Decision Support Systems 50 (2010) 152–163

Contents lists available at ScienceDirect

Decision Support Systems

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

Evaluating a model for cost-effective data quality management in a real-world CRM setting

Adir Even [a,*], G. Shankaranarayanan [b], Paul D. Berger [c]

[a] Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, P.O. Box 653, Beer-Sheva, 84105, Israel
[b] Technology, Operations, and Information Management, Babson College, Babson Park, MA 02457-0310, USA
[c] Marketing Department, Bentley University, Morison Hall, 175 Forest St., Waltham, MA 02452, USA

ARTICLE INFO

Article history:
Received 12 October 2009
Received in revised form 27 July 2010
Accepted 27 July 2010
Available online 6 August 2010

Keywords:
Data quality
Utility
Cost–benefit analysis
Data warehouse
CRM

ABSTRACT

Managing data resources at high quality is usually viewed as axiomatic. However, we suggest that, since the process of improving data quality should attempt to maximize economic benefits as well, high data quality is not necessarily economically-optimal. We demonstrate this argument by evaluating a microeconomic model that links the handling of data quality defects, such as outdated data and missing values, to economic outcomes: utility, cost, and net-benefit. The evaluation is set in the context of Customer Relationship Management (CRM) and uses large samples from a real-world data resource used for managing alumni relations. Within this context, our evaluation shows that all model parameters can be measured, and that all model-related assumptions are, largely, well supported. The evaluation confirms the assumption that the optimal quality level, in terms of maximizing net-benefits, is not necessarily the highest possible. Further, the evaluation process contributes some important insights for revising current data acquisition and maintenance policies.

1. Introduction

Maintaining data resources at a high quality level is a critical task in managing organizational information systems (IS). Data quality (DQ) significantly affects IS adoption and the success of data utilization [10,26]. Data quality management (DQM) has been examined from a variety of technical, functional, and organizational perspectives [22]. Achieving high quality is the primary objective of DQM efforts, and much research in DQM focuses on methodologies, tools and techniques for improving quality. Recent studies (e.g., [14,19]) have suggested that high DQ, although having clear merits, should not necessarily be the only objective to consider when assessing DQM alternatives, particularly in an IS that manages large datasets. As shown in these studies, maximizing economic benefits, based on the value gained from improving quality, and the costs involved in improving quality, may conflict with the target of achieving a high data quality level. Such findings inspire the need to link DQM decisions to economic outcomes and tradeoffs, with the goal of identifying more cost-effective DQM solutions.

The quality of organizational data is rarely perfect as data, when captured and stored, may suffer from such defects as inaccuracies and missing values [22]. Its quality may further deteriorate as the real-world items that the data describes may change over time (e.g., a customer changing address, profession, and/or marital status). A plethora of studies have underscored the negative effect of low DQ on decision performance (e.g., [7,9,16,29]) and have identified the need to develop data refreshing policies [23], to measure DQ ([13,19]), and to communicate DQ assessments to decision makers ([29,31]). However, maintaining data at a high quality level involves significant costs [12]. These costs are associated with efforts to detect and correct defects, set governance policies, redesign processes, and invest in monitoring tools. From an economic perspective, one would try to reach a certain quality level at a minimum possible cost. Targeting a higher DQ level improves utility of the data. (We use the term, "utility," as a synonym for "value" or "benefit", to be consistent with the use of this term in prominent prior literature. This has nothing to do with "utility theory"). Yet, at the same time, targeting a higher DQ level increases DQM costs [14]. However, although some DQM decisions involve significant utility/cost tradeoffs, economics-driven assessments of DQM alternatives are under-examined, barring a few exceptions. Some works (e.g., [3–5]) use utility-driven assessments to understand tradeoffs between different DQ dimensions, optimize their configuration accordingly, and use the results for improving data processes. An algorithm that minimizes the cost of retrieving data that meets certain quality requirements has been proposed in [2]. Policy for optimizing the cost for synchronizing the contents of a DW with the source systems from which data is retrieved has been examined in [11]. A similar issue is examined from the point of refreshing distributed data views [28] and from the point of the data retrieved by query execution in DW environments [15]. Other research has also

* Corresponding author. Tel.: +972 54 5536599; fax: +972 8 6472958.
E-mail addresses: adireven@bgu.ac.il (A. Even), gshankar@babson.edu (G. Shankaranarayanan), pberger@bentley.edu (P.D. Berger).

0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2010.07.011

used economic assessments for developing superior DQ measurements (e.g., [13,19]).

A framework for optimally configuring a tabular dataset, considering economic perspectives, has been described in [14]. In this study, we develop and evaluate that model further to examine two key questions for defining optimal quality improvement policies: a) within a large data resource, what subset of records (defined by the time-span coverage, as explained later) should be targeted for improvement? b) Within that chosen subset, what should be the targeted quality level? The model in [14] has been evaluated analytically, using closed-form solutions and numerical approximations to assess applicability, given certain assumptions and constraints. In this study, we describe a rigorous and comprehensive *empirical* evaluation, which examines the applicability and usefulness of the model in a real-world setting. We show that, within our evaluation context, all model variables can be operationalized and all parameters estimated. Further, our evaluation confirms our modeling assumptions about associations between decision variables (time span and quality level) and economic outcomes (utility, cost, and net-benefit). We show that improvements to current data acquisition and maintenance policies, identified from applying the model, can significantly increase the overall benefit. The evaluation also highlights enhancements to the model to address similar design decisions in other data management contexts. Our evaluation illustrates the importance of quantitatively assessing and understanding the cost–benefit tradeoffs, particularly in large datasets where such tradeoffs can be very significant.

