Big Data Quality – A Survey paper to Attain Data quality

Y. Anusha[1], R. Visalakshi[2], K. Srinivas[3]

Annamalai University[1,2]

CMR Technical Campus, Kandlakoya, Hyderabad[3]

Corresponding author: R. Visalakshi, Email: visalakshi_au@yahoo.in

Data, collection of data, and assessing the quality of data are crucial aspects. The World Federation of Hemophilia (WFH) collects data from its members, which helps it to keep track of the health details of the people. In such cases, if the available current data is accurate then the results will be correct. Ensuring high-quality data and assessing the quality of the data is also important in health systems. Quality data has a few measures to be checked. Timeliness, relevance, understandability, completeness, and reliability are some of the main measures of data quality. After reviewing data quality assessment methods, we found that institutions and health organizations review their data frequently to achieve data quality. Huge data with massive quantity and complicated nature is generally not structured well and is in transliterated language and it becomes an issue to deal with this big data. Big data comprises a 4V model which means huge volume, velocity, variety, and veracity. Huge volume is a chunk of gathered data that contains different formats like graphics, videos, text, images, etc. Data from all these areas will be collected and assessed. We will be working with the health databases and their outcomes. There are many limitations to assessing data quality; it is because of incomplete definitions of data quality and its measures. In the future, research can be done to improve data quality and its attributes. We can increase efforts to improvise the data collection process and define the attributes more clearly and precisely.

Keywords: Data quality, Data mining, Big data, Data assessment methods.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 703–711. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-71

1. Introduction

The evaluation of data quality requires definite rules for evaluating quality such as Validity, Completeness, Consistency, Accuracy, Uniqueness, and timeliness. When the data does not include the above rules, it is not good data with appropriate quality. The problems in data quality are the results of many aspects. Firstly, the data source levels include trust, inconsistency, data copying, unreliability, multisource along data orbit. Secondly, the generation level includes sensors devices readings, unstructured data, social media, human data entry with missing values. Thirdly, the process or application-level deals with duplication of data, ambiguous data, and data transformation errors. The methodology of quality assessment used for measuring the data quality included limited apparatus to diagnose information quality. Lastly, in the complete data quality management, the calibration of data quality along with calculations was not interpreted.

As per ISO 14005, Quality of Data is stated as the "characteristic of data that bears on their ability to satisfy stated requirements". Generally, according to empirical research, data from the right source from the right population with the required quality gives the answers to all the questions of research [1].

In 2006, a medical research network organization in Germany published the guideline for the quality of the data [1]. In short, it is TMF, Technology, Methods, and Infrastructure for networked medical research. In 2005, another review on relevant literature on data quality was published which also consists of some of these recommendations. A revision process was funded and started again by TMF, in the year 2011[2]. Fayyad stated that data mining is "a process of nontrivial extraction of implicit, previously unknown, and potentially useful information from the data stored in a database" [2].

According to the analysis, most of the researchers used classical data mining algorithms in their studies namely Bayesian networks, Logistic Regressions, Support Vector Machine, Decision Trees, and Naïve Bayes. Artificial Neural Networks, Rules stated by Fuzzy and the association rules, Genetic Algorithms, revealed satisfactory grades of correctness [3].

To achieve correct prediction and enriched projection, a lot of effort is made to frame and implement varied data mining techniques. Nevertheless, identifying treatment options for patients using data mining techniques has received less attention [3]. Even when we observe the results, efforts are done mostly in the areas related to cardiac arrest, hypertension, and other diseases on the cardiovascular system and cancer. [4]. Data mining techniques attend a great deal of the above-mentioned group diseases and the objective is focused on the follow-up of this study. [5] [6].

2. Big Data

An important role is performed by Big Data Analytics in medium-sized and small-sized industries to capitalize on their business; nowadays many organizations are storing, analyzing, and gathering a great quantity of data by using the analytics of Big Data. This data is called big data as it has volume, variety, and velocity [7]. According to Gartner 2012, by 2015 Big Data will create approximately 4.4 million IT jobs across the world and 1.9 million are especially from the USA [7].

