Scaling Effectivity of Research Contributions in Distributed Data mining over Grid Infrastructures

Shahina Parveen M.
Research Scholar
Dept of CS&E.,
JNTU, Hyderabad, India

G. Narsimha, PhD
Associate Professor
JNTUH, Kondagattu, Jagtial, Karimnagar,
Telangana, India

ABSTRACT

With the increasing need of data availability and cloud-based services, distributed database management has already gained a maximum momentum in technological advancement. With the data stored in distributed manner, performing distributed data mining is encountering challenges especially when the data is real-time, non-static, highly heterogeneous, unstructured, etc. Usually, such forms of distributed data management are only effective if it is managed over grid infrastructure, which offers a suitable arena to the technology to provide better performance. However, research work considering distributed data mining over grid interface has not been nurtured to the best point in the research community as compared to conventional data mining approaches. This manuscript discusses the research trends and its effectiveness about the techniques and adoptability of the prior literature towards distributed data mining.

Keywords

Distributed Data Mining, Grid Infrastructure, Mining, Cloud

1. INTRODUCTION

With the advancement of mobile internet and telecommunication, there is a drastic revolution in the communication system [1]. This advancement in communication technology has truly made the world a global village and has affected almost in any sector viz. Education, healthcare, industries, enterprises, defense, meteorology, social network, entertainment, etc. 30-40 years back, the data used to store in a standalone physical server, where a typical server-client relationship was used. However, with the advancement of database management system and ubiquitous computing, cloud computing has solved this problem [2]. Using cloud, the data is made available to the user at any point of time and place. Unfortunately, cloud offers a best and cost effective storage for any size of data, but question lies here – what to do with this data. The best answer to this question is data mining, which is a technique to extract a unique pattern of information hidden in the massive set of data [3]. Conventional data mining problem is also associated with some of the challenges e.g. clustering, classification, prediction, learning techniques, etc. [4]. Such problems also lie in any sophisticated machine learning approach too. The process of conventional data mining technique initiates by exploring the source data and then extracting the data points, which are required to be evaluated and analyzed. After pulling out the pertinent data, the significant step in data mining is to find out the key value from the already pulled out set of data. The final step is to perform interpretation of the data [5]. But the technology of the data management was subjected to certain amendments based on dynamic needs of our customers and users. The customer demands higher availability of data, virtual platforms, no downtime, no delay, better throughout, etc. This could be only achieved if the data are stored and retrieved in a highly distributed manner [6]. Now here comes the most confusing part to manage such forms of data i.e. distributed computing or grid computing. Interestingly distributed computing is all about managing more number of machines with lower computational capability. However, grid computing does the same thing with additional capabilities e.g. exploiting resource utilization of the heterogeneous system, manages workloads, etc. The positive fact about grid infrastructure is its capability to execute on numerous domains of administration and is more inclined towards optimization technique that is missing is distributed computing [7]. Hence, there is always a difference between carrying out data mining operation using distributed computing or grid computing, whereas the suitability of grid infrastructure holds more appropriate for distributed data mining. The input to distributed data mining is various forms of warehouses that already have historical data. Now re-performing mining operation on them is equivalent to performing optimization of the existing data mining technique to make it represent more like the distributed. However, the process is not that easy as it seems like. At present, the morphology of the data has entirely taken a shape of heterogeneity, unstructured, semi-structured, high-dimensional, etc. These all cases of complex and massive streams of data render conventional data mining technique ineffective. The problem thereby becomes stronger when it comes to grid infrastructure. The area of data mining has already gained enough paces in the area of research; however, distributed data mining over grid infrastructure is one of the less visited topics among the research communities globally. This paper discusses the effectiveness of the research contribution in last five years. Section 2 discusses the essential characteristics of grid infrastructure followed by a discussion of frequently used frameworks on grid infrastructure in Section 3. The discussion of data mining in the data grid is carried out in Section 4. Section 5 discusses existing review paper and scales its effectiveness followed by discussion for existing research contributions in Section 6. Discussion of research gap is made in Section 7, and Section 8 summarizes the entire review work.

2. INTRODUCTION TO GRID INFRASTRUCTURE

There are various fields of applications in the present day where a massive amount of data is being generated. The interesting point is such applications are capable of generating data from multiple distributed sources positioned in different geographical points. Some of such applications are meteorology, astronomy, and computational genomics.

Although we have different storage repositories, the biggest challenge in this point is to perform an analysis. To solve these issues of storage and retrieval, modern science has provided us with grid infrastructure. A grid infrastructure can be defined as an architecture that provides the user with a potential capability of accessing, editing, and transmitting a massive size of distributed data to be utilized in analysis (or research) purpose. The operation of such complicated process is assisted by middleware services and applications that are responsible for pulling the geographically distanced and distributed data as well as valuable resources against a typical query request of the user on the other end. The possibility of data localization could both on single or multiple places with its specific domain of administration and security protocols. To provide better data availability, the data grid system also makes numerous replicas of the data to distribute in multiple storage locations over the grid. The middleware system in grid infrastructure is responsible for carrying out integration between the data and the user. At present, the topology of the grid infrastructure supports federated topology, hybrid topology, hierarchical topology, and Monadic topology [8].

