certain customer (cp. e.g. Berger et al. 1998). It is obvious that a high CLV has to be assessed very critically, when the input data used for calculation are (possibly) outdated. Thus, quantifying DQ is an important issue in this context. Due to confidentiality, all data are anonymised and modified. But the principal results still hold.

The FSP acts as an independent company aiming at advising its customers (mostly academics) holistically during a large part of their life cycle. Thus, the CLV is a starting point for many decisions and plays an important role. CLV calculation is based on several input data like customer's *age*, current *professional status* (e.g. student) and *course* (e.g. engineering sciences). Such data attribute values – which were often acquired many years ago - are usually stored in the customer database.

In order to calculate CLVs, the FSP assumes that every customer passes through different phases of a typical customer life cycle. This customer life cycle starts at the instant of career entry because at this instant the FSP can start selling various products and services. The FSP then tries to determine a customer's (typical) demand for each of these phases (e.g. retirement). On this base, it is possible to estimate the cash flows resulting from selling products and services in each phase and to calculate the CLV as an expected net present value. In this way, the cash flows and CLVs have already been quantified for several target groups during a former project. The results are a starting point for our example.

However, the FSP also focuses on students in order to acquire them before they become career starters (then their demand is usually especially high). Therefore, the FSP has a very large number of customers being students. These are stored in the database with *professional status* "student", including *course* and *instant of enrolment*. But how should the CLVs of these customers be quantified? As most of the products cannot be sold to students, the staff of the FSP usually contact these customers quite sporadically (this finally leads to marginal, negative cash flows in this phase). As a result, a student's instant of career entry is typically not known by the FSP. Nevertheless, it would be much too time- and labour-intensive to contact each customer in order to verify the correctness of the stored *profes-*

sional status "student" (real world test). Therefore, it is necessary to quantify timeliness at a high level of automation and to avoid such real world tests. In the following, we describe two customer valuation procedures that were previously used by the FSP. Then, we discuss the application of the metric for timeliness.

According to the first procedure, the FSP calculated the CLV of customers with *professional status* "student" as follows: The attribute value for *instant of enrolment* was used to estimate the remaining duration of study by comparing the duration since the *instant of enrolment* with the average duration of study (in semesters). The latter was determined by using publicly available data about graduates and dropouts (Federal Statistical Office of Germany 2007, Heublein et al. 2003, Heublein et al. 2008): Accordingly, 64% of all students who finally graduated studied about 15.5 semesters on average, while the remaining 36% dropped out after 5.6 semesters on average. The FSP calculated an average duration of study of about 12 semesters (≈0.64*15.5 semesters+0.36*5.6 semesters). For this estimated remaining duration of study, the FSP calculated negative cash flows (note: as the FSP did not put DQ into question, the remaining duration of study was estimated with one semester, if the elapsed time since instant of enrolment was larger than or equal to the average duration of study). Starting from the instant of career entry, the FSP considered the CLV related to its life cycle approach. We briefly illustrate this simple procedure: Considering a customer, who enrolled in economics and social sciences in April 2003, a remaining duration of study of about 2 semesters is assumed by the FSP at the instant of customer valuation in April 2008 (computed by (average duration of study)-(duration since instant of enrolment)≈12-10=2). For this period of time negative cash flows (here: € -150 per semester) were calculated. On the one hand, the CLV of a graduate in economics and social sciences is € 4,000 (net present value according to the lifecycle approach). On the other hand, the CLV of a dropout is only about € 2,000. This leads to a weighted average CLV of 0.64*€ 4,000+0.36*€ 2,000=€ 3,280. Table 2 summarizes the results of this procedure:

Instant of enrolment	Instant of customer valuation	Duration since enrolment	Remaining duration of study (est.)	Cash flows of students (est.)	CLV at the instant of career entry (est.)	Calculated CLV at the instant of customer valuation
Apr. 2003	Apr. 2008	10 semesters	2 semesters	€ -150/semester	€ 3,280	€ 2,554

Table 2. Customer valuation in the example (first procedure)

The example illustrates that the FSP (with a discount rate of 5% per semester) calculated a CLV of $\sum_{i=1}^{2} \frac{€\,-150}{(1+0.05)^i} + \frac{€\,3{,}280}{(1+0.05)^3} \approx €\,2{,}554$ (net present value of the expected cash flows), which is quite low compared to the average CLV at the instant of career entry. The result is based on the implicit assumption that the customer will certainly be studying another two semesters before starting his/her career. All in all, this procedure is quite simple, but does by no means consider DQ aspects.

