Available online at www.sciencedirect.com
ScienceDirect
Decision Support Systems 42 (2006) 302–317
Decision Support Systems
www.elsevier.com/locate/dsw

Supporting data quality management in decision-making

G. Shankaranarayanan *, Yu Cai

Information Systems Department, Boston University, School of Management, 595 Commonwealth Ave., Boston, MA 02215, USA

Available online 19 February 2005

Abstract

In the complex decision-environments that characterize e-business settings, it is important to permit decision-makers to proactively manage data quality. In this paper we propose a decision-support framework that permits decision-makers to gauge quality both in an objective (context-independent) and in a context-dependent manner. The framework is based on the information product approach and uses the Information Product Map (IPMAP). We illustrate its application in evaluating data quality using completeness—a data quality dimension that is acknowledged as important. A decision-support tool (**IPView**) for managing data quality that incorporates the proposed framework is also described.

Keywords: Data quality; Completeness; Data quality dimensions; Information product; IPMAP

1. Introduction

The access to information in today's decision-environments is not restricted by business-unit or organizational boundaries. Decision-making in these environments involves large data volumes and includes a wide variety of decision-tasks. Decision-makers are forced to become more responsive as they have access to data anywhere and at anytime. In such environments it is important to assure decision-makers of the quality of data they use and allow them to gauge quality. Traditional methods for evaluating data quality dimensions do so objectively—without considering contextual factors such as the decision-task and the decision-maker's preferences. However, a classic definition of quality is fitness for use, or the extent to which a product successfully serves the purposes of customers [11]. Quality of the data, therefore, is dependent on the purpose (task). We believe that the perceived quality of the data is influenced by the decision-task and that the same data may be viewed through two or more different quality lenses depending on the decision-maker and the decision-task it is used for. For example, an instructor trying to place orders for course textbooks may find an approximate enrollment figure sufficiently accurate to decide the number of copies to order. The same instructor will not consider this enrollment figure an accurate-enough representation of his/her class-size when requesting a room (seating capacity) for class meetings. Decision-makers must

* Corresponding author. Tel.: +1 617 353 4605; fax: +1 617 353 5003.
E-mail addresses: gshankar@bu.edu (G. Shankaranarayanan), ycai@bu.edu (Y. Cai).

0167-9236/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2004.12.006

have the ability to evaluate data quality[1] based on the decision-task that the data is used for. It is therefore important to communicate data quality information to the decision-maker and offer the decision-maker the ability to gauge the quality of the data using task-dependent interpretations. The first objective in this paper is to propose a framework that communicates data quality information to the decision-maker and allows the decision-maker to gauge data quality by incorporating task-dependent factors.

The data quality management framework proposed here is based on the notion of managing information as a product—the information product (IP) approach [25]. The IP approach treats information as a product instead of a "by-product" of information systems [4]. Although research has focused on developing and implementing information systems that deliver the "right" data, the outputs often do not meet the consumer's expectations. One reason is the mismatch in specifications between the "output product" and the user's need. Another is the poor management of the raw materials and processing involved in creating this "output". Total Quality Management (TQM) methods were implemented to address similar problems in conventional manufacturing. Research in data quality suggests that the focus should shift from information systems to the output of such systems, the IP [4,25]. An IP such as a business report (inventory volume report or sales report) is the deliverable that corresponds to specific requirements of the consumers. TQM and other methods successfully employed to address quality issues in conventional manufacturing can be used to manage the processes that create the information product and implement Total Data Quality Management (TDQM) in information systems.

The information product map (IPMAP) is a representation scheme for representing the manufacture of an IP based on the IP approach [20]. The second objective of this paper is to propose the use of this visual representation for communicating quality-related metadata associated with an IP, informing decision-makers about the manufacturing processes used to create an IP, and for evaluating data quality of the IP at all manufacturing stages. A decision-support tool for data quality management (IPView) that incorporates the IPMAP is described. Methods for evaluating data quality, including the one proposed here for evaluating completeness, are implemented in IPView.

Past research has illustrated that data quality may be evaluated along several different quality dimensions [6,18,23]. The three important and commonly addressed data quality dimensions are accuracy, timeliness, and completeness [1,8,14,22]. Timeliness and accuracy have been addressed in depth [2,4]. However, completeness, acknowledged as an important data quality dimension, is addressed to a lesser extent [3]. Ballou and Pazer examine completeness by dividing it into structural completeness and content completeness and treating the two as independent [3]. In this paper, our context-dependent examination of completeness is based on the provision of data quality metadata (including measurement of structural or context-independent completeness). Kahn et al. identify that an important aspect of managing data quality is conformance to specification [12]. Without an appropriate measurement, it is difficult to determine the level of conformance to specification. The third objective of this paper is to provide an in-depth examination of completeness as a data quality dimension and propose a method for evaluating it. The paper also illustrates how completeness can be evaluated using the IPMAP.

The next section presents an overview of the relevant literature on data quality to differentiate this research and to define its scope. Section 3 describes the IPMAP, the method for evaluating completeness, and how this evaluation is done using the IPMAP. Section 4 describes the extensions to the IPMAP necessary for evaluating completeness and for implementing it as a decision-support tool (IPView) for communicating and evaluating data quality. Concluding remarks and the research directions are presented in Section 5.

2. Relevant literature

Conventional approaches to data quality management such as data cleansing [9], data tracking and statistical process control [18], data source calculus

[1] In the remainder of this paper, we refer to these two ways of examining data quality as objective or context-independent and context-dependent evaluation respectively.