Information & Management 50 (2013) 43–58

Contents lists available at SciVerse ScienceDirect

Information & Management

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

A multidimensional analysis of data quality for credit risk management: New insights and challenges

Helen-Tadesse Moges [a], Karel Dejaeger [a], Wilfried Lemahieu [a], Bart Baesens [a,b,c,*]

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

ARTICLE INFO

Article history:
Received 18 January 2012
Received in revised form 9 September 2012
Accepted 29 October 2012
Available online 16 November 2012

Keywords:
Data quality
Information quality
Credit risk
Data definition

ABSTRACT

Recent studies have indicated that companies are increasingly experiencing Data Quality (DQ) related problems as more complex data are being collected. To address such problems, the literature suggests the implementation of a Total Data Quality Management Program (TDQM) that should consist of the following phases: DQ definition, measurement, analysis and improvement. As such, this paper performs an empirical study using a questionnaire that was distributed to financial institutions worldwide to identify the most important DQ dimensions, to assess the DQ level of credit risk databases using the identified DQ dimensions, to analyze DQ issues and to suggest improvement actions in a credit risk assessment context. This questionnaire is structured according to the framework of Wang and Strong and incorporates three additional DQ dimensions that were found to be important to the current context (i.e., actionable, alignment and traceable). Additionally, this paper contributes to the literature by developing a scorecard index to assess the DQ level of credit risk databases using the DQ dimensions that were identified as most important. Finally, this study explores the key DQ challenges and causes of DQ problems and suggests improvement actions. The findings from the statistical analysis of the empirical study delineate the nine most important DQ dimensions, which include accuracy and security for assessing the DQ level.

1. Introduction

The risk of poor Data Quality (DQ) increases as larger and more complex information resources are collected and maintained [27,23]. Because most modern companies tend to collect increasing amounts of data, good data management is becoming increasingly important. In response, in the previous two decades, the aspect of DQ has received a lot of attention by both organizations worldwide and in academic literature. Several studies have explored DQ challenges and have focused on DQ measurement and improvement [3–11,19–27,30,32–34,39,37,40,41,43–48]. Fig. 1 illustrates this focus by plotting the increasing number of DQ related publications over the past ten years as reported by ISI Web of Knowledge.[1]

In practice, decision makers differentiate information from data intuitively and describe information as data that has been processed. Unless otherwise specified, this paper uses data interchangeably with information.

DQ is often defined by 'fitness for use' which implies the relative nature of the concept [30,21,4]. Quality data 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. Accounting tasks typically require the availability of all cash balances, e.g., when making up a balance sheet. Conversely, sales prediction tasks will always be possible irrespective of missing cash balances [30,37]. In addition to the task type, the contextuality of DQ can also be explained by the trade-offs between DQ dimensions where one dimension can be favored over other dimensions for a specific task. Data quality dimensions are not independent but are, in fact, correlated [22]. Moreover, if one dimension is considered more important than other dimensions for a specific application, then the choice of favoring this dimension may negatively affect other dimensions. For example, having accurate data may require checks that could negatively affect timeliness. Conversely, having timely data may result in less accuracy, completeness or consistency. A typical situation in which timeliness can be preferred to accuracy,

* Corresponding author at: Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. Tel. +32 16 32 68 84; fax +32 16 32 66 24.
E-mail addresses: Helen.Moges@econ.kuleuven.be (H.-T. Moges), Karel.Dejaeger@econ.kuleuven.be (K. Dejaeger), Wilfried.Lemahieu@econ.kuleuven.be (W. Lemahieu), Bart.Baesens@econ.kuleuven.be (B. Baesens).

[1] http://apps.isiknowledge.com

0378-7206/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.im.2012.10.001

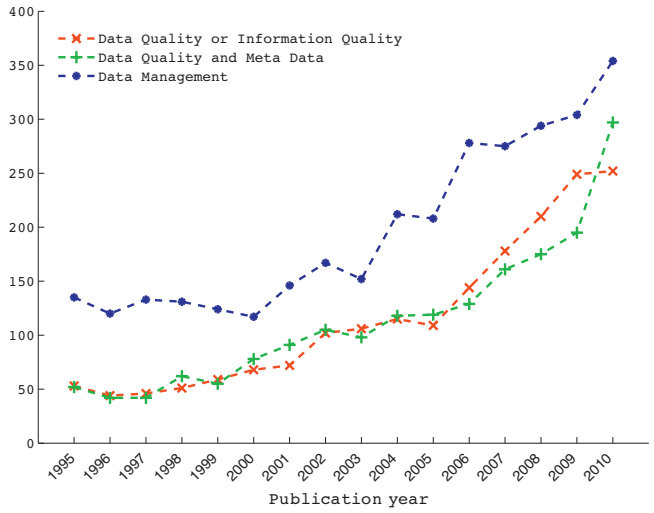

Fig. 1. Journal and conference proceedings from ISI web of knowledge.