Therefore, the FSP began to modify the procedure and extended it by probabilistic considerations. First, the assumption of a fixed duration of study (of 12 semesters) for all students was avoided. Instead, a customer's *professional status* "student" may remain or change (to "career entry/graduate" or "career entry/dropout") in each period (semester). All three possible transitions are assigned with probabilities, considering the customer's estimated remaining duration of study. By using this procedure, it is possible that the instant of career entry is before or after the average duration of study. Figure 2 illustrates the modified procedure, which is to large parts based on the concept of homogeneous Markov chains (cp. e.g. Pfeifer et al. 2000). Note: This concept implies that the transition probabilities (e.g. $p_{Stu.,Stu.}$) are identical for all periods considered (e.g. $p_{Stu.,Stu.}(i)=p_{Stu.,Stu.}(i+1)$ for each period i). The FSP calculated the probability $p_{Stu.,Stu.}$ for each customer so that the *expected* number of semesters, for

which a customer remains in *professional status* "student" ($\sum_{i=1}^{\infty} i \cdot p_{Stu.,Stu.}{}^{i} = \frac{p_{Stu.,Stu.}}{(1-p_{Stu.,Stu.})^2}$) was equal to the estimated remaining duration of study. Using this probability-based procedure and calculating the expected value to determine the CLV, it is considered that a customer may finish sooner or later than the average duration of study (variation).

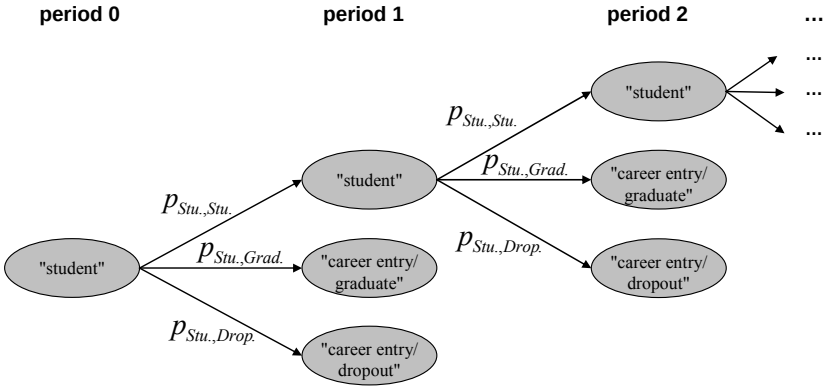

Figure 2. Possible transitions of the professional status "student" (modified procedure)

Taking into account the customer data in Table 2, the transition probability $p_{Stu.,Stu.}$ was determined to 50% (this goes along with an expected remaining duration of study of 2 semesters). Interpreting this probability means that 50 out of 100 customers studying in the 10th semester still remain students in the next semester and so on. The other 50 customers finish after the 10th semester. They were separated – based on the shares mentioned above – in 32 (≈0.64*50) graduates and 18 (≈0.36*50) dropouts to determine $p_{Stu.,Grad.}$ and $p_{Stu.,Drop.}$. Summing up, a CLV of € 2,845 was determined by means of the modified procedure: $\left(\sum_{i=1}^{\infty} \frac{0.50^{i-1} * (0.50 * € -150 + 0.32 * € 4{,}000 + 0.18 * € 2{,}000)}{(1+0.05)^i} \right)$. Though avoiding the assumption of a uniform fixed duration of study (of 12 semesters), the modified procedure does not address DQ issues of input data yet. What is particularly noteworthy is the underlying assumption that each customer with the attribute value "student" is still a student at the instant of customer valuation (in the example in April 2008). Both procedures are based on this assumption. Therefore, the FSP ignored that the stored customer data could already be outdated at the instant of customer valuation. Exactly this was observed by the FSP when some customers were asked in summer 2008 with respect to their *professional status* (cp. ex post analysis below). As a consequence, we analysed how the designed metric for timeliness can be used to meet this issue. The basic idea is as follows:

