International Journal of Advanced Science and Technology
Vol.117 (2018), pp.77-88
http://dx.doi.org/10.14257/ijast.2018.117.07

A Study on Data Profiling Based on the Statistical Analysis for Big Data Quality Diagnosis

Won-Jung Jang[1], Jong-Yoon Kim[2], Bum-Taek Lim[3] and Gwang-Yong Gim[4*]

[1,2,3]06978 Dept. IT Policy and Management, Graduate School Soongsil Univ., 369 Sangdo-ro, Dongjak-gu, Seoul, Republic of Korea
[4]06978 Dept. Business Administration, Soongsil Univ., 369 Sangdo-ro, Dongjak-gu, Seoul, Republic of Korea
[1]wjjang@cku.ac.kr, [2]kimjw@nice.co.kr, [3]neobtlim@gmail.com, [4]gygim@ssu.ac.kr

Abstract

The volume of digital information produced and distributed globally is expected to be around 90 zetabytes (ZB) by 2020. In the era of the Fourth Industrial Revolution, all things are connected to the Internet and various big data are being produced explosively. Industries are increasingly demanding big data to improve productivity of products, services, and factories using artificial intelligence, but systematic studies on the quality of industrial sites and public data are lacking. Artificial intelligence requires a good overview of data quality before demanding high quality data and taking advantage of big data. The purpose of this study is to propose a data profiling model using statistical analysis techniques to derive attributes for big data quality diagnosis. To do this, the R package and the Delhi Weather data set registered with Kaggle for empirical studies are used in this study. This study calculated property weights and attribute corrections by using empirical methods of statistical analysis for all attributes, and confirmed that the performance comparison study model is superior to the value (accuracy) error rate calculation model for attributes derived from the research model. It is expected that data profiling can be performed in a more scientific way rather than relying on the subjective judgment of the performer in the big data quality diagnosis.

Keywords: *data profiling, statistical analysis techniques, big data, data quality, data quality management*

1. Introduction

1.1. Research Background

It is expected that the amount of digital information produced and distributed globally will increase by 1.8 Zbytes (ZB) in 2011 and 50 times (90ZB) by 2020 [1]. In the era of the Fourth Industrial Revolution, more and more objects and people are connected to the Internet and various big data are explosively produced. Industries are increasingly demanding big data to improve productivity in their products, services, and factories, and the government also encourages the creation of new economic value-added and jobs by stimulating private sector openness and utilization of public data [17]. However, demand for Big Data is increasing, but systematic research on the quality of industrial sites and public data is lacking. Therefore, in this study, it is meaningful to suggest a data profiling method for extracting attributes as a more scientific method to diagnose big data quality using statistical analysis technique.

Received (January 5, 2018), Review Result (March 3, 2018), Accepted (March 7, 2018)
* Corresponding Author

ISSN: 2005-4238 IJAST

1.2. Purpose of the Study

Recently, artificial intelligence (AI)-related issues and machine learning technology, a core technology of artificial intelligence, are attracting attention all over the world. Machine learning is a technique that shows examples and teaches and trains machines on their own, instead of typing instructions into the program one by one. Artificial Intelligence Go 'AlphaGo' is a good example. AlphaGo has learned how to play Go, observing accumulated records of baduk. As such, machine learning is a way for a computer to learn data on its own without humans giving rules directly to the computer. In the end, it is important to get good quality data for the computer to learn on its own. Thus, the purpose of this study is to present a research model of data profiling using statistical analysis techniques for comprehensive review and data extraction of data quality prior to using the big data.

2. Theoretical Background

2.1. Big Data Value and Quality Factors

It is analyzed that the potential value of big data emphasizes the value of data rather than big volume [2, 18], and collecting and storing big data and finding new information based on it will have a significant value creation effect [3, 19]. In Table 1, the Economist, Gartner, and McKinsey also highlighted the potential and socioeconomic value of Big Data [4].

