Noname manuscript No.
(will be inserted by the editor)

Big Data Veracity: Methods and challenges

benabderrahmane MOUTASSEM · laouni DJAFRI · Abdel-Kader GAAFOUR

Received: date / Accepted: date

Abstract Today, the Internet has become the main source of information, a place where there are no restrictions on who can share information .This latter can play an important role in prediction, estimation and decision making processes. But, this role will not only be achieved through abundance, it will also be the result of data quality. Veracity refers to the assurance of quality or credibility of the data collected. The data can be incomplete, biased, vague or wrong. For this reason, automatic filtering mechanism has been developed. Moreover, due to the increasing velocity of information spread, manual assessment of information veracity became hard, a time consuming process and even the already existing automatic filtering mechanisms has to be improved to cope with the speed of information spread. In this paper, a literature review is established to highlight the recent methods and techniques which are exploited in computerized veracity assessment. The challenges and limitations of existing works will be discussed, and future research directions will be proposed to address critical issues of data veracity in the era of big data.

Keywords Big Data Analytics · Data Quality · Veracity · Prediction

Benabderrahmane Moutassem
Department of Computer Science, Ali Kafi University Center, Tindouf, Algeria
Laboratory of Environmental and Energy Systems, Ali Kafi University Center, Tindouf, Algeria
E-mail: moutassemabdo@gmail.com

Laouni DJAFRI
Department of Computer Science, Ibn Khaldoun University, Tiaret,Algeria
EEDIS laboratory, Djillali Liabes University, Sidi Bel Abbes, Algeria
E-mail: djaafri29tp@gmail.com

Abdel-Kader GAFOUR
Department of Computer Science, Djilali Liabes University, Sidi Bel Abes, Algeria
EEDIS laboratory, Djillali Liabes University, Sidi Bel Abbes, Algeria
E-mail: gafour1@yahoo.com

1 Introduction

Big data is used today as an effective tool in addressing the complex issues facing humanity such as global warming and climate change, treating intractable diseases and epidemics, as well as studying markets and public attitudes, and helping countries and governments to know the desires and trends of peopleand Control of complex phenomena by transforming digital and physical spaces. The deluge of data influences and directs economic, social and political practices. Thus, these new temporal and spatial scales became available for human interpretation and understanding despite their constant push towards more technological challenges. This data, which is generated either by users or automatically, allows us today to have more clarity on reality. Doug Laney first described Big Data in 2001 in three dimensions: Volume, Variety, and Velocity and they all named the three V (3V) [1]. Other dimensions that were later discovered are veracity, variability, and value. Hence, there are six dimensions of big data [2, 3] we will try to define it briefly:

1. Volume refers to the size or volume of data and the most obvious characteristics of big data. With the significant increase in the sources of data, the volume of the data keeps increasing exponentially. Big data sizes are reported in multiple Petabytes and Exabyte's;
2. Variety refers to different formats of the data. Structured data that resides in relational databases and file systems, unstructured data such as text documents, email, video and audio etc. Big data has to be connected and correlated during the analysis phase in order to extract useful information from it;
3. Velocity is defined as the sheer rate at which data is coming in, at a certain time and has multiple angles;
4. Veracity is coined by IBM and defined as the ambiguity of the data. Many companies which deal with the product review cannot trust the review made by the reviewer. Thus, there is an urge to find out a new tool that can deal with uncertainty;
5. Variability is introduced by SAS and refers to the variations in the data flow rates. Since data flow rates i.e. big data velocity are not consistent and keeps changing frequently, there is demanding need to pair, clean and remodel the data collected from various sources;
6. Value of the big data is considered to be nearly low value density as articulated by Oracle. The data that flows in original form has low value to its volume. If large volumes of aforesaid data are analyzed, high value could be attained.

The primary goal is our ability to absorb this big data and how we can convert it into expectations or use it to make decisions or build a forward-looking vision for the future. Thus, we can find effective solutions to our complex problems and build a clear vision in order to avoid obstacles. This will Not only be achieved through an abundance of data, but it will also be the result

of quality data: The veracity of the data is a very important condition in the whole process of exploitation and analysis.

