Asian Journal of Research in Computer Science

11(4): 1-18, 2021; Article no.AJRCOS.73759
ISSN: 2581-8260

A Survey of Data Mining Activities in Distributed Systems

Waleed A. Mohammad[1*], Hajar Maseeh Yasin[1], Azar Abid Salih[1], Adel AL-Zebari[1], Naaman Omar[1], Karwan Jameel Merceedi[1], Abdulraheem Jamil Ahmed[1], Nareen O. M. Salim[1], Sheren Sadiq Hasan[1], Shakir Fattah Kak[1] and Ibrahim Mahmood Ibrahim[1]

[1]Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2021/v11i430267
Editor(s):
(1) Dr. Dariusz Jacek Jakóbczak, Koszalin University of Technology, Poland.
Reviewers:
(1) Muhammad Ahmad Baballe, Kano state polytechnic, Nigeria.
(2) A. Meiappane, Pondicherry University, India.
Complete Peer review History: https://www.sdiarticle4.com/review-history/73759

Review Article

Received 02 July 2021
Accepted 06 September 2021
Published 07 September 2021

ABSTRACT

Distributed systems, which may be utilized to do computations, are being developed as a result of the fast growth of sharing resources. Data mining, which has a huge range of real applications, provides significant techniques for extracting meaningful and usable information from massive amounts of data. Traditional data mining methods, on the other hand, suppose that the data is gathered centrally, stored in memory, and is static. Managing massive amounts of data and processing them with limited resources is difficult. Large volumes of data, for instance, are swiftly generated and stored in many locations. This becomes increasingly costly to centralize them at a single location. Furthermore, traditional data mining methods typically have several issues and limitations, such as memory restrictions, limited processing ability, and insufficient hard drive space, among others. To overcome the following issues, distributed data mining's have emerged as a beneficial option in several applications According to several authors, this research provides a study of state-of-the-art distributed data mining methods, such as distributed common item-set mining, distributed frequent sequence mining, technical difficulties with distributed systems, distributed clustering, as well as privacy-protection distributed data mining. Furthermore, each work is evaluated and compared to the others.

**Corresponding author: E-mail: waleed.mohammad@uod.ac;*

Keywords: Distributed Systems; Data Mining; Parallel Platform; Distributed Clustering.

1. INTRODUCTION

With the rapid development of information technology and data collection, Knowledge Discovery in Databases (KDD), provides a powerful capability to discover meaningful and useful information coming from a collection of data [1]. KDD has numerous real-life applications and has resulted in several DM tasks, such as association rule mining (ARM), sequential pattern mining (SPM), clustering, classification, and outline detection, among others [2]. Depending on different requirements in various domains and applications, the discovered knowledge can be generally classified as frequent item sets and association rules, sequential patterns, sequential rules, graphs, high utility patterns, weight-based patterns, and other interesting patterns [3]. As an important task for a wide range of real-world applications, frequent item-set mining (FIM) or ARM has been extensively studied [4]. Due to the rapid growth of resource sharing, distributed systems are developed, which can be used to utilize the computations [5]. Data mining (DM) provides powerful techniques for finding meaningful and useful information from a very large amount of data, and has a wide range of real-world applications. However, traditional DM algorithms assume that the data is centrally collected, memory-resident, and static [6]. It is challenging to manage the large-scale data and process them with very limited resources. For example, large amounts of data are quickly produced and stored at multiple locations [7]. It becomes increasingly expensive to centralize them in a single place. Moreover, traditional DM algorithms generally have some problems and challenges, such as memory limits, low processing ability, and inadequate hard disk, and so on [8]. To solve the above problems, DM on distributed computing environment [also called distributed data mining (DDM)] has been emerging as a valuable alternative in many applications [9].

In general, computation and data distribution enable engineers/researchers to resolve a wide spectrum of problems, and it can be extended and implemented in various distributed applications [10]. The distributed systems denote that distributed processing units are interconnected and organized via networks to subtend the need for large-scale or great-performance computing, which has received significant attention in recent decades [1112]. peer-to-peer (P2P) systems, grids [13], ad-hoc networks[14], cloud computing systems [15], and social network systems have all received a lot of attention. Nowadays, distributed systems are used for a variety of purposes, including online services, file storage, and scientific computation [16]. Lately, centralized data mining approaches have become common for investigating large amounts of corporate either scientific data contained in databases [17]. The main issue in data mining is determining the relationship between data sets in a timely and accurate manner [18]. The emergence of massive and big data necessitates the use of a single computer to accomplish the computation operation [19]. However, the massive increase in data volume little by little pushes researchers to develop more sophisticated methods or strategies to address this challenge [20]. Distributed computing and parallel have gained popularity in recent years, with a focus on data analysis and knowledge extraction [21]. The introduction of distributed computing many years ago may have dealt with the present crisis, in which the mining data is now not just in Megabytes to Gigabytes scale, but also in the Terabytes to Petabytes range [22]. Every day, social media and network services generate massive amounts of data that reach the Petabyte scale [23]. Because of the presence of massive datasets and the need to access the information rapidly, the use of parallel or distributed computation is extremely relevant today [24]. Commodity hardware can now be conveniently linked to clusters for operating complex tasks in a distributed system [25]. The combination of distributed computing and data mining can enhance data mining algorithm performance, especially in distributed and large datasets [26]. The emergence of distributed data mining has recently become highly significant. It deals with data processing in the distributed system while also paying attention to many topics concerning computing, storage, data sharing, and interaction of human-computer [27]. Simultaneously, data mining has been thoroughly researched [26]. Organizations, corporations, industries, and scientific canters may use data mining techniques to uncover numerous types of secret yet valuable and significant trends and knowledge [28]. As previously reported, data mining techniques could be utilized to evaluate the distribution of collected data [29]. A significant data mining situation is where datasets are shared by two or more entities with each entity owning a component of the data

Historically, traditional methods assumed that the data were centralized and occupant in memory [30]. In distributed networks, this assumption is no longer valid [31]. Unfortunately, applying conventional mining algorithms straight to distributed databases is ineffective due to the high communication overhead [32]. Thus, implementing high-performance data mining in computing distributed systems has been a key advancement in enhancing a system's scalability [33].

A centralized method is essentially inappropriate in traditional data mining technologies for a variety of reasons, including the massive volume of data, the inability to centralize data processed at various locations, bandwidth constraints, energy limitations, including privacy issues [21,34]. To solve, these issues distributed data mining recently emerged as a significant research field. In the distributed data mining research, one of two hypotheses on how data are distributed throughout sites is widely used: heterogeneously (partitioned vertically) or homogeneously (partitioned horizontally) [35].

In general, distributed data mining solves several problems for evaluating distributed data and provides several algorithmic solutions to conduct various mining operations and data analysis in an essentially distributed manner that is sensitive to resource constraints [36]. Many researchers have various techniques to operate on a distributed system like cloud, grid computing, Hadoop [37, 38], and so on and spread the mining computation across more than a single node to increase the efficiency and scalability of data mining [39]. Previous research has demonstrated that distributed data mining is a valuable method for the end-user, government, or enterprise to review data and uncover various types of useful information [40]. It opens up new possibilities while still posing some difficulties for data mining. How do we summarize related researches in different forms of data mining of distributed systems and provide a general taxonomy for them? Thus, the purpose of this research is to investigate current research about data mining of distributed systems [41]. The main contributions of such research are a review of recent data mining works on distributed systems [42]. This is a medium-level study of distributed system approaches for data mining in many areas, including distributed clustering and privacy preservation of distributed data mining [43].