Table 1. Potential of Big Data and Social-Economic Value of Forecast

Institution name	Key forecasts
Economist	Data is about the same level of economic input as capital or labor, the new raw material for a business.
Gartner	Data is the crude oil of the 21st century, and will influence future competitive advantage, and companies must understand the upcoming data economy era and watch for information silos for success.
McKinsey	Big Data is a core component of innovation, competitiveness and productivity, creating more than $600 billion in value in five areas: healthcare and public administration.

Data quality is defined as a level ensuring the latest, accurate, and interrelated data and giving it useful value to users. Data quality management to maintain or enhance the quality of data from a user perspective is a set of activities such as setting data quality goals, diagnosing and improving quality, and tools to support it [6]. The quality indicator of the data is as follows [6].

Table 2. Data Quality Indicator

Index	Definition of indicator
Readiness	Indicator that defines the elements that should be basically managed for data management, and measures whether it is updated through continuous activities.
Completeness	Indicator that measure the logical design and physical structure of data warehouses that are designed to store data in accordance with business requirements.
Consistency	Indicator that measures whether an entity's attributes comply with a defined standard (standard terminology, code, domain) and there is consistency in the use or association of attributes.

Accuracy	Indicator that measures the level of action such as logic or method for preventing errors in the data input stage in order to obtain the prior quality, level at which the stored data is in the range and format of valid information according to the defined criteria, integrity of the cross-reference relationship level to be secured, whether the actual input value is stored according to business rules.
Security	Indicator that measures the level of management of who is responsible for the creation and management of the data operated to ensure continuous quality, the level of data access management by authority, and the level of database security management for minimizing the interruption of information service.
Timeliness	Indicator that measures whether the program performance is secured with the response time that the user satisfies, minimizes the work time from the receipt of the information requirement to the collection, processing and provision of the information requirement, and the level of the latest information.
Usefulness	Indicator that measures whether sufficient information is provided at the level that the user is satisfied with, whether the access to the information is convenient for the user, and the level at which the user is making useful use of the information.

It is desirable that the quality of the data is based on the big data processing technology and it is preferable to decide according to the usage data suitable for the usage purpose from the viewpoint of the user who utilizes it, and it is formed at the data value chain stage between the data producer, the data processor, and the data consumer [5]. From the user's point of view, the quality attribute classifies data quality types in terms of intrinsic, accessibility, contextual, and expressive aspects of data, and the detailed quality factors to guarantee the data quality are as follows [5].

Table 3. Data Quality Type and Detail Quality Factor

Type of data quality	Data quality factor
Intrinsic	Excellence in data itself, including accuracy, objectivity, and authenticity
Accessibility	Environmental excellence accessible to data, such as accessibility, and access security
Contextual	Relevance to circumstances users intended, such as relevancy, temporality, completeness, and the amount of data
Representational	Clarity of data representation, such as interpretability, conciseness, ease of use, and consistency

There are many ways to secure data quality, and it is required to make efforts for defining management items in consideration of data values, business rules, data standards as for data quality elements, and continuous data quality management through quality management infrastructure (tools/systems), quality control and policies and organization [15]. The components for big data quality control are as follows [15].