2.1 Big data sources

Satellites, as well as cell phones, produce data as per geospatial observation. Huge data is also created by minicomputers and sensors in the field of the Internet of Things. [8] as an output data is produced by the research projects. Over 200 petabytes [9] of data as an output form is generated in Switzerland and France by LHC, Large Hadron Collider at CERN.

2.1.1. Analytics of Big Data

The method of analyzing as well as exploring the huge bulk of data sets with the data type, like structured or unstructured to reveal unknown correlations, hidden patterns, unseen configurations, consumer priorities, and preferable business information is nothing but the analytics of big data. [10]. This analytics of big data is extracted from two different topics, one the big data, and the other the analytics, both of these constitute big data analytics.

2.2. Data Quality and the Review of the Literature

A study on the quality issue mostly about product quality had started in the early part of the 1950s. Based on this study there are few definitions given, for example, According to the GAQS, (General Administration of Quality Supervision), 2008, defines quality as “the degree to which a set of inherent characteristics fulfill the requirements”.

According to Wang & Strong, 1996, quality is stated as “fitness for use”; “conformance to requirements” (Crosby, 1988). After some years along with the growth of IT (Information Technology), these researches shaped into data quality [10]**.** Simultaneously, they characterized data quality dimensions as sequences of data quality features that personify one characteristic that is data quality.

Knight and Burn (2005) outlined familiar dimensions including the regularity that they are enclosed within various frameworks of data quality. At that moment, they had given a model covering i) Identity, ii)

Quantify, iii) Implement, and iv) Perfect, IQIP as a way to manage option as well as the implementation of the quality associated with algorithms of an Internet Wriggling Search Engine [10].

Improper quality of the data exhibited bad consequences. It cost a huge amount, around 600 billion US $ annually for US organizations, in addition to the outburst of the "challenger", the US space shettle. [11]

The important spot of data quality for its data set depends on the ambiance of its usage [12]. There are analogous numbers of means data that are benefited and correlated in the process of specifying quality has been intended in terms of dimensions [12][13]. The quality of the data attributes must be appraised, refined, and restrained throughout its lifecycle as it straight away results from the outcome of the stage of analysis [14].

Within the database system, Data quality is a common concept; it is a vital sector of research for a long time [15] [16]. Nevertheless, a straight approach of this type of abstraction to Big Data encounters vigorous trials in the matter of value besides the data pre-processing time. [17].

2.3. Measuring Data quality

Exclusively DQD's are necessary to be calibrated and calculated. DQD metrics epitomize various phases for assessing these magnitudes. Metrics provide the appropriateness of DQD's basic formulas to multi-variant complicated expressions. For citation, the calculation of disappeared quantities of any trait is investigated as the quantity to evaluate the DQD integrity.

2.4. The necessity of Data Quality

The emphasis on the quality of data in Big Data development delimits all the procedures of the surveillance of the data that is processed. Inspecting the quality of data comprises enhancing higher functionality in an exclusive phase with a developing control over quality along with checking to escape breakdown of quality throughout all aspects of the development. Big Data quality interpretation is implicated in characteristic features like cost, value, and performance [18].

2.5. Challenges of Data Quality

Data quality arguments arise when provisions on quality are not confronted with values of data [18]. All these arguments are given certain factors or procedures that occurred in divergent phases: 1) sources of the data, the level of reliability, copying of data, consistency levels, trust, multisource, and the domain of the data, 2) the level of generation: entry of data by humans, media source, sensors devices readings, data that is not well structured and values which are missed 3) the level of processing or application. Pre-processing of data revises the quality of data by enforcing countless assignments together with tasks like normalization, conversion, and assimilation of data along with fusion.

2.6. Evaluation of Data Quality

Any scheme is driven by data desires approaching quality valuation on the existing accomplished data. Consequently, computing and calibrating the DQD is compulsory. Concerning the type of data, whether structured or partially structured the data is applicable as a bunch of traits interpreted in vertical columns or else in horizontal rows. Jointly the standards of all the attributes are registered. All metrics (data quality) must stipulate whether the standard of data regards the quality of the attributes [19].

