Identifying Quantifiable and Automatable Data Quality Parameters for Data Exchange

Authors: Debarun Sengupta, Anjula Gurtoo, Minnu Malieckal, Jyotirmoy Dutta

Introduction about the paper

Data quality (or data fitness) gets defined as the level of capabilities that satisfy the needs of the user. According to Redman (2001), a particular dataset is considered high-quality if the dataset fits the intended use for operation, decision-making, and planning. Consequently, data quality is a multi-dimensional concept which can be grouped into intrinsic and contextual data qualities. Intrinsic data quality indicates quality aspects independent of the perception or belief of the data user, whereas contextual quality is a relative concept determined by the perception of the data user (Mahanti, 2019). For example, accuracy, consistency, and completeness are regarded as intrinsic data quality parameters, while accessibility and interpretability can be considered contextual data quality depending upon the use of the data.

With rapid digitization the world is producing about 2.5 quintillion bytes of data per day in 2019 (Price waterhouseCoopers, 2019). To give a structured format to this exponential volume of data and to build a foundation for the formal data market, the concept of data exchange is being introduced worldwide. For instance, the Shanghai Data Exchange (SDE) is the leader in data advertising, profiling, and compiling in China. Under the SDE framework, a data buyer who opts for a particular dataset is connected to the data provider based on a specific commission. Data Exchange Platforms (DEPs) also help to break the existing silos of data. The COVID-19 data exchange platforms and the European Radiological Data Exchange Platform (EURDEP) are the two most common examples of DEPs that helped break data silos in healthcare services.

DEPs, hence, address several challenges with respect to data usage by transmitting data without altering the inherent meaning, normalize transmitted data for easier consumption by the receiving system, and facilitate data-as-a-service acquisition. The platforms facilitate smooth and seamless data transmission of large data using technology and standardized formats. Data is transmitted in such a way to be easily consumed by a receiving system(s), many times normalizing the data as well. Further, the reporting and visualization value added to the raw data enriches the data without altering its inherent meaning. These exchanges often

provide data-as-a-service capabilities where subject matter data can be acquired to draw data driven insights. The integration provided by the platforms through amalgamation of variety of data can support marketing, and strategy solutions for organizations.

Against this backdrop on the identification and assessment of data quality dimensions, the present study tries to make the following contribution towards the existing body of literature, practice, and industrial uses, 1) this study is probably the first to assess data quality from the data exchange perspective. In contrast to the previously available literature on data quality, the current assessment is based on quantifiable and programmable (or mathematical) formulations of quality dimensions, 2) compared to the subject context of measures of quality, these mathematical (or objective) formulations deliver a numerical value independent of any subjective context. This evaluation technique provides a ranking of quality dimensions and helps to identify the laggard dimension that needs attention, and 3) furthermore, our formula-based approach permits the development of a practical, usable, automated evaluation technique that might improve the efficiency and minimise the cost of data quality assessment. Thus, identification and fixing data quality issues can be addressed appropriately.