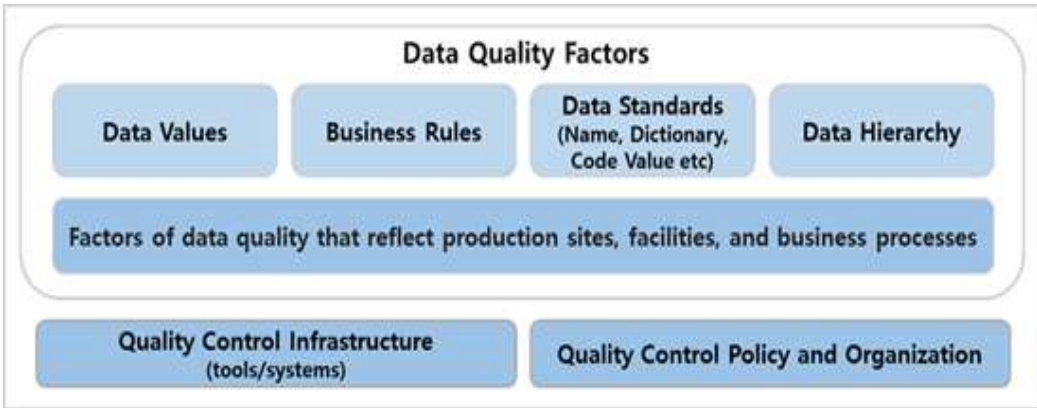

Figure 1. Components for Big Data Quality Management

2.2. Data Quality Error Calculation

The data quality error rate is a value obtained by quantifying the quality level of the database from the viewpoint of value, structure, and standardization and quantifying the result, and the quality diagnosis result of value, structure and standardization is classified into three quality indicators (accuracy, completeness, and consistency) and scored by error rate (indicator) [6]. The quality error rate of the database is calculated as the sum of the error rates for each quality factor (value, structure, and standardization), and the quality error rate is calculated as follows [6, 20].

$$QE(\%) = \sum_{i=1}^{n} (E_i \times W_i) \quad (1)$$

Where, i is a quality factor, E is an error rate per quality factor, and W is a weight per quality factor. The definition of error rate by quality factor is as follows.

Table 4. Definition of Error Rate by Quality Factor

Quality factor error rate	Definition of quality factor error rate
Value(accuracy) error rate(E_1)	Meaning error level for values in database
Structure(Completeness) error rate(E_2)	Meaning the extent to which the structure of the database is not faithful
Standard(Consistency) error rate(E_3)	Meaning compliance of database standards is insufficient

The criteria for applying weight by each quality factor is as follows.

Table 5. Weighting Criteria by Quality Factor

Quality factor(E_i)	Related indicator	Weight(W_i)
Standard	Consistency	0.2
Structure	Completeness	0.1
Value	Accuracy	0.7

Value (accuracy) error rate (E_1) calculation formula is as follows.

$$E_1 = \frac{e}{S} \times 100, s = \sum_{i=1}^{n} S_i, e = \sum_{i=1}^{n} e_i \quad (2)$$

Where, i is a value diagnostic item, S is the total number of data, and e is the number of error data. Structure (completeness) error rate (E_2) calculation formula is as follows.

$$E_2 = \frac{1}{n} \times \sum_{i=1}^{n} e_i \quad (3)$$

Where, n is the number of structural diagnosis items, i is the structural diagnosis item, and e is the error rate of each structural diagnosis item. Standard (consistency) error rate (E_3) calculation formula is as follows.

$$E_3 = \frac{1}{n} \times \sum_{i=1}^{n} e_i \quad (4)$$

Where, n is the number of standard diagnostic items, i is the standard diagnostic item, and e is the error rate per standard diagnostic item. The criterion for data value (Accuracy) quality diagnosis is defined, but selection of attribute for quality diagnosis is performed by subjective judgment of performer or data quality diagnosis is performed for all items. Especially, in the case of large-capacity data, it is more inefficient and the data error rate (%) might be inversely proportional to the number of data.

2.3. Data Quality Measurement Metrics

The data quality measurement metric is obtained by applying different weights to the columns according to the ratio of the error data in entire database and the purpose of data use, and the calculation formula of the data quality metric is as follows [7].

$$Q = 1 - \frac{T}{N}, \quad T = \sum_{j=1}^{m_i} C_{ij} \times W_j, \quad C_{ij} = \sum_{k=1}^{a_{ij}} n_{ijk} \quad (5)$$

Where, Q is data quality measurement value, N is the number of total quality measurement object, T is the number of total error data granted with weight by column, C_{ij} is the number of error data corresponding to each column, W_j is the weight corresponding to each column, a_{ij} is the number of total error data item required for data quality measurement, n_{ijk} is the number of a_{ijk} error data, and a_{ijk} is error data item k required for data quality characteristics. The data quality measurement metric is useful for extracting the rate of occurrence of the entire database error but the quality is measured only by the importance (=2), middle (=1) and lower (=0), and there is a possibility that there will be many variations in the measurement result depending on the subjective judgment of the performer who performs the measurement.

