Computational Intelligence Techniques for Assessing Data Quality: Towards Knowledge-Driven Processing

Nunik Afriliana[1,2][0000−0001−6314−6128], Dariusz Król[1][0000−0002−2715−6000], and Ford Lumban Gaol[3][0000−0002−5116−5708]

[1] Wroclaw University of Science and Technology, Wroclaw, Poland
{nunik.afriliana,dariusz.krol}@pwr.edu.pl
[2] Universitas Mutimedia Nusantara, Jakarta, Indonesia
[3] Computer Science Department, Bina Nusantara University, Jakarta, Indonesia
fgaol@binus.edu

Abstract. Since the right decision is made from the correct data, assessing data quality is an important process in computational science when working in a data-driven environment. Appropriate data quality ensures the validity of decisions made by any decision-maker. A very promising area to overcome common data quality issues is computational intelligence. This paper examines from past to current intelligence techniques used for assessing data quality, reflecting the trend for the last two decades. Results of a bibliometric analysis are derived and summarized based on the embedded clustered themes in the data quality field. In addition, a network visualization map and strategic diagrams based on keyword co-occurrence are presented. These reports demonstrate that computational intelligence, such as machine and deep learning, fuzzy set theory, evolutionary computing is essential for uncovering and solving data quality issues.

Keywords: Big Data · Data Quality · Computational Intelligence · Knowledge Engineering · SciMAT · Uncertainty Processing · VOSviewer.

1 Introduction

Data plays a significant role in every organization. High-quality data are those that can be quickly analyzed to reveal valuable information [19]. Therefore, data quality assessment is essential to be performed to ensure the quality of data. Improving data quality is the most crucial process of a today's zettabyte era [4].

Nowadays, data shows explosive growth, with collaborative, heterogeneous, and multi-source characters, which increases the complexity and difficulty of data assessment. Assessing data quality is a challenging task. Data quality (DQ) has been investigated extensively, however, only a few research looks at the actual data quality level within organizations [22]. Along with the rapid development of computer science, computational intelligence techniques become noticeably promising.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

In this paper, a complete bibliometric analysis is developed, by retrieving articles from the following databases: IEEE Xplore, ProQuest, ScienceDirect, Scopus, Springer Link. The time window to be analyzed is 2001–2020. The keywords for this searching were "data quality" or "data quality assessment" or "data quality improvement" or "assessing data quality". The number of documents retrieved was 354 articles including 351 unique documents. These articles were screened and analyzed based on their title and abstract following the PRISMA flow diagram guideline [18] shown in Fig. 1. According to the screening process, 190 articles were excluded based on the title. We then evaluate the abstract of 161 articles. 42 articles were excluded based on the abstract resulting in 119 articles eligible to be reviewed. Eventually, 93 articles were included in this paper based on the full-text review.

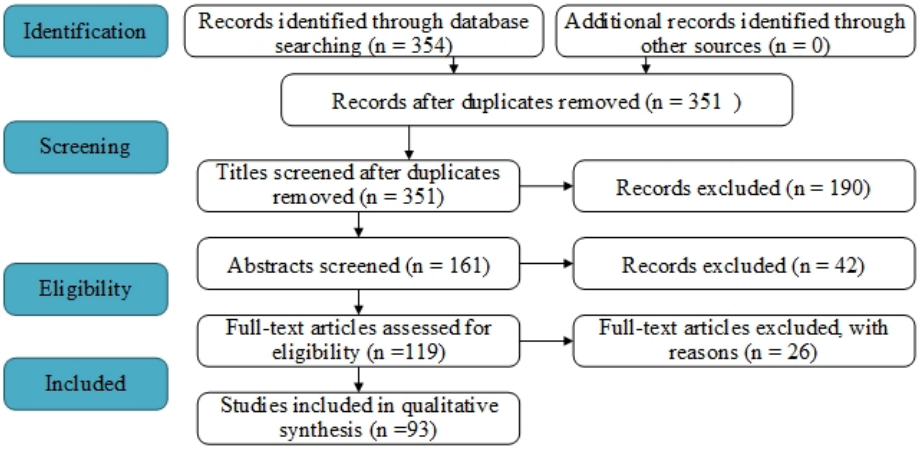


Fig. 1. The article screening process flow based on PRISMA guideline.

An analysis was performed using two bibliometric software tools, namely VOSviewer ver. 1.6.16 and SciMAT ver. 1.1.04. VOSviewer is a freely available computer program for constructing and viewing bibliometric maps [10]. VOSviewer was utilized to generate a network and density visualization. SciMAT is a performance analysis and science mapping software tool for detecting and visualizing conceptual subdomains [7]. In this research, the SciMAT was utilized to build a strategic map for data quality assessment trends. Both analyses were conducted based on the article's keywords in order to show growing patterns in DQ-related techniques.