A grid infrastructure facilitates the user with a sophisticated architecture for storing massive distributed data as well as architecture to operate it. The prime responsibility of such grid architecture will be:

1. To visualize a problem regarding a design for enhancing the performance of a system controlled by the user.
2. To facilitate a model for the encapsulating low-level process of repositing, data transmission, etc.
3. To facilitate a proper management of resources, access policies, security, etc.

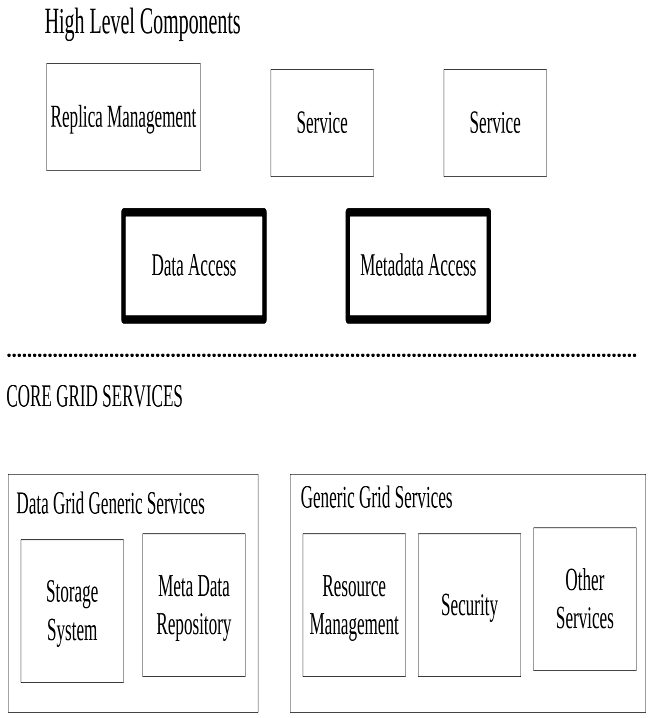

Fig 1: Conventional Architecture of Grid Infrastructure

The main services of the grid are used for facilitating the fundamental process of managing resources as well as security. The top of the main grid service layer consists of high-level components, which is responsible for management and selection of replica. The prime feature of the high level component is to consider Meta data as well as data access management as the primary essential services required for maintaining grid infrastructure. The data access component is responsible for managing data transfer to external entities as well as accessing data. On the other hand, the services about metadata are responsible for managing the information related to the data about its origination, about its usage; about its mapping characteristics of file instances towards reposit location. Apart from this, grid infrastructure also performs functioning of authentication/authorization of the grid infrastructure. The major focus is on the security. The biggest beneficial point of grid infrastructure is the extensibility of the design, which means that the architecture can be highly customized based on the original requirements of the enterprise.

It consists of Automatic Storage Management (ASM) which is responsible for controlling the volume of the system as well as the file system for managing database files of Oracle. Usually, the database files support normal database of Oracle as well as database about configuration of Real Application Clusters (RAC). There is another functionality of Oracle Grid called as Oracle Restart which enhances the data performance while storing and retrieving. It starts in providing data management services even in the worst case of system failure. Fig.2 showcases the standard architecture of Oracle Grid Infrastructure. It allows a client to store and retrieve the data in highly distributed fashion [9].

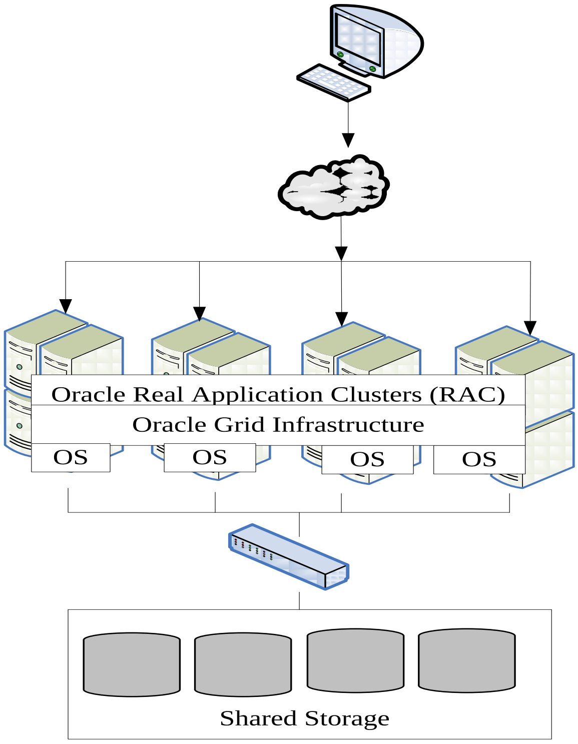
Oracle Real Application Clusters (RAC)
Oracle Grid Infrastructure
OS
OS
OS
OS
Shared Storage

Fig 2: Oracle Grid Infrastructures

At present European Union has a funded project by the name Data Grid [10]. The prime purpose of this funded project is to develop a next generation infrastructure for computing with sophisticated data. It also targets to facilitate cost-effective computation, developing high-level analytics for large-scale and high-dimensional distributed data with a size of petabytes.