2.4. Study on Statistical Analysis Technique

Outlier means the data points that are not well explained by a given regression model, and Externally Studentized Residual is used for outlier detection [8]. Studentized residuals are the residuals divided by the standard deviation of the residuals, and generally the standard deviation is obtained for the whole data [9]. If the value of Bonferroni p observed from the data is smaller than 0.05, it is detected as an outlier value [8, 10].

Regression analysis is a method of modeling the functional relationship between continuous output variable Y and input variable X as follows [11].

$$Y = f(X) + \varepsilon \tag{6}$$

Where, ε is error item, expectation result is 0 and diffusion is σ^2. Therefore, in regression analysis the function is estimated from the data assuming the conditional expectation result of Y when X is given. Linear regression analysis assumes that the relationship between input and output variables is linear and that f is a linear function. Linear models use the formula operator and the general interpretation of the formula operator is as follows [8].

Table 6. Formula Operator

Operator	Formula(example)	Meaning
+	Y~X1+X2	Model Y as X1, X2. Constant term is implicitly allowed. Thus, using this formula for linear regression implies Y=a*X1+b*X2+c.
-	Y~X1-X2	Y is modeled as X1, but X2 is excluded. In particular, in the linear regression, Y~X1+X2-1 means that Y is modeled as X1 and X2, but the constant term is excluded. That is, Y=a*X1+b*X2.
|	Y~X1|X2	Data is grouped according to the value of X2, and then Y~X1 is applied to each group.
:	Y~X1:X2	Y is modeled according to the interaction of X1 and X2. Interaction refers to a situation where X1 and X2 simultaneously affect the Y value, such as Y=a*X1*X2+b.
*	Y~X1*X2	An abbreviated expression of Y~X1+X2+X1:X2.

3. Research Model

3.1. Method to Extract Data Attributes

A flowchart for data attribute value extraction, data attribute value extraction method, empirical method of statistical analysis, attribute weight extraction criterion, and attribute correction value extraction criterion are presented. A flowchart for extracting data attributes is as follows.

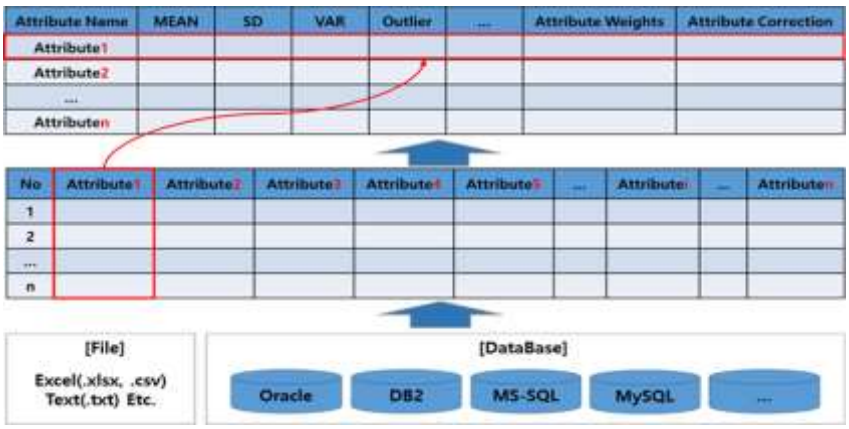

Figure 2. Flowchart for Extracting Data Attribute

Data profiling is performed using statistical analysis technique to extract attribute combination for data quality diagnosis. The statistical analysis technique extracts attribute values by applying missing values (NA), variance, standard deviation, five numerical summaries, outliers, and near zero variance. And it extracts attribute weights and attribute correction values. The collected data processing uses the data analysis tool R package. The missing value is different from the blank value, while the null values are nothing but the missing values, but there is something unknown [12]. The variance is number that measures how far the random variable is from the expected value, and $\mu = \mathrm{E(X)}$ is expectation result of random variable, dispersion (var(X)) is as follows [13].