1.1 Problems and desired goals

The goal of this paper is to review recent works to summarize existing technical solutions, which has the capacity to improve the veracity of the data. We list the challenges of applying them on the ground and identify future research, for this, the paper research will be as follows:

- " List the technical solutions that were used to improve data veracity for different applications according to the existing literature;
- " Extract challenges or short comings of the solutions presented and extract potential future research directions to improve data veracity.

Thus, the main contributions of this paper are the classification of existing work, the description of the technical solutions that have been used to improve data veracity and the discussion of challenges and future directions to improve the veracity of big data.

1.2 Research proposal

The paper organized as follows. Section 2 explains data veracity definitions from the perspective of the applications domains, and Section 3 explains the technical solutions that have been proposed. The synthesis and discussion are explained in Section 4. While the last section sums up the work and presents conclusions

2 Definition

Data veracity as defined as the assurance of quality of the collected data. Data veracity is related to source reliability, information credibility and content validity. It has been defined as the fourth dimension of big data. There are several definitions of data veracity in the literature. Wibowo et al. [4] say veracity means that the data can be sure so it can be trusted. Lin et al. [7-5] say that veracity in big data must ensures that the data used are trusted, authentic, accurate and protected from unauthorized access and modification. Others authors have identified several dimensions such as accuracy, confidence, completeness, data volume, a timeliness [6, 7]. Suthanthira et al. [8] studied Big data vercity in social media and they say that veracity is the degree to which the information is accurate and trusted. Shukla et al. [9] say veracity refers to the existing uncertainty in the dataset. However, Laure et al. [10] gives a list of causes that frequently affect data veracity can we summarize with: Measurement system limits, Limits of features extraction, Data integration limits, Data ambiguity and uncertainty, Data falsification and source collusion.

3 Methods and application

Our work is a state of the art that focuses on data veracity in big data applications. There are several areas where big data applications are widely used and for different purposes. The veracity determines the importance of the data. While processing uncertain data causes malfunctions in all associated systems and causes deviations in making decisions based on this interpretation. For better review, Big Data veracity articles, its contribution and the techniques used, the corresponding domain and data format processed are summarized in the table 1. At first, and before addressing the various studies, we cite three papers whose authors reviewed, summarized, examined and compared all previous studies like [11, 12, 13]. In the first paper [11] the authors classify existing works into three main domains: social media, the web, and IoT application. They compare data veracity techniques in order to discuss current challenges and provide future directions to improve data veracity. Authors reviewed existing techniques that were proposed to assess, predict, and improve data veracity. to compare between the proposed approaches for assessing, predicting, and improving data veracity, they defined two metrics: computational cost and communication cost: computational costs are the amount of time an algorithm/process takes to run, Communication costs are the amount of communication that is needed between parties to solve a problem. A standard definition is needed since veracity definitions have been defined differently based on the related field. For example, in social media applications, veracity has been measured in terms of data trustworthiness and they focused on the identification of rumors and fake news. On the other hand, security specialists have studied data veracity in terms of access authorization for web data. These are totally different perspectives; thus, they believe that data veracity should have one general standard definition that covers the related aspects and should be used to direct research in the area. The second paper [12] research carried out by (Lozano et al) reviewed the approaches, methods, algorithms, and tools which are used or proposed by research papers that focused on the assessments of the veracity of online data in social media and open sources. Open source data we refer to information published online such as social media posts, blog posts, forum entries, newspaper articles, whether on the shallow or the deep web. The authors conclude that the researchers have not reached consensus on a veracity (assessment) definition. Although there is some convergence in the methods used to assess it. The last paper [13], authors say that veracity helps to make accurate and precise decisions from data and select relevant information from given data with noise removal. They study the use of the veracity in order to deal with big data in order to make decision in different fields. Data Veracity relates to the following terms according to them: Source Selection, Data Quality, Data Ambiguity, False and Biased Information. They gived three dimension to varacity are: Objectivity, Credibility and Trustworthiness.