3. FRAMEWORK OF GRID INFRASTRUCTURE

The basic motive of grid infrastructure is to perform integration of various sophisticated resources along with its operating system, software, hardware, encryption protocols, etc. These all supportability is given by existing frameworks or architectures of grid infrastructure. The standard supportability of the architectures for grid infrastructure is discussed below:

3.1 Open Grid Service Architecture (OGSA)

OGSA is used for scientific and the business utilities. It is a form of open source service-oriented architecture specifically designed for grid environments. OGSA assists in establishing multiple for of interactions as well as interoperability of distributed nature to establish internal communication among the resources. The design of OGSA is carried out over Simple Object Access Protocol and Web Services Description Language, which is commonly used in the commercial market. OGSA primarily targets to free itself from any dependency toward data management from transport level. It can be said that OGSA is an advanced framework to support architecture of web services. OGSA can offer substantial benefits of data services, encryption, the operation of management services, information services, resource management, etc.

3.2 Open Grid Services Infrastructure (OGSI)

The similar community who framed OGSA has also developed OGSI i.e. Global Grid Forum. It is responsible for adding up the infrastructure layer to the OGSA in grid computing. It also extends the web services for obliging the resources for grid computing by using essential components of web services model for representing an integration of numerous protocols, messaging formats, encoding patterns, etc.

3.3 Grid FTP

Grid FTP is another standard framework from Globus toolkit that offers a potential protocol for safer and reliable transmission of data for a communication network with high channel capacity. Designed on the concept of data transmission over FTP, it is considered as most reputed protocol over grid environment. One of the essential characteristics of Grid FTP is the usage of higher channel capacity accomplished by using several streams of TCP. Another interesting feature of Grid FTP is its supportability of transfer of the partial file over the grid and thereby overcome the issue of unavailability of network or services.

3.4 Open Grid Service Architecture-Data Access & Integration (OGSA-DAI)

It is of the frequently used framework for managing distributed data access that significantly permits federation of data resources along with accessibility using web services over grid infrastructure. These web services can be used for updating process, querying, transforming, and integrating the data. It overcomes the problem of data location (which plays a critical role in cloud data center), data transfer from the multiple sites. The applications of OGSA-DAI can be seen on various sectors e.g. geographical information system, astronomy, medical research, transport, meteorology, computer-aided design, etc. The standard architecture of OGSA-DAI can be seen in Fig.3.

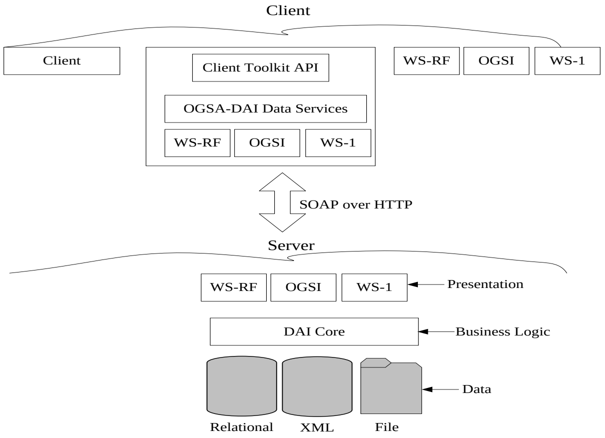

Fig 3: Standard Architecture of OGSA-DAI

Data layer is the bottom later of OGSA-DAI that makes use of conventional database system e.g. XML, SQL Server, Oracle, etc. Business layer interface is responsible for establishing interaction between the data layer and business layer. The business logic layer is responsible for managing queries, data transformation, delivery management, as well as management of data streams. In the above figure, the business logic is referred to as DAI core. The presentation layer performs communication between business and presentation layer. The presentation layer performs encapsulation of functionality associated with using OGSA-DAI. The best part of this framework is its extensive supportability to accessing from any client that complies with OGSI and web service resource framework. It also depends on the type of the server where the presentation layer of OGSA-DAI resides. It also facilitates different forms of client toolkit that can furnish the superior level of interaction with services of OGSA-DAI. However, supportability of the various versions is still a problem.

4. DATAMINING IN DATA GRID

The massiveness of the data over the distributed grids is of no use if it is not subjected to data analysis or data mining technique. Data mining is a technique that applies the analytical concept to extract certain hidden and unique patterns in the massive dataset called as knowledge.

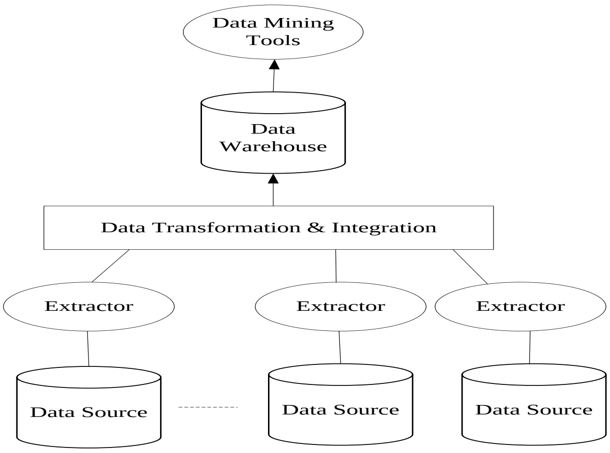

Fig 4: Data Mining Process

However, data mining process differs from distributed data mining process. Fig.4 shows the conventional data mining process, where the first step is to extract the unique data from the data source that is again subjected to data transformation and integration process. The processed data is stored in another repository called as the data warehouse, which is again subjected to various data mining approaches or tools (based on end-user application). Distributed data mining differs in the process of storage. The first step is to access the data reposits (data warehouse) from multiple geographical location, which can be subjected to homogeneous or heterogeneous data mining algorithms to extract knowledge. The extracted knowledge is then subjected to a local model of knowledge discovery, where the outcomes are further aggregated to accomplish the finally extracted knowledge.