$$\mathrm{var(X)} = \mathrm{E}((X-\mu)^2) \tag{7}$$

That is, it is equal to the average of squared distances from the mean of X. The standard deviation is a measure of the scattering of the data, defined as the square root of the variance, not the negative. The smaller the standard deviation, the closer the variance is. The standard deviation of the random variable X (σ) is as follows [13].

$$\sigma = \sqrt{E(X-E(X))^2} = \sqrt{E(x^2)-(E(X))^2} \tag{8}$$

The five numerical summary extraction means that the whole data is quadrupled when the whole data is arranged in the order of magnitude. One quarter (Q1), two quarters (Q2 or median), and three quarters (Q3). The median (Q2), the quartiles (Q1, Q3), the minimum value, and the maximum value correspond to the positions of 25%, 50%, and 75% of the total. The outlier range is as follows.

$$\text{Outlier Range} = [Q_1 - \mathrm{k}(Q_3 - Q_1), Q_3 + \mathrm{k}(Q_3 - Q_1)] \tag{9}$$

Where, Q_1 is a quarter of the whole, Q_3 is three-quarter of the whole, and k is 1.5. Regression models are used to extract outliers and the formula applied is as follows.

$$\mathrm{F} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim X + X^2 = \bar{X} \sim X + X^2 \tag{10}$$

Where, i is 1, …, n and X is input variable and $\bar{\bar{X}}$ is mean of X. The linear model is used for the idea extraction and the fit model using the lm() function of the R package is as follows.

$$\mathrm{FM} = \mathrm{lm}\big(\mathrm{mean(x)} \sim \mathrm{X} + \mathrm{I}(X^2)\big) \tag{11}$$

In the extracted attribute value, the attribute weight extraction criterion is as follows.

Table 7. Attribute Weight Extraction Criteria

Data type	Weighting criteria	Weight
All data types	Missing value(NA)>0	0.1
Integer or Numeric	Near Zero Variance is TRUE	0.1
Integer or Numeric	SD(Standard Deviation) $\geq$ 100	0.1
Integer or Numeric	Outlier Bonferroni p < 0.05	0.1
Factor	Space > 0	0.1
Date	(LastDate - FistDate)>(CurrentDate - FirstDate)	0.1

The criterion for extracting attribute correction values from extracted attribute values is as follows.

Table 8. Attribute Correction Value Extraction Criteria

Data type	Criteria for applying correction values	Correction value
All data types	The number of missing values (NA) is 1% or more	0.1
Integer or Numeric	Outlier Bonferroni p 0.1e-5	0.1

3.2. Architecture for Model Derivation

The model derivation is implemented using R package and R-Studio, and the architecture for research is as follows.