Our metric for timeliness represents the probability that a customer with the attribute value "student" is still a student in the real world at the instant of customer valuation. This in mind, the values of the metric are appropriate for calculating the transition probabilities $p_{Stu.,Stu.}$, $p_{Stu.,Grad.}$, and $p_{Stu.,Drop.}$. Neglecting the limiting assumption of both former procedures, we first calculated the probability ($Q^{\omega}_{Time.}(t,0) = Q^{\omega}_{Time.}(t)$ that a customer is still studying at the instant of customer valuation (April 2008 = period 0). We did not consider supplemental data in the first step. The value of the metric had to be calculated at the instant of customer valuation considering the probabilities for graduation $P_{Grad.}(0)$ as well as for dropout $P_{Drop.}(0)$ up to that point in time: $Q^{\omega}_{Time.}(t,0) = 1 - (P_{Grad.}(t,0) + P_{Drop.}(t,0))$. We similarly determined the transition probabilities for the following periods (semester). Since conditional probabilities were needed, we calculated $p_{Stu.,Stu.}(t,i)$ for each period i with $i \geq 1$ as follows: $p_{Stu.,Stu.}(t,i) = \frac{Q^{\omega}_{Time.}(t,i)}{Q^{\omega}_{Time.}(t,i-1)}$. In the next step, we determined the transition probabilities $p_{Stu.,Grad.}(t,i)$ and $p_{Stu.,Drop.}(t,i)$ taking into account the values of the metric, too. They represent conditional probabili-

ties for graduation and dropout respectively regarding period i $\left(p_{Stu.,Grad.}(t,i) = \frac{P_{Grad.}(t,i) - P_{Grad.}(t,i-1)}{Q^{\omega}_{Time.}(t,i-1)} \text{ and } p_{Stu.,Drop.}(t,i) = \frac{P_{Drop.}(t,i) - P_{Drop.}(t,i-1)}{Q^{\omega}_{Time.}(t,i-1)} \right)$. Summing up, we get $p_{Stu.,Stu.}(t,i)+p_{Stu.,Grad.}(t,i)+p_{Stu.,Drop.}(t,i)=1$ for each period i (equivalent to $Q^{\omega}_{Time.}(t,i) = 1-(P_{Grad.}(t,i)+P_{Drop.}(t,i))$. Figure 3 illustrates the procedure considering the metric:

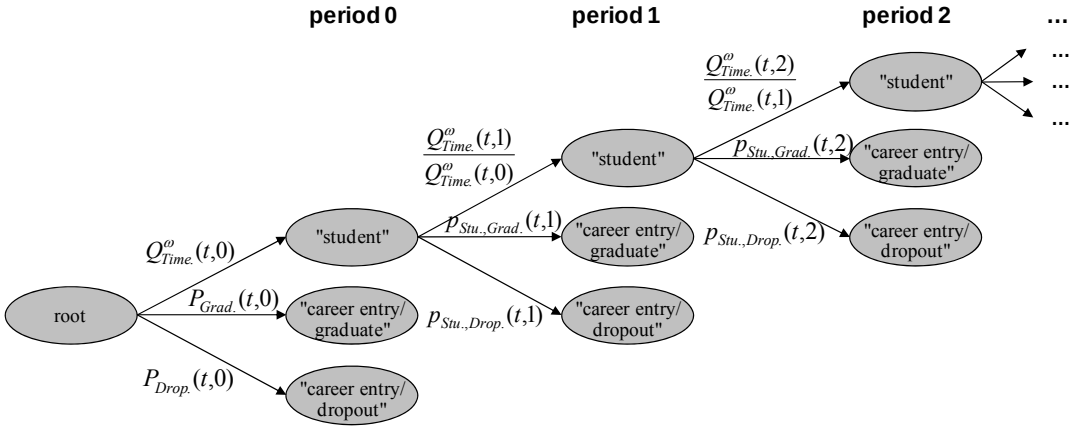

Figure 3. Possible transitions of the professional status "student" (new procedure)

With this procedure, it is possible to consider DQ issues within the FSP's customer valuation.