Distributed systems allow productivity to increase through enormous parallelism [44]. Unfortunately, increasing the number of the computers available to customers will also increase troubleshooting in cases of failure [45]. Data mining enables patterns to be removed within enormous data volumes, thereby forming the basis for a viable debugging approach, especially for such distributed systems [46].

To informing judgments about future actions, data mining provides tools for determining correlations, trends and knowledge in huge datasets. Applications from several fields have embraced this method for efficient data processing [47]. When data of a size big and geographically spread across several locations are used for such approaches, several difficulties need to be addressed [48].

A large volume of data was given through the widespread usage of computers and advances in database technology [49]. Touchy rise of data in databases has led to the necessity to find information and expertise in effective data mining approaches [50]. On the other hand, a natural demand for scalable data mining technologies has been established throughout the creation of Network Distribution Computing, such as reserved intranet, internet besides wireless networks [51]. Distribution Data Mining (DDM) seeks to find and combine information from many data sources, geographically scattered across numerous locations. There are, however, several problems with the use of data mining techniques for such systems [52]. A distributed computing system is more complicated than a central or host-based system. It may be necessary to handle heterogeneous systems, several databases and perhaps various schemes [53]. The communications protocol between nodes should be scalable and effective, as well as how information collected from various nodes may be used selectively [54].

DDM faces a major problem in developing mining algorithms without needlessly communicating data [55]. For reasons of efficiency, precision with privacy, this capability is necessary. Furthermore, the mining of dispersed data requires proper protocols, languages, and network services to manage the necessary information and mapping [56].

The study is organized as follows: Distributed Systems and its technical difficulties section introduce the definitions, some important features of distributed systems. Data Mining Techniques on Distributed Environment section highlights and discusses the state-of-the-art

research on data mining in distributed computing resources. The opportunity for distributed data mining section briefly summarizes some opportunities for data mining job in a distributed environment. Finally, conclusions are given in the end.

2. CONCEPTS OF DISTRIBUTED SYSTEMS AND DATA MINING

Distributed computing is a field of computer science that studies distributed systems [54]. A distributed system is a system whose components are located on different networked computers, which communicate and coordinate their actions by passing messages to one another from any system [57]. The components interact with one another for achieving a common goal. Three significant characteristics of distributed systems: concurrency of components, lack of a global clock, and independent failure of components [58]. Examples of distributed systems vary from SOA-based systems to massively multiplayer online games to peer-to-peer applications [59]. A computer program that runs within a distributed system is called a distributed program (and distributed programming is the process of writing such programs) [59]. There are many different types of implementations for the message passing mechanism, including pure Hypertext Transfer Protocol (HTTP), RPC-like connectors and message queues [60]. Distributed computing also refers to the use of distributed systems to solve computational problems. In distributed computing, a problem is divided into many tasks, each of which is solved by one or more computers, which communicate with each other via message passing [39].

The word distributed in terms such as "distributed system", "distributed programming", and "distributed algorithm" originally referred to computer networks where individual computers were physically distributed within some geographical area [61]. The terms are nowadays used in a much wider sense, even referring to autonomous processes that run on the same physical computer and interact with each other by message passing. While there is no single definition of a distributed system, the following defining properties are commonly used such as:

- There are several autonomous computational entities (computers or nodes), each of which has its private local memory [62].
- The entities communicate with each other by message passing.

A distributed system may have a common goal, such as solving a large computational problem; the user then perceives the collection of autonomous processors as a unit. Alternatively, each computer may have its private user with individual needs, and the purpose of the distributed system is to coordinate the use of shared resources or provide communication services to the users [63]. Other typical properties of distributed systems include the following:

- The system has to tolerate failures in individual computers.
- The structure of the system (network topology, network latency, number of computers) is not known in advance, the system may consist of different kinds of computers and network links, and the system may change during the execution of a distributed program [64].
- Each computer has only a limited, incomplete view of the system. Each computer may know only one part of the input.

Various hardware and software architectures are used for distributed computing. At a lower level, it is necessary to interconnect multiple CPUs with some sort of network, regardless of whether that network is printed onto a circuit board or made up of loosely coupled devices and cables [65]. At a higher level, it is necessary to interconnect processes running on those CPUs with some sort of communication system. Distributed programming typically falls into one of several basic architectures: client/server, three-tier, n-tier, or peer-to-peer; or categories: loose coupling, or tight coupling [66].

- Client/server: architectures where smart clients contact the server for data then format and display it to the users. Input at the client is committed back to the server when it represents a permanent change [67].
- Three-tier: architectures that move the client intelligence to a middle-stage so that stateless clients can be used. This simplifies application deployment. Most web applications are three-tier [68].
- n-tier: architectures that refer typically to web applications that extra forward their requests to other enterprise services. This type of application is the one most responsible for the success of application servers [69].
- Peer-to-peer: architectures where there are no special machines that provide a service or manage the network resources. As a

replacement for all responsibilities are uniformly divided among all machines, known as peers. Peers can serve both as clients and as servers. Examples of this architecture include BitTorrent and the bitcoin network [70].

Another basic aspect of distributed computing architecture is the method of communicating and coordinating work among concurrent processes [71]. Through various message passing protocols, processes may communicate directly with one another, typically in a master/slave relationship [72]. Alternatively, a "database-centric" architecture can enable distributed computing to be done without any form of direct inter-process communication, by utilizing a shared database [73]. Database-centric architecture in particular provides relational processing analytics in a schematic architecture allowing for live environment relay [74]. This enables distributed computing functions both within and beyond the parameters of a networked database [75].

Data mining is a process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems [76, 77]. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use [78,79]. Data mining is the analysis step of the "knowledge discovery in databases" procedure, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating [80].

The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself [81]. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence [82, 83]. The book Data mining: Practical machine learning tools and techniques with Java (which covers mostly machine learning material) was initially to be named just Practical machine learning, and the term data mining was only added for marketing reasons [84]. Often the more general terms (large scale) data analysis and analytics—or, when referring to actual methods, artificial intelligence besides machine learning—are more appropriate [85].

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining) [86]. This usually involves using database techniques such as spatial indices [87]. These patterns can then be seen as a kind of summary of the input information, and may be used in further analysis or, for example, in machine learning and predictive analytics [88]. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system [89]. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but organise belong to the overall KDD process as additional steps [90]. The difference between data analysis and data mining is that data analysis is used to test models and hypotheses on the dataset, e.g., investigating the effectiveness of a marketing campaign, regardless of the amount of data; in contrast, data mining uses machine learning and statistical models to uncover clandestine or hidden patterns in a large volume of data [91].

3. LITERATURE REVIEW

In reality, computation and data distribution aid in the resolution of issues and can be adopted and implemented in a wide range of distributed applications. Distributed systems are networks that connect and coordinate distributed computational units that subtend the large and high-performance computing demands that have received a lot of coverage in recent decades [92]. There are many forms of distributed systems that have been thoroughly studied, including peer-to-peer (P2P) systems, grids, computational systems ad-hoc networks, and systems of online social networks. Distributed networks are currently used for a variety of uses, including online servers, data storage, and massive science computing. data mining has been thoroughly researched in order.

In the past, previous methods presumed the data is clustered and memory-resident. This argument is no longer true of distributed networks. Worse, because of involving the high coordinating overhead, implementing traditional mining algorithms immediately to distributed datasets becomes ineffective [93]. From a two-dimensional data stream, [94]presented the parallel algorithms of mining Correlated Hitters. They designed and applied a message-passing, a shared memory also a hybrid algorithm in particular. Web crawlers were maturely contributed with major search engines including search fields during a period of the rapid growth of Internet technologies and rising social desires of people [21]. F. Liu and W. Xin [95] demonstrates the architecture incorporation of the Spark-based distributed crawler system, provides a corresponding framework diagram, which presents the distributed framework platform in depth utilizing Spark's RDD elastic computational model and task assignment algorithm. However, we can fix the issue of insufficient resource consumption and poor collection performance using this Spark-based distributed crawler method, and then resolve the contradiction between the present exponential growth of data size as well as the speed of collecting information [96,97].

