VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL
VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Handling uncertainty in the big data processing

Hitashri Dinesh Sankhe[1],Suman Jai Prakash Barai[2]
[1](MCA, VIVA Institute of Technology / University of Mumbai, India)
[2](MCA, VIVA Institute of Technology / University of Mumbai, India)

***Abstract**: Big data analytics has gained wide attention from both academics and industry as the demands for understanding trends in massive datasets increase. In recent developments in sensor networks, IoT has increased the collection of data, cyber-physical systems to an enormous scale. the analysis of such massive amounts of data requires advanced analytical techniques for efficiency or predicting future courses of action with high precision. Previous research and survey conducted on big data analytics tend to focus on one or two techniques. However, little work had been done in the field of uncertainty when applied to big data analytics.*

***Keywords:** Big data, Big data analytic, data , Measuring uncertainty data , Uncertainty, Uncertainty elimination, Uncertain Data Due to Statistics Analysis*

I. INTRODUCTION

According to the National Security Agency, the Internet processes 1826 petabytes (PB) data per day [1]. In 2018, the amount of data generated daily was 2.5 quintillion bytes [2]. Previously, the International Data Corporation (IDC) estimated that the amount of data produced would double every 2 years, yet 90% of all data in the world was produced in the last 2 years, and moreover, Google is now processing more than -40,000. searches every second or 3.5 billion searches per day [2]. Facebook users upload 300 million photos, 510,000 comments, and 293,000 status updates per day [2,4]. Needless to say, the amount of data produced on a daily basis is astounding. As a result, strategies are needed to analyze and understand this huge amount of data, as it is a great source of useful information. Advanced data analysis methods can be used to convert big data into intelligent data for the purpose of obtaining sensitive information about large data sets [5,6]. Thus, intelligent data provides useful information and improves the decision-making skills of organizations and companies. For example, in the field of health care, analyses performed on large data sets (provided by applications such as Electronic Health Records and Clinical Decision Systems) may allow health professionals to deliver effective and affordable solutions to patients by examining trends throughout patient history, as opposed to relying on evidence provided local or current data. Big data analysis is difficult to perform using traditional data analysis [7] as it can lose efficiency due to the five V characteristics of big data: high volume, low reliability, high speed, high variability, and high value [2,8,9]. In addition, many other factors exist for large data, such as variability, viscosity, suitability, and efficiency [10]. A number of artificial intelligence (AI) techniques, such as machine learning (ML), natural language processing (NLP), computer intelligence (CI), and data mining are designed to provide greater data analysis solutions as they can be faster, more accurate, and more accurate in large data volumes [8]. The purpose of these advanced analytical methods is to obtain information, hidden patterns, and anonymous links to large databases [7]. For example, a detailed analysis of a patient's historical data can lead to early detection of a devastating disease, thus enabling the best treatment or treatment program [11, 12]. Additionally, risky business decisions (e.g., entering a new market or introducing a new product) can benefit from simulations with better decision-making skills [13].

Although large data sets using AI have many promises, many different challenges are presented when such strategies are under uncertainty. For example, each V element presents multiple sources of uncertainty, such as random, incomplete, or noisy data. In addition, uncertainty can be embedded in the entire mathematical process (e.g., collecting, editing, and analyzing big data). For example, dealing with incomplete and accurate information is a critical challenge for most data mining and ML strategies. In addition, the ML algorithm may not receive full results if training data are biased in any way [14,15][16] presented six important challenges in the analysis of big data, including uncertainty.

They focus more on how uncertainty affects learning performance over big data, while distinct concern is about reducing the uncertainty that exists within big data. These challenges are often present in ML data

mining and strategy. Raising these concerns to a greater level of data will effectively cover any errors or shortcomings of the entire mathematical process. Therefore, reducing uncertainty in big data analysis should be at the forefront of any automated approach, as uncertainty can have a significant impact on the accuracy of its results.