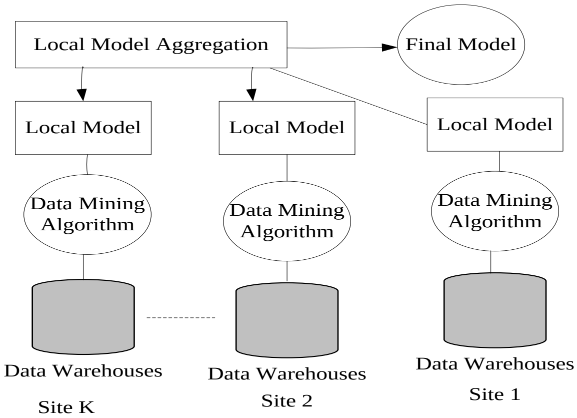

Fig 5: Distributed Data Mining Process

It should be understood that both knowledge discovery process and its algorithms are bit computationally intensive, which means the demands of computational capability should be extremely high. The good news is grid computing facilitates such computational and data-intensive requirements of processing data mining algorithms of grid infrastructure. Therefore, distributed data mining approaches will permit the organization or any specific user to analyze their data in most cost-efficient manner.

4.1 Benefits of Distributed Data mining over Grid

The significant benefits of encouraging the usage of distributed data mining techniques will be quite outstanding if implicated in reality. The design of data mining algorithms over distributed grid system will facilitate various mechanism and tools for aiding analytical process, a system to infer knowledge outcomes, exploration of the complex and distributed nature of data for research purpose. An effective design of the superior performance distributed data mining algorithm requires an effective design of computational grid as well as knowledge grid to facilitate a precise process of knowledge discovery. Therefore, adoption of distributed data mining approach will extensively enhance the computational capability required for catering up the challenging issues of dynamic queries. The concept can meet the demands of the industry as well as the user to process the massive size of the data that are stored in a distributive manner thereby assisting the corporate for faster knowledge delivery process. The layered-form of grid architecture can be benefitted by the design of the distributed data mining that enables the lower level to facilitate further supportability of middleware services.

4.2 Challenges in Distributed Data mining over Grid

The complications of the data mining process are directly proportional to the dimensionality of the data. Higher the dimensionality of the data bigger is the complexity in analyzing data. One of the biggest challenges in implementing distributed data mining over the grid is to design an integrated hypothesis of data mining algorithm. As the sources and types of data warehouse differs, so it is quite common the data mining algorithms will be majorly heterogeneous in nature over different data warehouse. This results in generations of knowledge to local model with a difference in its values. Exploring the error in the knowledge generated from local data mining algorithm from aggregation view is one of the most challenging tasks over grid infrastructure. Another significant problem is the streaming of high-speed data resulting to complex high dimensional data. Problems also lie in extracting knowledge from time-series data, sequenced data, unstructured data, semi-structured data. As the data is generated from multiple data warehouses, the security protocols may be different on specific warehouses that significant result in a delay in authentication and authorization problems for real-time data. Heterogeneous nature of data even from the same source will pose a significant problem in implementing distributed data mining. Although distributed data mining process has the significant potential to perform knowledge discovery over a grid, it is not that easy to accomplish a precise amount of knowledge is a lesser duration of time with high data quality.

4.3 Tools used in Distributed Data mining over Grid

There are various tools for performing data mining, but this section will explicitly discuss the tools used for analyzing or implementing distributed data mining over Grid.

4.3.1 Weka4WS

It is one of the open source tools for applying distributed data mining techniques over grid infrastructure. It does so be using standard Weka library in the form of web services resource framework. The standard architecture of Weka4WS consists of three nodes viz. i) storage node, ii) compute node, and iii) user node (Fig.6). The data to be subjected for mining operation are kept in the storage node while the algorithm for knowledge discovery runs on compute node.

The local system used by the client to make a query is called as user node. The task of the local extraction of knowledge is carried out at grid node with an aid of Weka library. The tool also assists in remote extraction of knowledge. Weka4WS also consists of Grid FTP that allows transmission of distributed data.

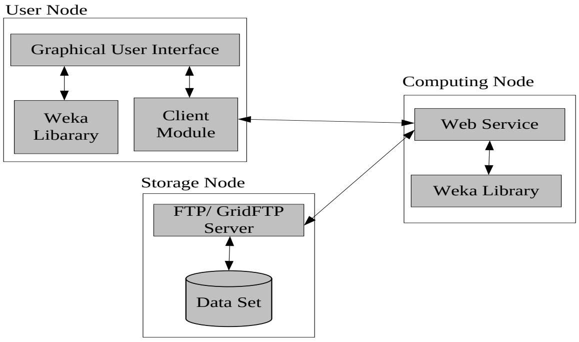

Fig 6: Principle of Weka4WS

4.3.2 GridWeka2

It is an enhancement of conventional data mining tool Weka. GridWeka2 posse's two significant components e.g. The Server and client. The design principle of server is completely based on conventional Weka whereas client is designed for providing necessary input and carrying out the task related to data distribution.

