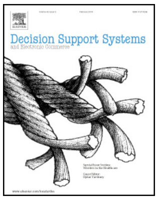
Decision Support Systems xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

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



Development and evaluation of a continuous-time Markov chain model for detecting and handling data currency declines

Yuval Zak, Adir Even *

The Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, P.O.B. 653, Beer-Sheva 8410501, Israel

ARTICLE INFO

Article history:
Received 14 July 2016
Received in revised form 9 September 2017
Accepted 20 September 2017
Available online xxxx

Keywords:
Data quality management
Data currency
Continuous-time Markov chain

ABSTRACT

Data currency declines, caused by recorded data values becoming outdated, can damage the usability and accountability of data resources. Detecting and updating outdated values may improve data currency and reduce the associated damage, but such efforts may be costly and cannot always be justified. This study models currency decline scenarios using a continuous-time Markov chain stochastic process with a finite number of states, each reflecting a valid data value. The model considers state transition probabilities, transition time distributions, and the tradeoff between the damage associated with outdated data and the cost of reacquisition. The proposed formulation permits the currency level to be estimated without having to rely on a baseline for comparison, as well as the prediction of future currency declines, assessment of their accumulated damage, and optimization of the timing of cost-effective data auditing and reacquisition. The study introduces a comprehensive evaluation of the proposed model, using a large real-world dataset relating to the handling of insurance claims over multiple time periods. The evaluation results highlight the applicability of the model, and its potential contribution to proactive data quality management and cost-effective handling of currency declines.

1. Introduction

In the context of Data Quality (DQ) management, *currency* reflects the degree to which data values are recent and up to date, considering the time-lag since their acquisition. Currency declines, and their hazardous impact on data accountability and usability, have long captured the attention of DQ management research and practice. Detecting and correcting discrepancies between previously acquired data and the correct real-world state often requires a baseline for comparison, e.g., the actual real-world state, if known, or another reliable data source. Obtaining such a correct baseline, if possible at all, may involve major efforts and costs, often beyond the organization's resource constraints and budget limitations.

Motivated by that challenge, the aim of this study is to develop a continuous-time Markov chain (CTMC) model that reflects a common mechanism behind currency declines: a failure to detect real-world state transitions and update the data accordingly, due to the high costs of doing so. The model targets data acquisition scenarios in which a real-world entity may reside in one of a finite number of states, each described by a set of data-attribute values. With some necessary adaptations, the proposed model treats state transitions as a stochastic process that has some typical CTMC characteristics. An assessment of whether a certain data record must be revisited, and possibly reacquired, considers state-transition probabilities, transition-time distributions, and tradeoffs between the damage associated with outdated data and the cost of data reacquisition.

The solutions derived from the proposed model support some key DQ-management tasks, which can be associated with the broadly-accepted Total Data Quality Management (TDQM) framework [1]. TDQM promotes proactive DQ management: rather than a one-time data correction effort, DQ improvement should be managed as an ongoing cycle of definition, measurement, analysis and improvement stages. Through in-depth understanding of root causes behind DQ defects, DQ management must aim at predicting their future formation, taking preventive measures, and optimizing DQ improvement policies accordingly. While adhering to TDQM principles and stage definitions, this study makes a few contributions to that end:

a) *Definition*: Currency is often defined in a reactive manner – *To what extent is the data under evaluation still up-to-date*, *considering the time-lag since data acquisition*? Adopting and extending the approach taken in some previous work, this study defines currency in probabilistic, rather than deterministic, terms. Further, it argues that proactive DQ management should redirect the scope of assessment, asking instead, *What is the likelihood that the data under evaluation will remain up-to-date in the future*?
b) *Measurement*: Currency, similar to other DQ metrics, is often measured as a [0,1] ratio of non-defective data values. Using the proposed CTMC model, the likelihood of currency decline is formulated

[*] Corresponding author.
E-mail address: adireven@bgu.ac.il (A. Even).

http://dx.doi.org/10.1016/j.dss.2017.09.006
0167-9236/© 2017 Elsevier B.V. All rights reserved.

as a function of the time-gap since data acquisition. The proposed formulation permits convenient currency estimation, without necessarily having to rely on a baseline for comparison.

c) *Analysis*: Considering a newly acquired value, the proposed formulation estimates the likelihood that, at a certain later point in time, this value will still reflect the correct real-world state. Given that prediction, the model determines the expected time until data will become outdated, due to possible real-world state transitions.

d) *Improvement*: Beyond prediction, the model also links time-lag effects and currency declines to potential benefits and costs. Thus, the model can guide cost-effective data reacquisition policies, prescribe DQ improvement actions, and recommend optimal timings for them.

To ascertain the feasibility and validity of the proposed model, the study evaluated it using a large-scale real-world dataset relating to claim-handling for insurants who suffered work-related injuries. An optimal claim-handling process requires strict maintenance of correct insurant data at all times; hence, the issue of data currency can be associated with substantial cost–benefit tradeoffs. Ensuring the currency of this dataset may turn out to be expensive and time-consuming, as claim-handling representatives must often call insurants or even meet them in person to verify their data. Given their inherent time and workload constraints, representatives must often prioritize their contact efforts and, as stated by their managers, their heuristics-based prioritization practices are far from being optimal. The costs associated with contacting an insurant are often wasteful, if no updates to the data are required. On the other hand, failures to record and reflect real-world transitions in insurants' states often lead to major revenue losses. Evaluation of this business scenario and the associated datasets shows that the proposed model can help to optimize the timing of a call to an insurant, thereby making substantial cost savings.