2 Performance Analysis

The mapping analysis based on the co-occurrence of keywords from articles was carried out using VOSviewer. There were 1430 keywords identified from those

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

articles. We used the threshold of at least three co-occurrence of the keywords to be represented in a node. It has clustered the research trend into 7 clusters. These clusters consist of 91 items and 851 links. Every cluster is represented by different colors. Therefore, it can be easily investigated, see Fig. 2. The total items from cluster 1 to cluster 7 are 21, 15, 15, 13, 10, 9, and 8, respectively. The density of every item is visualized by the size of every node in the network. In cluster 1, the densest item is data quality with 59 occurrences, followed by data quality assessment and data reduction with 30 and 17, respectively. The distribution of the four densest items and the related techniques or algorithms that occurred in every cluster is presented in Table 1.

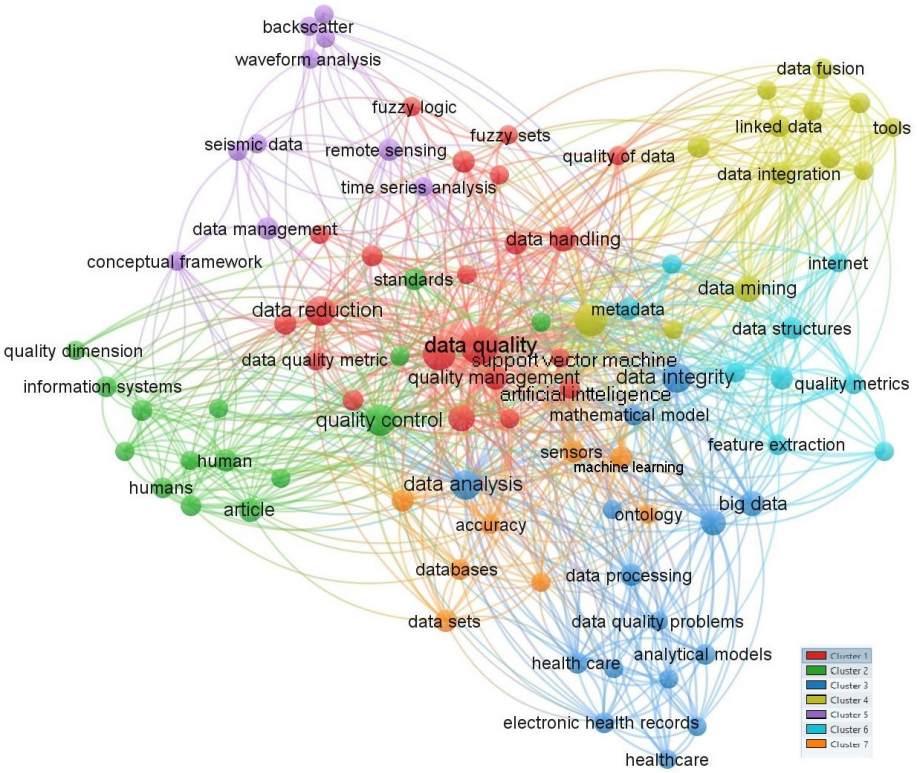


Fig. 2. Network map based on keywords co-occurrence generated by VOSviewer.

The second analysis was performed using SciMAT science mapping software tool. It generated the strategic diagram in the selected period. We divided the whole period (2001–2020) into four consecutive slides listed in Table 2.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

Table 1. Four densest items on the network.

Cluster	Four densest items	Techniques
Cluster 1	data quality, data quality assessment, data reduction, decision making	fuzzy set theory, fuzzy logic, artificial intelligence, support vector machine
Cluster 2	quality control, article, standard, quality dimensions	qualitative analysis
Cluster 3	data analysis, data integrity, data model, big data	analytic hierarchy process, mathematical model
Cluster 4	quality assessment, data mining, data integration, linked data	
Cluster 5	data management, conceptual framework, remote sensing, time series analysis	time series analysis
Cluster 6	measurement, data structure, meta data, feature extraction	feature extraction
Cluster 7	sensor, machine learning, quality, data sets	machine learning

Table 2. Performance distribution divided into four consecutive slides.