Through constructing a database system of an engineering standard to enhance the efficiency for querying engineering specifications, an engineering system of the standard database depends on a Web crawler was suggested to provide a public service portal for intellectual construction drawing analysis [94]. The system primarily obtains unstructured engineering appropriate information mostly on Web site via the crawler module, collects information in the graphic on Web site via image recognition module, filters non-critical information into the organized information via the data cleaning module, and eventually guarantees real-time data changes via the data updating module.

Formal Concept Analysis (FCA) is used in a spectrum of fields including data processing, artificial intelligence with software engineering. Algorithms of FCA are computationally costly, with an irregular recursion tree. To control the FCA computational complexity, some parallel algorithms have been introduced [98],[20]. As a result, it is important to create a more scalable and adaptable mining framework to discover unseen yet significant and valid trends and information from distributed and complex datasets instead of centralized databases. To address these issues, data mining of distributed networks has appeared as a critical research field [99].

One of the fundamental techniques for discovering information from big data produced by modern applications is matrix decomposition. However, processing very large amounts of data on a single computer is still unreliable or impossible. Furthermore, big data is most often distributed, compiled, and maintained on several devices. As a result, such data usually contain a high level of heterogeneous noise. Developing distributed matrix decomposition through big data analytics becomes critical and useful. The distributed Bayesian matrix of decomposition model (DBMD) is proposed for clustering and large data mining [100,101].

An engineer charged with evolution or maintenance will benefit from changeability and evolvability research.[102] utilizes evolution mining and change mining in the context of emerging distributed networks. First, we propose the Service Transition Classified dependent Interface Slicing algorithm which mines update information from two iterations of the distributed system of evolution. Second, four Metrics of Service Evolution are proposed to capture the evolution of the system. As the two are combined, they form the foundation of our current Service Evolution Analytics method that involves learning during the implementation process.

Most large-scale training approaches have been introduced, with a personalized design through top to bottom and sophisticated synchronization support being preferred. Introduce ZenLDA, an LDA training framework with a generalized design only for distributed data system [99].

The information uncovered in the cloud-computing context is concerned about their private data by the data owners without authorization.

The notion of distributed databases is a way to solving this problem where several parties have vertical or horizontal data divisions. Cluster analysis is a regularly utilized data mining task aimed at degrading or splitting a normally multivariate group of data, to make the data items in one group more similar [35,15,17] dedicated to overcoming the challenge of data security for data mining systems distributed.

Previous research has demonstrated that distributed data mining is an impactful tool for end-users, governments, and enterprises to investigate data and find different forms of useful knowledge. It expands data mining's options, but it also introduces new challenges. However, [92] proposed a system infrastructure for constructing data of big scholars a broad information graph, discovering meta paths among entities also calculating the relevance of entities in the database.

Lately, [103] looked at the issue of identifying large-scale flexibility patterns. The fact that an enormous amount of information is spread out temporally and spatially over multiple tracking sensors is a common problem of mobility tracking systems. As a result, they develop a spatial testing and knowledge sharing protocol that offers probabilistic guarantees of detecting prominent patterns.

Pooled mining, at the other hand, allows blockchains to transform into clustered networks when pool members transfer decision-making authority to pool managers [104]. Although some related studies have been conducted in the past, the plurality of them provides a very preliminary analysis of a specific category of the distributed system, such as the grid load balancing study [11,105], the load balancing article in cloud computing, or the survey of capacity balancing in (P2P) networks.

4. THE TECHNICAL DIFFICULTIES WITH DISTRIBUTED SYSTEMS

In opposed to traditional centralized systems, the concept of the distributed system relates to a massive collection of resources distributed by computers connected by a network. Software sharing, Hardware sharing, service sharing, data sharing, are just a few examples. The growth of parallel computing, collaborative computing, and distributed computing encouraged the creation of distributed systems. it is one in which elements at networked computers interact and coordinate their operations only through message transmission [15,106].

A distributed system, in other terms, is a combination of independent computer parts (subsystems) that seem to consumers as a single cohesive system. The distributed system is now a sophisticated system that necessitates powerful technology and complex algorithms, as illustrated in Fig. 1.

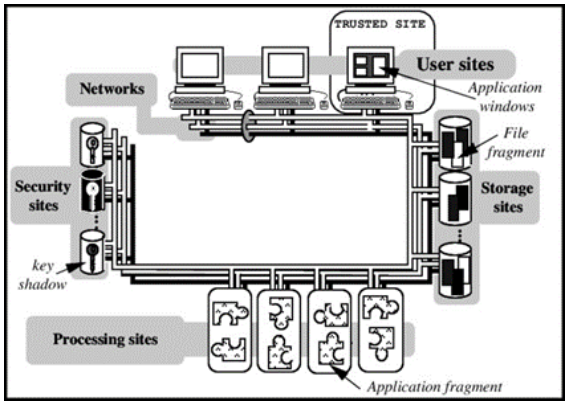

Fig. 1. Distributed System Architecture

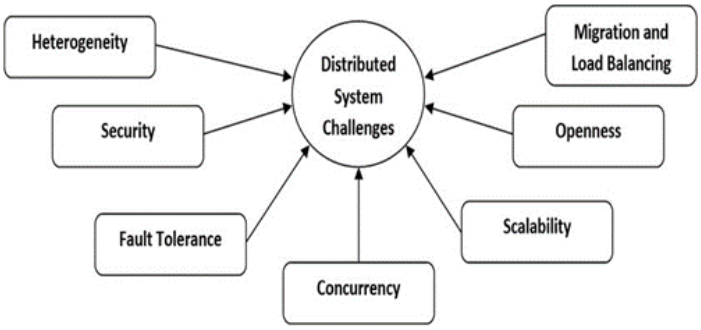

Fig. 2. Technical challenges in the distributed system

Distributed systems, in which distributed computational components are linked and arranged by networks to achieve the demand for huge and high-performance computing, have received a lot of interest in recent decades. Many different forms of distributed system applications are being extensively researched. Currently, distributed systems have a variety of applications such as web servicing, data mining, scientific computation, and data storage. Despite significant advancements in distributed systems, there are also some technical challenges. The major goal in distributed systems, as seen in Figure 2, can be categorized into eight aspects: heterogeneity, security, openness, failure handling, scalability, concurrency, quality of service, transparency [106].

5. DIFFICULTIES IN DISTRIBUTED DATA MINING

Traditional data mining algorithms are predicated on the assumption that data is memory-resident, centralized, and static [23]. Two challenges have arisen as a result of the enormous growth of huge data in the past few decades. First, massive quantities of data are generated in a short period. Second, the data is stored in numerous places, and centralizing it in one location is becoming increasingly expensive. As a result, distributed data mining is a significant issue in a wide range of complex network datasets. In such a distributed system, probes are distributed throughout the network at strategic places particularly in areas with restricted energy and memory [107]. Distributed data mining investigates ways for applying data mining in a noncentralized manner using data acquired from distributed places. The objective here would be to reduce the quantity of data transferred between the various locations. Some significant challenges of the distributed data mining issue are being investigated, like how to essentially minimize the communication expense, how to mine over various heterogeneous data, e.g, multisource databases, and how to do multi relationship mining in the distributed system [14]. As seen in Figure 2, six technical issues in the distributed system, which include heterogeneity, security, scalability, concurrency, quality of service, and transparency, are the same while implementing distributed data mining particularly heterogeneity, scalability, and security. The distributed data mining addresses these issues in the study of distributed data by providing multiple algorithmic methods to perform various mining operations and data analysis in a distributed manner while keeping resource constraints in mind.