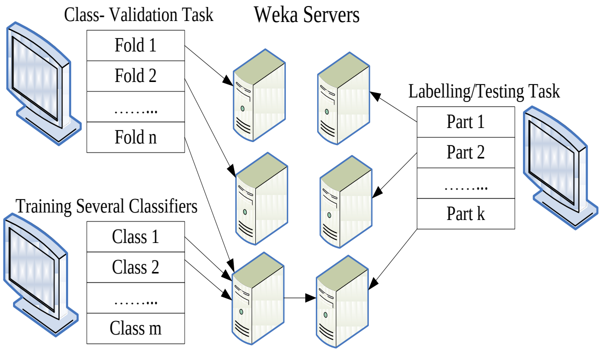

Fig 7: Operation of GridWeka2

The client component in GridWeka2 is also responsible for carrying out task scheduling followed by recovery and monitoring of faults. The analysis of the mining as well as job allocation towards servers can be carried out by specifying resource constraints. The Weka functions are used for translating the request of the client in the form of calls along with providing various valuable recovery of a local storage system in the circumstances of a system crash. GridWeka2 supports resource sharing that allows multiple clients to access the different data from the same server.

4.3.3 WekaG

This is another version of Weka, which uses similar service client approach in analyzing data. The data mining algorithm is implemented over the server side while client implements WekaG for developing an instance of services towards grid infrastructure to act itself as an interface.

4.3.2.1 Research Trends in DM and DDM

The existing literature has significant studies towards data mining algorithms. There are more than 4500 journals in IEEE Xplore digital library about the data mining techniques, and there are less than 282 Journals in same IEEE Xplore digital library associated with distributed data mining techniques. We found around 717 journals for data mining-related research in Springer whereas for distributed data mining there are 391 journals during 2005-2015. For Science Direct, there are only 64 journals on data mining published during same last 10 years. However, we couldn't find any significant research manuscript about grid computing and data mining together or distributed data mining. We find that the work carried out over data mining is quite large and high in comparison to distributed data mining technique. A specific pattern could be found in the research studies. Following trend have been observed for the research work pertaining to Data mining and distributed data mining:

- Studies on data mining are quite higher than studies towards distributed data mining.
- Studies towards grid computing and its data mining are quite less.
- Simulation-based study is more in comparison to real-time dataset-based distributed data mining.
- Majority of the outstanding research work towards distributed mining has been carried out 2000 and 2002.
- Year 2002-2009 have infrequent research manuscript with quite a lot deviation in distributed data mining.
- Weka4WS, OGSA-DIA, Knowledge Grid are some of the frequently used tools explored and used using 2005-2015.
- Problems associated with Data mining are still highly searched and adopted research topic in comparison to grid-based data mining.
- Less amount of research work have discussed or emphasized about complexities associated with grid-based data mining.

5. EXISTING SURVEY PAPERS

The visualization of the trends towards distributed data mining technique could be assessed by reviewing some of the review papers being discussed in the present system. Table 1 shows the most significant review works towards distributed data mining approaches only till date. Our investigation shows that study conducted by Park and Kargupta [11] is one of the best review works till date in comparison to all the review work being introduced till date. Researchers like Xiao et al. [12] have discussed both the parallel and distributed techniques of data mining with a good balance. Ahamed & Hariharan [13] have investigated some of the relevant techniques on the grid system. The authors discussed that Globus toolkit was the appropriate choice of evaluating distributed data mining along with compliance on OGSA. The authors have also discussed predictive apriori and normal apriori algorithm. Sunny and Thampi et al. [14] have reviewed the similar algorithms but considering the case study of the peer-to-peer network. The research has also discussed clustering technique, sophisticated data mining technique e.g. association rule mining, classification, primitive operation, etc. The study has also reviewed the problems and significant issues in a centralized approach. The investigation also outlined the challenges in distributed data mining technique e.g. emerging applications of complex data sources, loosely coupled data in distributed system, security issues, etc. Sawant et al. [15] have carried out theoretical discussion of distributed data mining techniques with emphasis on agent components in knowledge discovery process. Devi et al. [16] have also discussed the trends of distributed data mining. The authors have raised a discussion for classifier learning approach, ensemble learning techniques, and association rules mining. Masih and Tanwani [17] have discussed similar trends of distributed data mining techniques. The authors have discussed various tools of parallel data mining technique along with distributed techniques. Various techniques of the data mining and its algorithm, association rule, classification rules, techniques on cloud, etc.

Table 1: Summary of Existing Review Papers

Authors	Year of Publication	Literatures		Strength	Weakness
		Total	Relevant		
Park and Kargupta [11]	2002	100	78	Elaborated discussion of prior techniques between 1986-2001	No discussion of research gap or benchmarked study
Xiao [12]	2010	89	53	1. Elaborated discussion of prior techniques between 1986-2001 2. Comparison of Techniques	1. No discussion of research gap. 2. Generalized comparative analysis
Ahamed & Hari [13]	2011	21	11	Discussed about knowledge grid, Globus toolkit services, prediction & apriority algorithm	1. No discussion of research gap. 2. Generalized comparative analysis
Sunny & Thampi [14]	2012	54	37	Discussed prior studied related to distributed data mining in P2P	1. More Emphasis on Theory and less on research attempts 1. No discussion of research gap.
Sawant et al. [15]	2013	7	4	Discussed importance of agents in distributed data mining	Less focus on prior research contribution
Devi [16]	2014	48	27	Elaborated discussion of theory involved in trend of distributed data mining	1. Paper discusses about stale information (older than 2007)
Masih & Tanwani [17]	2014	83	53	Discussed trends of parallel and distributed data mining	1. More Emphasis on Theory and less on research attempts 2. No discussion of research gap or benchmarked study
Srinivasulu et al. [18]	2014	23	19	Discussion of research issues	1. Few discussion on prior research work. 2. Discussion of quite back-dated data