Period	Number of documents	Number of citations
2001-2005	5	28
2006-2010	15	802
2011-2015	30	620
2016-2020	43	159

A strategic diagram is a two-dimensional space built by plotting themes according to their centrality and density rank value [7]. It consists of four quadrants with the classification as follows:

- The upper-right quadrant is a motor theme with strong centrality and high density. It is important for structuring a research.
- The upper-left quadrant includes highly developed internal ties, very specialized and isolated themes.
- The lower-left quadrant shows emerging or declining themes, weakly developed (low density) and marginal (low centrality).
- The lower-right quadrant presents transversal and general themes. These themes are important for the research field but are not developed.

There are 4 strategic diagrams have been generated by the SciMAT that represent 4 periods from 2001 to 2020. Due to limitation of the space we only presented one strategic diagram from 2016-2020 period shown in Fig. 3, however Fig. 4 gives a more straightforward overview and summarizes the essential themes for structuring research in data quality assessment from all slides in two decades. We considered only those themes that resided in two quadrants: motor themes and basic and transversal themes.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

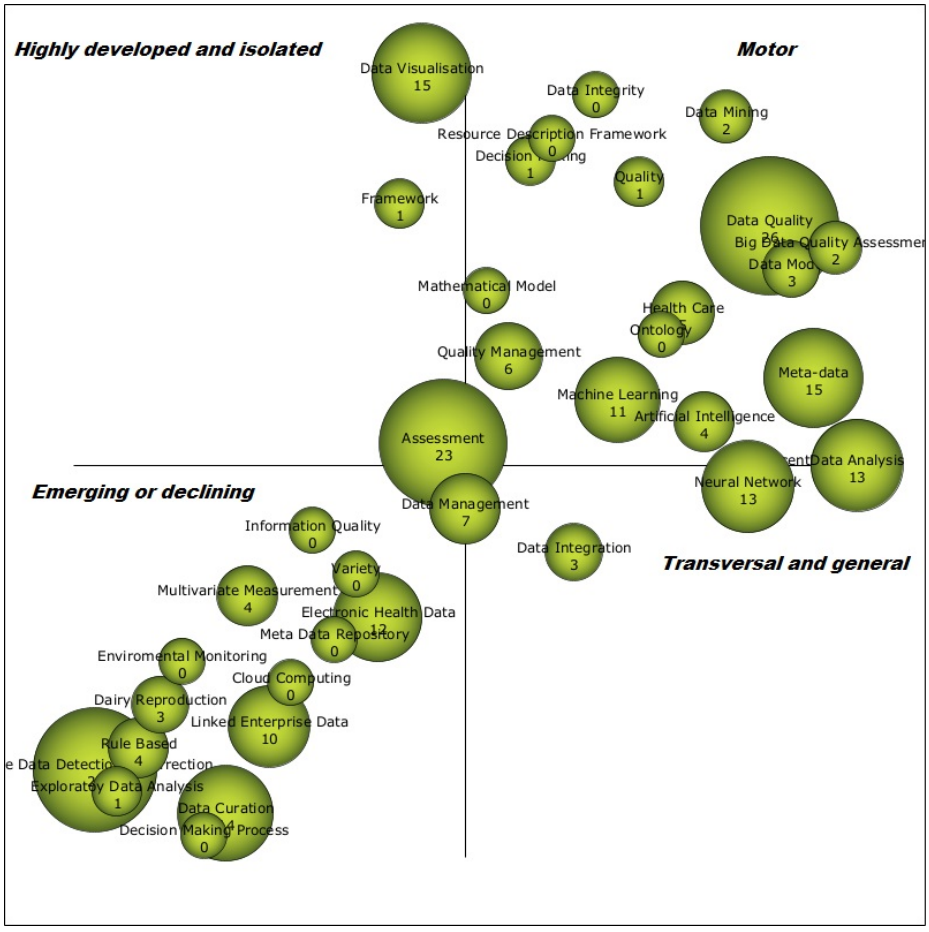


Fig. 3. Strategic diagram for the 2016–2020 period based on the sum of citations.

Based on the strategic diagram for the first period (2001–2005), the research motors are mathematical themes and some themes that are not specified in any particular field. Quality control theme is transversal and general, which means it is essential for the study but not well developed. The highest citation is quality control themes with 28 citations from 3 documents. Therefore, it can be concluded that there was no sufficient research concerning the intelligence techniques for assessing data quality during this period.

