Decision Support Systems 72 (2015) 82–96

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Metric-based data quality assessment — Developing and evaluating a probability-based currency metric

Bernd Heinrich *, Mathias Klier [1]

Department of Management Information Systems, University of Regensburg, Universitätsstraße 31, D-93040 Regensburg, Germany

ARTICLE INFO

Article history:
Received 3 January 2014
Received in revised form 8 February 2015
Accepted 9 February 2015
Available online 16 February 2015

Keywords:
Data quality
Data quality assessment
Data quality metric
Currency of data

ABSTRACT

Data quality assessment has been discussed intensively in the literature and is critical in business. The importance of using up-to-date data in business, innovation, and decision-making processes has revealed the need for adequate metrics to assess the currency of data in information systems. In this paper, we propose a data quality metric for currency that is based on probability theory. Our metric allows for a reproducible configuration and a high level of automation when assessing the currency of attribute values. The metric values represent probabilities and can be integrated into a decision calculus (e.g., based on decision theory) to support decision-making. The evaluation of our metric consists of two main steps: (1) we define an instantiation of the metric for a real-use situation of a German mobile services provider to demonstrate both the applicability and the practical benefit of the approach; (2) we use publicly available real world data provided by the Federal Statistical Office of Germany and the German Institute of Economic Research to demonstrate its feasibility by defining an instantiation of the metric and to evaluate its strength (compared to existing approaches).

1. Introduction

Data quality issues are discussed intensively in the literature and are critical in business. High-quality data in information systems (IS) are needed as a basis for business, innovation, and decision-making processes [1,38]. Thus, poor data quality often results in bad decisions and economic losses [12,15,18,43]. In addition, great effort is required to ease or solve data quality problems [11,30]. The growing relevance of data quality has also revealed the need for adequate assessment (e.g., [30,43]). Quantifying data quality (e.g., quality of customer data) is essential for taking into account data quality aspects in decision-making (e.g., to select the customers to be addressed in a mailing campaign considering the quality of the address data stored). Moreover, assessing data quality constitutes an indispensable step toward the ability to decide whether a data quality measure (e.g., address data cleansing) should be taken from an economic perspective. In this context, it is necessary to quantify and consider the effects of measures with respect to the data quality level — a fact that is often illustrated as part of the data quality loop (for details cf. e.g., [22]).

A recent report (cf. [43]) has revealed that one of the most common data defects is outdated data, which primarily results in wasted budgets, loss of potential customers, and reduced customer satisfaction. For example, two-thirds of surveyed organizations observe several problems in the context of customer relationship management, such as sending mailings to the wrong address or sending the same mailing to the customer multiple times; this indicates that outdated address and customer data negatively affect customer perceptions. Indeed, several investigations have shown that time-related aspects (e.g., up-to-date data) are particularly important in data quality management [31,47].

Despite their relevance for theory and practice, however, there is still a lack of well-founded and applicable data quality metrics to assess the currency of data in IS. Therefore, we state the following research question:

How should a metric be defined to assess the currency of data in IS?

To contribute to this question, we propose a probability-based currency metric (PBCM). By means of this metric, information about the currency of the assessed data can be considered in decision-making and add value in terms of better decisions. Indeed, the PBCM can also be seen as a possible basis for integrating data quality aspects in the theoretical framework of the value of information and particularly its probability-based normative concept [7,28,32,36,45].

The remainder of the paper is organized as follows. Section 2 illustrates the problem context and provides an overview of prior works. In Section 3, we develop a metric that is based on probability theory. The evaluation in Section 4 consists of two steps. First, we instantiate the metric for a real-use situation at a mobile service provider to demonstrate both its applicability and practical benefit. Second, we

* Corresponding author. Tel.: +49 941 943 6100; fax: +49 941 943 6120.
E-mail addresses: bernd.heinrich@wiwi.uni-regensburg.de (B. Heinrich), mathias.klier@wiwi.uni-regensburg.de (M. Klier).
[1] Tel.: +49 941 943 6102; fax: +49 941 943 3211.

http://dx.doi.org/10.1016/j.dss.2015.02.009
0167-9236/

use publicly available real world data to demonstrate feasibility by defining an instantiation of the metric and to evaluate its strength. Finally, we summarize, reflect on the results, and provide an outlook on future research.

2. Background

First, we provide some basic definitions and present the problem context. We then discuss existing contributions with respect to assessing the data quality dimension currency and identify the research gap.

2.1. Basic definitions and problem context

Parssian et al. [40, p. 967] use the terms information quality and data quality to "characterize mismatches between the view of the world provided by an IS and the true state of the world" (for a similar definition cf. [39]). We take this definition as a basis. Data quality is a multi-dimensional construct [34,44] comprising several dimensions such as accuracy, completeness, currency, and consistency (for an overview cf. [48]). Each dimension provides a particular view on the quality of attribute values in IS. We focus on currency and investigate how to assess this dimension by means of a metric.