There are different techniques has been used to improved Data veracity. Traditional techniques [8, 5] introduce overhead, which can be inefficient when applied to big data. A study by Suthanthira et al. [8] proposed a technique,

known as the hybrid rumour source-detector approach. They studied Big data vercity in social media and they say that veracity is the degree to which the information is accurate and trusted. to identify the veracity of the rumour and find the source of the rumour authors analyse the PHEME data set(This dataset contains a collection of Twitter rumours and non-rumours posted during breaking news) in which the Sydneysiege event is chosen for analysis. they combines the merits of graph and tree data structures to proposes a novel and hybrid rumour source-detector approach. The study was performed on Sydneysieg events. All of the tweets and their replies and retweets were collected. Conversation threads were classified as rumor/nonrumor. Furthermore, rumor threads were classified based on the support type in which each follower responded by either supporting the information of the source tweet or not. Both manual labeling and the proposed method detected the same threads to be sources of rumors. To predict the possibility of rumor spreading, a network of rumors with infected parts was developed. A breadth-first search (BFS) tree structure along with a spanning tree were created to calculate the probability for each node to be a source of rumors. The predicted nodes as the source of rumors matched the actual rumors. Hui Lin et al. in [5] for the goal to defend against internal attacks (The internal attacks are launched by an inside attacker who is a legal and certified user, may compromise certain users and gain full control of them) and enhance data veracity in Mobile Cloud Computing the authors proposed a Category-based Context-aware and Recommendation incentive-based reputation Mechanism. This approach based on the Vickrey-Clarke-Groves (VCG) mechanism. The VCG mechanism belong a category of mechanism design, that can achieve ex-post incentive compatibility, has truth as a dominant strategy and makes efficient choices.

It is difficult to apply traditional encryption techniques to big data to ensure the veracity. [14] proposed a new technique named Computing on Masked Data (CMD), that improves veracity by doing computation directly on masked data using the AES256 algorithm. For checking the performance of system, Twitter corpus was used on CMD database operations like inserting and querying. The result shows inserting and querying of tweets data is quite reasonable time.

The blockchain technique are used to improve data veracity [4, 15]. However, blockchain allows a user to transfer data securely using the concept of consensus without third parties. Authors in [4] use Blockchain to improve data interoperability, security, and veracity of Local Tax Big Data. Big data analytics using advanced techniques to analyse very large and diverse data sets that include structured, semistructured and unstructured data. It can be collected from different sources, in different sizes from terabytes to zettabytes [4]. Blockchain is a technology behind cryptocurrency, that allows parties to transfer assets to each other without intermediaries securely moreover enables transparency, immutable records, and autonomous execution of business rules. Blockchain actually is a part of database technology and has a distinctive working concept such as distributed database, peer-to-peer transmission, transparency, irreversible notes, and computational logic. they conclude that

Local Tax Big Data is a suitable case for Blockchain implementation. Several methods can be used to improve veracity, such as token-based crowdsourcing, source identification, and the special application of token-based value change and achievement. When Authors explore the role of technology such as blockchain, satellite imaging and others to improving the veracity and timeliness of social and environmental audits in supply chains [15]. For example, satellite imaging can provide information on land usage patterns deep in the agricultural supply chain, blockchain technology can provide a comprehensive accounting of a product's history, and environmental sensors can provide continuous monitoring of discharges to the air, water, or land. In addition to enhancing veracity, the authors relate the work's value and novelty in its exploration of how technology-enhanced auditing relates to (1) the audit's scope/focus, firm's motivations for SEA, governance of the audit function, (2) possible contextual factors, (3) the 4 V's of data (volume, variety, velocity and veracity); and (4) possible mechanisms by which veracity and timeliness affect social and environmental performance in the supply chain beyond what is possible with traditional SEA practices[15].

Dering et al. [16] introduce another technique named Generative Adversarial Network (GAN) . They describe the use of a GAN to generate sketch data, as part of automated data generation that fits into a Big Data pipeline. Such as those that might be used in a human verification task. This approach are verified as recognizable using a crowdsourcing methodology, and finds correctly recognized 43.8 % of the time, in contrast to human drawn sketches which were 87.7% accurate. A generative model is one which models data as a distribution, or combination of distributions, which can then be sampled. the authors of this work describe how generating large amounts of human quality data could be used for training a big data platform before deployment. Several studies have shown that Adversarial training of models will be helpful in providing a model that is more robust to error. This approach is valuable in assuring the veracity of data, which has long proved to be a substantial challenge when drawing inferences from high dimensional data such as text.

In [17] Herrera et al. use Failure Mode and Effect Analysis method to improve data veracity and validity in the context of big data analysis. The authors extend the usage of the FMEA method to improve data veracity and validity. The proposed extension (DVV-FMEA: Data Veracity and Validity Failure Mode and Effect Analysis) is illustrated with an electronics manufacturing application for quality assurance. The authors conclude that by using this method, experts can transfer knowledge, understand data, and business priorities for the success of further analysis.