We evaluate the model in a CRM context. Several studies (e.g., [8,17,21,27]) have underscored the importance of managing customer data at a high quality level. DQ defects (e.g., missing, inaccurate, and/or outdated data values) might prevent managers and analysts from having the right picture of customers and their purchase preferences and, hence, might damage marketing efforts significantly. Some studies (e.g., [19,23]) have also discussed methodologies and techniques for improving the quality of customer data. For our evaluation, we use large data samples from a real-world system that helps manage alumni relationships in a large university. This system helps segment and categorize donors, predict donor behavior, and manage solicitation campaigns, much like how a traditional CRM helps manage customers [6,23,27]. Though we focus on CRM, our model and evaluation methodology applies, in general, to data environments that manage large data resources, such as data warehouses (DW) and enterprise resource planning (ERP) systems. Such environments execute business processes, support decision making, and generate revenue through the sale of data products (e.g., [18,20,32]). We see the plethora of data usages as ways of gaining benefits from the data resource. Such benefits can be conceptualized as "utility" [1] — a measure for the value gained through enhancements to business performance, improvements to decision outcomes, or the data consumer's willingness to pay. We posit that assessing utility-cost tradeoffs toward the maximization of the net-benefit gained from using data resources must be an important goal for managing these resources.

In the remainder of this paper, we first briefly review the dataset optimization model and state our evaluation objectives. We then describe our process for evaluating the model with the alumni data, present and analyze the results, and highlight important insights gained through such analyses. To conclude, we restate the contributions of this study, discuss implications for DQM research and practice, and suggest directions for future research.

2. Evaluating the dataset optimization model

Our evaluation of the dataset optimization model proposed in [14] has two important goals. First, we aim at validating and demonstrating the usability of the model within a real-world context. Second, we wish to gain important insights towards improving data quality within the specific evaluation context — managing alumni data. The Total Data Quality Management (TDQM) approach [30] promotes the notion that data quality improvement is not a one-time effort, but rather an on-going cycle of incremental improvements (Fig. 1), which consists of four main stages: a) *Define* — identifying the evaluation objectives and scope, the set of feasible actions, and a model that describes the anticipated effect of these actions, b) *Measure* — assigning quantitative values to the model variables and parameters, c) *Analyze* — using the model for assessing the different alternatives toward identifying the optimal solution, and d) *Improve* — translating the analysis results to recommendation of a set of actions that should be taken toward data quality improvement.

Fig. 1. The data quality improvement process.

In this study, we demonstrate one full cycle of the evaluation process, along the stages of the TDQM cycle. As we discuss later, this cycle should be followed by others, which look into other possible data quality improvement and process enhancements.

2.1. Model overview

The design-optimization framework in [14] suggests that certain design characteristics of information systems affect *utility*, a measure of business benefits gained from using data resources, and affect the *cost* of implementing and maintaining the resources. It views design characteristics as decision variables in a deterministic model that the designer configures, where the goal is maximizing *net-benefit* — the difference between utility and cost.

The tabular dataset model, derived from that framework, addresses two key decisions that can be interpreted as being associated with managing the quality of large datasets: (a) *Time Span* (T) is typically defined by a "cut off" record age. Data administrators may consider managing records that are older than T differently (e.g., discard or archive them). The time-span variable T ranges between 0 and T^S, the maximum time-span coverage available. Increasing T broadens the range of data records covered by the quality improvement efforts; hence, it increases the potential for gaining utility. However, it also increases the associated costs. The preliminary model assumes that the marginal utility of data records declines exponentially with age. It can, hence, be shown that the overall utility of a dataset increases with T, but at an exponentially-decreasing rate — i.e., $U \propto (1-e^{-\alpha T})$, where $\alpha>0$. Further, the model assumes that the number of records (and cost), grows linearly with T. (b) *Targeted Quality Level* (Q): the presence of defects in a data resource reduces its utility, and the model uses an objective quality measurement (a [0,1] ratio) that reflects the presence of defects [25]. Given a certain quality level (i.e., the proportion of non-defective dataset records), one may decide to reduce the presence of defects, which improves quality and usability of the evaluated data resource; hence, the associated utility. The model assumes that dataset utility grows with quality to a certain power (i.e., $U \propto Q^{\lambda}$, where $\lambda>0$). However, the higher the quality level targeted, the higher are the associated improvement and maintenance costs. The model assumes that a certain quality level (Q^S) is guaranteed by the data source, and that the cost increases with quality to a certain power when a higher quality is targeted (i.e., $C \propto Q^{\delta}$, where $\delta>0$). It is likely, although not mandated, that $\delta>1$ and that the cost-mapping function is convex with Q. The model allows "quality" to be defined in