6. EXISTING RESEARCH TRENDS

This section discusses the existing research attempts towards analyzing the contribution. Although, the survey papers discussed in the prior section have already highlighted some of the prior research contributions, we feel that the discussion of the review papers are quite theoretical and have quite old data. We re-investigate this fact to find that studies towards distributed data mining considering grid infrastructure are quite less not even more than 50 number of research publications. We found very few research journals or transaction paper that has focused on challenges towards mining over grid infrastructure. In this paper, we avoid repeated discussion of the research attempts. Only significant work under taken for distributed data mining was considered in discussion in this section for the research manuscript published between 2005-2015. This section chooses to discuss the most relevant 30 research papers published in last 10 years pertaining to the problem of distributed data mining over grid infrastructure.

Studies towards distributed data mining over grid interface are not new, and it dates back more than a decade old. One of the potential works has been carried out by Cannataro et al. [19] in 2002. The author has discussed extensively the usage of Knowledge Grid system that is based on mobile agents. The same author in the year 2005 has presented another technique on Knowledge Grid highlighting more on the application part of it [20]. Luo et al. [21] have presented a technique of distributed data mining by reusing the presented used data mining approaches in the form of mobile agents. The authors used the data collected from oil rigs for carrying out the analysis. The authors have also presented a scheduling model where the outcome is assessed using response time, inter-arrival time, and throughput. Cardona et al. [22] have used the open source software framework (Map Reduce) to enhance the efficiency towards storage as well as processing of distributed data. The optimization of the mining is also carried out by using a probabilistic neural network. Huo et al. [23] have implemented a service-oriented architecture and specifically developed a process for replica management. The authors have developed a grid structure for carrying out data mining.

Talia et al. [24] [25] is one of the significant researchers who have published many research manuscript based on distributed mining process. His work has introduced the process of knowledge discovery as a significant service. He has constantly advocated the usage of the tool Weka4WS along with usage of service oriented architecture. Huang et al. [26] have used Knowledge Grid framework and enhanced the knowledge discovery process using semantics and association rule. The outcome of the study is evaluated using execution time parameters. Atkinson et al. [27] have jointly introduced the concept of knowledge discovery along with access/integration characteristics in distributed data mining using OGSA-DAI. Lackovic et al. [28] have used Weka4WS architecture as well as service-oriented architecture for carrying out distributed data mining technique. Brescia et al. [29] have discussed a project called as DAME or Data Mining Exploration which mainly targets to develop a distributed grid infrastructure. Hmida and Slimani [30] have presented as Weka4GML architecture for performing distributed data mining. Kantarcioglu and Nix [31] have applied game theory along with Vickrey-Clarke-Groves process for the purpose of validating the resultant data of distributed data mining.

Oyana [32] have presented a new technique of clustering for distributed data. Experimented over synthetic and real data, the presented technique shows an efficient query processing. The outcome of the study was evaluated on mean squared error and response time on increasing percentage of data. Rao and Vidyavathi [33] have used multi-agent approach to carrying out distributed data mining using game theory. The outcome of the study was evaluated using gain as a performance parameter. Tlili and Slimani [34] have used association rule to perform distributed data mining as well as dynamic load balancing. However, the study is more inclined to load balancing. Prusiewicz and Zieba [35] have designed a data mining technique using service-oriented architecture and semantics. Santos et al. [36] have developed a distributed data mining considering a case study of the healthcare sector. Mallik et al. [37] have presented an analysis for usage of an asynchronous algorithm. Usage of the multi-agent system was also found in the work of Pandey et al. [38], Prajapati and Menaria [39]. Zhang et al. [40] have presented an efficient learning process for enhancing distributed data mining. The authors have used Big Data approach for optimizing the accuracy of learning. Belbachir et al. [41] have presented a sequential algorithm for generating association rule. Vishvapathi et al. [42] have developed a distributed mining algorithm for grid system using supervised learning algorithm using Weka tool.

A unique form of study is carried out by Amini et al. [43] where the authors have used density-based clustering process over heterogeneous data of Internet-of-Things. The outcome is faster processing time with data quality. Maab et al. [44] have presented a study for processing data from the smart grid using Big Data Analytics. Ogunde et al. [45] have applied association rule mining that suits better in distributed data storage. Rebbah et al. [46] have presented a technique for extracting association rule about grid computing. Srinivasan and Palanisamy [47] have used swarm intelligence-based concept to perform optimization of clustering process in high-dimensional data. The study outcomes were tested on outliers. Zhou [48] has presented a data mining approach over cloud platform. The author has presented a simple mathematical modeling of establishing the linear and non-linear relationship. The outcome of the study was shown to enhance the performance of distributed data mining on running time, memory occupancy, and average clustering quality.

Hence, it can be seen that there are various approaches being presented by the researchers in the last decade out of which the multi-agent approach, game theory, service-oriented architecture, usage of open source machine learning and data mining e.g. Weka, Weka4WS, etc. are quite frequently and repetitively exercised. All the techniques have their advantage as well as a limitation. Table 2 summarizes the prior research contributions in a more compact manner on outcomes and limitations observed in the prior studies.

