International Advance Journal of Engineering, Science and Management (IAJESM)
ISSN -2393-8048, January-June 2021, Submitted in April 2021, iajesm2014@gmail.com

Literature Study on the Artificial Intelligence Related to Big Data using Techniques

Anju Research Scholar, Department of Computer Science, Monad University, Hapur, Uttar Pradesh (India)
Dr. Kailash Kumar Assistant Professor, Department of Computer Science, Monad University, Hapur, Uttar Pradesh (India)

Abstract:

Artificial Intelligence (AI) has vast potential in marketing. It aids in proliferating information and data sources, improving software's data management capabilities, and designing intricate and advanced algorithms. AI is changing the way brands and users interact with one another. The application of this technology is highly dependent on the nature of the website and the type of business. Marketers can now focus more on the customer and meet their needs in real time. By using AI, they can quickly determine what content to target customers and which channel to employ at what moment, thanks to the data collected and generated by its algorithms. Data analytics using artificial intelligence is the process of leveraging advanced AI techniques to extract insights and knowledge from large and complex datasets. This involves utilizing machine learning algorithms, deep learning models, and natural language processing techniques to uncover patterns and relationships within big data that can inform decision making and drive innovation. The goal of big data analytics using AI is to automate data analysis and make the process faster, more accurate, and more scalable, enabling organizations to harness the full potential of their data and gain a competitive advantage.

Keywords: Big Data, Clustering Algorithms, MapReduce, Swarm Optimization Techniques.

Introduction: Digital era with its opportunity and complexity overwhelms industries and markets that are faced with a huge amount of potential information in each transaction. Being aware of the value of gathered data and benefitting from hidden knowledge create a new paradigm in this era, which redefines the meaning of power for corporation. The power of information leads organizations toward being agile and to hit the goals. Big data analytics (BDA) enforces industries to describe, diagnose, predict, prescribe, and cognate the hidden growth opportunities and leads them toward gaining business value [8]. BDA deploys advanced analytical techniques to create knowledge from exponentially increasing amount of data, which will affect the decision-making process in decreasing complexity of the process [3]. BDA needs novel and sophisticated algorithms that process and analyze real-time data and result in high-accuracy analytics. Machine and deep learning allocate their complex algorithms in this process considering the problem approach [28]. Clustering is an unsupervised learning technology, and it groups information (observations or datasets) according to similarity measures. Developing clustering algorithms is a hot topic in recent years, and this area develops rapidly with the increasing complexity of data and the volume of datasets. In general, big data clustering methods can become categorized into two main groups: single-machine clustering techniques and multiple-machine clustering methods [12]. Lately multiple machine clustering techniques offers drawn more interest because they are even more versatile in scalability and provide faster response time to the users. Although the intricacy and velocity of clustering algorithms is definitely related to the quantity of situations in the dataset, but at the various other hands dimensionality of the dataset can be other important element [13]. In truth the more sizes data possess, the even more is complexity and it means the longer performance period. Sampling methods decreases the dataset size however they perform not really provide a answer for high dimensional datasets [14].

Although sampling and dimensions decrease strategies utilized in single-machine clustering algorithms enhances the scalability and velocity of the algorithms, but today the development of data size is usually method very much quicker than memory and processor developments, as a result one machine with a solitary processor and a memory cannot deal with terabytes of data and it underlines the want algorithms that can become operate on multiple machines [15].

One of the vital consequences of the digital world is creating a collection of bulk of raw data. Managing such valuable capital with different shape and size on the basis of organizations' needs the manager's attention. Big data has the power to affect all parts of society from social aspect to education and all in between. As the amount of data increases especially in technology-based companies, the matter of managing raw data becomes much more important. Facing with features of raw data like variety, velocity, and volume of big data entitles advanced tools to overcome the complexity and hidden body of them. So, big data analytics has been proposed for “experimentation,” “simulations,” “data analysis,” and “monitoring.” Machine learning as one of the BDA tools creates a ground to have predictive analysis on the basis of supervised and unsupervised data input. In fact, a reciprocal relation has existed between the power of machine learning analytics and data input; the more exact and accurate data input, the more effective the analytical performance. Also, deep learning as a subfield of machine learning is deployed to extract knowledge from hidden trends of data [8].