Based on our review of existing research, little work has been done on how uncertainty significantly affects the integration of big data and the analytical methods used. To address these shortcomings, this article presents an overview of existing AI methods for analyzing big data, including ML, NLP, and CI in view of the uncertain challenges, as well as the appropriate guidelines for future research in these domains. The contributions to this project are as follows. First, we consider the uncertainty challenges in each 5 V big data aspect. Second, we review several major data analysis strategies that influence uncertainty with each system, and we review the impact of uncertainty on a few major data analysis strategies. Third, we discuss the strategies available to deal with each challenge raised by uncertainty.

To the best of our knowledge, this is the first article that explores the uncertainty in large-scale data analysis. The rest of the paper is arranged as follows. The "Back" section introduces background data for big data, uncertainty, and big data analysis. The "view of big data uncertainty" takes into account the challenges and opportunities associated with uncertainty in the various AI strategies for data analysis. "Summary of mitigation strategies" links survey activities with its uncertainty. Finally, the "Discussion" section summarizes this paper and presents future research guidelines.

II. BACKGROUND

In this section reviews background information on key data sources, uncertainties, and statistical processes that address existing uncertainty in big data. Big data definition data containing high variability, coming with increasing volumes and additional speed. ... Simply put, big data is big, complex data sets, especially for new data sources. These data sets are so powerful that conventional data processing software simply cannot manage them.

III. METHODOLOGY

Big data

In May 2011, big data was announced as the next frontier of production, innovation, and competition. In 2018, the number of internet users grew by 7.5% from 2016 to more than 3.7 billion people. In 2010, more than 1 zettabyte (ZB) of data was produced worldwide and increased to 7 ZB in 2014 as per the survey. In 2001, the emerging features of big data were defined by three Vs (Volume, Velocity, and Variety). Similarly, the IDC described big data using four Vs (Volume, Variety, Speed, and Value) in 2011. In 2012, Veracity was introduced as the fifth major data element [20,21,22]. Although many other Vs exist, we focus on the five most common aspects of big data. Fig1

Big data analysis

Big data analysis describes the process of analyzing large data sets to detect patterns, anonymous relationships, market trends, user preferences, and other important information that could not previously be analyzed by traditional tools. With the formalization of the five elements of big V data, analytical strategies need to be revised to overcome their limitations in time and space analysis [19]. The possibilities for using big data are growing in the modern world of digital data. The global annual growth rate of big data technology and services is projected to increase by about 36% between 2014 and 2019, while global revenue for big data and business figures is expected to increase by more than 60%.[21] Several advanced data analysis techniques (i.e., ML, data mining, NLP, and CI) and possible techniques such as similar, split-and-win, additional learning, samples, granular computing, feature selection, and choosing an example can turn big problems into smaller problems and can be used to make better decisions, reduce costs, and enable more efficient processing.[22]In the case of large-scale data analysis, simulation reduces the calculation time by breaking down large problems into smaller ones themselves and performing smaller tasks simultaneously (e.g., distributing small tasks to multiple threads, cores, or processors).[22] Matching does not reduce the amount of work done but reduces the calculation time as smaller tasks are completed at the same time instead of one after another in a row.

Uncertainty

Big data statistics explain the process of analyzing large databases for pat- finding Terms, anonymous links, market styles, user preferences, and more information that could not have been previously analyzed by traditional tools. With the Formalization of the five elements of V data, analytical methods are required to be re-evaluated in order to overcome their limitations in time analysis once space. The possibilities for using big data are growing in today's world of digital data. The global annual growth rate of big data technology and services is projected to increase

by about 36% between 2014 and 2019, with global revenue
big data and business statistics are expected to rise by more than 60%. Several advanced data analysis techniques (i.e., ML, data mining, NLP, and CI) and possible strategies such as uniformity, split-and-win, growing learning, samples, granular computing, feature selection, and sample selection can turn big problems into smaller problems and can be used to make better decisions, reduces costs, and enables more efficient processing. When it comes to analyzing big data, comparisons reduce the calculation time to divide big problems into smaller ones and make smaller one's simultaneous activities (e.g., distributing small, multi-thread operations, cores, or processors). Matching does not reduce the amount of work built but instead reduces calculation time as small tasks are completed the same point in time instead of sequence in sequence. The divide and conquer strategy play an important role in processing big data.