The constant rise of accessible data volumes from numerous sources presents new problems for comprehend them effectively. Discovery of knowledge in big data repositories includes computer-intensive, collaborative besides dispersed procedures and activities. The Grid is a lucrative infrastructure for managing data mining and the discovery of information. For this, the creation of KDD applications requires modern software tools and services. The KN is a high-level framework that provides tools and services for the development of awareness based on the grid. These services allow users to develop and manage sophisticated applications to find awareness that incorporate data sources and tools for data mining as distributed grid services. The new design and implementation of all such services as WSRF-compliant Grid Services are now taking place.

Databases are growing so huge that traditional data analysis and ways of display become outdated. Data mining and knowledge research in databases are concentrated in the extraction of models and interest patterns from huge databases (KDD). The methods of data mining are based on statistical approaches, patterns, and database recognition. All these technologies are instances of artificial intelligence, high-performance computation with parallel visualization.

6. PRIVACY PRESERVATION IN DISTRIBUTED DATA MINING

The Internet has been the most frequent communication channel for people from various groups as well as various business organizations throughout the world. The protection of information and private data is critical for the proper maintenance of every networking system. In today's world, the cellular network is among the most dependable and popular network technologies. Shared networks can handle massive amounts of databases and data sets [14].

In [14]suggested a method for mining route intelligence in cellular networks with high data volumes. This approach enables scalability by lowering sequence data through distributed clustering while maintaining privacy. This methodology aggregates raw data and then uses a statistical survey to collect further privacy

protection and preservation. This is a well-organized planning for preserving privacy while also securing the necessary database in the distributed data mining [17].

One typical strategy for preserving privacy is to provide anonymous IDs to different nodes. Several research publications and researchers have used anonymous Id assignments of various nodes to protect their privacy. They employed an Algorithm in [11] to share private data including anonymous ID assignment among many nodes.
For the sake of privacy and security, they allocated 1 to N ID codes in this study for N nodes.
When data sharing is the duty of more than a node in a network, privacy may be preserved and is more secure. As a result, privacy protection through secured multiparty computing is an excellent technique for data mining and data sharing. While a certain node is in charge of data sharing and data receiving from a adjacent, nodes maintain track of the transferring activities. When a node receives data in the network it uses the Linear approach to be a faithful node. In [15] provide a strategy for preserving privacy in distributed data mining utilizing secured multiparty computing.

In this research, they used the well-known linear methodology to locate friendly nodes for a specific node that receives data in a distributed database sharing network. In today's networking systems, cloud computing gives a better platform for exchanging information and data on a shared platform. Data mining has to gain popularity among academics in contemporary culture as a means of exchanging huge amounts of data [31].

Several approaches have been established to date for the privacy protection of databases in a dispersed network or any network. Several hard computing approaches have a wide range of applications in the protection of privacy. However, soft computing techniques such as artificial neural networks and fuzzy logic may be used in this context to secure data in a network. In several situations, soft computing approaches confirmed to be low-cost models [108].

This approach protects the privacy of people and nodes while without impacting the final output result of the Neural Network. Although data may be lost in this research, privacy is maintained [15]. The protection of one's privacy is a major concern in any networking system. Maintaining privacy and protecting sensitive information in an ad-hoc network is a major challenge. In the case of an ad-hoc network, the contact between nodes is made only when necessary, and it may be lost after the data exchange. As a result, data and information loss may be irreversible once the link is severed [109]. As a result, if the node data and node number are shared with adjacent nodes, it is simple to detect fraudulent activities inside the network. [17] describe an approach to ensure privacy and maintain track of data lost in an ad-hoc network. In this research, they used a mechanism to maintain track of the nodes, and adjacent nodes play an important role in data exchange and privacy preservation.

7. DISTRIBUTED CLUSTERING AND DYNAMIC PARALLEL

Clustering techniques are very attractive for identifying and extracting patterns of interests from datasets. However, their application to very large spatial datasets presents numerous challenges such as high-dimensionality, heterogeneity, and high complexity of some algorithms. Distributed clustering techniques constitute a very good alternative to the Big Data challenges. The standard DP (Dynamic Programming) algorithms are limited by the substantial computational demands they put on contemporary serial computers. PTIMIZATION methods are a constant source of interest when applied to solving problems of a practical type. Minimization of a cost function is fundamental for obtaining an optimal decision policy to achieve the desired specifications. Furthermore, in real problems, it is generally necessary to consider constraints, in order to set limits to state and decision variables. This fact adds a considerable degree of complexity to the problem of optimization. Combinatorial search and optimization techniques are characterized by looking for a solution to a problem among many potential ones. For many search and optimization problems, exhaustive search is unfeasible, so some form of directed search is proposed. In addition, rather than only the best (optimal) solution, a good nonoptimal solution is often sought. Dynamic programming (DP), based on Bellman’s Principle of Optimality, is a classical, powerful, and well-known technique for solving large kinds of optimization problems under very general conditions. There are many and well-known DP applications: laying out circuits in VLSI to minimize the area dedicated to wires, scheduling, string-editing, packaging, inventory management, automatic control, artificial intelligence, economics, etc. However, this

method has not been widely used due to a combinatorial explosion drawback. Although, for some applications DP can be applied analytically, in general, the solution has to be found numerically and the problem of dimension plays a very important role.

The distributed clustering and dynamic parallel model are estimated to address the limitations of existing distributed clustering and parallel techniques while efficiently handling big data concerns. It combines the properties of both hierarchical and partitioning methods, and more significantly, it does not inherit the issue of dividing clustering's set number of partitions, nor the problems of hierarchical clustering's ending conditions. The number of ultimate clusters is dynamically estimated and generated hierarchically. All of these characteristics appear to be highly promising, but some of these have been examined and evaluated on the small scale [21]. distributed dynamic clustering method is separated into two stages:(i) the global, (ii) the local model. Therefore, switching local clusters with nodes in the network will result in considerable overheads and a major slowing of the operation. This is among the most serious issues with the majority in distributed clustering algorithms. To solve this challenge, exchange a minimal number of points between nodes.

Instead of providing all the clusters' data sets, exchange only their sample points, which account for 1.2 percent of the whole dataset size. The form and density of a spatial cluster are the best ways to portray it. The border points of a cluster indicate its form (see Fig. 3).

Many algorithms for identifying cluster boundaries may be discovered in various literature. To produce the borders of the clusters, the shape method uses triangulation. It is a fast algorithm for creating non-convex borders. Distributed clustering and dynamic parallel analyze data in a parallel distributed manner while reducing node communications. As in the current version, the HDFS sends data to various nodes at random. The Hadoop system manages the algorithm's parallel and distributed characteristics.