3. Literature Review on Methods and Materials

According to Le Yao et.al [1], we can predict the data quality by applying the Gaussian mixture model, by adding some semi-supervised inputs. The main advantage of this method is we can handle the bid data modeling issues. Algorithms can be handled easily. The most important advantage is we can reuse the unlabeled data, for further predictions. The main disadvantage is when we deal with huge datasets. The computational burden will be more dealing with huge datasets.

Concerning Eslam.M.Hassib et.al [2], for attaining data quality we need to apply an established whale optimization algorithm, with a combination with bidirectional recurrent neural network algorithm, to classify the Big Data. An important advantage of using this method is the developed framework finds the optimal subset of features and improves the accuracy of classification. We can also get the balance of minority and majority classes with the method. The main disadvantage of this method arrives when we deal with large data sets. Either it will take a long time to run the algorithm or it will stop working.

In the opinion of Le Yao et.al [3], advanced distributed parallel intensive learning along with the combination of hierarchal extreme learning machines will help the users to attain data quality. With this method, multi-mode data quality prediction is possible for big data processing. The key advantage is that this method is more efficient for model training. This method is most suitable for multi-mode data quality prediction. The disadvantage of this model is that for a huge measurement of a dataset, two variables are portrayed. This is because of the fact that the cluster midpoints are indiscriminately initialized. In the end, it meets the mean value.

S.K. Lakshmanaprabu et.al [4] developed a Random Forest classifier and map-reduce process for achieving data quality. This method is more beneficial when we deal with the data of the Internet of Things. This method helps in classifying Big Data. This method is advantageous as IDA (Improved

Dragonfly Algorithm) was used to choose the best features from the medical dataset. To predict the disease, it helps to develop a model which is cost-effective. There is a limitation of the developed random forest and the map-reduce algorithm and for the big database, it was very slow.

Mikel Elkano et.al [5] produced a compressed fuzzy model with distributed rule induction algorithm. This model enables handling big data classification. This approach is advantageous as the computing cluster is not required to administer the algorithm. Regarding the computational cost, each bit is equal to the small data classification. The only disadvantage of this method is, when the initial distribution (of the training set) was not known, the exact cumulative distribution function cannot be computed.

W. Fan et.al. [6] Developed AIMQ. (AIMQ) AIM model consists of the "Product and Service Performance model" (PSP). This is used for Information quality (IQ). For measuring information quality, an instrument and an analysis technique are used in this model. The instrument is named IQA and an analysis technique, named information gap. Based on a questionnaire, the quality of information is assessed in this model. Later on, in this model, for identifying the problem in information quality, statistical analysis will be applied. This model aims to achieve high-quality information with the help of IQ or PSP, which is guided by attributes such as representational, accessibility, intrinsic along with contextual. The advantage of this method is it calibrates data quality dimensions in the attributes of contextual, representational, and intrinsic along with accessibility. Limited apparatus to diagnose information quality dispute stretches is the main disadvantage of this method.

Bill Hamilton et.al [7] developed TDQM for attaining data quality. Total Data Quality Management (TDQM) has been scheduled before strengthening the perception of 'data as a product'. By this means of perception, the immense quality of data is accomplished by imitating the objective fabrication of an immense quality product. Total Data Quality Management (TDQM) elongated the framework of Total Quality Management (TQM) that is recycled in material production. This procedure begins with the interpretation of Information Product (IP). The information product, at this stage, has intrinsic peculiarities along with prerequisites to attain a high-quality state. At that instant, the IQ, (information quality) metrics are refined as well as recycled to estimate the IP. Then the estimated outcome is interpreted applying pattern recognition, statistical process control along with the Pareto chart [7]. While using this method, data might be distributed with users whereas crude evidence authorized to particular merchandise and this is the main disadvantage of this method.