De-duplication [16,17,18] can become divided into four measures: data chunking, chunk computation, chunk index search, and exclusive data shop. Resource de-duplication can be a well-known plan that works the 1st two guidelines of the de-duplication process at the customer aspect and chooses whether a chunk is certainly a duplicate before data transfer to conserve network bandwidth by staying away from the transfer of redundant data, which varies from target de-duplication that performs all de-duplication techniques at the focus on side [19].

In parallel clustering [20], designers are included with not only parallel clustering difficulties, but also with information in data distribution procedure between different machines obtainable in the network as well, which makes it extremely difficult and time consuming. Difference between parallel algorithms and the MapReduce [21,22] framework is normally in the comfortless that MapReduce provides for developers and discloses them type unneeded networking complications and ideas such as weight handling, data distribution, fault tolerance and etc. by managing them instantly. This feature enables huge parallelism and less difficult and faster scalability of the parallel program.

The field of data analysis and big data processing has seen a significant increase in the amount of huge data being generated and stored in recent years. Some studies argue that handling and using this huge data could become a new pillar of economics, scientific research, experimentation, and simulation. Indeed numerous chances of big data appearing in different areas similar to health (Enhancing the effectiveness of some treatments), transportation (reducing costs), finance (minimizing pitfalls), administration (decision stuff with high effectiveness and speed), social media, and government services. However, in today's era, big data is also fraught with problems and has some quality issues like issues of scale, heterogeneity, privacy, timeliness, and visualization, at all stages of the analysis pipeline from data acquisition to result in interpretation. To improve data processing's effectiveness and usefulness, the most recent techniques and technologies are used to deal with this large data [1]. Another crucial data analysis technique is cluster analysis, which aims to categorize physical or abstract sets into related object classes so that items within the same group share a high degree of similarity and differ significantly from one another. There are different clustering algorithms used to manage large sets of data. But no clustering algorithm can solve all the Big Data issues [2]. Among them, the K-means algorithm is widely used because of its simplicity, but how to make it more compatible with the development of the era of big data still faces very big challenges like how to reduce the time complexity of the K-means algorithm and improve our clustering effect still needs further optimization [3]. In this research work, we propose K-Means Clustering Algorithm with Artificial Bee Colony (ABC) algorithm and MapReduce Framework. It is a powerful approach for solving large-scale clustering problems.

Literature Survey

Big data in business context can be managed and analyzed through big data analytics, which is known as a specific application of this field. Also, big data gained from social media can be managed efficiently through big data analytics process. In this way, customer behavior can be understood and five features of big data, which are enumerated as volume, velocity, value, variety, and veracity, can be handled. Big data analytics not only helps business to create a comprehensive view toward consumer behavior but also helps organizations to be more innovative and effective in deploying strategies [4]. Small and medium size company use big data analytics to mine their semistructured big data, which results in better quality of product recommendation systems and improved website design [9]. As Ref. [9] cited, big data analytics gains advantages of deploying technology and techniques on their massive data to improve a firm's performance. This quick growth is certainly sped up by the dramatic boost in approval of social networking applications, such as Facebook, Twitter, etc., that enable users to produce material openly and enhance the currently huge Web volume [9].

Working with Big Data, the amount of space required to shop it is normally extremely relevant. There are two primary methods: compression where we do not lose anything or sampling where we select what is the data that is usually more relations [10].

AI is a computer science technology that teaches computers to comprehend and emulate human communication and behaviour. Based on the data provided, AI has created a new intelligent machine that thinks, responds, and performs jobs the same way people do. AI can do highly technical and specialised activities such as robotics, speech and picture recognition, natural language processing, problem-solving, etc. AI is a collection of several technologies capable of executing tasks that need human intelligence. When applied to standard commercial processes, these technologies can learn, act, and perform with human-like intelligence. It simulates human intelligence in machines, saving us time and money in business transactions [[19], [20], [21], [22]].

AI is concerned with creating intelligent machines that can think and act like humans. It provides exceptional opportunities for a wide range of industries. Every industry mentioned is either terrified or enthralled by the arrival of AI. AI creates intelligent machines and devices that can think and react like humans. This technology has been dubbed the "next step" in the industrial revolution. It is believed that AI and ML hold solutions to most of today's problems.