8. DISCUSSION AND COMPARISON

Each author has a unique opinion but the same goals in creating this article on distributed data mining systems performance. The purpose of applying data mining in distributed systems is to accelerate the storage, processing, analysis, and administration of massive amounts of data. Each author uniquely describes the distributed system; what is data mining? The distributed system architecture, as well as the data mining functions of a distributed system. Furthermore, the data mining performance in distributed systems vs other traditional methodologies. Table 1 demonstrates the functions and techniques of data mining utilized in distributed systems for massive clusters

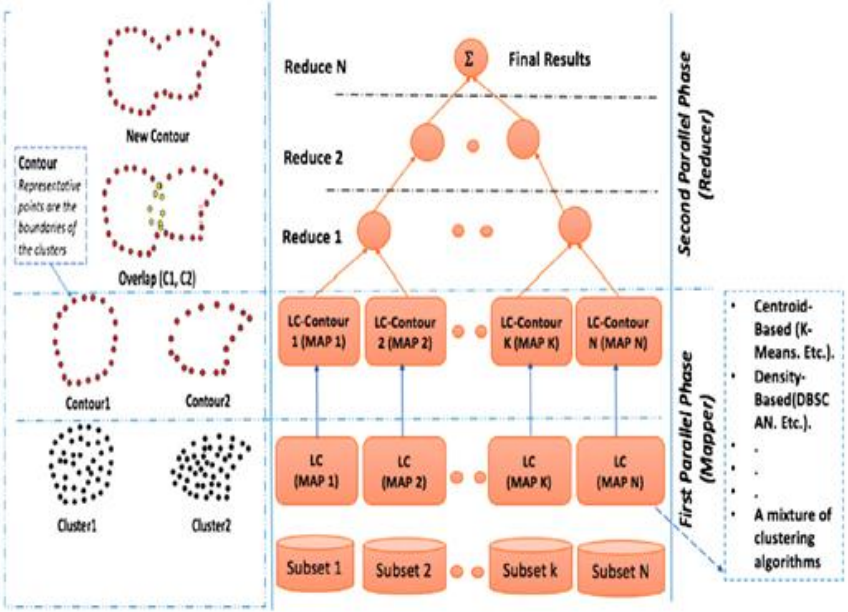

Fig. 3. An overview of the DPDC Approach

Table 1. Performance evaluation in distributed systems

Author(s)	Year	Author's objectives	Description
C. Zhang et al. [100]	2020	Introduce a model of distributed Bayesian matrices decomposition for clustering and big data mining.	One of the main methods for discovering knowledge from huge data created by current applications is matrices decomposition.
F. Liu and W. Xin [92]	2020	Integrates architecture of the crawler system spark based distributed as well as a detailed introduction to distributed framework system	We can overcome the problem of inadequate resource utilization with a crawler system of spark-based distribution.
Chaturvedi et al. [95]	2020	evolution mining and change mining are used to changing distributed systems in this study.	An engineer dealing with evolution or maintenance might benefit from evolvability and changeability analysis.
Y. Shen et al. [94]	2019	implementation and Design of an engineering standardized database system dependent on data mining	To build a public service portal for intelligent construction drawing evaluation.
P. Lekshmy and M. A. Rahiman [15]	2019	It is proposed to develop a unique method-based of Privacy-Preserving of Distributed Data Mining.	The privacy-based objective variable will be constructed to expand on the privacy notion.
I. Anikin and R. Gazimov [17]	2019	Considering the issue of privacy-protected data mining in distributed networks.	Privacy-protection DBSCAN clustering algorithm protects data stored in distributed repositories.
L. Ren and P. A. Ward [104]	2019	Created and validated a Proof of performed Works' mining model.	They proposed and empirically determined the idea of equivalent blocks depending on the Poisson process model.
B. Zhao et al. [99]	2018	Present ZenLDA, an LDA training system with a generalized approach for the distributed data-parallel system.	Upon on distributed data-parallel system, the ZenLDA algorithm achieves scalable and efficient LDA training.
Da Zhang [110]	2018	Formulating a distributed software and hardware architecture to manage massive scholar data.	A system framework for constructing massive scholarly data into large knowledge graphs and discovering relationships between entities.
M. Pulimeno et al. [93]	2018	Describe parallel approaches for mining connected heavy hitters from a double data stream.	Parallel algorithms focused on the passage of the message, common memory, and hybrid mining technique.
S. Patel et al. [98]	2018	Due to computational skew, parallelizing LCM is challenging.	It makes the finding of busy workers easier and gives a technique to identify terminations.
P. Katsikouli et al. [103]	2018	Described a distributed approach for locating popular paths.	Even with a limited application of sensors and a limited user sample, fairly reliable findings may be obtained.
J. Jin et al. [31]	2018	Using HBase, electricity data is stored	How can electrical power data be controlled using HBase, Apache's distributed database

Lichen Zhao et al. [108]	2018	Design a data protection framework for the distribution amongst numerous data owners without any third parties of collaborative data mining.	The solution enables individual data proprietors to preserve their privacy while preserving the predictability of an original model created.
Malika Bendechache et al. [21]	2018	Implement a dynamic parallel approach and distributed clustering	Clustering algorithms are extremely interesting for the identification of data set interest patterns.
M. Idhammad et al. [110]	2018	For cloud environments, a distributed intrusion detecting system is presented	This enables the physical layer's edge network routers to intercept forthcoming network traffic.
E. Trunzer et al. [96]	2017	This work proposes a generic design to overcome the present obstacles.	It makes the integration and aggregation of data also communication across systems easier.
Galina V. Rybina et al. [97]	2017	Models and methods for distributed knowledge acquisitions from databases as alternative knowledge sources.	The issue of the distributed acquisition of information for consistency in knowledge in integrated skilled systems
R. Raman et al. [111]	2017	It is possible to generate implement and rules software agents in such a distributed environment.	Demonstrate the maximum usage of available computing resources for data mining activities in academic institutions.
M. Bendechache [101]	2017	A Dynamic Parallel and Clustering Distributed (DPDC) approach has been developed.	Can analyze huge data and give accurate results within an acceptable reaction time.

9. CONCLUSION

Data mining algorithms often seek to cover desirable patterns or perform classification, clustering, outline detecting, and so on. Generally, data analysis executed applications and collected characteristically distributed in nature. Data mining in distributed computing settings has arisen as an essential research area due to the issues and challenges related to classic data mining techniques while processing distributed data. However, few researchers have synthesized the associated developments in many forms of distributed data mining systems and instead created a broad taxonomy on them. Thus, in this work, we discuss the definitions, general architectures, and some significant properties of a distributed system, before highlighting the challenges of data mining functions in distributed systems. We study current developments in distributed data mining, presenting state-of-the-art information as our main contributions.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Ibrahim BR, Khalifa FM, Zeebaree SR, Othman NA, Alkhayyat A, Zebari RR, et al. "Embedded System for Eye Blink Detection Using Machine Learning Technique," in 2021 1st Babylon International Conference on Information Technology and Science (BICITS), 2021; 58-62.
2. Hasan DA, Zeebaree SR, Sadeeq MA, Shukur HM, Zebari RR, Alkhayyat AH. "Machine Learning-based Diabetic Retinopathy Early Detection and Classification Systems-A Survey," in 2021 1st Babylon International Conference on Information Technology and Science (BICITS), 2021;16-21.
3. Jijo BT, Zeebaree SR, Zebari RR, Sadeeq MA, Sallow AB, Mohsin S, et al. A comprehensive survey of 5G mm-wave technology design challenges. Asian Journal of Research in Computer Science. 2021;1-20.
4. Kareem FQ, Zeebaree SR, Dino HI, Sadeeq MA, Rashid ZN, Hasan DA, et al. A survey of optical fiber communications: challenges and processing time influences.

