Chapter 4
Big Data Quality for Data Mining in Business Intelligence Applications: Current State and Research Directions

Arun Thotapalli Sundararaman
Accenture, India

ABSTRACT

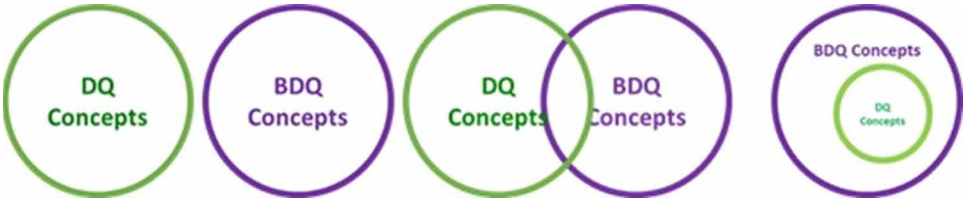
Study of data quality for data mining application has always been a complex topic; in the recent years, this topic has gained further complexity with the advent of big data as the source for data mining and business intelligence (BI) applications. In a big data environment, data is consumed in various states and various forms serving as input for data mining, and this is the main source of added complexity. These new complexities and challenges arise from the underlying dimensions of big data (volume, variety, velocity, and value) together with the ability to consume data at various stages of transition from raw data to standardized datasets. These have created a need for expanding the traditional data quality (DQ) factors into BDQ (big data quality) factors besides the need for new BDQ assessment and measurement frameworks for data mining and BI applications. However, very limited advancement has been made in research and industry in the topic of BDQ and their relevance and criticality for data mining and BI applications. Data quality 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 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 BDQ definitions and measurement for data mining for BI, analyzes the gaps therein, and provides a direction for future research and practice in this area.

DOI: 10.4018/978-1-7998-5781-5.ch004

INTRODUCTION

This Chapter is primarily focused on current challenges, research progress and directions in Data Quality (DQ) in Big Data environments where Big Data is used for Data Mining and consumed through Business Intelligence (BI) Applications, including Artificial Intelligence applications. We introduce a new term and concept, BDQ i.e. Big Data Quality to refer DQ issues related to use of Big Data. It is important to evaluate if the current knowledge of DQ concepts and definitions are all applicable to BDQ and if new concepts are added or if a new set of concepts are applicable for BDQ. This chapter seeks to address which of the below representations (Figure 1) holds good in the study of BDQ when applied to Data Mining for BI Applications.

Figure 1. BDQ Vs. DQ



Let's start the discussions with a very brief definition of the 2 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 & 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 should exhibit some essential attributes: *it should be understandable, valid, novel and useful*.

Many research publications have used the terms Data Quality 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 Information Quality or DQ 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 model to be appropriate for use in 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, actionability, etc. Extending this theory of knowledge to Big Data environments, new factors that influence quality of decisions and confidence of the consumers of data emerge. The essential characteristics of Big Data viz., Volume, Variety, Velocity, Veracity introduce these new DQ factors. Each of these characteristics changes the consumption paradigm which constitutes the need for new DQ factors.

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