Furthermore, AI may aid in the prediction of future problems. AI can create new technologies, industries, and environments. In a nutshell, AI simulates human intelligence processes by machines. This may include learning, reasoning, and, most importantly, the ability to self-correct [[23], [24], [25]].

AI can analyse, comprehend, and make decisions. It is for existing user data and is used to make market predictions and predict user behaviour. It is also known as data forecast, and organisations worldwide use it to fine-tune their sales and marketing strategies to increase sales. Most AI applications in marketing nowadays employ ML, from personalising product suggestions to assisting in discovering the most successful promotion channels, estimating churn rate or customer lifetime value, and building superior customer groups [26,27].

Using compression, we may consider even more time and much less space, so we can consider it as a change from period to space. Using sampling, we are dropping info, but the benefits in space may end up being in orders of degree. Using merge-reduce the little units can after that be utilized for resolving hard machine learning complications in parallel processing [11]. Despite that the info found out by data mining can become extremely useful to many applications; people possess demonstrated raising concern about the additional part of the gold coin, specifically the privacy risks presented by data mining.

This section covers a broad review of big data difficulties, clustering algorithms, in particular, the KMeans Clustering Algorithm, the Artificial Bee Colony Algorithm, and the MapReduce Framework and big data applications. The development of big data has led to the analysis of a wide variety of data formats, most of which are streaming in nature. As a result, conventional techniques have a difficult time meeting Big Data needs.

Inspiring from hierarchical structure of human brain, deep learning algorithms extract complex hidden features with a high level of abstraction. When massive amounts of unstructured data represent, the layered architecture of deep learning algorithms works effectively. The goal of deep learning is to deploy multiple transformation layers where in every layer output representation is occurred. Big data analytics comprises the whole learnt untapped knowledge gained from deep learning. The main feature of big data analytics, which is extracting underlying features in huge amounts of data, makes it a beneficial tool for big data analytics [42].

Big data are generated through internal and external sources of data; thus, existing systems fail to handle the unprecedented data. High-performance, highly scalable systems with advanced techniques are required to process valuable information. The study shows that the current tool and technology must be updated with time as the data is continuously growing [4]. The term "big data" describes a collection of numerical data generated by applying new technologies for either personal or professional usage. Big data analytics is used to analyze large amounts of data to find hidden patterns. The complexity of the analysis of this data, however, varied depending on the process that was needed [5], from traditional data analysis to the more current big data analysis and data analytics. The KDD process serves as the study's framework from a systems perspective. The unresolved problems with computing are discussed, resulting in quality, security, and privacy [6]. By grouping data using a variety of clustering algorithms, we set out to identify the day of the year with the greatest heart rate. A more effective clustering technique with improved accuracy, recall, and F-measure is produced via hybrid methodology. The hybrid technique produces the most clusters and includes each data point in each cluster [32]. EM and FCM clustering algorithms exhibit good performance in terms of the quality of the clustering outputs. Future research should address each clustering algorithm's shortcomings because none performs well for all evaluation criteria [8]. K-means clustering is a highly traditional clustering algorithm, and its use will increase over time. Future research may enhance the capability to handle large or multidimensional data sets. An area of study is the clustering of exponential data using K-Means [9]. A popular clustering method that is frequently used for clustering massive amounts of data is K-means. An effective method for clustering data points is presented in this research. The suggested approach guarantees that clustering is completed in O(nk) time [10]. However, Kmeans requires initial data point selection and nearest cluster assignment. This study explains how to more accurately assign data points to their nearest clusters and determine initial centroids using improved methodologies [11]. An analysis of previous work on artificial bee colony algorithm (ABC), ABC variations, and data clustering applications. ABC is a straightforward and adaptable method that requires less parameter tuning than other algorithms. The efficiency, precision, and usefulness of ABC in solving various optimization issues are demonstrated by numerous tests conducted in the pertinent literature [12]. ABC works on position updating formula and objective function. The iterative optimization procedure is more effective by using a position update formula based on local better and global best [13]. An artificial bee colony algorithm based on information learning (ILABC) could be useful for data structuring and data probation. The design of wireless telecommunications networks and the flow scheduling problem illustrate difficult optimization problems that can be solved with ILABC. Applying ILABC to more difficult issues may be worthwhile [14]. Our dataset's size has constantly been growing, making it challenging to cluster the data using conventional clustering algorithms. The fastest execution time is provided by the ABC system, which is also more effective for all sorts of data. To discover the optimal fitness value, the mapper phase simulates the behavior of an employed bee. In the reducer mode, the behavior of an observer bee is simulated to optimize the clusters [15].