Quality assessment, quality improvement, unsupervised classification, medical audit are mainly the motor themes within the 2006–2010 period. In this period, data quality assessment gained more attention along with the intelligence techniques such as unsupervised classification. These themes were playing essential roles in the research field as they were in the motor theme quadrant. The most cited theme with 622 citations from 8 documents are data quality,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

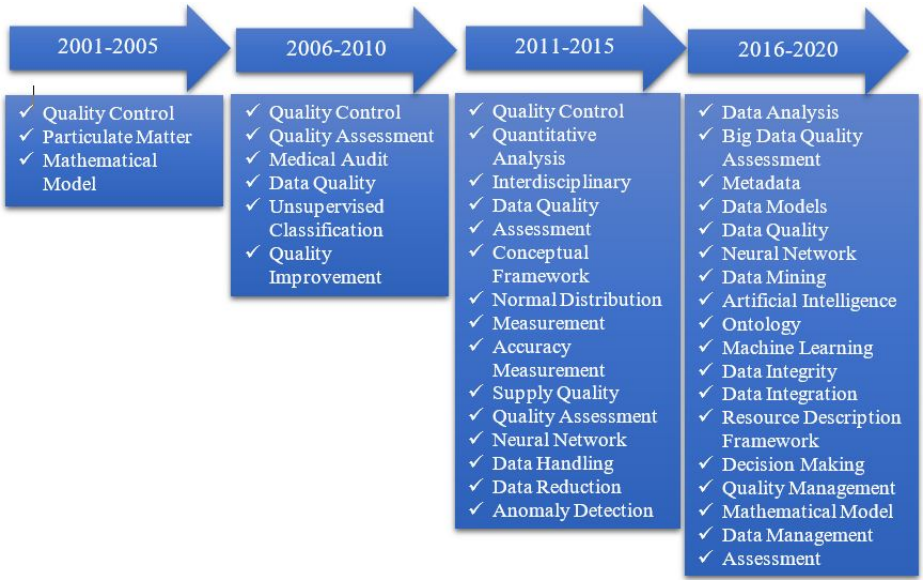


Fig. 4. Essential themes for structuring research in the whole period.

which belonged to the transversal and general quadrant. In this period, data quality and quality control were important for structuring research, although they are general themes.

Quality assessment, anomaly detection, measurement, qualitative analysis, conceptual framework, and neural network are mainly the most cited themes from the motor quadrant within 2011–2015 period. The number of citations of these themes is 460. Artificial intelligence and machine learning themes are placed in the highly developed and isolated quadrant. In the transversal and general quadrant, there are data quality, assessment, quality control, accuracy measurement, and data reductions. These themes were also crucial but not well developed as they belong to the transversal and general quadrant.

Fig. 3 shows the recent themes for the 2016-2020 period. They are mainly distributed into two quadrants: motor and emerging or declining quadrant. Only a few themes were plotted in the highly developed and isolated, or transversal and general quadrant. Thus, it would be easier to conclude what are the trends and important themes. From this strategic diagram, we observe that intelligence techniques represented by artificial intelligence, machine learning, and neural network are important for structuring the research field. Artificial intelligence and machine learning are both plotted in motor quadrant, which is well-developed and important for a research, while neural network is at the border between transversal and general quadrant.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

3 Top 3 Most Common Data Quality Issues

3.1 Data quality dimensions

Data quality dimension is intended to categorize types of data quality measurement. It consists of accuracy, completeness, concordance, consistency, currency, redundancy, or any other dimension [28] [34]. For instance, the World Health Organization identifies the completeness and timeliness of the data, the internal coherence of data, the external coherence of data, and data comparisons on the entire population as data quality indicators [24]. Therefore, stating a set of requirements for data quality is crucial in establishing the quality of the data despite standards and methodology used [25]. The authors of [6] concluded that the most common data quality dimensions are completeness, timeliness, and accuracy, followed by consistency and accessibility.

3.2 Outlier detection

Outlier detection is a process to identify objects that are different than the majority of data, resulting from contamination, error, or fraud [15]. Outliers are either errors or mistakes that are counterproductive. Identifying outliers can lead to better anomaly detection performance [36]. Detecting outliers can also be utilized to give insight into the data quality [26]. In [17], an outlier is used to measure the consistency of climate change data, while in [20], an outlier detection method based on time-relevant k-means was used to detect the voltage, curve, and power data quality issues in electricity data.

In the strategic diagram presented in Fig. 5, outlier or anomaly detection was a motor theme for research on data quality within the 2011-2015 period. Therefore, it is crucial to identify the outlier in order to assess and improve the data quality.

3.3 Data cleaning and data integration

Data cleaning is the backbone of data quality assessment. It purposes to clean the raw data into new data that meet the quality criteria. It includes cleaning the missing values, typos, mixed formats, replicated entries, outliers, and business rules violations [5].

