Decision Support Systems 83 (2016) 32–46

Contents lists available at ScienceDirect

Decision Support Systems

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

Determining the use of data quality metadata (DQM) for decision making purposes and its impact on decision outcomes — An exploratory study

Helen-Tadesse Moges[a], Véronique Van Vlasselaer[a], Wilfried Lemahieu[a], Bart Baesens[a,b,c,*]

[a]Department of Decision Sciences and Information Management, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[b]School of Management, University of Southampton, Southampton, SO17 1BJ, United Kingdom
[c]Vlerick Business School, Leuven, Belgium

ARTICLE INFO

Article history:
Received 28 April 2014
Received in revised form 30 October 2015
Accepted 23 December 2015
Available online 31 December 2015

Keywords:
Data quality
Decision strategy
Decision support systems
Data quality metadata (DQM)

ABSTRACT

Decision making processes and their outcomes can be affected by a number of factors. Among them, the quality of the data is critical. Poor quality data cause poor decisions. Although this fact is widely known, data quality (DQ) is still a critical issue in organizations because of the huge data volumes available in their systems. Therefore, literature suggests that communicating the DQ level of a specific data set to decision makers in the form of DQ metadata (DQM) is essential. However, the presence of DQM may overload or demand cognitive resources beyond decision makers' capacities, which can adversely impact the decision outcomes. To address this issue, we have conducted an experiment to explore the impact of DQM on decision outcomes, to identify different groups of decision makers who benefit from DQM and to explore different factors which enhance or otherwise hinder the use of DQM. Findings of a statistical analysis suggest that the use of DQM can be enhanced by data quality training or education. Decision makers with a certain level of data quality awareness used DQM more to solve a decision task than those with no data quality awareness. Moreover, those with data quality awareness reached a higher decision accuracy. However, the efficiency of decision makers suffers when DQM is used. Our suggestion would be that DQM can have a positive impact on decision outcomes if it is associated with some characteristics of decision makers, such as a high data quality knowledge. However, the results do not confirm that DQM should be included in data warehouses as a general business practice, instead organizations should first investigate the use and impact of DQM in their setting before maintaining DQM in data warehouses.

1. Introduction

Although the importance of DQ has been recognized for more than decades, different DQ problems continue to exist even in simple traditional systems because of huge data volumes and their complexity [1]. The problem is exacerbated by the fact that decision support systems are becoming vital to support decision making processes. The DQ level in decision support systems may not be good for different reasons. One reason is that DQ problems can be aggravated when data are merged or integrated from different sources which is typically the case in decision support systems or data warehouses. The other reason can be that soft data analysis is needed for strategic planning. Soft data are subjective assessment or future trend forecast which can be used for decision making [2]. For example, decision makers need to utilize soft data, such as the marketing strategies of competitors in order to change or adapt the marketing strategy of the company accordingly. Most of the time, managers make decisions without considering the DQ level of the data. Decision makers who are familiar with the data have an intuitive knowledge about the data. However, this intuitive knowledge can be missed when data are used by different decision makers for purposes other than the original purpose for which the data were created, which is becoming more and more the case with the increasing use of data warehouses. Decision makers who do not have prior experience with the data may avoid using them because they can't verify the quality of the data [2]. Because of such and other reasons, DQ is very important for decision making processes, but organizational data warehouses are still facing different DQ problems [2].

As one of the different ideas to reduce the impact of poor DQ on decision outcomes, the literature suggests the inclusion of metadata about the quality of data (DQM) for two reasons [1–4]. First, decision makers must adjust their decision making processes accordingly by

* Corresponding author at: Department of Decision Sciences and Information Management, KU Leuven, Naamsestraat 69, B-300 Leuven, Belgium. Tel.: +32 16 32 68 84; fax: +32 16 32 66 24
E-mail addresses: Helen.Moges@kuleuven.be (H. Moges), Veronique.VanVlasselaer@kuleuven.be (V. Vlasselaer), Wilfried.Lemahieu@kuleuven.be (W. Lemahieu), Bart.Baesens@kuleuven.be (B. Baesens).

http://dx.doi.org/10.1016/j.dss.2015.12.006
0167-9236/

recognizing the DQ level of the given data [5]. Second, DQ is context-dependent, meaning that data with good quality for one use may not be appropriate for other uses. For instance, the extent to which data are required to be complete for accounting tasks may not be required for sales prediction tasks. Therefore, DQM can help decision makers to determine the appropriateness of the DQ level in the context of the task at hand [6]. Additionally, DQ practitioners have acknowledged the importance of providing DQM to facilitate the decision making process [7,8].