The ease of use and quick convergence, the clustering algorithm has become a popular technique for cluster analysis. The IABC algorithm is suggested to solve the issues with the K-means clustering algorithm's randomly chosen initial centre points and poor global search capability [16]. The k-means algorithm challenges selecting an appropriate set of parameters, such as the number of clusters k and initial centroids. For the ABC algorithm, they have not discovered any attempts to date. A novel method to generate variable-length food sources for the ABC algorithm with a variable length (ABCVL) to supply the system with an appropriate level of diversity [17]. The ABC-based cluster has improved the influence of the initial center value and increased inter-group variation and similarity in the clustering [18]. A hybrid clustering algorithm based on modified ABC and K-Means algorithms. The relative fitness of

each person - the ratio between their individual and overall fitness is used to create a roulette wheel. In the onlooker bee phase, variable tournament selection is used instead of roulette wheel selection [33].
This study aimed to provide an overview of the MapReduce ideas used in big data analytics. To analyze large data, which is unstructured data like web data, Google developed Map Reduce [20]. Big data and related technologies can positively impact the company's operations. A few guidelines must be followed to acquire fast and beneficial results from big data. Programming MapReduce using the Hadoop framework, which is an open-source system, accelerates the processing of massive amounts of data [21]. Without any prior programming knowledge, programmers can simply grasp the MapReduce framework. Load balancing, fault tolerance, serialization, and parallelization are no longer required [22]. The data mining environment of the Hadoop cluster is used to study the K-means method. With the help of the improved algorithm, catering decision-makers may identify highvalue consumer segments and provide superior service. The k-Means algorithm for processing data mining has superior expansion performance and mining efficiency in a cluster of cloud computing platforms, which has been demonstrated [23].
The K-Means Clustering Algorithm offers a reliable and effective method for classifying data that have similar features. It lowers the implementation costs associated with handling such massive data volumes via a distributed network. Reducing the number of iterations needed to finish a task allows for improvements [2]. A parallel Kmeans method based on Hadoop is given in work with quite good findings for data processing effectiveness and convergence. As the amount of data increases, the acceleration effect is better for processing huge amounts of data, especially in the MapReduce architecture [25]. The standard K-means method has been enhanced. The problem of the K-Means initial center point sensitivity was resolved by the modified approach, which successfully identified the initial clustering centers. Large data processing was made possible by better algorithm parallelization. The performance of the K-means algorithm has been increased, and both techniques significantly improve results [26]. K-means algorithm improves MapReduce design using an iteration-saving technique. They illustrate that this keeps 80% of the clustering accuracy while reducing the number of iterations and execution time in clustering techniques [2].
An effective artificial bee colony for MapReducebased large-scale data clustering is developed. In the Hadoop system, the ABC could be used to streamline the clustering of enormous amounts of data. It provides an adequate level of grouping and performance in comparison to more current methods [28]. The novel optimization method has effective search capabilities in the solution space, and a pattern is applied to achieve the best outcomes with fewer iterations. Many methods enhance the search quality and fast local search time in global search by integrating and extracting the features of both MapReduce and a specific method [29]. MapReduce's parallelization capabilities make using the Artificial Bee Colony technique simple. Each member of the population just needs to look in a very small area, which allows them to find the answer more quickly. Because the particles continually update themselves after each iteration, the proposed model for parallel ABC can use a huge population but cannot be used with a large dataset [30]. The Modified Artificial Bee Colony Algorithm is the optimization algorithm we used (MABC). A method for utilizing the map-reducing algorithm to solve resource issues in clouds. With the aid of the optimization algorithm, the MapReduce algorithm creates a further improved solution. The suggested approach to resource problem reduction works better because it requires less space for data storage [31].
With the rapid innovations of digital technologies, the volume of digital data is growing fast (Klein, 2017). Consequently, large quantities of data are created from lots of sources such as social networks, smartphones, sensors, etc. Such huge amounts of data that conventional relational databases and analytical techniques are unable to store and process is called Big Data. Development of novel tools and analytical techniques are therefore required to discover patterns from large datasets. Big data is produced quickly from numerous sources in multiple formats. Henceforth, the novel analytical tools should be able to detect correlations between

rapidly changing data to better exploit them. As mentioned, traditional processing techniques have problems coping with a huge amount of data. It's necessary to develop effective ways for data analysis in big data problems. Various big data frameworks such as Hadoop and Spark have allowed a lot of data to be distributed and analyzed (Oussous et al., 2018). Furthermore, different types of Artificial Intelligence (AI) techniques, such as Machine Learning (ML) and search-based methods were introduced to deliver faster and more precise results for large data analytics. The combination of big data tools and AI techniques has created new opportunities in big data analysis.