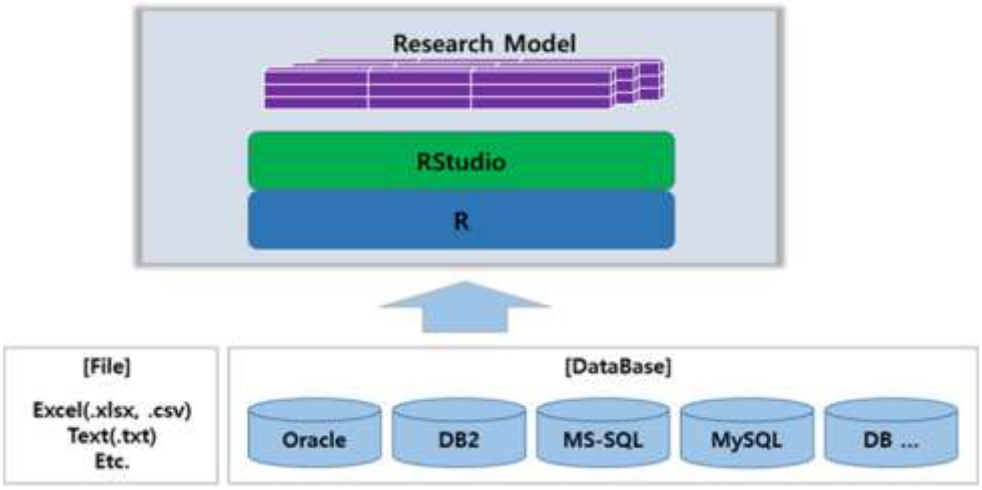

Figure 3. Architecture for Research

3.3. Research Model

This study presents a research model using the geometric mean of the Pythagorean mean in the attribute combination extraction model. The data matrix of n data attributes and attribute extraction values is as follows.

$$\text{Attribute}(a)_{n\times p} = \begin{bmatrix} a_{11}a_{12}\dots a_{1p} \\ a_{21}a_{22}\dots a_{2p} \\ \dots a_{ij}\dots\dots \\ a_{n1}a_{n2}\dots a_{np} \end{bmatrix} \quad (12)$$

The extracted attribute weights and attribute correction data are used to find the optimal combination of error test columns, and the formula for calculating the geometric mean (GM(weight+correction)) of the attribute weights and attribute correction values summed from the attribute values is as follows.

$$\text{GM}(\text{weight} + \text{correction}) = \sqrt[n]{\prod_{a_i \in S}^{n} (a_{iw} + a_{ic})} \quad (13)$$

Where, S is attribute set ($a_1, a_2, \ldots, a_n$), n is the number of attribute selected among S group (But, n>1), a_i is ith attribute and subset of S, a_{iw} is ith attribute weight, and a_{ic} is ith attribute weight.

4. Empirical Analysis

4.1. Data Collection and Analysis Method

In order to evaluate the performance of this study, we use the Delhi Weather data registered in Kaggle [14]. The attribute description of the collected data is as follows.

Table 9. Collection Data Attributes

Attribute description		Attribute description	
datetime_utc	String	_snow	Numeric
_conds	String	_tempm	Numeric
_dewptm	Numeric	_thunder	Numeric
_fog	Numeric	_tornado	Numeric
_hail	Numeric	_vism	Numeric
_heatindicatorm	String	_wdird	Numeric
_hum	Numeric	_wdire	String
_precipm	String	_wgustm	String
_pressurem	Numeric	_windchillm	String
_rain	Numeric	_wspdm	Numeric

The datetime_utc attribute of the Delhi Weather data is a total of 21 attributes separated by date and time for data value profiling, and the entire data (100,990 total) is used for attribute value extraction. We used the statistical analysis function provided by R package for analysis.

4.2. Performance Evaluation Method

In this paper, we use the data value (accuracy) error rate (%) formula (2) and the attribute result derived from equation (13). In this paper, we compare and evaluate the data quality (efficiency) measurement value (data quality efficiency measurement value) to compare the data value (accuracy) error rate (%) and the performance of the research model, and calculation formula is as follows [16].

$$\mathrm{DQEM}(\%) = \left(1 - \frac{m}{S}\right) \times 100 \tag{14}$$

Where, S is a product of the total number of attributes and the total number of data (records), and m is a product of the number of attributes subject to error verification and the total number of data (records). The data quality efficiency measurement value is closer to 100%, and the performance (efficiency) is excellent. Generally, data values (accuracy) are used to diagnose all data attributes for error rate (%) diagnosis. For example, for a total of 100,990 Delhi Weather data and 21 attributes, 100,990×21= 2,120,790 Diagnose error of value (accuracy) by number of times. In the study, we are going to assume the prerequisites in Table 10 for performance evaluation.