Another technique based on fuzzy logic was used in [9]. The continuous flow of unstructured data with unwanted noise may bring abnormality in the dataset making them unusable. where they use a novel method to handle the veracity characteristic of the big data using the concept of footprint of uncertainty (FOU) in interval type-2 fuzzy sets (IT2 FSs). The proposed method helps in handling the veracity issue in big data and reduces the instances to a manageable extent. they compared the results with the existing clustering

Table 1 Big Data veracity articles, its contribution and the techniques used, the corresponding domain and data format processed

Article (Authors, year,ref)	Contribution: Problem to be solved	Trchnique	Domaine	Data format
Wibowo et al. 2019 [4]	Improve data interoperability, security, and veracity of Local Tax Big Data.	Blockchain	Web application (Big data analytics)	Numeric, text
Lin et al. 2016 [5]	A Category-based Context-aware and Recommendation incentive-based reputation Mechanism against the internal attack	Vickrey-Clarke-Groves (VCG) mechanism	IoT application (Mobile Cloud Computing)	Numeric, text (Personel private data)
Suthanthira et al. 2020 [8]	Identify the veracity of the rumour and find the source of the rumour	Combines the merits of graph and tree data structures	Social media	Text
Amit et al. 2020 [9]	Handling the veracity issue in big data and reduces the instances to a manageable extent	Footprint of uncertainty in interval type-2 fuzzy sets and K-means clustering algorithm	Statistic dataset	Numerical attributs
Castla et al. 2020 [15]	Improve the veracity and timeliness of Social and environmental audits	Blockchain, satellite imaging	Web application	Numeric, text, image
Dering et al. 2017 [16]	Use GAN to generate sketch data, as part of automated data generation that fits into a Big Data pipeline	GANs are special type of neural networks that are used for data generation	Web application	Sketches data (Images)
Herrera et al. 2020 [17]	Improve data veracity and validity in the context of big data analysis	Failure Mode Effect Analysis	Web application	Numerical attributs
Kim et al 2020 [18]	Assessment of the reproducibility and veracity by statistical machine learning models using high dimensional TCGA data for classification.	Supervised machine learning methods	medical research applications	Numeric, text
Rath et al 2017 [19]	A machine learning framework based LINE (Large-scale Information Network Embedding)	Algorithm LINE algorithm with the 2-hop distance for generating user embeddings from the trust network	Social media	Text

based methods and examined the relationship between the clusters and the FOUs by comparing their centroids and defuzzified values.

4 Synthesis and discussion

According to the papers studied previously, we conclude that the veracity of big data is of great importance in obtaining accurate predictions from our data

and that it is related to the quality and credibility of the data. There is no consensus on a unified definition of the veracity of the data, each definition takes into account the field of study. This paper discusses the use of veracity in different areas of big data, where appropriate techniques are selected according to the field of study. On the other hand, it deals with different types of data such as numbers, text, images, etc.

When we discuss the use of veracity in different domains of Big Data, we especially mention the social media networks, internet of thing (IoT) and web application. Each domain has its appropriate techniques. For example, most papers study social media use machine learning algorithms. While crowdsourcing techniques is suitable to improve veracity for web applications. Whereas relying on reputable sources helps improve the veracity of data collection for IoT applications.

We have studied the latest work that touched on the veracity of data. We have found that there are variations in terms of the techniques used, although there is almost unanimity in the use of supervised machine learning for prediction in social media application.

Future research is directed at using machine learning-based models, especially deep learning, to handle low-quality real-time data to find the reliable source. These models remain limited due to the large deluge of big data and the multiplicity of its sources as well as its diversity and heterogeneity. Which makes the field open for its development and improvement of its work.

5 Conclusion

As one of the most important V's in big data (the fourth V) which is related to the accuracy and truthfulness of the data, the data veracity impacts the decision-making and prediction process in many domains. Like social media where the data is untrustworthy, or the Internet of things whose data is sometimes biased, as well web application. This made it imperative for users of big data to evaluate its veracity and quality before using it in prediction and decision-making applications. Most of the studies in the literature have focused on veracity aspects that are related to social media applications where utilized machine learning (ML) algorithms for data veracity prediction. Some of the future directions is to built incremental deep learning based models to handle low-quality real-time data to find the reliable source.