Asian Journal of Research in Computer Science. 2021;48-58.
5. Abdullah SMSA, Ameen SYA, Sadeeq, S. Zeebaree MA. Multimodal emotion recognition using deep learning. Journal of Applied Science and Technology Trends. 2021; 2:52-58.
6. Sadeeq MA, Zeebaree S. Energy management for internet of things via distributed systems. Journal of Applied Science and Technology Trends. 2021;2:59-71.
7. Omer MA, Zeebaree SR, Sadeeq MA, Salim BW, Mohsin SX, Rashid ZN, et al. "Efficiency of malware detection in android system: A survey," Asian Journal of Research in Computer Science. 2021;59-69.
8. Maulud DH, Zeebaree SR, Jacksi K, Sadeeq MAM, Sharif KH. State of art for semantic analysis of natural language processing," Qubahan Academic Journal. 2021;1:21-28.
9. Sadeeq MM, Abdulkareem NM, Zeebaree SR, Ahmed DM, Sami AS, Zebari RR. IoT and Cloud computing issues, challenges and opportunities: A review," Qubahan Academic Journal. 2021;1:1-7.
10. Abdullah RM, Ameen SY, Ahmed DM, Kak SF, Yasin HM, Ibrahim IM, et al. Paralinguistic Speech Processing: An Overview. Asian Journal of Research in Computer Science. 2021;34-46.
11. Bagavathi A, Mummoju P, Tarnowska K, Tzacheva AA, Ras ZW. Sargs method for distributed actionable pattern mining using spark. In 2017 IEEE international conference on big data (big data), 2017; 4272-4281.
12. He P. An end-to-end log management framework for distributed systems. In 2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS). 2017;266-267.
13. Lv Z, Deng W, Zhang Z, Guo N, Yan G. A Data Fusion and Data Cleaning System for Smart Grids Big Data," in 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), 2019;802-807.
14. Swain DK, Mishra S, Rout SB. Privacy preservation in distributed data mining for protein secondary structure prediction. In 2017 2nd International Conference on Communication and Electronics Systems (ICCES). 2017;122-127.
15. Lekshmy P, Rahiman MA. A sanitization approach for privacy preserving data mining on social distributed environment," Journal of Ambient Intelligence and Humanized Computing. 2020;11:2761-2777, 2020.
16. Ibrahim IM, Ameen SY., Yasin H. M., Omar N, Kak SF, Rashid ZN, et al., "Web Server Performance Improvement Using Dynamic Load Balancing Techniques: A Review," Asian Journal of Research in Computer Science. 2021;47-62.
17. Anikin I, Gazimov R. Approach to Privacy Preserved Data Mining in Distributed Systems," in 2019 International Russian Automation Conference (RusAutoCon), 2019;1-5.
18. Ahmed DM, Ameen SY, Omar N, Kak SF, Rashid ZN, Yasin HM, et al. A State of Art for Survey of Combined Iris and Fingerprint Recognition Systems.Asian Journal of Research in Computer Science. 2021;18-33.
19. Maulud DH, Ameen SY, Omar N, Kak SF, Rashid ZN, Yasin HM, et al. Review on Natural Language Processing Based on Different Techniques," Asian Journal of Research in Computer Science. 2021;1-17.
20. Salih AA, Ameen SY, Zeebaree SR, Sadeeq MA, Kak SF, Omar N, et al. Deep Learning Approaches for Intrusion Detection. Asian Journal of Research in Computer Science. 2021;50-64.
21. Bendechache M, Tari A-K, Kechadi M-T, Parallel and distributed clustering framework for big spatial data mining," International Journal of Parallel, Emergent and Distributed Systems. 2019;34:671-689.
22. Hassan RJ, Zeebaree SR, Ameen SY, Kak SF, Sadeeq MA, Ageed ZS, et al. State of art survey for iot effects on smart city technology: challenges, opportunities, and solutions. Asian Journal of Research in Computer Science. 2021;32-48.
23. Triantafillou P. Towards intelligent distributed data systems for scalable efficient and accurate analytics," in 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS), 2018;1192-1202.
24. Yahia HS, Zeebaree SR, Sadeeq MA, Salim NO, Kak SF, Adel A-Z, et al.

Comprehensive survey for cloud computing based nature-inspired algorithms optimization scheduling. Asian Journal of Research in Computer Science. 2021;1-16.
25. Ageed ZS, S. R. Zeebaree, M. M. Sadeeq, S. F. Kak, Z. N. Rashid, A. A. Salih, et al., "A survey of data mining implementation in smart city applications," Qubahan Academic Journal, vol. 1, pp. 91-99, 2021.
26. Ageed ZS, Zeebaree SR, Sadeeq MA, Abdulrazzaq MB, Salim BW, Salih AA, et al. A state of art survey for intelligent energy monitoring systems. Asian Journal of Research in Computer Science. 2021; 46-61.
27. Anikin IV, Gazimov RM. Privacy preserving DBSCAN clustering algorithm for vertically partitioned data in distributed systems," in 2017 International Siberian Conference on Control and Communications (SIBCON), 2017;1-4.
28. Salih A, Zeebaree ST, Ameen S, Alkhyyat A, Shukur HM. A Survey on the Role of Artificial Intelligence, Machine Learning and Deep Learning for Cybersecurity Attack Detection. in 2021 7th International Engineering Conference “Research & Innovation amid Global Pandemic"(IEC), 2021;61-66.
29. Abdullah DM, Ameen SY, Omar N, Salih AA, Ahmed DM, Kak SF, et al. Secure data transfer over internet using image steganography. Asian Journal of Research in Computer Science. 2021;33-52.
30. Kareem FQ, Ameen SY, Salih AA, Ahmed DM, Kak SF, Yasin HM, et al. SQL injection attacks prevention system technology. Asian Journal of Research in Computer Science. 2021;13-32.
31. Jin J, Song A, Gong H, Xue Y, Du M, Dong F, et al. Distributed storage system for electric power data based on hbase," Big Data Mining and Analytics. 2018;1:324-334.
32. Ismael HR, Ameen SY, Kak SF, Yasin H. M, Ibrahim IM, Ahmed AM, et al. Reliable communications for vehicular networks," Asian Journal of Research in Computer Science. 2021;33-49.
33. Abdulla AI, Abdulraheem AS, Salih AA, Sadeeq M, Ahmed AJ, Ferzor BM, et al. Internet of things and smart home security. Technol. Rep. Kansai Univ, 2020;62:2465-2476.
34. Abdulraheem AS, Salih AA, Abdulla AI, Sadeeq M, Salim N, Abdullah H, et al. Home automation system based on IoT; 2020.
35. Harikrishnasairaj K, Prasad VK. Secure frequent itemset mining from horizontally distributed databases. In 2017 International Conference on Intelligent Computing and Control (I2C2). 2017;1-4.
36. Agrawal N, Kaur A. An algorithmic approach for text recognition from printed/typed text images. in 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2018;876-879.
37. Lin C-Y, Lin Y-C. An overall approach to achieve load balancing for Hadoop Distributed File System. International Journal of Web and Grid Services. 2017; 13:448-466.
38. Salih AA, Zeebaree S, Abdulraheem AS, Zebari RR, Sadeeq M, Ahmed OM. Evolution of mobile wireless communication to 5G revolution. Technology Reports of Kansai University, 62:2139-2151.
39. Dino HI, Zeebaree S, Salih AA, Zebari RR, Ageed ZS, Shukur HM, et al. Impact of Process Execution and Physical Memory-Spaces on OS Performance," Technology Reports of Kansai University. 2020; 62:2391-2401, 2020.
40. Hamdi SJ, Ibrahim IM, Omar N, Ahmed OM, Rashid ZN, Ahmed AM, et al. A Comprehensive Study of Malware Detection in Android Operating Systems.
41. Ageed ZS, Ahmed AM, Omar N, Kak SF, IM. Ibrahim HM. Yasin, et al. A State of Art Survey of Nano Technology: Implementation, Challenges, and Future Trends.
42. Abdulqadir MM, Salih AA, Ahmed OM, Hasan DA, Haji LM, Ahmed SH, et al. A Comprehensive Study of Caching Effects on Fog Computing Performance.
43. Yazdeen AA, SR, Zeebaree MM. Sadeeq SF, Kak OM, Ahmed, Zebari RR. FPGA implementations for data encryption and decryption via concurrent and parallel computation: A review. Qubahan Academic Journal. 2021;1:8-16.
44. Ageed ZS, Zeebaree SR, Sadeeq MM, Kak SF, Yahia HS, Mahmood MR, et al. Comprehensive survey of big data mining approaches in cloud systems," Qubahan Academic Journal. 2021;1:29-38.
45. Abdulrahman LM, Zeebaree SR, Kak SF, Sadeeq MA, Adel A-Z, Salim BW, et al. A state of art for smart gateways issues and