Maintaining DQM into databases means maintaining the level of DQ measured along DQ dimensions such as accuracy, completeness and timeliness. However, the advantage of providing DQM to decision makers along with the actual data should be fully studied because it would be expensive to collect, maintain and manipulate DQM. Additionally, DQM can be difficult to capture and measure, and may require training and software tools. Moreover, the impact of DQM on decision outcomes can be negative. In response, prior DQM research investigated the use of DQM for decision making processes, although there is no full consensus on the results [1,2,9]. Some researchers have found that DQM is used in certain situations [2], and others didn't find any statistical evidence that DQM is actually used, even when it is available [10]. The difference in the results may be attributed to the different approaches used by prior researchers. In addition, the impact of DQM on the effectiveness of decision outcomes is not studied adequately. To fill this gap, this paper investigates the impact of DQM on decision outcomes in a different setting from previous research. We have developed a critical decision task (bankruptcy prediction) based on an Altman-Z model [11] to understand the impact of DQM on the effectiveness of decision outcomes, to identify different groups of decision makers who benefit from DQM and to explore different factors which enhance or otherwise hinder the use of DQM. This study aims to provide a concise set of guidelines for system designers to determine the importance of DQM for their specific case and to justify the associated cost of capturing and maintaining it. The study incorporated all the variables studied in previous DQM research in addition to novel variables such as DQA which makes the study inclusive. This, in turn, helped to measure the effect of the variables on the use of DQM in a similar environment where similar subjects are used, consequently removing the impact of an experimental design. The main contribution of this study, apart from the inclusion of the DQA variable, is the way in which the decision outcome measures were defined.

The paper is structured as follows. The next section reviews previous research in DQM. The third section discusses the research design and the fourth section explains the results. Finally, the paper ends by giving concluding remarks and indicating future research ideas.

2. Literature review

2.1. Data quality

Recently, data quality (DQ) is becoming a concern to organizations where plenty of data are available. Similarly, DQ is constantly growing as a crucial research topic in academic world. DQ research can be categorized into two broad types, *intrinsic* and *contextual* DQ studies. The *intrinsic* DQ research concerns about the intrinsic value of the data. It depends on the data themselves without considering the context in which the data is used. The contextual DQ study considers factors such as the purposes for which the data are used and the characteristics of the data users. Prior research has indicated that these contextual factors can strongly affect the way DQ is assessed for daily use. For example, Wang and Strong [12] have indicated the importance of recognizing the multi-dimensionality nature of DQ and measure data items accordingly using users' perceptions. However, the importance of considering contextual DQ assessment may increase the complexity level of DQ management. For example, consider a production company sales sheet which shows "item codes", "quantities", "cost" and "selling prices" where some values for the "cost" column are missing. For decisions regarding production efficiency, the sheet with missing "cost" data would be considered incomplete. However, the same sales sheet can be considered as complete for making inventory decisions (reconciling the amount of quantities on the sheet and the physical quantities in a store) because all the values for the "quantities" column are present. Although not easy, considering the contextual nature of DQ can improve DQ management in databases. In line with this, it is important that decision makers can determine the level of DQ for the task at hand. This is also one of the reasons why recent DQ research has suggested the integration of DQM along with the data in decision support systems [10].

2.2. Data quality metadata (DQM)

Data quality metadata (DQM) is information about the quality level of stored data in organization databases, and is measured along different dimensions such as accuracy, currency, and completeness. Also, DQM is considered to be intrinsic to the data because the metadata is usually produced objectively. DQ tagging is the process by which DQM is created [13]. There are different types of metadata in information systems which are maintained and managed, such as data dictionary metadata, administrative metadata, and metadata about the system infrastructure (see Table 1).

However, there are different issues in DQ tagging. First, there are no established rules, to the best of our knowledge, at which level DQM should be maintained in databases. It is possible to have DQM at the level of the individual data item, at an attribute/column level and at the level of a relational table [1,2]. However, the merits and demerits of these levels of DQM representations are not fully discussed in the literature. The most common level of DQM representation used by previous researchers is at the data item level [1,2,3,13].

Second, determining the DQ dimension(s) for which quality measures should be stored as DQM is context-dependent. The most commonly used DQ dimension in the literature is the accuracy DQ dimension [1,2,3,13]. This may acknowledge the importance of the accuracy dimension for different tasks [6]. This paper also uses the measure of the accuracy dimension as DQM in order to facilitate comparison with prior studies.

The third important consideration is the format of DQM, in particular how DQM is created, maintained and represented to the end users. The format in which DQM is represented can affect the decision making process and should be designed to facilitate the process [3,17]. There are different DQM representations used in previous

Table 1
Different types of metadata as discussed in literature [14–16].

Types of metadata	Description
Data quality metadata	It indicates the quality level of specific data in databases. For example, it can be indicated that sales data are 90% complete for the month of January 2014.
Descriptive metadata	It describes the data in terms of e.g. purpose, author, and title.
Terms and conditions metadata	It describes the conditions under which the data can (not) be used, e.g. intellectual property rights.
Administrative metadata	It indicates when and how the data are created, and who can access them.
Data dictionary metadata	It indicates the meaning of and relationships within the data.
Structural metadata	It describes the syntactical aspects of the data, e.g. the structure and base type of the data records.