The next section sets out the background for the model development, and underscores its contribution alongside previous work that has addressed the challenges associated with currency declines from different perspectives. This is followed by the development of the proposed model and its evaluation within real-world settings. Finally, the concluding section summarizes the study, states its key contributions, highlights its limitations, and proposes possible directions for future research.

2. Background

A plethora of DQ studies have explored currency decline from different perspectives, which can be associated with the different stages of the TDQM framework, namely definition, measurement, analysis and improvement. Similarly to this study, some have applied Markov chain modeling techniques to understand the root causes behind currency declines and assess their impact. This section reviews their influences on the concepts and tools used by this study, and highlights its added contribution.

2.1. Currency definition and measurement

Currency is often discussed within the broader terminology of *DQ dimensions*, which reflect the differences between various forms of DQ defects and their implications for DQ management [2]. The *Currency* dimension (also termed *Timeliness*, *Recency* or *Freshness* in some DQ work) reflects the potential impact of time-lags between real-world transitions, data acquisition, and/or data usage [3,4]. A growing time-lag increases the chance of currency decline, i.e., a greater likelihood that a data value no longer reflects the real-world state correctly [5,6]. The lower the currency of data resources, the lesser is their usefulness and relevance for organizations and decision-makers ([6,7,8,17]).

Heinrich and Klier [4] point out the need to differentiate between *Currency* and *Accuracy*, as the definitions of those two dimensions tend to be inconsistent across DQ research. While the former reflects temporal decline effects, the latter refers to discrepancies between the data-in-hand and the correct real-world state, not necessarily associated with temporal effects. This study adopts Heinrich and Klier's [4] view of currency, as a DQ dimension that expresses whether or not an attribute value, which has been previously acquired correctly, still reflects the corresponding real-world value correctly at the time of evaluation. Notably, currency might decline even if data has been acquired accurately, e.g., due to a failure to reflect real-world transitions correctly by updating and/or reacquiring existing data [8,9], or due to unexpected latencies in data integration processes [10].

DQ measurement scores, reflecting levels of currency and other DQ dimensions, are used for assessing DQ state and informing IT personnel and end-users accordingly [2,11,12]. In conformance with common approaches for defining DQ metrics, a currency metric can be defined as the [0,1] ratio between the number of data items that still reflect the real-world state correctly and the total number of values [2]. The use of such metrics mandates the verification of data values against a reliable baseline, e.g., the actual real-world state, if can be obtained, or another data source that has been validated to be correct. Obtaining a reliable baseline and evaluating data against it might involve major efforts and costs (e.g., [7,13,17]). Rather than relying on a baseline, some studies (e.g., [3,5]) have derived currency-decline proxies, using time-lags measured from database logs or record-level timestamps. However, such proxies fail to capture the inherent uncertainty and complexity of currency decline patterns, and hence might bias currency assessment substantially [6].

Acknowledging those limitations, other studies (e.g., [6,11,14]) have proposed alternative techniques for developing DQ metrics, which consider assessments of uncertainty, hazard probabilities, and/or information value. Arguing that currency, and possibly other DQ dimensions, should be defined and assessed in probabilistic terms, Heinrich and Klier [4] propose probability-based currency metrics (PBCM) that assign a [0,1] probability of data still being current, given the associated distribution functions and the time-lag since acquisition. Heinrich and Hristova [8] develop PBCM that consider multiple state transitions between data acquisition and assessment, and can be applied to discrete-state as well as to continuous-state variables. Similarly, Wechsler and Even [9] estimate the likelihood of future currency decline given the value recorded at the time of acquisition and the number of transition stages.

Adopting Heinrich and Klier's [4] view, this study adheres to the notion of currency as the likelihood of data values still reflecting the correct real-world state, considering the time-lag since data acquisition and possible state transitions occurring during that time period. Similarly to Wechsler and Even [9] and Heinrich and Hristova [8], it uses a Markov chain model as a baseline for developing a metric that estimates a [0,1] likelihood of currency decline (Eq. (2)). However, the decline is expressed in this study in an explicit functional form that reflects a continuous-time variable.

2.2. Currency analysis and improvement

In association with the TDQM framework's stages of analysis and improvement, studies have highlighted possible root causes for currency declines, assessed their negative impact on decision-making and business success, and offered solutions and guidelines for data reacquisition and updating.

In general, the benefits gained by DQ improvement cannot always justify the associated costs, and aiming at perfect DQ is not necessarily optimal from an economic viewpoint [6,11]. One ought to consider the inherent cost–benefit tradeoffs while driving toward cost-effective DQ management, which requires an equilibrium to be found between the benefits associated with DQ improvement (i.e., preventing the damages