For data quality, F. Sidi et.al [8] Used DQMMM. Data Quality Management Maturity Model (DQMMM) enhances the quality of data structure along with the outcome; it will contribute to the great data quality [8]. According to this, an integrated database structure was achieved by regularizing owned metadata. Metadata in the database is standardized by segregating it into different stages namely, logical, physical along with mapping metadata information. In general, while considering with present data quality methodologies and data quality management models, data quality will not be managed at the integration level, when we deal with various databases in the organization, data quality will not be maintained. To avoid that problem, this model mostly stresses the importance of data integration which further helps to maintain data consistency along with accuracy. Assuring high data quality while the integration of database, is the main advantage of this method. This method helps to maintain data quality in the course of the database integration procedure and this is the advantage of this method. Convenience for the sake of a relational database is the disadvantage of this method.

Caballero et.al [9] implemented CDQMM. Complete Data Quality Management (CDQM).is appropriate for unstructured or semi-structured data categories. CDQM suggests an intuitive, empirical along with a theoretical approach that assesses the quality of the data. [9]. this embodied 3 phases; reconstruction state, assessment along a selection of the best process for improvement. The dominance of CDQM is the elasticity of the technique that helps to strengthen unstructured, semi-structured along with unstructured data. Anyhow, specific methods of measurement or estimation are not in existence to calibrate the dimensions of data quality in CDQM. Consequently, CDQM usage in an institution is hampering.

This method strengthens unstructured, semi-structured along with the structured data type and this is this method's advantage. Data quality dimensions calibrations along with calculations were not interpreted in CDQM, which is the disadvantage of this method.

4. Results & Discussions

The only motto is to attain data quality, where all the data attributes perform well. We have to remove the duplicates, find accurate data and extract knowledge from the database. Quality data gives answers to all the questions, which in the future helps the researchers to perform their tasks.

According to Le Yao et.al [1], by applying the Gaussian mixture model, we can attain accurate data and even the time of the test will be low. Concerning, Eslam. M. Hassib et.al [2], by using WOA with BRNN method, an error will be negligible and the accuracy of the data will be high. In the opinion of Le Yao et.al [3], by implementing “Distributed Parallel Deep Learning of Hierarchical Extreme Learning Machine”, the time utilized is very less and the completeness of the data is high and sufficient.

According to S. K. Lakshmanaprabu et.al [4], by performing random forest classifier and map-reduce process, sensitivity, specificity, accuracy, and performance metric of the database are up to the mark. Concerning Mikel Elkano et.al [5], by using the compact fuzzy model with distributed rule induction algorithm, we will meet all the data quality metrics and especially accuracy will be more than 95%. Knowledge should be extracted from the datasets and quality data should be the only outcome.

5. Conclusion

Quality assessment of data is a crucial part of the world in this situation. Large data is being produced, which forces to check the quality of the data. Data quality assessment methods and algorithms are to be applied for quality checks. Applying appropriate methods to achieve data quality is the main objective of this paper. Finally, it is concluded that applying the mining algorithms with few changes will yield the necessary results. Algorithms along with the methods are to be recycled according to the application and data quality will be achieved.