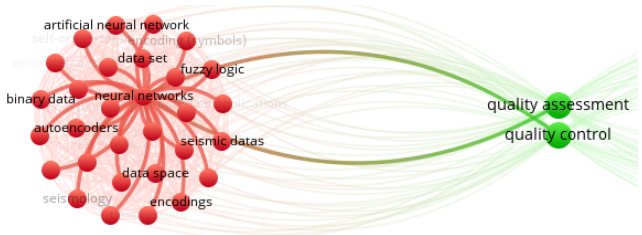
When we take a look at the strategic diagram in Fig. 3, data integration was plotted in the transversal and general quadrant. This theme has the highest citation among others. It means that data integration has an essential role in data quality research although it was classified as not developed. The study in [23] gives an example of a framework for data cleaning and integration in the clinical domain of interest. The framework includes data standardization to finally enable data integration.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

4 Computational Intelligence Techniques

4.1 Neural Networks for Assessing Data Quality

The neural network was plotted at the border between the motor and transversal quadrant (Fig. 3), while its bibliometric network is shown in Fig. 5(a). To be able to display all the networks, the analysis was made slightly different from Fig. 2. We changed the network reduction parameter from 3 to 1. The neural network has relations with items such as artificial neural network, autoencoders, data set, fuzzy logic, etc. It has also links with quality assessment and quality control.



(a) Neural network

(b) Genetic algorithm

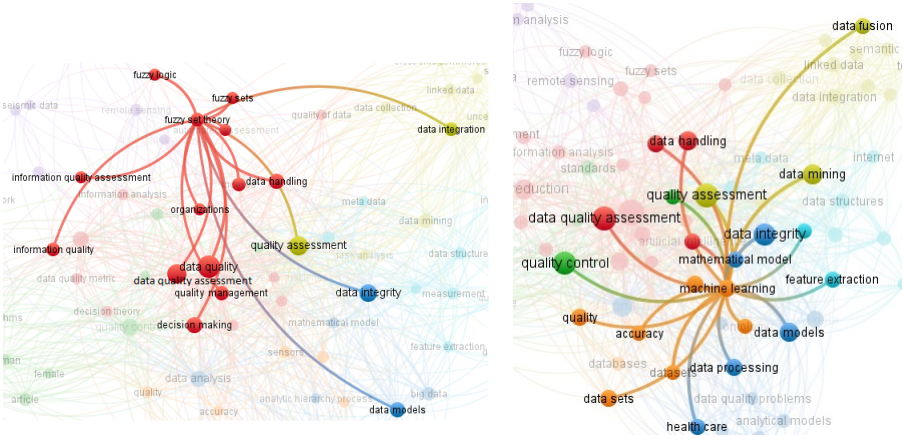
Fig. 5. The bibliometric network visualization.

Based on the literature, here are some examples of neural network implementation for assessing data quality. The VGG-16 and the Bidirectional Encoder Representations from Transformers are employed in a task-oriented data quality assessment framework proposed in [16]. Another usage of a neural network model is presented in [8]. Multi-Layer Perceptron was utilized to evaluate the framework of data cleaning in the regression model. The result shows that the models give a better result after the cleaning process. Long Short-Term Memory Autoencoder based on RNNs architecture was used in the DQ assessment model for semantic data in [14]. This framework presents the web contextual data quality assessment model with enhanced classification metric parameters. This contextual data contains metadata with various threshold values for different types of data.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

4.2 Fuzzy Logic for Assessing Data Quality

The network of fuzzy set theory in the data quality research is shown in Fig. 6(a). It is a part of the network created by VOSviewer in Fig. 2. It has 16 links, including the link to the data quality, data quality assessment, data integrity, decision making, and data models. Some fuzzy set theories were identified during this study. The Choquet integral in a fuzzy logic principle is utilized by data, and information quality assessment from the framework proposed in [3]. It is used in the context of fuzzy logic and as part of the multi-criteria decision aid system.



(a) Fuzzy set theory (b) Machine learning

Fig. 6. The bibliometric network visualization.

Fuzzy TOPSIS was used to measure the results of the systematic literature review approach to identify critical and challenging factors in DevOps data quality assessment [27]. It was implemented to prioritize the investigated challenging factors concerning the DevOps data quality assessment.