References

1. Douglas Laney, 3D data management: Controlling data volume, velocity and variety. META Group Res. Note 6 (70), (2001)
2. Chen, H., Chiang, R.H., Storey, V.C.,. Business intelligence and analytics: From big data to big impact. MIS Q. 36 (4), (2012)
3. Zikopoulos, P., Parasuraman, K., Deutsch, T., Giles, J., Corrigan, D, Harness the Power of Big Data the IBM Big Data Platform. McGraw Hill Professional. Zadeh, L.A., 1965. Fuzzy sets. Inf. Control 8 (3), 338-353, (2012)

4. Wibowo, S. and Sandikapura , Improving Data Security, Interoperability, and Veracity using Blockchain for One Data Governance, Case Study of Local Tax Big Data, IEEE, International Conference on ICT for Smart Society (ICISS), volume 7, pages 1-6, (2019)
5. Hui Lin, Jia Hu, Youliang Tianc, Li Yang, Li Xua, Toward better data veracity in mobile cloud computing: A context-aware and incentive-based reputation mechanism, Elsevier, Information Sciences 1-16, (2016)
6. Batini, C., Cappiello, C., Francalanci, C., and Maurino, A., Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3):1-52, (2009)
7. Klein, A. and Lehner, W., 2009. Representing data quality in sensor data streaming environments. Journal of Data and Information Quality (JDIQ), 1(2):1-28, (2009)
8. P. Suthanthira Devi, S. Karthika, P. Venugopal and R. Geetha, Veracity Analysis and Prediction in Social Big Data, Springer, Information and Communication Technology for Sustainable Development, pages 289- 298, (2020)
9. Amit K. Shukla, Megha Yadav, Sandeep Kumar, Pranab K. Muhuri, Veracity handling and instance reduction in big data using interval type-2 fuzzy sets, Elsevier, Engineering Applications of Artificial Intelligence 88, (2020)
10. Laure B., Javier B. Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics, Morgan and Claypool, (2015)
11. Fatmah Y. Assiri, Methods for Assessing, Predicting, and Improving Data Veracity: A survey, ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal Vol. 9 N. 4, 5-30, (2020)
12. Marianela Garcia Lozano, Joel Brynielsson, Ulrik Franke, Magnus Rosell, Edward Tjornhammar, Stefan Varga, Vladimir Vlassov, Veracity assessment of online data, Elsevier, Decision Support Systems, 129, (2020)
13. Muhamamd Junaid Ali,Dr Muhammad Naeem, a review of big data veracity, opportunities and challenges, Vol. 1, No. 1, (Preprint)
14. J. Kepner, V. Gadepally, P. Michaleas, N. Schear, M. Varia, A. Yerukhimovich, and R. K. Cunningham, "Computing on masked data: a high performance method for improving big data veracity," in 2014 IEEE High Performance Extreme Computing Conference (HPEC), IEEE, pp. 1-6, (2014)
15. Pavel Castka , Cory Searcy , Jakki Mohr, Technology-enhanced auditing: Improving veracity and timeliness in social and environmental audits of supply chains, Elsevier, Journal of Cleaner Production 258 , (2020)
16. Matthew L. Dering, Conrad S. TuckerGenerative Adversarial Networks for Increasing the Veracity of Big Data, IEEE, International Conference on Big Data (BIGDATA), 2595 -2602, (2017)
17. Ana Elsa Hinojosa Herrera, Chris Walshaw and Chris Bailey, Failure Mode and Effect Analysis and another Methodology for Improving Data Veracity and Validity, International Association of Educators and Researchers (IAER), Annals of Emerging Technologies in Computing (AETiC), Vol. 4, No. 3, , pp. 9-16, (2020)
18. Ahyoung Amy Kim, Samir Rachid Zaim, Vignesh Subbian, Assessing reproducibility and veracity across machine learning techniques in biomedicine: A case study using TCGA data, , Elsevier, International Journal of Medical Informatics, 141, (2020)
19. Bhavtosh Rath, Wei Gao, Jing Ma, and Jaideep Srivastava, From Retweet to Believability: Utilizing Trust to Identify Rumor Spreaders on Twitter, Association for Computing Machinery, 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 179-186, (2017)