Table 2: Summary of Existing Research towards Distributed Data mining in Grid Infrastructure

Authors	Technique	Performance Outcome	Limitation
Cannataro et al. [19]	Knowledge Grid	-Nil-	Outcomes not discussed
Cannataro et al. [20]	Computational model on Knowledge Grid	Supports distributed data analysis	Outcomes not discussed
Luo et al. [21]	Scheduling, distributed datamining,	Reduction in response time and increment in throughput	-Lacks benchmarking -Lacks data complexity
Cardona et al. [22]	MapReduce, Neural network	High availability Better than round robin	Doesn’t evaluate algorithm complexity
Huo et al. [23]	Service-oriented architecture	Lower latency	-Poor outcome analysis -Lacks benchmarking
Talia et al. [24] [25]	Weka4WS	Lower overhead for invocation of web service	-Lacks benchmarking -Lacks data complexity
Huang et al. [26]	Knowledge Grid, Association rule	Lower execution time	-Lacks benchmarking

			-Lacks data complexity -Poor outcome analysis
Atkinson et al. [27]	Mining & Integration	High Throughput	-Do-
Lackovic et al. [28]	Weka4WS	Can execute parallel datamining	-Do-
Brescia et al. [29]	Data Mining Exploration	Easy to use,	-Do-
Hmida and Slimani [30]	Weka4GML	Can port parallel and distributed mining	-Do-
Kantarcioglu and Nix [31]	Game theory model	0.6% classifier accuracy	-Do-
Oyana [32]	k-means approach	Faster processing compared to existing k-means algorithm	Doesn't address high-dimensional data complexity
Rao and Vidyavathi [33]	Game theory model	Enhanced gain	-Lacks benchmarking -Lacks data complexity
Tlili and Slimani [34]	Load balancing, association rule, apriori algorithm	Enhanced processing time	-Less extensive analysis of outcome
Prusiewicz and Zieba [35]	Service-oriented architecture	Can solve regression problem	-less evidence of supportability to high-dimensional & heterogeneous data
Santos et al. [36]	Distributed Data mining over grid	Faster processing	-less evidence of supportability to high-dimensional & heterogeneous data
Mallik et al. [37]	Prediction of data	Lower communication cost	-Do-
Pandey et al. [38]	Multi-agent system	Enhanced performance	-Do-
Prajapati and Menaria [39]	Multi-agent system	Enhanced performance	-Do-
Zhang et al. [40]	Big Data Approach	Computational efficient, stable precision rate	-Lacks benchmarking
Belbachir et al. [41]	Association rule	Lesser dependency on communication	-less evidence of supportability to high-dimensional & heterogeneous data
Vishvapathi et al. [42]	Support vector machine in mining	Lesser rate of error	-Do-
Amini et al. [43]	Density-based clustering	Faster processing rate	-Nil-
Maab et al. [44]	BigData Analytics	Can effectively process dynamic data of smart grid.	-Lacks benchmarking
Ogunde et al. [45]	Association rule mining	Lesser response time	-Lacks benchmarking -Doesn't address high-dimensional data complexity
Rebbah et al. [46]	Association rule mining	Lesser extraction time	-Lacks benchmarking -Doesn't address high-dimensional data complexity
Srinivasan and Palanisamy [47]	Swarm intelligence	Addressing clustering in high-dimensional data.	-Lacks benchmarking -Less extensive analysis of outcome
Zhou [48]	Cloud-based data mining	Lesser time of running, reduced memory usage	-Doesn't address high-dimensional data complexity

7. RESEARCH GAP

This section discusses the significant research gap explored in the area of distributed data mining techniques over grid infrastructure. The comments for this section are based after reviewing the existing research trends and contributions discussion in prior sections.

7.1 Less Emphasis on Grid Infrastructure

Although there are research papers on distributed data mining techniques, but quite a less number of papers have implemented the grid infrastructure. Various complexities associated with the grid, forms of data generated from the warehouses, heterogeneity, data volume, etc are some of the overlooked factors in the past research work. Although the studies on data mining is quite a large compared to distributed data mining, we didn't find any much applicability of data mining techniques in grid other than using Weka4WS, GridFTP, OGSA-DIA, etc. There no novel model being introduced in the past which is independent of such frequently practiced frameworks/tools.

7.2 Lower emphasis on data complexity:

In reality, there are different data warehouses in the concept of distributed data mining, which should further give rise to high-dimensional data, sub-space clustering, and higher extent of data heterogeneity. These facts give rise to highly complex data. Hence, the existing studies didn't mention how such micro level problems are being solved or addressed. Many of the existing studies are symptomatic i.e. they solve one problem leaving the other associated problem unsolved. Hence, data complexity problem is not considered in the existing study.

7.3 Less Study towards Unified Architecture

It is a common fact that distributed data mining protocols depend on the exact need of different organization. Even to use Oracle Grid Infrastructure, it is essential that organizational need for current and future to be aligned with the grid cluster design and connectivity. The biggest complexity lies in understanding the technique to design an effective and extensible unified protocol and architectural design that can cater up the storage and analytical needs of an organization. There is no such study being carried out towards a generic architecture or protocol towards distributed data mining over grid infrastructure.