4.3 Evolutionary Computation for Assessing Data Quality

Two works in our set have been identified which use a genetic algorithm. The first study was using the genetic algorithm to measure the data quality in dimensions of accuracy, comprehensibility, interestingness, and completeness [21]. The second study [9] employed the association rule for the purpose of quality measurement. A multi-objective genetic algorithm approach for data quality with categorical attributes was utilized. The bibliometric network of the genetic algorithm is presented in Fig. 5(b). It has eight links, one of them for data mining.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

4.4 Computational Learning Theory for Assessing Data Quality

The computational learning theory is a fundamental building block of a mathematical formal representation of a cognitive process widely adapted by machine learning. The overview of machine learning's bibliometric map is presented in Fig. 6(b). It has 22 links connected to some items such as data quality assessment, data integrity, data fusion, mathematical model, etc. Random forests were utilized in [13] to predict the accuracy of an early-stage data cleaning, which was used to assess the quality of fertility data stored in dairy herd management software. Time-relevant k-means was employed to measure data accuracy by detecting the electricity data outlier [20].

Local Outlier Factor (LOF), an algorithm for identifying distance-based local outliers, was used in [8]. With LOF, the local density of a particular point is compared with its neighbors. An outlier is suggested if the point is in a sparser region than its neighbors. Another outlier detection used with respect to DQ is Density-Based Spatial Clustering of Applications with Noise (DBSCAN). It chooses an arbitrary unassigned object from the dataset. If it is a core object, DBSCAN finds all connected objects, and these objects are assigned to a new cluster. If it is not a core object, then it is considered as an outlier object [8].

Two clustering techniques, namely t-SNE and PCA, were utilized in a framework that aims to evaluate the precision and accuracy of experimental data [30]. Using t-SNE and PCA gives the dimensional reduction while retaining the most of the information to detect outlier eventually. For unstructured data, the study in [31] suggests to combine techniques such as machine learning, natural language processing, information retrieval, and knowledge management to map the data into a schema.

4.5 Probabilistic and Statistical Methods for Assessing Data Quality

The probabilistic and statistical approach was rarely being included in the abstract or articles' keywords, therefore its appearance in the network map or strategic diagram was not identified. However, this approach was identified varying from many methods and purposes, therefore human-oriented analysis on this approach was made. Univariate and multivariate methods for outlier detection are utilized in [23]. The framework of DQ Assessment for Smart Sensor Network presented in [1] used the interquartile statistical approach to detect outliers. If the data received is lower or greater than the boundaries, it is considered as an outlier. To find a mislabeled data, Shannon Index was utilized in [33]. Shannon Index is a quantitative measure that reflects how many different data there are in a dataset. If the Shannon index is lower than the threshold, it can be predicted as a mislabeled data.

In [2], the completeness of data is measured using logistic regression, while the timeline is measured using binomial regression. Chi-square of patient's age and sex in the medical health record was utilized in [32]. It computes mean

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

and median age by sex for the database population and compares to an external/standard population. Another utilization of the statistical approach is shown in [29]. An ANOVA test was performed to validate the selected characteristics and dimensions for assessing data quality.

In [11, 12], a probability-based metric for semantic consistency and assessing data currency is performed using a set of uncertain rules. This metric would allow conditions that are supposed to be satisfied with unique probabilities to be considered. The last approach was found in a medical big data quality evaluation model based on credibility analysis and analytic hierarchy process (AHP) in [35]. Firstly, data credibility is evaluated. It calculates the data quality dimensions after excluding the inaccurate data. Then by combining all dimensions with AHP, it obtains the data quality assessment outcome.

5 Final Remarks

Regarding uncertainty and a vast amount of data, computational intelligence such as machine learning, deep learning, fuzzy set theory, etc., are potent approaches for DQ assessment problems. The use of these methods has been identified in this work. However, the best practice still varies depending on the characteristics and goals of the assessment. The increasing number of successful implementations of these approaches demonstrates the versatility of computational intelligence techniques in assessing DQ.