modification. Asian Journal of Research in Computer Science. 2021;1-13.
46. Abdulqadir HR, Zeebaree SR, Shukur HM, Sadeeq MM, Salim BW, Salih AA, et al., A study of moving from cloud computing to fog computing, Qubahan Academic Journal. 2021;1: 60-70.
47. AL-Zebari A, Zeebaree S, Jacksi K, Selamat A. ELMS–DPU ontology visualization with Protégé VOWL and Web VOWL," Journal of Advanced Research in Dynamic and Control Systems, 2019; 11:478-85.
48. Zeebaree A, Adel A, Jacksi K., and Selamat A. Designing an ontology of E-learning system for duhok polytechnic university using protégé OWL tool," J Adv Res Dyn Control Syst. 2019;11:24-37, 2019.
49. Adel A-Z, Zebari S, Jacksi K. Football Ontology Construction using Oriented Programming," Journal of Applied Science and Technology Trends. 2020;1:24-30.
50. SA, A-Z, Selamat A. Electronic Learning Management System Based on Semantic Web Technology: A Review," Int. J. Adv. Electron. Comput. Sci. 2017;4:1-6.
51. Abdullah RM, Abdulazeez AM, and A. Al-Zebari. Machine learning Algorithm of Intrusion Detection System. Asian Journal of Research in Computer Science, pp. 1-12, 2021.
52. Shukur H, Zeebaree SR, Ahmed AJ, Zebari RR, Ahmed O, Tahir BSA, et al. A state of art survey for concurrent computation and clustering of parallel computing for distributed systems," Journal of Applied Science and Technology Trends, 2020;1:148-154.
53. Tahir B, Ali J, Saktioto M, Fadhali R, Rahman, Ahmed A. A study of FBG sensor and electrical strain gauge for strain measurements," Journal of optoelectronics and advanced materials. 2008;10:2564-2568.
54. Harki N, Ahmed A, Haji L. CPU scheduling techniques: A review on novel approaches strategy and performance assessment. Journal of Applied Science and Technology Trends. 2020;1:48-55.
55. Ahmed A, Ahmed O. Correlation pattern among morphological and biochemical traits in relation to tillering capacity in sugarcane (Saccharum Spp). Acad J Plant Sci. 2012;5:119-122.
56. Ahmed AJ, Mohammed FH, Majedkan NA. An Evaluation Study of an E-Learning Course at the Duhok Polytechnic University: A Case Study," Journal of Cases on Information Technology (JCIT), 2022;24:1-11.
57. Ahmed O, Geraldes R, Ahmed A., DeLuca G, Palace J. Multiple sclerosis and the risk of venous thrombosis: a systematic review," in MULTIPLE SCLEROSIS JOURNAL, 2017;757-758.
58. Salim NO, Abdulazeez AM. Human diseases detection based on machine learning algorithms: A review," International Journal of Science and Business. 2021; 5:102-113.
59. Salim NO, Zeebaree SR, Sadeeq MA, Radie A, Shukur HM, Rashid ZN. Study for Food Recognition System Using Deep Learning. In Journal of Physics: Conference Series, 2021, p. 012014.
60. Salim NO, Abdulazeez AM. Science and Business," International Journal, 5;102-113.
61. Eesa AS, Sadiq S, HASSAN M, Orman Z. Rule generation based on modified cuttlefish algorithm for intrusion detection system," Uludağ University Journal of The Faculty of Engineering, vol. 26, pp. 253-268, 2021.
62. Eesa AS. Optimization Algorithms For Intrusion Detection System: A Review. International Journal of Research-Granthaalayah,2020; 8:217-225.
63. Haji SH, Zeebaree SR, Saeed RH, Ameen SY, Shukur HM, Omar N, et al., "Comparison of software defined networking with traditional networking," Asian Journal of Research in Computer Science, 2021;1-18.
64. Zebari S, Yaseen NO. Effects of parallel processing implementation on balanced load-division depending on distributed memory systems," J. Univ. Anbar Pure Sci. 2011;5:50-56.
65. Malallah H, Zeebaree SR, Zebari RR, Sadeeq MA, Ageed ZS, Ibrahim IM, et al. "A comprehensive study of kernel (issues and concepts) in different operating systems," Asian Journal of Research in Computer Science. 2021;16-31.
66. Yasin HM, Zeebaree SR, Sadeeq M. A., Ameen S. Y., Ibrahim I. M., Zebari R. R., et al. IoT and ICT based smart water management, monitoring and controlling system: A review," Asian Journal of Research in Computer Science. 2021;42-56.