Due to its relation to accuracy, we briefly discuss this data quality dimension in a first step. Afterwards, we define currency and delimit it from accuracy. Many authors (e.g., [4,44]) define accuracy as the closeness of an attribute value stored in an IS to its real world counterpart. Usually, comparison or distance functions are used to determine the closeness of the attribute value with respect to its real world counterpart [4]. The assessment of accuracy involves a real world test that constitutes a direct evaluation, for example by means of a survey or interview (cf. e.g., [49]). Thus, both the stored attribute value and its real world counterpart are known when assessing accuracy. In contrast to the widely accepted definition of accuracy, the definitions of time-related data quality dimensions are much less uniform in the literature (cf. [3,4,8]). To express and specify time-related aspects, a number of different terms are used such as currency, timeliness, staleness, up-to-date, freshness, and temporal validity. Some contributions use different terms to define very similar or equal concepts while others use the same term describing different concepts. Ballou et al. [2], for instance, refer to currency as the age of an attribute value at the instant of assessment. They use the term timeliness to describe whether "the recorded value is not out of date" [2, p. 153]. In contrast, Batini and Scannapieco [4, p. 29] highlight that "currency concerns how promptly data are updated". Other authors such as Redman [44, p. 258] state that currency "refers to a 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". A similar definition is proposed by Nelson et al. [37]. Cho and Garcia-Molina [9, p. 3] address an analog concept but use the term up-to-date to express that previously stored values "equal those of their real-world counterparts". Xiong et al. [51, p. 952] also refer to a similar concept as that discussed by Redman [44] and Nelson et al. [37] but use the terms fresh and freshness, stating that "a real-time data object is fresh (or temporally valid) if its value truly reflects the current status of the corresponding entity in the system environment". This brief discussion illustrates that there is no widely accepted definition of such time-related data quality dimensions. As we primarily build upon the definitions of Redman [44] and Nelson et al. [37], we also use the term currency and clearly define the concept behind it for our context.

At its heart, currency expresses whether an attribute value that was stored in an IS in the past is still the same as the value of that attribute in the real world at the instant of assessment (i.e., in the present). This means that the attribute value, which was accurate when it was initially captured (Scenario A), updated (Scenario B), or acknowledged (Scenario C), is still the same as the current value of that attribute in the real world at the instant when its data quality is assessed. Currency explicitly focuses on the temporal decline of a stored attribute value. To illustrate this focus, we clarify the Scenarios A to C (cf. Fig. 1).

Scenario A shows the basic case: An attribute value was initially captured at the instant t_0' (i.e. the accurate value was stored in the IS). The instant of creation of its real world counterpart is represented by t_0. Here, it must be assessed whether the stored attribute value is still the same as the value of that attribute in the real world at the instant of assessment t_1. Thus, the question arises whether the real world counterpart has changed (which is unknown) since the attribute value was captured at t_0'. In Scenario B, it is known that the real world counterpart changed (e.g., at the instant t_{-0}). The stored attribute value was therefore updated accordingly at the instant t_0''. In Scenario C, the stored attribute value was acknowledged at t_0'', as no changes had been made to the real world counterpart. Both additional known instants (update and acknowledgement) can be useful (see below) when assessing the currency of the attribute value at the instant t_1.

When assessing the accuracy at the instant t_1, a real world test is needed. The result represents a statement under certainty. In contrast to assessing accuracy, assessing currency does not involve a real world test. Instead, a metric for currency delivers an indication, not a verified statement, as to whether an attribute value has changed in the real world since the instant it was captured, updated, or acknowledged.

An assessment of currency seems to be helpful in the following settings:

(a) Unknown shelf life of the considered attribute value: Assessing currency is helpful if the shelf life of the considered attribute value is unknown. Otherwise, the attribute value's currency can trivially be determined under certainty. The shelf life is defined as the length of time the stored attribute is still the same as the value of that attribute in the real world. In Scenario B, for example, the attribute value was created at the instant t_0 and changed at the instant t_{-0}. Hence, the attribute value's shelf life was t_{-0}–t_0. Possible application settings include customer master data such as name, address, phone number, marital status, number of children, profession, educational background, employer, and income, as the shelf life of the respective stored attribute values is usually unknown. However, even in the case of real world objects with a rather fixed shelf life such as credit cards that have a validity period of two years, for example, the shelf life of single attribute values may be unknown as the credit cards can become invalid earlier due to events such as withdrawal, theft, or loss of creditworthiness. Therefore, assessing currency can even be helpful in such cases. In addition to customer master data, the shelf life of product data, transaction data, and project data may also be unknown, which leads to further promising fields of application. The processes in production planning, for example, are typically based on data from a variety of internal and external sources (e.g., from suppliers or manufacturing partners) with the results strongly depending on the quality of the data used. In this context, the shelf life of the attribute values is unknown as well and assessing currency can provide helpful indications of whether the stored data are still the same as in the real world.

(b) Real world test not possible or time-consuming or cost-intensive: In the event the shelf life of an attribute value is unknown (cf. (a)), one can propose a real world test that directly compares the stored attribute value and the value of that attribute in the real world. However, such a real world test is often not practicable or too time-consuming and cost-intensive, for example, when customers have to be surveyed. For instance, analyses of data from a firm with more than 20 million customers show that every year about 2 million customers change their place of residence, 230,000 die, and 60,000 get divorced [46]. In this case, it would be very cost-intensive and impractical to regularly survey all customers in order to assess accuracy and