References

1. L. Yao and Z. Ge, "Scalable Semi-supervised GMM for Big Data Quality Prediction in Multimode Processes," in IEEE Transactions on Industrial Electronics, vol. 66, no. 5, pp. 3681-3692, May 2019, DOI: 10.1109/TIE.2018.2856200.
2. Hassib, E.M., El-Desouky, A.I., Labib, L.M. and El-kenawy, E.S.M., 2020. WOA+ BRNN: An imbalanced big data classification framework using Whale optimization and deep neural network. Soft computing, 24(8), pp.5573-5592, https://doi.org/10.1007/s00500-019-03901-y, 11 March 2019, Issue Date April 2020.
3. Yao, L. and Ge, Z., 2019. Distributed parallel deep learning of hierarchical extreme learning machine for multimode quality prediction with big process data. Engineering Applications of Artificial Intelligence, 81, pp.450-465, 2019
4. Lakshmanaprabu, S.K., Shankar, K., Ilayaraja, M., Nasir, A.W., Vijayakumar, V. and Chilamkurti, N., 2019. Random forest for big data classification in the internet of things using optimal features. International journal of machine learning and cybernetics, 10(10), pp.2609-2618, " " 2019
5. Elkano, M., Sanz, J.A., Barrenechea, E., Bustince, H. and Galar, M., 2019. CFM-BD: a distributed rule induction algorithm for building Compact Fuzzy Models in Big Data classification problems. IEEE Transactions on Fuzzy Systems, 28(1), pp.163-177, 2019
6. W. Fan, F. Geerts, and J. Wijsen, "Determining the currency of data," ACM Trans. Database Syst. TODS, vol. 37, no. 4, p. 25, 2012.
7. Bill Hamilton, Big Data Is the Future of Healthcare, Cognizant white paper, September 2012. Cognizant 20-20 insights.
8. F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data quality: A survey of data quality dimensions," in 2012 International Conference on Information Retrieval Knowledge Management (CAMP),2012, pp. 300–304, doi: 10.1109/InfRKM.2012.6204995.
9. Caballero and M. Piattini, "CALDEA: a data quality model based on maturity levels," in Third International Conference on Quality Software, 2003. Proceedings, 2003, pp. 380–387. 2003.
10. A Survey Of Big Data Analytics in Healthcare and Government.J.Archenaa1 and E.A.Mary Anita,1Research Scholar, MET University, Chennai, 2S.A.Engineering College, Chennai, Email: archulect@gmail.com, PY - 2015/12/31, SP - 408, EP - 413,, VL - 50, DOI - 10.1016/j.procs.2015.04.021
11. Yanglin Ren, Monitoring patients via a secure and mobile healthcare system, IEEE Symposium on wireless communication, Vol 17, No 1, pp 59-65, February 2010. DOI: 10.1109/MWC.2010.5416351
12. Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, Vol. 51, No 1, Communications of the ACM, Jan2008.
13. Bill Hamilton, Big Data Is the Future of Healthcare, Cognizant white paper, 2010.
14. J. Archenaa, E.A. Mary Anita, A Survey of Big Data Analytics in Healthcare and Government, Procedia Computer Science, Volume 50, 2015, Pages 408-413, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.04.021.(https://www.sciencedirect.com/science/article/pii/S1877050915005220)
15. A review of data quality research in achieving high data quality within organization izham Jaya, Fatimah Sidi, Iskandar Ishak, Lilly suriani affendey, marzanah a. Jabar, PY - 2017/06/30, SP -2647, EP - 2657, VL - 95, DO - 10.5281/zenodo.5374545, Journal of Theoretical and Applied Information Technology

16. A Big Data revolution in Health Care Sector: Opportunities, Challenges, and Technical Advancements. Sanskruti Patel and Atul Patel, Faculty of Computer Science & Applications, CHARUSAT, Changa, India, PY - 2016/03/31, SP - 155, EP - 162, VL - 6, DO - 10.5121/ijist.2016.6216, JO - International Journal of Information Sciences and Techniques
17. Hugh J. Watson, Tutorial: Big Data Analytics: Concepts, Technologies, and Applications, Communications of the Association for Information Systems, Volume 34, Article 65, pp. 1247-1268, April 2014
18. Using Open Source to Distribute Big Data from the Large Hadron Collider, retrieved from https://www.linux.com/news/enterprise/networking/873403-using-open-source-to-distribute-big-datafrom-the-large-hadron-collider on December 15, 2015
19. Cai, L, and Zhu, Y 2015 The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14: 2, pp. 1-10. DOI: http://doi.org/10.5334/dsj-2015-002.
20. Anon., 2006. DataCentric Systems and Applications. In: Data Quality .s.l.: Springer. Electric ISSN 2197-974X. Print ISSN 2197-9723.
21. Anon., n.d. data quality. [Online] .Available at: http://searchdatamanagement.techtarget.com/definition/data-quality [Accessed 4 May 2015].
22. Caballero, I., Serrano, M. & Piattinni, M., 2014. A data quality in Use model for Big Data. ER workshops, pp. 65-74. ER 2014. Lecture notes in Computer Science, vol 8823, Springer, and Charm. DOI: http://doi.org/10.1007/978-3-319-12256-4_7
23. Soares, S., 2012. Big Data quality. In: Big Data Governance: An emerging imperative. s.l.: MC Press, pp. 101-112, 2012. Big Data Governance: An Emerging Imperative Kindle Edition ISBN-13: 978-1583473771, ISBN-10: 1583473777