So far, we did not consider supplemental data - like *type of university*. Figure 1 illustrates that supplemental data are crucial for quantifying timeliness. We therefore adapted the procedure for customer valuation by using the novel metric (cp. term (3)). Considering these supplemental data implies that the probabilities $P_{Grad.}(t,i)$ and $P_{Drop.}(t,i)$ have to be calculated depending on both costumer's *type of university* and *course*. Such *individual* supplemental data have to be used when calculating the conditional probabilities. Not only these probabilities are customer-specific, but also the value of the metric, which is now represented by a conditional probability using the supplemental data as conditions. Based on this, we computed the transition probabilities – see Figure 3 – using the conditional probabilities for $P_{Grad.}(t,i)$ and $P_{Drop.}(t,i)$ as well as the new metric $Q^{\omega}_{Time.}(t,w_1,...,w_n)$. In this context using this metric has the following advantages:

1.) Timeliness of customer data (*professional status* "student") is determined individually for each customer considering *supplemental data* (e.g. *type of university* and *course*). This allows to calculate the CLV systematically und methodically well-founded.

2.) It is possible to determine a customer-specific probability of whether a customer with the professional status "student" is still a student at the instant of customer valuation (in the example in April 2008). This reduces the risk of valuing customers incorrectly.

3.) We can avoid the assumption that all transition probabilities are constant over time (see above). That is, they can be determined individually (and automatically) for each period as well as for each customer. This is very important for customer valuation because the probabilities of a dropout after the 5th and the 9th semester obviously differ in reality.

In the following, we discuss the *development* of the new metric by means of the example in more detail. The attribute value "student" can lose its validity in two ways. Studies are either completed successfully or aborted. For both alternatives, we had to determine the decline rate of the data value "student". This indicates how many data values become outdated on average within one period of time. For the attribute value "student", the decline rates can be determined statistically based on publicly available data (Federal Statistical Office of Germany 2007, Heublein et al. 2003, Heublein et al. 2008).

Considering dropouts for example, the cumulative relative frequencies by Heublein et al. (2008) for each *type of university* and each *course* can be used. Figure 4 shows the cumulative relative frequency distribution of dropouts (in relation to all dropouts) for economics and social sciences at universities. For example, approx. 25% of all dropouts occur within the first two semesters. All in all, the figure shows that the dropout rates with respect to all students of a semester are approx. constant (contrary to the absolute number of dropouts which is obviously decreasing). Hence, we can assume a constant relative decline rate and apply an exponential distribution. Using a least square estimation, we determined an exponential distribution with a decline rate of 0.165 (see Figure 4). The value of the coefficient of determination R^2 was calculated to 0.98, which shows a very good approximation. Therefore, $P_{Drop.}(x)$=0.19*(1-exp(-0.165*x)) denotes the probability that a student has already aborted his/her studies after x semesters. Here, the factor 0.19 corresponds to the fraction of dropouts in relation to all students who have enrolled at universities in economics and social sciences.

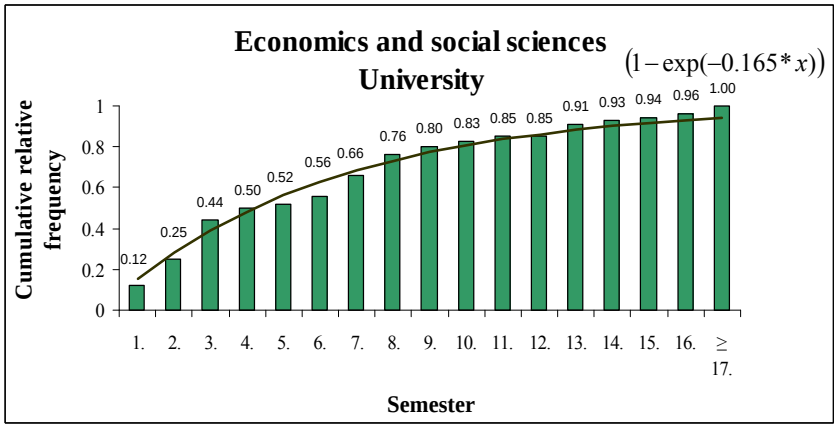

Figure 4. Cumulative relative frequency distribution of dropouts

If we compare this result to the function 0.05*(1-exp(-0.145*x)), which was determined equally for students enrolled at universities in medical sciences, significant differences become apparent. Considering the first five semesters (i.e., x=5) for example, the probability $P_{Drop.}(x)$ for economics and social sciences is 10.7% in contrast to just 2.6% for medical sciences. Such differences completely depend on the different supplemental data with respect to the data attribute *course*. We similarly computed the probability $P_{Grad.}(x)$, which represents the cumulative probability that a customer with *professional status* "student" graduated after x semesters. Here, it was necessary to assume a Weibull distribution (for this distribution see Heinrich et al. 2009). Then, we defined the metric $Q^{\omega}_{Time.}(t, w_1,...w_n)$ based on both probabilities and calculated its values for each customer and period automatically.

