World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:14, No:01, 2020

Dataset Quality Index:Development of Composite Indicator Based on Standard Data Quality Indicators

Authors : Sakda Loetpiparwanich, Preecha Vichitthamaros

Abstract : Nowadays, poor data quality is considered one of the majority costs for a data project. The data project with data quality awareness almost as much time to data quality processes while data project without data quality awareness negatively impacts financial resources, efficiency, productivity, and credibility. One of the processes that take a long time is defining the expectations and measurements of data quality because the expectation is different up to the purpose of each data project. Especially, big data project that maybe involves with many datasets and stakeholders, that take a long time to discuss and define quality expectations and measurements. Therefore, this study aimed at developing meaningful indicators to describe overall data quality for each dataset to quick comparison and priority. The objectives of this study were to: (1) Develop a practical data quality indicators and measurements, (2) Develop data quality dimensions based on statistical characteristics and (3) Develop Composite Indicator that can describe overall data quality for each dataset. The sample consisted of more than 500 datasets from public sources obtained by random sampling. After datasets were collected, there are five steps to develop the Dataset Quality Index (SDQI). First, we define standard data quality expectations. Second, we find any indicators that can measure directly to data within datasets. Thirdly, each indicator aggregates to dimension using factor analysis. Next, the indicators and dimensions were weighted by an effort for data preparing process and usability. Finally, the dimensions aggregate to Composite Indicator. The results of these analyses showed that: (1) The developed useful indicators and measurements contained ten indicators. (2) the developed data quality dimension based on statistical characteristics, we found that ten indicators can be reduced to 4 dimensions. (3) The developed Composite Indicator, we found that the SDQI can describe overall datasets quality of each dataset and can separate into 3 Level as Good Quality, Acceptable Quality, and Poor Quality. The conclusion, the SDQI provide an overall description of data quality within datasets and meaningful composition. We can use SQDI to assess for all data in the data project, effort estimation, and priority. The SDQI also work well with Agile Method by using SDQI to assessment in the first sprint. After passing the initial evaluation, we can add more specific data quality indicators into the next sprint.

Keywords : data quality, dataset quality, data quality management, composite indicator, factor analysis, principal component analysis

Conference Title : ICBDAA 2020 : International Conference on Big Data Applications and Analytics

Conference Location : Tokyo, Japan

Conference Dates : January 06-07, 2020

Open Science Index, Computer and Information Engineering Vol:14, No:01, 2020 publications.waset.org/abstracts/111833.pdf