completeness, or consistency is given by most web applications. As time constraints are often very stringent for web data, it is possible that such data are deficient with respect to other quality dimensions. For instance, a list of courses published on a university web site must be timely though there could be accuracy or consistency errors, and some fields specifying additional course details could be missing. Conversely, when considering administrative applications, accuracy, consistency and completeness requirements are more essential than timeliness, and therefore, delays are mostly permissible. Another example can be a trade-off between completeness and consistency. A statistical data analysis typically requires a significant and representative set of data, and in this case, the approach will be to favor completeness while tolerating inconsistencies or adopting techniques to address these inconsistencies. Conversely, when publishing a list of student scores on an exam, it is crucial to check the list for consistency, which may possibly defer the publication of the complete list [30,4]. Accordingly, studying the DQ in the context of a specific task is a recognized method [11,25–27,30,47,48].

1.1. Credit risk assessment task

DQ is of special interest and relevance in a credit risk setting because of the introduction of compliance guidelines, such as Basel II and Basel III [2,15]. Because the latter has a direct impact on the capital buffers and, hence, on the safety of financial institutions, special regulatory attention is being given to addressing DQ issues and concerns in this context. Therefore, given its immediate strategic impact, DQ in a credit risk setting is more closely monitored than in most other settings and/or business units [42,34].

The credit risk assessment task is primarily concerned with quantifying the risk of the loss of principal or interest stemming from a borrower's failure to repay a loan or meet a contractual obligation. Therefore, financial institutions are obliged to assess the credit risk that may arise from their investment. These institutions may estimate this risk by taking into account information concerning the loan and the loan applicant.

The quality of the credit approval process from a risk perspective is determined by the best possible identification and evaluation of the credit risk that results from a possible default on a loan. Credit risk can be decomposed into four risk parameters as described in the Basel II documentation [42]. These parameters are the Probability of Default (PD), the Loss Given Default (LGD), the Exposure at Default (EaD) and the Maturity (M). These parameters are used to calculate the Buffer Capital (BC), which also referred to as regulatory capital and is the money set aside to anticipate future unexpected losses due to loan defaults.

$$BC = f(PD, LGD, EaD, M)$$

The correct estimation of these parameters and the appropriateness of the function or algorithm used to calculate the risk concentration are crucial because incorrect parameters or inappropriate algorithms may result in a loss or even bankruptcy of the institution. The Risk Concentration (RC) refers to an exposure with the potential to produce losses large enough to threaten a financial institution's health or ability to maintain its core operations [1]. Improving the quality of the data used for calculating these parameters is one way of improving the precision of the parameter estimates and, consequently, of improving the correctness of the credit approval decisions [2,14].

1.2. Total Data Quality Management Program

Poor DQ impacts organizations in many ways. At the operational level, poor DQ has an impact on customer satisfaction, increases operational expenses and can lead to lowered employee job satisfaction. Similarly, at the strategic level, poor DQ affects the quality of the decision making process. An enterprise may experience various DQ problems [21,34]. However, no improvement can be made without knowing and measuring the problems. It is argued in the literature that organizations should implement a Total Data Quality Management (TDQM) program that includes *DQ definition*, *measurement*, *analysis* and *improvement*. This enables them to achieve a suitable DQ level [39,28].

The *DQ definition* phase is the starting point for a TDQM program. In this phase, all the necessary DQ dimensions to be measured, evaluated and analyzed are identified. Next, the *measurement* process is implemented. The results from the measurement process are *analyzed*, and DQ issues are detected. These issues will be taken into account during the *improvement phase*. In this phase, the collection of poor quality data cases is thoroughly investigated, and improvement actions are suggested. The four phases are iterated in this order over time, as shown in Fig. 2. In fact, the primary goal of DQ assurance is the continuous control of data values and possibly their improvement [44,4].

The identification of DQ dimensions from a user perspective defines the list of important DQ dimensions for a specific task that need to be assessed, analyzed and improved [44,4]. Therefore, the first aim of this paper is to identify the DQ dimensions that are considered relevant to assess the DQ in the context of credit risk assessment. Second, the paper investigates the impact of different factors, such as the existence of DQ teams and the size of financial institutions, on the importance of DQ dimensions. Third, the DQ level of a credit risk databases is assessed by incorporating the DQ dimensions categorized as relevant, and finally, frequent recurring DQ challenges and their causes in a credit risk assessment context are also explored.

The remainder of the paper is structured as follows. The next section explores the related literature, while the third section explains our research methodology. The fourth section elaborates on the key findings, while the final section presents the conclusions and lists topics for further research.

2. Related research

2.1. Identification and definition of DQ dimensions

DQ problems cannot be addressed effectively without identifying the relevant DQ dimensions. Therefore, the first objective of DQ research is to determine the characteristics of the data that are