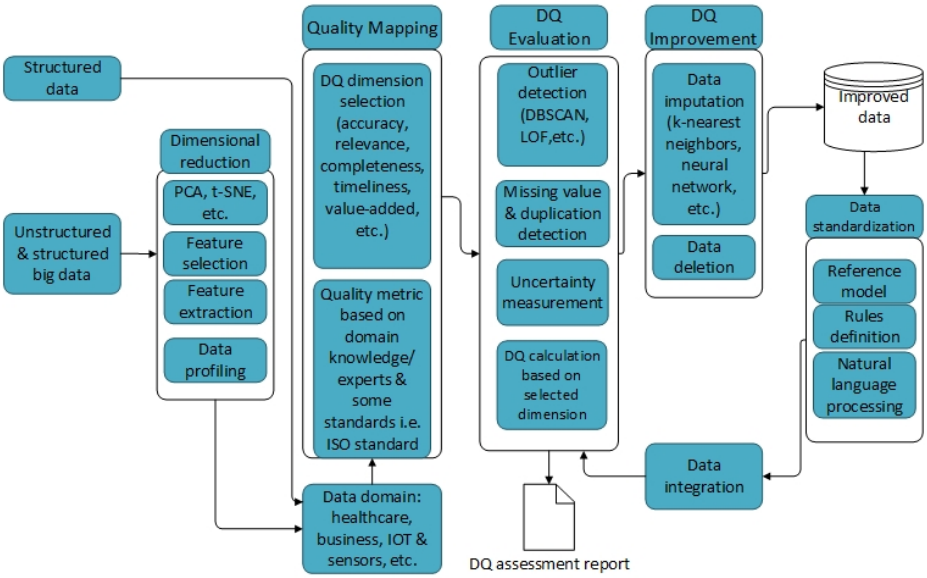


Fig. 7. Proposed data quality assessment framework.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

As a final note, we propose a conceptual framework for data quality evaluation, improvement, and possibly data integration, as shown in Fig. 10. This proposed framework will mainly consist of the data model, quality mapping, DQ evaluation, DQ improvement, data standardization, and data integration. In the future, we will evaluate the suitable computational intelligence techniques to be implemented into this framework. From our point of view, developing an intelligent data quality framework would require further advancement in computational science.

Acknowledgments. Part of the work presented in this paper was received financial support from the statutory funds at the Wrocław University of Science and Technology.

References

1. de Aquino, G.R.C., de Farias, C.M., Pirmez, L.: Hygieia: Data Quality Assessment for Smart Sensor Network. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. pp. 889–891. SAC '19, Association for Computing Machinery, New York (2019)
2. Boes, L., Houareau, C., Altmann, D., der Heiden, M.A., Bremer, V., Diercke, M., Dudareva, S., Neumeyer-Gromen, A., Zimmermann, R.: Evaluation of the German surveillance system for hepatitis B regarding timeliness, data quality, and simplicity, from 2005 to 2014. Public Health **180**, 141 (2020)
3. Bouhamed, S.A., Dardouri, H., Kallel, I.K., Bossé, E., Solaiman, B.: Data and information quality assessment in a possibilistic framework based on the Choquet Integral. In: 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). pp. 1–6 (2020)
4. Cai, L., Zhu, Y.: The challenges of data quality and data quality assessment in the big data era. Data Science Journal **14**, 1–10 (2015)
5. Chu, X., Ilyas, I.F., Krishnan, S., Wang, J.: Data cleaning: Overview and emerging challenges. Proceedings of the ACM SIGMOD International Conference on Management of Data pp. 2201–2206 (2016)
6. Cichy, C., Rass, S.: An overview of data quality frameworks. IEEE Access **7**, 24634–24648 (2019)
7. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. Journal of Informetrics **5**(1), 146–166 (2011)
8. Corrales, D.C., Corrales, J.C., Ledezma, A.: How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning. Symmetry **10**(4), 99 (2018)
9. Das, S., Saha, B.: Data Quality Mining using Genetic Algorithm. International Journal of Computer Science and Security IJCSS **3**(2), 105–112 (2009)
10. van Eck, N.J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics **84**(2), 523–538 (2010)
11. Heinrich, B., Klier, M.: Metric-based data quality assessment — developing and evaluating a probability-based currency metric. Decision Support Systems **72**, 82–96 (2015)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