Table 3 lists some selected customers with the supplemental data (*type of university* and *course*) from the FSP's customer database as well as the corresponding metric values $Q^{\omega}_{Time.}(t, w_1,...w_n)$.

Customer	Professional status	Instant of enrolment	Type of university	Course	$Q^{\omega}_{Time.}(t, w_1,...w_n)$
A	"Student"	Apr. 2004	University	Engineering sciences	0.58
B	"Student"	Apr. 2004	University of applied sciences	Engineering sciences	0.24
C	"Student"	Apr. 2004	University	Medical sciences	0.92
D	"Student"	Apr. 2004	University	Economics and social sciences	0.46

Table 3. Supplemental data for selected customers and corresponding value of the metric

The differences between the values of the metric highlight the impact of supplemental data (since all other data are the same). For our example customer (see above), we get the following results when applying the procedure based on the new metric: The probability that the customer, who enrolled at a university in economics and social sciences in April 2004, is still a student at the instant of customer valuation (April 2008) is only 46%. In contrast, the probabilities that he/she has already graduated or aborted were calculated to 38% and 16% respectively. When using the previous procedures, it was not possible to determine such probabilities at all. Instead, it was assumed that the customer is still studying in April 2008. A further advantage of the metric is that we do not need to assume that the transition probabilities are constant over time (this is unrealistic). But, it is possible to determine customer-specific transition probabilities for each period automatically. Appling the new metric within the example, we got a CLV of € 3,173 which is more than before due to 1), 2) and 3) (cp. above).

The importance of supplemental data can also be demonstrated by an ex post analysis. After the valuation in April 2008, the FSP instructed its sales staff to ask about the current *professional status* of customers who were stored as "student" in the database. For all customers who had already graduated or dropped out the staff had to acquire the instant of graduation or dropout. Otherwise *professional status* "student" was confirmed. Similar information was requested by the FSP in a campaign starting in May 2008. The FSP wanted to know from customers with *professional status* "student" whether and when their status had changed. All in all, the FSP and its staff changed the attribute value "student" for 1,510 customers until the end of August 2008. We analysed these customers by comparing their actual instant of graduation or dropout with the results and estimations of each procedure accomplished. According to the first and simple procedure the FSP assumed an average duration of study of 12 semesters. Thus, for each of the 1,510 customers we could determine when he/she would have been expected to finish his/her studies. Comparing these instants with the actual semester of graduation or dropout, we found that these conformed for only 130 out of 1,510 customers. In other words, in 91.4% of all cases, the estimation was actually incorrect. We also analysed the other probability-based procedures. For every instant of enrolment, we determined the corresponding number of students. On this basis, we calculated how many students would have been expected to graduate and dropout in each of the following semesters using the transition probabilities (see Figures 2 and 3). An example illustrates this: 157 customers out of all 1,510 customers enrolled in October 2002. With the transition probabilities $p_{Stu.,Stu.}(0,1)=0.86$, $p_{Stu.,Grad.}(0,1)=0$, and $p_{Stu.,Drop.}(0,1)=0.14$, we get no customer who expectedly graduates in the first semester (until Feb. 2003), 22 customers who dropout and 135 customers who continue their studies. We did such calculations for each probability-based procedure and compared the expected numbers with the actual numbers acquired by the FSP. This way we calculated the difference between the expected frequency distribution of each procedure and the actual frequency distribution of all 1,510 selected customers (number of faults=1,510-number of customers where actual and expected semester corresponded).

The results of the ex post analysis were the following: The second procedure, which does not consider DQ issues, had 1,136 faults (75.2%). For the third procedure, which includes the metric for timeliness without supplemental data, these faults could be reduced to 892 (59.1%). Finally, by using the last procedure, which is based on the novel metric for timeliness considering supplemental data (*type of university* and *course*), these results could be further improved to 710 faults (47.0%). Table 4 summarizes the findings of the ex post analysis (the number of faults is quite high for all procedures because we count every "minor" difference in terms of one semester as a fault):

Number of analysed customers	Number of faults first procedure without DQ	Number of faults second procedure without DQ	Number of faults existing DQ metric	Number of faults novel DQ metric
1,510 customers (100.0%)	1.380 customers (91.4%)	1,136 customers (75.2%)	892 customers (59.1%)	710 customers (47.0%)

Table 4. Ex post analysis

The analysis shows that using the metric for timeliness and considering supplemental data obviously improve the results. Focusing on DQ issues, we did not evaluate the CLVs, but the metrics for timeliness. Here, using the novel metric (instead of existing approaches) allows estimating transition probabilities better and therefore creates practical benefit to a significant extent.