Table 10. Performance Evaluation Prerequisites

Division	Precondition
Value (Accuracy) error rate (%)	All Attributes are measured, but date and time are record key attributes, so they are excluded from measurement
Research model	All the attributes extracted by Eq. (13) is object of measurement

4.3. Data Attribute Combination Experiment Result

The experiment result of the data attribute combination by Eq. (13) is as follows.

Table 11. Attribute Combination Experiment Result

Attribute weighting classification table	Result of attribute combination test	Experiment result value
Attribute weight 0.1	_dewptm, _fog, _hail, _hum, _pressurem, _rain, _snow, _tempm, _thunder, _tornado, _vism, _wdird, _wspdm	0.2970103
Attribute weight 0.2	_dewptm, _hail, _hum, _pressurem, _rain, _snow, _tempm, _tornado, _vism, _wdird, _wspdm	0.3399097
Attribute weight 0.3	_pressurem, _wdird	0.4472136

As a result of attribute combination experiment, _pressurem, _wdird attribute is the most highly likely to make error, but in this paper, we evaluate the performance of all attributes derived by Eq. (13).

4.4. Performance Evaluation Results

The results of performance evaluation using Table 10 and Table 11 for extracting data quality efficiency measure (DQEM) are as follows.

Table 12. Performance Evaluation based on Preconditions

Division	Data quality efficiency measurement value (%)
Value (Accuracy) error rate (%)	$\left(1-\frac{19\times 100{,}990}{21\times 100{,}990}\right)\times 100 = 9.524$
Research model	$\left(1-\frac{13\times 100{,}990}{21\times 100{,}990}\right)\times 100 = 38.095$

Table 12 shows the experimental results based on the preconditions of Table 10 and it is confirmed that the research model is superior to the value (accuracy) error rate by 28.571%.

5. Conclusion

In this paper, we present a Big Data Profiling research model using statistical analysis for attribute extraction for data quality diagnosis. We calculated the statistic value by using empirical method of statistical analysis for all attributes, calculated the attribute weight by weighting the criterion, and calculated the attribute correction value based on the attribute correction value extraction criterion in the calculated statistic. The attributes for the data quality diagnosis were derived by using the geometric mean of the extracted

attribute weights and the sum of the attribute correction values. Generally, for data quality diagnosis, data quality diagnosis is performed by performing data quality diagnosis on all attributes and data, or weighting important attributes by subjective judgment of the performer. In this paper, we confirmed that the performance evaluation result study model is 28.571% better than the value (accuracy) error rate calculation model.

The scientific implications and meaning of this study is that the research model for data profiling by the more scientific method, rather than depending on the subjective judgment of the practitioner for the data quality diagnosis and the excellence of the research model is clarified through the performance comparison. As a limitation of this study, data quality analysis should be performed considering data values, business rules, data standards, data structures, and characteristics of industrial sites, but a data profiling model based on data values is suggested. Future research could be applied to various data quality factors.