12. Heinrich, B., Klier, M., Schiller, A., Wagner, G.: Assessing data quality – A probability-based metric for semantic consistency. Decision Support Systems **110**, 95–106 (2018)
13. Hermans, K., Waegeman, W., Opsomer, G., Van Ranst, B., De Koster, J., Van Eetvelde, M., Hostens, M.: Novel approaches to assess the quality of fertility data stored in dairy herd management software. Journal of Dairy Science **100**(5), 4078–4089 (2017)
14. Jarwar, M.A., Chong, I.: Web Objects Based Contextual Data Quality Assessment Model for Semantic Data Application. Applied Sciences **10**(6), 2181 (2020)
15. Larson, S., Mahendran, A., Lee, A., Kummerfeld, J.K., Hill, P., Laurenzano, M.A., Hauswald, J., Tang, L., Mars, J.: Outlier detection for improved data quality and diversity in dialog systems. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies **1**, 517–527 (2019)
16. Li, A., Zhang, L., Qian, J., Xiao, X., Li, X.Y., Xie, Y.: TODQA: Efficient Task-Oriented Data Quality Assessment. In: 15th International Conference on Mobile Ad-Hoc and Sensor Networks. pp. 81–88. IEEE (2019)
17. Li, J.S., Hamann, A., Beaubien, E.: Outlier detection methods to improve the quality of citizen science data. International Journal of Biometeorology **64**, 1825–1833 (2020)
18. Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration, vol. 62. PubMed (2009)
19. Liu, H., Ashwin Kumar, T.K., Thomas, J.P.: Cleaning framework for big data - object identification and linkage. In: 2015 IEEE International Congress on Big Data. pp. 215–221 (2015)
20. Liu, H., Wang, X., Lei, S., Zhang, X., Liu, W., Qin, M.: A Rule Based Data Quality Assessment Architecture and Application for Electrical Data. In: Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing. Association for Computing Machinery (2019)
21. Malar Vizhi, J.: Data Quality Measurement on Categorical Data Using Genetic Algorithm. International Journal of Data Mining & Knowledge Management Process **2**(1), 33–42 (2012)
22. Nagle, T., Redman, T., Sammon, D.: Assessing data quality: A managerial call to action. Business Horizons **63**(3), 325–337 (2020)
23. Pezoulas, V.C., Kourou, K.D., Kalatzis, F., Exarchos, T.P., Venetsanopoulou, A., Zampeli, E., Gandolfo, S., Skopouli, F., De Vita, S., Tzioufas, A.G., Fotiadis, D.I.: Medical data quality assessment: On the development of an automated framework for medical data curation. Computers in Biology and Medicine **107**, 270–283 (2019)
24. Pietro Biancone, P., Secinaro, S., Brescia, V., Calandra, D.: Data Quality Methods and Applications in Health Care System: A Systematic Literature Review. International Journal of Business and Management **14**(4), 35 (2019)
25. Plotkin, D.: Important Roles of Data Stewards. In: Data Stewardship, pp. 127–162. Morgan Kaufmann (2014)
26. Pucher, S., Król, D.: A quality assessment tool for koblenz datasets using metrics-driven approach. In: Fujita, H., Fournier-Viger, P., Ali, M., Sasaki, J. (eds.) Trends in Artificial Intelligence Theory and Applications. pp. 747–758. Springer (2020)
27. Rafi, S., Yu, W., Akbar, M.A., Alsanad, A., Gumaei, A.: Multicriteria based decision making of DevOps data quality assessment challenges using fuzzy TOPSIS. IEEE Access **8**, 46958–46980 (2020)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33

28. Rajan, N.S., Gouripeddi, R., Mo, P., Madsen, R.K., Facelli, J.C.: Towards a content agnostic computable knowledge repository for data quality assessment. Computer Methods and Programs in Biomedicine **177**, 193–201 (2019)
29. Simard, V., Rönnqvist, M., Lebel, L., Lehoux, N.: A general framework for data uncertainty and quality classification. IFAC PapersOnLine **52**(13), 277–282 (2019)
30. Symoens, S.H., Aravindakshan, S.U., Vermeire, F.H., De Ras, K., Djokic, M.R., Marin, G.B., Reyniers, M.F., Van Geem, K.M.: QUANTIS: Data quality assessment tool by clustering analysis. International Journal of Chemical Kinetics **51**(11), 872–885 (2019)
31. Taleb, I., Serhani, M.A., Dssouli, R.: Big Data Quality Assessment Model for Unstructured Data. Proceedings of the 13th International Conference on Innovations in Information Technology pp. 69–74 (2018)
32. Terry, A.L., Stewart, M., Cejic, S., et al.: A basic model for assessing primary health care electronic medical record data quality. BMC Med Inform Decis Mak **19**, 30 (2019)
33. Udeshi, S., Jiang, X., Chattopadhyay, S.: Callisto: Entropy-based Test Generation and Data Quality Assessment for Machine Learning Systems. In: 13th International Conference on Software Testing, Verification and Validation. pp. 448–453 (2020)
34. Valencia-Parra, Á., Parody, L., Varela-Vaca, Á.J., Caballero, I., Gómez-López, M.T.: DMN4DQ: When data quality meets DMN. Decision Support Systems **141**, 113450 (2021)
35. Zan, S., Zhang, X.: Medical data quality assessment model based on credibility analysis. In: 4th Information Technology and Mechatronics Engineering Conference. pp. 940–944 (2018)
36. Zimek, A., Schubert, E.: Outlier detection. In: Liu, L., Öszu, T. (eds.) Encyclopedia of Database Systems. Springer (2017)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_33