Divide-and-conquest has three stages:
(1) To reduce one major problem into Minor problems
(2) To complete minor problems, in which each is solved a small problem contributes to solving a big problem, and
(3) Inclusive solutions to small problems into one big solution so big the problem is considered solved. For many years the strategy of division and conquest has been used on the largest website for the use of records by most groups Data at once. Increased learning is a learning algorithm that is widely used to spread data Trained only with new data than trained with existing data only.

Increase Mental learning adjusts the parameters to a learning algorithm over timing to each new input data and each input is used for training only once. Sampling can be used as a data reduction method for large derivative data patterns on large data sets by selecting, manipulating, and analyzing the subset set data. Some studies show that achieving effective results using sampling depends on the sampling factor of the data used. Computer parts from a large area to make things easier into subsets, or granules. Granular computing is an effective way to do this explains the uncertainty of objects in the search field as it reduces large objects into a small search space. Feature selection is a common way of handling large data for the purpose of selecting a smaller set of related features to compile aggregated but more accurate data shipping. Feature selection is a very useful strategy for data mining before compiling high-quality data.[22] Selecting situations applies to many ML or data mining operations as a major factor in pre-processing data. By using the example option, it is possible to reduce the train sets and working time in the dividing or training stages. Costs of uncertainty (both financially and statistically) and challenges
in producing effective models of uncertainty in large-scale data analysis are the keys to finding strong and efficient systems. Thus, we explore several openings problems of the implications of uncertainty in the analysis of big data in the next section [Fig.2]

The uncertainty stems from the fact that his agent has a straightforward opinion about the true truth, which I do not know certain. This lack of knowledge does it is impossible to determine what certain statements are about the world is true or false, all that can be done is measure the tendency of the statement to be true or lie using a certain amount of value the degree to which one can be sure.

Big Data analytics is ubiquitous from advertising to search and distribution of organizational units. In supply chains, Big Data helps organizations predict the future. An organization can then change plans in such a way that the results are reversed, and success is achieved. This is a feature that movie-makers and artists use when bringing their products to market. Manufacturers evaluate the market, obtain data from systems, understand what consumers want, create models and metrics to test solutions, and apply results in real-time.

IV. FIGURES AND TABLES

In this paper, we have discussed how uncertainty can affect big data, both mathematically and in the database, itself. Our aim was to discuss the state of the art in relation to big data analysis strategies, how uncertainty can adversely affect those strategies, and testing with the remaining open problems. For each standard edition, we summarize the research to help others in the community as they develop their strategies.
We've discussed the issues surrounding V's five of big data, yet many others.[Fig.1]

V is there to look up for the issue to resolve the uncertainty and for handling purposes. According to available research, the focus is on volume, variety,Measurement, speed, and authenticity of data, with less-available function by value (e.g., related data Business interests and decision-making in a particular domain)