Summing up, it can be stated that the example was intentionally kept simple in order to illustrate the practical benefit. It is possible to represent much more difficult issues considering further statuses besides "student", "career entry/graduate", and "career entry/dropout". For customers whose *professional status* was acquired a long time ago other statuses seem to be reasonable (e.g. "junior consultant" or "senior consultant" instead of "career entry/graduate"). In addition, the new metric is valuable for determining the instant of career entry because such information can be used to contact customers in a more goal-oriented way. Figures like average duration of study are bad estimates here, especially considering good students. These students who are often characterised by a short duration of study would be contacted too late, though being more attractive for the FSP in many cases. Nevertheless, the simple example illustrates that quantifying DQ helps to improve customer valuation.

5 SUMMARY AND CONCLUSIONS

In this paper, we presented a novel metric for timeliness that is based on probabilistic considerations. Extending an existing approach, we take supplemental data into account and define the metric as the conditional probability that a considered data value is still up-to-date at the instant of quantifying DQ. Hence, quantifying timeliness can be done in a more accurate and methodically well-founded way. The metric's practical benefits as well as its applicability were illustrated by a real use situation.

In contrast to previous approaches, the new metric allows us to consider supplemental data. Here we work with conditional probabilities using supplemental data as conditions of the probability density function. The example illustrates the importance of supplemental data like *type of university* and *course*. Moreover, when designing the metric, we were able to avoid limiting assumptions concerning the probability distribution (a number of possible probability distributions is presented by Heinrich et al. 2009). This assures that the metric is appropriate for many attributes and their individual characteristics (such as constant, increasing or decreasing decline rates). In practice, the metric is valuable for calculating decision variables like the CLV. If companies rely on the CLV in order to manage customer relationships, outdated customer data may result in wrong decisions. The same holds for other fields of application (e.g. mobile services provider (cp. Heinrich et al. 2007a, b, 2009), financial services provider (cp. Heinrich et al. 2008), where such metrics for Quality of Conformance could be applied successfully.

Besides these findings, calculating the probability distribution function can be difficult in some cases: On the one hand, publicly available data (e.g. from Federal Statistical Offices or scientific institutions) can be applied to define the metric. On the other hand, internal data (e.g. from the data warehouse) may be analysed using statistical software such as SPSS to derive the probability distribution function. Moreover, interviews (as the FSP from the example did) and experts' estimations are further instruments. However, quantifying correctness by means of a real world test for every single attribute value – which is an alternative to quantifying timeliness – is usually much more cost-intensive. Additionally, it has to be considered that a metric, which was developed once, can be reused frequently or adapted to several fields of application. The authors are working currently on a model-based economic approach for planning DQ measures. For implementing such a model, adequate DQ metrics are necessary. The approach presented here provides a basis for those purposes. Nevertheless, further metrics for other DQ dimensions should be developed and further research in this area is encouraged.

References

Al-Hakim, L. (2007). Information quality factors affecting innovation process. International Journal of Information Quality, 1 (2), 162-176.

Ballou, D.P. and Pazer, H.L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. Management Science, 31 (2), 150-162.

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

Ballou, D.P., Wang, R.Y., Pazer, H.L. and Tayi, G.K. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality. Management Science, 44 (4), 462-484.

Batini, C. and Scannapieco M. (2006). Data Quality. Concepts, Methodologies and Techniques. Springer, Berlin.

Berger, P.D. and Nasr-Bechwati, N. (1998). Customer Lifetime Value: Marketing Models and Applications. Journal of Interactive Marketing, 12 (1), 17-30.

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

Federal Statistical Office of Germany (2007). Education and Culture - Statistics of Exams of Higher Education, series 11, part 4.2. Accessed: 2008/11/01, http://www.destatis.de

Fisher, C.W., Chengalur-Smith, I.N. and Ballou, D.P. (2003). The Impact of Experience and Time on the Use of Data Quality Information in Decision Making. Information Systems Research, 14 (2), 170-188.