References

[1] D. H. Choi, J. H. Lee and U. M. Kim, "Survey on Issue and Utilizing of Big Data", Proceedings of Korea Society of IT Services, vol. 2015, **(2015)**, pp. 141-144.
[2] IDC, Editor, "Worldwide Big Data Technology and Service 2012-2015 Forecast", IDC #233485, **(2012)** March.
[3] McKinsey & Company, Editor, "Clouds, big data, and smart assets: Ten tech-enabled business trends to watch(2010)", **(2010)**.
[4] J. S. Jung, Editor, "New value creation engine, new possibilities in Big Data and Strategies", IT & Future Strategy, NIA, **(2011)** December.
[5] S. K. Choi and S. C. Jeon, "A propose of Big data quality elements", Journal of Korea Navigation Institute., vol. 17, no. 1, **(2013)**, pp. 09-15.
[6] NIA, Editor, "Public Information Quality Management Manual (2014)", **(2014)**.
[7] J. Y. Yang and B. J. Choi, "A Methodology of Measuring Data Quality from Viewpoint of Software user", Proceedings of the 28th KISS Fall Conference, vol. 28, no. 2, **(2001)** October, pp. 436-438.
[8] M. G. Seo, Editor, "Data Processing and analysys using R", gilbut, Seoul, **(2014)**.
[9] http://en.wikipedia.org/wiki/Studentized_residual, **(2017)** January 17.
[10] E. Schubert, A. Zimek and H. P. Kriegel, "Local outlier detection reconsidered: A generalized view on locality with application to spatial, video, and network outlier detection", Journal of Data mining and knowledge discovery, vol. 28, no. 1, **(2014)**, pp. 190-237.
[11] C. L. Park, Y. D. Kim, J. S. Kim, J. W. Song and H. S. Choi, Editor, "Data mining using R", kyowoo, Seoul, **(2011)**, pp. 17-28.
[12] D. S. Lee, "Design of an Inference Control Process in OLAP Data Cubes", Doctoral dissertation, Soongsil University, **(2009)**, pp. 190.
[13] S. M. You and H. J. Park, Editor, "Basic Statistics Learned with Minitab", Eratec, Gyenggi, **(2006)**, pp. 442-473.
[14] https://www.kaggle.com/mahirkukreja/delhi-weather-data/data, **(2017)** December 1.
[15] H. G. Park, H. G. Song, W. J. Jang, S. R. Lee and C. S. Lim, Editor, "Fourth Industrial Revolution, era of New Manufacturing", HeuteBooks, Gyenggi, **(2017)**, pp. 23-108.
[16] W. J. Jang, "Big Data Profiling Statistical analysis based on the R language", Master dissertation, Sogang University, **(2015)**.
[17] NIPA, Editor, "Ride the wind Big Data, business analytics software market is the fastest growing", IDC & Info World, **(2012)** June.
[18] Measuring the Economics, Editor, "Big Data", Working Party on Indicators for the Information Society, DSTI/ICCP/IIS, OECD, **(2011)**.
[19] NIPA, Editor, "Future society and Big data technologies", IT & Future Strategy, **(2012)** April.
[20] W. J. Jang, J. Y. Kim, B. T. Lim and G. Y. Gim, "A Study on Data Profiling based on Data Attribute Value Analysis", Advanced Science and Technology Letters ASTL 150, Ho Chi Minh, Vietnam, **(2018)** February 1-3, pp. 206-209.

Authors

Won-Jung Jang, (wjjang@cku.ac.kr)

He works as a professor at the Department of Global Start-up Consulting Department at Catholic Kwandong University and he is studying in Soongsil University Graduate School of IT Policy and Management. His research areas are Big Data, Machine Learning, Artifical Inteligence, and Software Engineering, etc. His books include 4th Industrial Revolution, How to start and 4th Industrial Revolution, the era of new manufacturing.

Jong-Yoon Kim, (kimjw@nice.co.kr)

He is working for National Information & Credit Evaluation Inc as executive director. And he is studying is Soonsil University Graduate School of IT Policy and Management. His interests include CB(Credit Buraeu) business, Consumer credit risk modeling and big data analysis.

Bum-Taek Lim, (neobtlim@gmail.com)

He worked for Defense Integrated Data Center and was a System Operation Department Manager. And he works at DIDC as a technical advisor on defense IT Policies, Big data and the cloud. Currently his major is IT Service Management, hi is studying in Soongsil University Graduate School of IT Policy and Management. His research areas include Big data, The cloud and system operations on based AI.

Gwang-Yong Gim, (gygim@ssu.ac.kr)

He works as a professor at the Department of Business Administration of Soongsil University. Dr. Gim has been interested in research such as 4th Industry Revolution, ICT ODA, intellectual property rights, service science, big data analysis, S/W industrial policy, and open innovation. He published a number papers on journals such as Information Science, Fuzzy sets and System, journals of society of management information systems, and journals of management science.