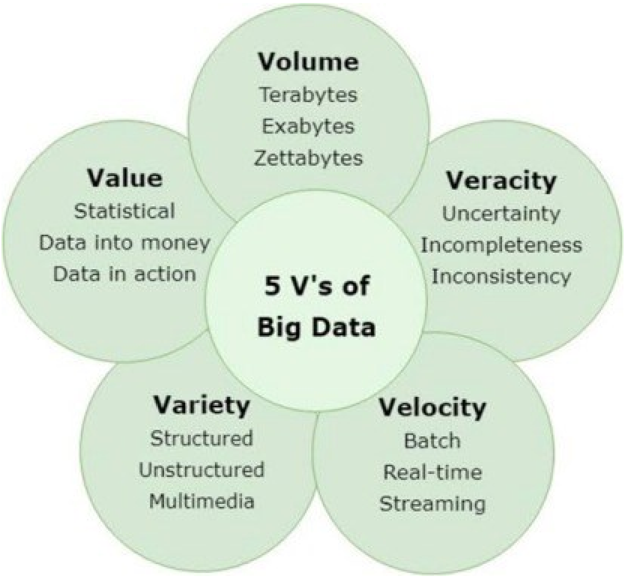

Fig.1

Assessing the level of uncertainty is an important step in analyzing big data. Although there are various strategies for analyzing large data, the accuracy of the analysis may be negatively affected if the data uncertainty or the process itself is ignored. Models of uncertainty such as chance theory, confusion, established theory, etc. can be used to develop larger data analysis techniques to deliver more accurate and meaningful results. Based on previous research, the Bayesian model and the vague set theory are common in modeling uncertainty and decision-making. Table 1 compares and summarizes the strategies we have identified as appropriate, which include comparisons between different uncertainty strategies, focusing on probability theory, Shannon entropy, ambiguous set theory, and complex set theory.[Fig.2]
The uncertainty is due to the fact that his agent has a positive view of the truth, which I do not know for sure. This lack of information cannot determine whether certain statements in the world are true or false, all that can be done is to measure the inclination of a statement to be true or to lie using a certain amount of value to one's attainment. you can be sure

-**Uncertain Data Due to Statistics Analysis**
Some data is recorded statistically so not naturally. This type of data often occurs when research is taken from a scientific context for evaluation.
Uncertain Data for Security Reasons Some data is intentional and uncertain of security reasons. Some data may not be accurately measured, for one reason or another, and will involve unavoidable uncertainty. In such cases, the best thing we can do is try to measure the tendency of the statement to be true (or false). This can be done with the help of an incomprehensible set and by providing a membership level in a statement whether it is true or false. Such data comes from public surveys.

- **Shannon's Entropy** it's just a "amount of information" in the variable. Typically, that translates to the amount of storage (e.g., the number of bits) required for the storage of variables, which can be accurately understood to correspond to the amount of information in that variable.
The odds are easy for the event to happen and always take a value between 0 and 1 (including 0 and 1).

-**A fuzzy concept** an idea where usage limits can vary greatly depending on context or circumstances, rather than permanent. This means that the concept is somewhat vague, has no fixed meaning, is accurate, unless it is still vague or completely meaningless.
- In the theory of the wrong set, the training data set is called the information table or information system. Represents a table where lines represent objects or situations and columns representing attributes or features

-**Vague or ambiguous data** means not clearly expressed or not clear in meaning or application. Vague data contains some vague predicate such as “tall” or “cloudy day”.
Ambiguous means doubtful, uncertain, or capable of being understood in either of two or more possible senses.

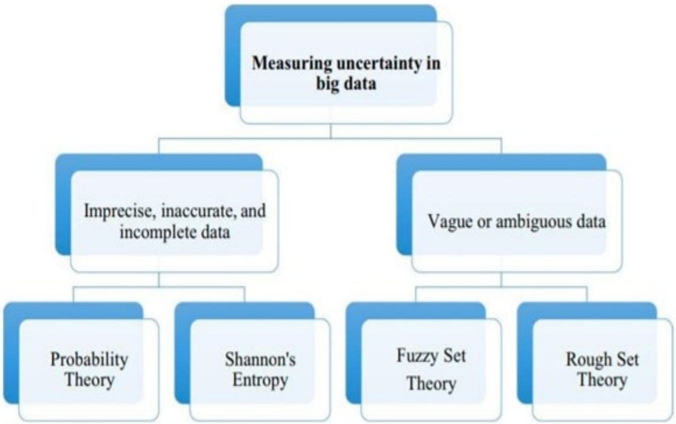

Fig.2

V. CONCLUSION

Here we would like to conclude our Research that big data has great potential for finding the most effective learning. It stimulates new research questions and designs, applies new technologies and tools to data collection and analysis, and eventually becomes a common research concept However, it is still novel and uncommon for many researchers. In this paper, we describe the general background, contextual concepts, and recent developments of this rapidly growing domain. As well as emerging opportunities, we have highlighted key challenges and emerging trends in big data use in education, reflected in academic research, policy making, and industry. Here it summarizes the major challenges and potential solutions for big data in education and many more sectors and how uncertainty can be handled.

