Chapter 3
Data Quality for Data Mining in Business Intelligence Applications: Current State and Research Directions

Arun Thotapalli Sundararaman
Accenture, India

ABSTRACT

Data Quality (DQ) in data mining refers to the quality of the patterns or results of the models built using mining algorithms. DQ for data mining in Business Intelligence (BI) applications should be aligned with the objectives of the BI application. Objective measures, training/modeling approaches, and subjective measures are three major approaches that exist to measure DQ for data mining. However, there is no agreement yet on definitions or measurements or interpretations of DQ for data mining. Defining the factors of DQ for data mining and their measurement for a BI System has been one of the major challenges for researchers as well as practitioners. This chapter provides an overview of existing research in the area of DQ definition and measurement for data mining for BI, analyzes the gaps therein, besides reviewing proposed solutions and providing a direction for future research and practice in this area.

INTRODUCTION

This chapter is intended to primarily cover research aspects and directions in data Quality (DQ) measurement for data mining in Business Intelligence (BI) applications Let's start the discussion with a very brief definition of the two terms that are so extremely critical for this chapter, namely data mining and DQ. A complete list of definitions of all other key terms is provided at the end of this chapter. A frequently cited definition for data mining is given by Decker and Focardi (1995) as "Data mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data". According to Cios, Pedrycz, Swiniarski, & Kurgan, (2007), the goal (of data mining) is to efficiently and effectively extract information and knowledge from data that should make sense of the data, i.e., this knowledge

DOI: 10.4018/978-1-4666-6477-7.ch003

should exhibit some essential attributes: it should be understandable, valid, novel and useful. Many research publications have used the terms DQ and Information Quality (IQ) interchangeably, although certain differences exist between the two from the perspective of users of data/information. In the absence of a single definition of DQ or IQ as it pertains to data mining, we may resort to refer the standard global definition of quality that describes as "fit for use". DQ for data mining would encompass those factors that render the underlying data and the insights derived from data mining models to be appropriate for use in the decision making process, enabled through a BI System. Thus, the factors or dimensions that constitute DQ for data mining in BI applications may be derived from these definitions as understandingness, validity, novelty, usefulness, and actionability, etc., discussed in more detail in subsequent sections of this chapter.

The central theme of this chapter revolves around DQ measurement approaches for data mining. This chapter is focused on presenting a comprehensive view of existing frameworks for measurement of DQ in data mining and analyzing them with a view to present the gaps in existing frameworks. The main contribution of this chapter lies in proposing appropriate frameworks and specific directions for future research in the field of DQ for data mining, based on the proposed frameworks.

The objectives of this chapter are as follows:

- Identify potential reasons for DQ issues in data mining for BI;
- Identify the factors that constitute DQ and DQ measures for data mining;
- Analyze existing DQ measurement frameworks in light of their applicability to data mining for BI;
- Review state of current research and future research directions in the study of measurement of DQ for data mining in BI.

There has been a growing interest among researchers and industry practitioners in DQ, DQ measurement, DQ assessment and improvement, more so, with respect to data mining. DQ measure in data mining should be one that supports achieving the goal of data mining, i.e., to build a model that most accurately predicts the desired target value for new data. Use of this measure in an incorrect form may lead to quality issues of the model. However, even the basic principles of DQ assessment in data mining and its improvement still remain largely open.

Research interest in DQ remains an enduring subject, which is reiterated even by very recent literature. For example, Fehrenbacher et al., (2012) state that the importance of DQ is ever increasing and that research in this field focuses mainly on two aspects criteria and assessment. This recent work observes that while researchers have developed a number of frameworks, criteria lists and approaches for assessing and measuring DQ, still research in this discipline indicates that assessing DQ remains to be challenging. This work argues that although DQ is subjective, most of the existing frameworks and assessment methodologies do not often consider the context in which the assessment is performed. Through empirical data research this cited work suggests that the perceived importance of DQ criteria has changed over the last decade.

Direct references to DQ in BI Systems in published literature are limited. The same may be explored through references to related types of Information Systems such as DW, DSS, EIS or data mining. Similarly, the terms DQ and IQ have been used interchangeably in literature and differentiation, if any, have been very thin. There has been no consensus about the distinction between DQ and IQ – DQ may be referred to technical issues or quality in sources or 'raw data' and IQ may be referred to non-technical issues or quality in processed data stores or insights presented from data for decision making. In this chapter the term DQ refers to both these aspects.