Organizations can extract valuable information and patterns that may affect business through big data analytics (Gandomi & Haider, 2015). Thus, advanced data analysis is needed to identify the relations between features and forecast future observations. Big data analytics refers to techniques applied to achieve insights from huge datasets (Labrinidis & Jagadish, 2012). The big data analytics results can improve decision-making and increase organizational efficiency

References:

[1] Guma Abdulkhader Lakshen, Sanja Vranes, and Valentina Janev, "Big Data & Quality- A Literature Review," 24th Telecommunications forum TELFOR, pp. 1-4, 2016.

[2] Prajesh P. Anchalia, Anjan K. Koundinya, and Srinath N. K, "Map Reduce Design of K-means Clustering Algorithm," IEEE International Conference on Information Science and Applications (ICISA), pp. 1-5, 2013. [3] Chen Jie et al., "Review on the Research of K-means Clustering Algorithm in Big Data," IEEE, International Conference on Electronics and Communication Engineering, pp. 107-111, 2020.

[4] R Rawat and R Yadav, "Big Data: Big Data Analysis, Issues and Challenges and Technologies," IOP Conference Series Materials Science and Engineering, vol. 1022, 2021.

[5] Gandomi A, Haider M. 2015. Beyond the hype: big data concepts, methods, and analytics. International Journal of Information Management 35(2):137–144 DOI 10.1016/j.ijinfomgt.2014.10.007.

[6] Gantz J, Reinsel D. 2012. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future 2007(2012):1–16.

[7] Ghani NA, Hamida S, Hashem IAT, Ahmed E. 2019. Social media big data analytics: a survey. Computers in Human Behavior 101:417–428.

[8] Glossary GI. 2014. Big Data (definition). Gartner.com. Available at http://www.gartner. com/it-glossary/ big-data (accessed on 17 November 2014).

[9] Hammou BA, Lahcen AA, Mouline S. 2020. Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics. Information Processing & Management 57(1):102122 x

[10] Abdulbaset S. Albaour, and Yousof A. Aburawe, "Big Data: Review Paper," International Journal Of Advance Research And Innovative Ideas In Education, vol. 7, no. 1, 2021.

[11] Chun-Wei Tsai et al., "Big Data Analytics: A Survey," Journal of Big Data, vol. 2, no. 20, 2015.

[12] Fatema Jamnagarwala, and P.A.Tijare "Implementation of Data Mining With lustering of Big data for Shopping mall's data using SOM and K-means Algorithm," International Journal of Computer Trends and Technology, vol. 67, no. 12, pp. 3-7, 2019.

[13] Adil Fahad et al., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 267-279, 2013.

[14] Bao Chong, "K-Means Clustering Algorithm: A Brief Review," Academic Journal of Computing & Information Science, vol. 4, no. 5, 2021.

[15] Shi Na, Liu Xumin, and Guan Yong "Research on k-means Clustering Algorithm", 3 rd Intl Symposium on Intelligent Information Technology and Security Informatics, pp. 63-67, 2010.

[16] Unnati R. Raval, and Chaita Jani, "Implementing & Improvisation of K-means Clustering Algorithm," International Journal of Computer Science & Mobile Computing, vol. 5, no. 5, pp. 191-203, 2016.

[17] Ajit Kumar, Dharmender Kumar, and S. K. Jarial, "A Review on Artificial Bee Colony Algorithms and Their Applications to Data Clustering," Cybermetics and Information Technologies, vol. 17, no. 3, pp. 3-28, 2017.

[18] Yi Yang, and Ke Luo, "An Artificial Bee Colony Algorithm Based on Improved Search Strategy," 2nd International Conference on Artificial Intelligence and Information, no. 191, pp. 1-4, 2021.