Acknowledgments

We would like to express our sincere thanks to our guide Prof. Neha Lodhe for taking time from her busy schedule to provide us with a great deal of help, support and encourage us to work diligently at every aspect of our research paper. Her views have always been equitably providing a perfect balance between encouragement and constructive criticism. Her tips and suggestions helped us to decide the correct approach to the research paper. We attempted to find help from a variety of individuals at various stages of the research paper. We would like to thank everyone for their guidance.

REFERENCES

1. Jaseena KU, David JM. Issues, challenges, and solutions: big data mining. Computer Sci. Info. Technology (CS & IT). 2014;4:131–40.
Google Scholar
2. Marr B. Forbes. How much data do we create every day? 2018. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#4146a89b60ba.
3.McAfee A, Brynjolfsson E, Davenport TH, Patil DJ, Barton D. Big data: the management revolution. Harvard Bus Rev. 2012;90(10):60–8.
Google Scholar
4. Zephoria. Digital Marketing. The top 20 valuable Facebook statistics—updated November 2018. 2018. https://zephoria.com/top-15-valuable-facebook-statistics/.
5. Iafrate F. A journey from big data to smart data. In: Digital enterprise design and management. Cham: Springer; p. 25–33. 2014.
Google Scholar
6. Lenk A, Bonorden L, Hellmanns A, Roedder N, Jaehnichen S. Towards a taxonomy of standards in smart data. In: IEEE international conference on big data (Big Data), 2015. Piscataway: IEEE. p. 1749–54. 2015.
7.Tsai CW, Lai CF, Chao HC, Vasilakos AV. Big data analytics: a survey. J Big Data. 2015;2(1):21.
Article Google Scholar
8. Chen M, Mao S, Liu Y. Big data: a survey. Mobile Netw Appl. 2014;19(2):171–209.
Article
Google Scholar
9. Ma C, Zhang HH, Wang X. Machine learning for big data analytics in plants. Trends Plant Sci. 2014;19(12):798–808.
Article
Google Scholar
10. Borne K. Top 10 big data challenges a serious look at 10 big data v's. Recuperate de. 2014. https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs. Accessed 11 Apr 2014.
11. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: the next frontier for innovation, competition, and productivity. 2011.
12. Pouyanfar S, Yang Y, Chen SC, Shyu ML, Iyengar SS. Multimedia big data analytics: a survey. ACM Comput Surv (CSUR). 2018;51(1):10.
Google Scholar
13. Cimaglobal. Using big data to reduce uncertainty in decision making. 2015. http://www.cimaglobal.com/Pages-that-we-will-need-to-bring-back/velocity-archive/Student-e-magazine/Velocity-December-2015/P2-using-big-data-to-reduce-uncertainty-in-decision-making/.
14. Maugis PA. Big data uncertainties. J Forensic Legal Med. 2018;57:7–11.
Article-Google Scholar
15. Saidulu D, Sasikala R. Machine learning and statistical approaches for Big Data: issues, challenges and research directions. Int J Appl Eng Res. 2017;12(21):11691–9.-Google Scholar
16. Wang X, He Y. Learning from uncertainty for big data: future analytical challenges and strategies. IEEE Syst Man Cybern Mag. 2016;2(2):26–31.

17. https://journalofbigdata.springeropen.com/articles
18. https://www.xenonstack.com/insights/big-data-challenges
19. https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-019-0206-3.pdf
20. https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.researchgate.net/publication
21.Uncertainty_in_big_data_analytics_survey_opportunities_and_challenges&ved=https://www.semanticscholar.org/paper/Uncertainty-in-big-data-analytics
22.https://journalofbigdata.springeropen.com/articles
23.www.google.com