ICIC Express Letters
Part B: Applications
Volume **11**, Number **2**, February **2020**

ICIC International ©2020 ISSN 2185-2766
pp. 145–149

DATA PROFILING PROCEDURE TO ASSESS DATA QUALITY

Sungjoon Lim, Changhan Lee, Hyungjun Kim and Youngkuk Kim

Data Quality and Standardization Team
Korea Data Agency
42, Sejong-daero 9-gil, Jung-gu, Seoul 04513, Korea
{joon; leech; khj; kyk9885}@kdata.or.kr

Received August 2019; accepted November 2019

Abstract. *The purpose of data profiling is known as identifying the properties of the values contained in the columns of a data set and the properties of the structures of the columns of a data set, and deriving data quality issues based on the data set. This paper describes normalized way to carry out data profiling to get objective result. The value domain and dependencies of the data set are arranged into the data profile, which forms the basis of deriving data rules to measure data errors in data quality assessment. The procedure of data profiling is suggested that the processes consist of analysis such as structure, column and relationship between columns. Also value domain classification is provided as a result of extracting data element of database systems. This profiling methodology is limited to DBMS until now. However, in the future, attempts to improve quality using machine learning on big data are underway in various areas related to this data profiling work.*

Keywords: Data quality, Data quality assessment, Data profiling, Value domain

1. **Introduction.** Data profiling is considered as one of data analysis techniques to data in actual operation to determine the values and structural properties of data. Furthermore, it is based on the surmise that the existing metadata of the data set to be evaluated are incomplete and inaccurate. In data profiling, metadata are reverse-engineered for actual data, and compared with the existing metadata to generate perfect and correct metadata. The repetition of this process improves the accuracy and completeness of the metadata. Furthermore, the metadata determined through the process are arranged in the form of a data profile, which forms the basis of establishing data rules. This paper describes the data profiling procedure in which an organization implements in data quality assessment in terms of data values, and the method for deriving value domains and dependencies of target data sets that is able to form the basis of data rules. Data quality assessment is considered a general task nowadays in organizations, also there are various academic work and technical literature. However, there is no normalized assessment reference model for data quality. That is the reason why data quality has been considered that it depends on the goal of business or project. This paper provides an objective way to assess data quality.

2. **Literature Review.** Data quality is defined as a level of securing the up-to-date, accuracy, and interconnectivity of data, and providing users with a useful value by using it. The purpose of data quality assessment is to maintain and improve data quality systematically and continuously. In particular, a lot of resources are being invested in efforts to assess and improve the quality of data in conjunction with data releases from public institutions. Authentication means such as DQC-V (Data Quality Certificate – Value) and DQC-M (Data Quality Certificate – Management) are implemented as means for evaluating and authenticating data quality, and as an authentication domain for data

DOI: 10.24507/icicelb.11.02.145

values, it is a system to deliberate and authenticate the quality effects factor of the data itself based on domain and business rule [1]. Business rules mean the rules of all data related to business. Business rules are data measurement rules that an organization uses to continuously manage data quality and are conditional expressions for the correct data values [2]. To conduct business rule-based quality assessment, business rules are identified, conditions or constraints are set, and rules are applied to the actual operational database using SQL (Structured Query Language) or the like. It diagnoses the quality by extracting error data from the applied database and checking the error rate.

A classification of data profiling tasks is provided and the state of the art for each class is comprehensively reviewed in [3]. Also an information quality framework based on semiotic theory, the linguistic theory of sign-based communication, to describe the form-, meaning-, and use-related aspects of information is presented in [4], which provides a sound theoretical basis both for defining quality categories, based on these different information aspects and for integrating the different research approaches required to derive quality criteria for each category. In case of international standards, ISO 8000 defines the characteristics of information and data that determine its quality, and provides methods to manage, measure and improve the quality of information and data in [5]. When assessing the quality of data, it is useful to perform an assessment according to the documented methods. Moreover, it is important to document the tailoring of standardized methods with respect to the expectations and requirements pertinent to the business case at hand. This research specifies the method for deriving data properties through data profiling, summarizing the properties as data rules, and assessing data quality.