67. Ibrahim I. M., Task scheduling algorithms in cloud computing: A review," Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021;12:1041-1053.
68. Zebari I. M., S. R. Zeebaree, and H. M. Yasin, "Real time video streaming from multi-source using client-server for video distribution," in 2019 4th Scientific International Conference Najaf (SICN), 2019;109-114.
69. Yasin H. M., S. R. Zeebaree, and I. M. Zebari, "Arduino based automatic irrigation system: Monitoring and SMS controlling," in 2019 4th Scientific International Conference Najaf (SICN), 2019;109-114.
70. Zeebaree S, Yasin H. M. Arduino based remote controlling for home: power saving, security and protection," International Journal of Scientific & Engineering Research, 2014;5:266-272.
71. Hasan DA, B. K. Hussan, S. R. Zeebaree, D. M. Ahmed, O. S. Kareem, and M. A. Sadeeq, "The impact of test case generation methods on the software performance: A review," International Journal of Science and Business, 2021;5:33-44.
72. Jacksi K, R. K. Ibrahim, S. R. Zeebaree, R. R. Zebari, and M. A. Sadeeq, "Clustering documents based on semantic similarity using HAC and K-mean algorithms," in 2020 International Conference on Advanced Science and Engineering (ICOASE), 2020;205-210.
73. Sadeeq MA, A. M. Abdulazeez, "Neural networks architectures design, and applications: A review," in 2020 International Conference on Advanced Science and Engineering (ICOASE), 2020; 199-204.
74. S. Zeebaree and I. Zebari, "Multilevel client/server peer-to-peer video broadcasting system," International Journal of Scientific & Engineering Research. 2014;5: 260-265.
75. Ageed Z. S., R. K. Ibrahim, and M. Sadeeq, "Unified ontology implementation of cloud computing for distributed systems," Current Journal of Applied Science and Technology, 2020;82-97.
76. Zeebaree S., S. Ameen, and M. Sadeeq, "Social media networks security threats, risks and recommendation: A case study in the kurdistan region," International Journal of Innovation, Creativity and Change 2020; 13:349-365.
77. Zebari D. A., H. Haron, S. R. Zeebaree, and D. Q. Zeebaree, "Multi-Level of DNA Encryption Technique Based on DNA Arithmetic and Biological Operations," in 2018 International Conference on Advanced Science and Engineering (ICOASE), 2018, pp. 312-317.
78. Sulaiman M. A., M. Sadeeq, A. S. Abdulraheem, and A. I. Abdulla, "Analyzation study for gamification examination fields," Technol. Rep. Kansai Univ, vol. 62, pp. 2319-2328, 2020.
79. Zeebaree S. R., A. B. Sallow, B. K. Hussan, and S. M. Ali, "Design and simulation of high-speed parallel/sequential simplified DES code breaking based on FPGA," in 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019;76-81.
80. Sadeeq M., A. I. Abdulla, A. S. Abdulraheem, and Z. S. Ageed, "Impact of electronic commerce on enterprise business," Technol. Rep. Kansai Univ. 2020; 62: 2365-2378.
81. Alzakholi O., H. Shukur, R. Zebari, S. Abas, and M. Sadeeq, "Comparison among cloud technologies and cloud performance," Journal of Applied Science and Technology Trends, 2020;1:40-47.
82. Ageed Z., M. R. Mahmood, M. Sadeeq, M. B. Abdulrazzaq, and H. Dino, "Cloud computing resources impacts on heavy-load parallel processing approaches," IOSR Journal of Computer Engineering (IOSR-JCE), 2020;22:30-41.
83. Ibrahim BR, SR. Zeebaree, and B. K. Hussan, "Performance Measurement for Distributed Systems using 2TA and 3TA based on OPNET Principles," Science Journal of University of Zakho, 2019;7:65-69.
84. Sallow A., S. Zeebaree, R. Zebari, M. Mahmood, M. Abdulrazzaq, and M. Sadeeq, "Vaccine tracker," SMS reminder system: Design and implementation; 2020.
85. Sadeeq M. A., S. R. Zeebaree, R. Qashi, S. H. Ahmed, and K. Jacksi, "Internet of Things security: a survey," in 2018 International Conference on Advanced Science and Engineering (ICOASE). 2018; 162-166.
86. Abdulazeez A. M., S. R. Zeebaree, and M. A. Sadeeq, "Design and implementation of electronic student affairs system," Academic Journal of Nawroz University, 2018;7: 66-73.

87. Zeebaree S, R. R. Zebari, K. Jacksi, and D. A. Hasan, "Security approaches for integrated enterprise systems performance: A Review," Int. J. Sci. Technol. Res, 2019;8.
88. Sallow AB, M. Sadeeq, RR. Zebari, M. B. Abdulrazzaq, M. R. Mahmood, H. M. Shukur, et al., "An investigation for mobile malware behavioral and detection techniques based on android platform," IOSR Journal of Computer Engineering (IOSR-JCE), 2020;22:14-20.
89. Dino H, M. B. Abdulrazzaq, S. Zeebaree, A. B. Sallow, R. R. Zebari, H. M. Shukur, et al., "Facial expression recognition based on hybrid feature extraction techniques with different classifiers," TEST Engineering & Management. 2020;83: 22319-22329.
90. Zeebaree S, R. R. Zebari, and K. Jacksi, "Performance analysis of IIS10. 0 and Apache2 Cluster-based Web Servers under SYN DDoS Attack," TEST Engineering & Management, 2020;83: 5854-5863.
91. Jader OH, S. Zeebaree, and R. R. Zebari, "A state of art survey for web server performance measurement and load balancing mechanisms," International Journal of Scientific & Technology Research,2019;8:535-543.
92. Zhang D, Kabuka MR. Distributed relationship mining over big scholar data," IEEE Transactions on Emerging Topics in Computing, 2018; 9:354-365.
93. Pulimeno M., I. Epicoco, M. Cafaro, C. Melle, and G. Aloisio, "Parallel mining of correlated heavy hitters on distributed and shared-memory architectures," in 2018 ieee international conference on big data (big data), 2018;5111-5118.
94. Shen Y., M. Wang, H. Zhou, Q. Zhu, S. Ma, M. Cao, et al., "Design and Implementation of Engineering Standard Database System Based on Data Mining," in 2019 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), 2019;124-127.
95. Liu F, Xin W. Implementation of Distributed Crawler System Based on Spark for Massive Data Mining. in 2020 5th International Conference on Computer and Communication Systems (ICCCS), 2020;482-485.
96. Trunzer E, I. Kirchen, J. Folmer, G. Koltun, and B. Vogel-Heuser, "A flexible architecture for data mining from heterogeneous data sources in automated production systems," in 2017 IEEE International Conference on Industrial Technology (ICIT), 2017;1106-1111.
97. Rybina GV, Y. M. Blokhin, and E. S. Sergienko, "Distributed knowledge acquisition basing on integration of Data Mining and Text Mining methods and their usage with AT-TECHNOLOGY workbench," in 2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), 2017;1-6.
98. Patel S, Agarwal U, Kailasam S. A Dynamic Load Balancing Scheme for Distributed Formal Concept Analysis, in 2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS), 2018;489-496.
99. Zhao B, Zhou H, Li G, Huang Y. ZenLDA: Large-scale topic model training on distributed data-parallel platform," Big Data Mining and Analytics, 2018;1:57-74.
100. Zhang C, Yang Y, Zhang W, Zhang S, "Distributed Bayesian Matrix Decomposition for Big Data Mining and Clustering," arXiv preprint arXiv:2002.03703, 2020.
101. Bendechache M, Le-Khac N-A, Kechadi M-T. "Performance evaluation of a distributed clustering approach for spatial datasets," in Australasian Conference on Data Mining, 2017;38-56.
102. Chaturvedi A, Tiwari A, Binkley D, Chaturvedi S. Service evolution analytics: change and evolution mining of a distributed system," IEEE Transactions on Engineering Management. 2020;68:137-148.
103. Katsikouli P, Astefanoaei MS, Sarkar R, "Distributed mining of popular paths in road networks," in 2018 14th International Conference on Distributed Computing in Sensor Systems (DCOSS), 2018;1-8.
104. Ren L, Ward PA. Pooled mining is driving blockchains toward centralized systems," in 2019 38th International Symposium on Reliable Distributed Systems Workshops (SRDSW), 2019;43-48.
105. Carrillo GE, Abad, CL. Inferring Workflows with Job Dependencies from Distributed Processing Systems Logs," in 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on

Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2017;1025-1030.

106. Mole PV. Empirical Research on the Challenges of Distributed Databases: A Literature Review."
107. Rana MS, Sohel MK, Arman MS. "Distributed Database Problems Approaches and Solutions-A Study," in International Journal of Machine Learning and Computing (IJMLC), ed; 2018.
108. Zhao L, Ni L, Hu S, Chen Y, Zhou P, Xiao FF, et al. "Inprivate digging: Enabling tree-based distributed data mining with differential privacy," in IEEE INFOCOM 2018-IEEE Conference on Computer Communications, 2018;2087-2095.
109. Rathnayake RS, Poravi G. Review on Textual Data Mining for Reviewer Recommendation in Pull-Based Distributed Software Development," in 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019;1-5.
110. Idhammad M, Afdel K, Belouch M. "Distributed intrusion detection system for cloud environments based on data mining techniques," Procedia Computer Science, 2018;127:35-41.
111. Raman R, Vadivel S, Raj BE. A framework for cost-effective distributed data mining in academic institutions using intelligent agents," in 2017 Fourth HCT Information Technology Trends (ITT), 2017;13-18.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle4.com/review-history/73759