Harris Interactive (2006). Information Workers Beware: Your Business Data Can't Be Trusted. Accessed: 2008/11/01, http://www.businessobjects.com/news/press/press2006/20060626_data_quality_survey_comp.asp

Heinrich, B. and Helfert, M. (2003). Analyzing Data Quality Investments in CRM - A model-based approach. Proceedings of the 8th International Conference on Information Quality, Cambridge/Boston, S. 80-95.

Heinrich, B., Kaiser, M. and Klier, M. (2007a). How to measure data quality? – a metric based approach. Proceedings of the 28th International Conference on Information Systems (ICIS), Montreal.

Heinrich, B., Kaiser, M. and Klier, M. (2007b). DQ metrics: a novel approach to quantify timeliness and its application in CRM. Proceedings of the 12th International Conference on Information Quality (ICIQ), Cambridge/Boston, 431-445.

Heinrich, B., Kaiser, M. and Klier, M. (2008). Does the EU Insurance Mediation Directive help to improve Data Quality? - A metric-based analysis. Proceedings of the 16th European Conference on Information Systems (ECIS), Galway.

Heinrich, B., Kaiser, M., Klier, M. (2009). A Procedure to Develop Metrics for Currency and its Application in CRM. Accepted for ACM Journal of Data and Information Quality.

Heublein, U., Schmelzer R. and Sommer D. (2008). Die Entwicklung der Studienabbruchquote an den deutschen Hochschulen: Ergebnisse einer Berechnung des Studienabbruchs auf der Basis des Absolventenjahrgangs 2006. Accessed: 2008/11/01, http://www.his.de/publikation/archiv/X_Pub

Heublein, U., Spangenberg, H. and Sommer, D. (2003). Ursachen des Studienabbruchs, Hamburg.

Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, 28 (1), 75-105.

Hinrichs, H. (2002). Datenqualitätsmanagement in Data Warehouse-Systemen. Oldenburg.

Jiang, Z., Sarkar, S., De, P. and Dey, D. (2007). A Framework for Reconciling Attribute Values from Multiple Data Sources. Management Science, 53 (12), 1946-1963.

Juran, J.M. (1998). How to think about Quality. Juran's Quality Handbook. McGraw-Hill, New York.

Klein, B.D. and Callahan, T.J. (2007). A comparison of information technology professionals' and data consumers' perceptions of the importance of the dimensions of information quality. International Journal of Information Quality, 1 (4), 392-411.

Knight, S. and Burn, J. (2005). Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science Journal, 8 (3), 159-172.

Lee, Y.W., Strong, D.M., Kahn, B.K. and Wang, R.Y. (2002). AIMQ: a methodology for information quality assessment. Information & Management, 40 (2), 133-146.

Parssian, A., Sarkar, S. and Jacob, V.S. (2004). Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. Management Science, 50 (7), 967-982.

Pfeifer, P.E. and Carraway, R. L. (2000). Modelling Customer Relationships as Markov Chains. Journal of Interactive Marketing, 14 (2), 43-55.

Pipino, L., Lee, Y.W. and Wang, R.Y. (2002). Data Quality Assessment. Communications of the ACM, 45 (4), 211-218.

Redman, T.C. (1996). Data Quality for the Information Age. Artech House, Boston.

Price, R., Shanks, G. (2005). A semiotic information quality framework: development and comparative analysis. Journal of Information Technology, 20 (2), 88-102.
Teboul, J. (1991). Managing Quality Dynamics. Prentice Hall, New York.
Wang, R.Y. (1998). A Product Perspective on Total Data Quality Management. Communications of the ACM, 41 (2), 58-65.
Wang, R.Y., Storey, V.C. and Firth, C.P. (1995). A Framework for analysis of data quality research. IEEE Transactions on Knowledge and Data Engineering, 7 (4), 623-640.
Wang, R.Y. and Strong, D.M. (1996). Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems, 12 (4), 5-33.
Wand, Y. and Wang, R.Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39 (11), 86-95.
Yu, Z. and Wang, Y. (2007). An empirical research on non-technical factors related to statistical data quality in China's enterprises. International Journal of Information Quality, 1 (2), 193-208.