3. **Processes of Data Profiling.** The purpose of data profiling is to identify structure, column and relationship of a data set, and to derive data quality issues from the result of data profiling. The constraints of value domain and dependency are arranged into a result of data profiling, which form a basis of deriving data rules to measure data errors in data quality assessment.

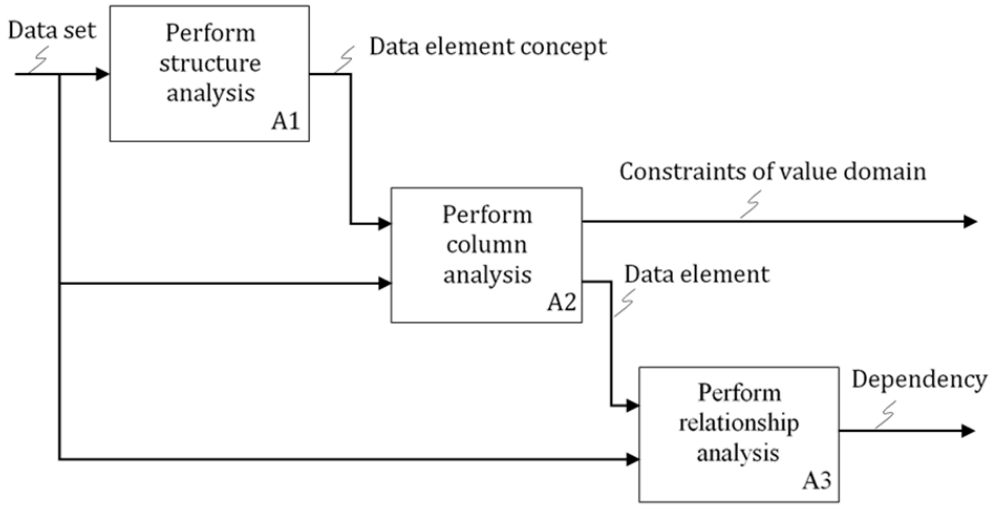

FIGURE 1. Model of data profiling

3.1. **Structure analysis.** Input of structure analysis is data set. Data set means logically meaningful grouping of data [6]. This includes value of data and optionally some basic information such as column name and descriptions. Structure analysis consists of extracting conceptual domain from the value and basic information and determining data element concept to be used for the column analysis and relationship analysis. Output of the structure analysis is data element concept.

3.2. **Column analysis.** Inputs for column analysis are the data set and the data element concept from the structure analysis. Column analysis consists of extracting data elements from the data element concept, comparing the data elements with value of data set and determining value domain. The methods for extracting data elements include discovery, assertion testing, and visual inspection, and are supported by automated tools. The outputs of column analysis are list of constraints of the value domain. The constraints are grouped by categories such as cardinalities, storage and valid values.

Example of categorized list of the constraints of the value domain:

- Cardinalities: number of rows, size of values, null values, number of disparate values and uniqueness;
- Storage: data type, length of values and decimals;
- Valid values: discrete value list, range of values, skip-over rules, pattern and domain.

3.3. **Relationship analysis.** The inputs for relationship analysis are the data set and data element from the column analysis. The relationship analysis extracts relationships between the columns within not only single table but also multiple tables. Relationship analysis consists of comparing the extracted data element with the basic information from data set and determining dependency. The discordance of the data structure leads to basic business definitions relating to data structure. Tables have correlations with real-world entities, and identifying relationships between tables requires business knowledge. As structural discordance requires semantic interpretation, the correct structural information has to be determined in collaboration with business and profiling experts. The outputs of relationship analysis are a list of dependencies. The list is grouped by categories such as dependencies of columns and synonyms.

Example of categorized list of the dependencies:

- Dependencies of columns: primary key, foreign key, functional dependency and derived columns;
- Synonyms: primary/foreign key synonyms and redundant data synonyms.

4. **Value Domain Classification.** The value domain-based data quality diagnosis is a task to use profiling technique for data, and it carries out structure analysis, column analysis, and relationship analysis according to characteristics of each domain by using classified domain column. To do this, the value domain for the column must be categorized. Value domains are like Table 1. In the case of the domain classification, it has a problem of taking much time since the user checked the data one by one and proceeded manually, human error frequently occurred, and the working time also took a long time.

Research is also underway to tune the DBMS (DatBase Management System) using machine learning. It collects indexes and schemes, variablizes, regularizes, and tunes them from meta DBMS in which tools that allow users to directly view or visualize data are installed, such as OLAP [7]. However, the studies on detailed and individual items of data standards and quality are not actively realized. Particularly, it is more so in the field that requires classification through human thought, like value domains.

5. **Preliminary Results.** Data quality is considered as a key factor for businesses in industry and commerce. Good data quality effects make benefits such as increasing revenues, reducing time costs to reconcile data, and increasing customer satisfaction. This data profiling procedure can be applied to understand the current status of data quality for some organization or systems, and it will be used for the effort to improve data quality.

Data profiling is known as a process to collect statistics or informative summaries from an existing information source such as a database or file. Also data profiling can be used to extract metadata from information source especially there is not proper specification

TABLE 1. Examples of value domain

Domain classification	Value domain example	Contents checked
No.	Resident registration number, business registration number, postal code, customer number, account number	Pattern of number related data and diagnosis of check bit
Amount	Amount, Tax, Price, Unit Price, Cost, Charge, Balance, Total amount	Diagnose tolerance range of money related data
Name	Name, address, ID, place, customer name, English customer name, URL, e-mail, IP	Pattern and length diagnosis of name related data
Quantity	Case, number, number of times, number, distance, scale, length, weight, speed, number of times, balance, side	Diagnose tolerance range of quantity-related data
Classification	Whether or not, presence, division, status	Diagnose standard definition values of classification-related data
Date	Month, year, year, d/m/y, hour, minute, second, day, semiannually, quarterly	Diagnose tolerance range and valid values for date-related data
Rate	Money rates, interest rate, ratio, exchange rate, percentage	Diagnose tolerance range of ratio (%) related data
Contents	Content, remarks, description, information, summary	Diagnose the language pattern applied to the content-related data
Code	Individual code, integrated code	Diagnose code values for code-related data
Key	Primary key, foreign key	Diagnose referential integrity of key-related data
Common	Data standardization	Diagnose compliance with data standards

or description of information source. Extracted metadata can be applicable to find non-conformity data and discover relationship between columns. This non-conformity data and relationship between columns can be used for defining data rules and form the basis of data quality assessment. In the future, attempts to improve quality using machine learning on big data are underway in various areas related to this data profiling work. Efforts are made to increase the quality of all automated tests and solve problems through machine learning from database and open source data to the report standardization and product production [7].

Acknowledgment. This work was supported by IITP grant funded by the Korean government (MSIT) (No. 2017-0-00163, Big Data Quality Evaluation Tool Development).

REFERENCES

[1] S.-G. Lee et al., Quality management model of atypical science and technology big data based on data profiling and regular expression, *The Journal of the Korea Contents Association*, vol.14, no.12, pp.486-493, 2014.
[2] *A Guideline for Data Quality Assessment*, Korea Data Agency, 2011.
[3] Z. Abedjan, L. Golab and F. Naumann, Profiling relational data: A survey, *The VLDB Journal*, vol.24, no.4, pp.557-581, 2015.

[4] R. Price and G. Shanks, Empirical refinement of a semiotic information quality framework, *Proc. of the 38th Hawaii International Conference on System Sciences*, 2005.
[5] *https://en.wikipedia.org/wiki/ISO_8000*, 2019.
[6] *ISO 8000-2:2018 Data Quality – Part 2: Vocabulary*, 2018.
[7] D. Van Aken et al., Automatic database management system tuning through large-scale machine learning, *Proc. of the 2017 ACM International Conference on Management of Data*, 2017.