7.4 Selection of Less Effective Performance Parameters

With the increasing size and complexity of data over the data warehouses, the algorithm's computational complexity also increases. At present, we found that various researchers have used throughput, error rate and response time mainly as the performance parameters. However, we strongly feel that performance of distributed data mining also depends on how the network connectivity is designed. Hence, it is quite imperative to check for certain network parameters too e.g. bandwidth, latency, end-to-end delay, algorithm processing time, memory consumption, etc. To assess the Quality-of-Experience (QoE). As the data analysis depends on the availability of the data, hence, it is also imperative to understand how the existing algorithms of distributed data mining over grid addresses the problems of data volume, veracity, velocity, and variety for large and massive data. The existing study doesn't focus much on algorithm performance and is more inclined on prototype overall performance.

8. CONCLUSION

This paper has discussed an importance of distributed data mining system and indirectly highlights that there is a massive difference between the research work and reality. We believe more the number of research work there is a better possibility of rolling out a prototype in the commercial market for customer's benefit. At present, we have high-end software e.g. Oracle Grid Infrastructure, which can manage very effectively about the data in grids and clusters, however, carrying out mining operation is quite complicated one. We strongly believe that present day data is very different that conventional data. The data in the present day is quite massive, heterogeneous, and unstructured in size, which makes even an SQL-based approach impossible to store the data in a conventional relational database system. Therefore, when such forms of complicated data cannot be stored in SQL, than it cannot be subjected to conventional data mining approach for analysis purposes. To solve such problem, we have a cloud for storage and various open source frameworks e.g. Hadoop and Map Reduce for storage and retrieval of distributed data. Although, Map Reduce offers some mining operations, it cannot cater up the real requirements of faster and highly efficient mining. For the complicated and bigger data like astronomy data and genomics data, there is a need to develop a faster-distributed data mining technique over grid clusters. The study finding suggests that studies towards distributed data mining are quite less than standalone data mining techniques. There is also prominent research gap, which motivates us to carry the study in future direction. Our work towards future direction will be to develop a novel protocol and architecture that can address the complications in distributed data mining system.

9. REFERENCES

[1] T. Jain.2013. Technology Advancement in Wireless Communication. International Journal of Scientific & Technology Research, vol. 2, Issue. 8

[2] R.S. Sengall.2015. Research and Applications in Global Supercomputing. IGI Global, Computers, pp. 672

[3] M.J. Shaw., C. Subramaniam., G. W. Tan., and M. E. Welge.2001.Knowledge management and data mining for marketing. Decision support systems, vol. 31, no. 1, pp.127-137

[4] D.T. Larose.2014. Discovering knowledge in data: an introduction to data mining, John Wiley & Sons

[5] H.R. Rollinson.2014. Using geochemical data: evaluation.presentation, interpretation. Rutledge

[6] G. Shi., M. Mortazavi., J.Chen., and V.G.R. Kotha.2015.Method and apparatus for providing highly-scalable network storage for well-gridded objects. U.S. Patent, vol. 8, pp.996-803

[7] G. Weichhart., A. Molina., D.Chen., L.E. Whitman., and F. Vernadat.2015. Challenges and current developments for Sensing, Smart and Sustainable Enterprise Systems. Computers in Industry

[8] S. Venugopal., R. Buyya., and K. Ramamohanarao.2006. A taxonomy of data grids for distributed data sharing, management, and processing. ACM Computing Surveys (CSUR), vol. 38, no. 1

[9] D.B. Keator., J.S. Grethe., D. Marcus., B. Ozyurt., S. Gadde., S.Murphy., S. Pieper.2008. A national human

neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). Information Technology in Biomedicine, IEEE Transactions, vol.12, no. 2, pp. 162-172

[10] R.J. Wilson. 2001. The European DataGrid Project.Institute de Fisica d'Altes Energies and Colorado State University, Barcelona, Spain and Fort Collins, Colorado, USA

[11] B-H. Park and H. Kargupta.2002. Distributed data mining: Algorithms, systems, and applications

[12] H. Xiao.2010. Towards parallel and distributed computing in large-scale data mining: A survey. Technical University of Munich, Tech. Rep

[13] B.B. Ahamed., and S. Hariharan.2011. A Survey On Distributed Data Mining Process Via Grid. International Journal of Database Theory and Application, vol. 4, no. 3, pp.77-90

[14] S.M. Thampi.2012. Survey on Distributed Data Mining in P2P Networks. arXiv preprint arXiv, pp.1205-3231

[15] V.Sawant, D. J. Sanghvi, K. Shah.2015. A Review of Distributed Data Mining using Agents. International Journal of Advanced Technology & Engineering Research (IJATER)

[16] S. G. Devi.2015.A Survey on Distributed Data Mining and Its Trends. Retrieved, 23rd Nov, 2015

[17] S. Masih., and S.Tanwani.2014. Data Mining Techniques in Parallel and Distributed Environment-A Comprehensive Survey.

[18] D.L. Srinivasulu., B. S. Kumar., and V. G. Akula.2015. A Survey on Research Problems in Distributed Data Mining. Retrieved, 23rd Nov, 2015

[19] M. Cannataro., D. Talia., and P.Trunfio.2002. Distributed data mining on the grid. Future Generation Computer Systems, vol. 18, no. 8, pp.1101-1112

[20] M. Cannataro., D.Talia., and P.Trunfio.2002. Design of distributed data mining applications on the knowledge grid. In Proceedings NSF Workshop on Next Generation Data Mining, pp. 191-195

[21] P. Luo., K. Lü., Z. Shi., and Q. He.2007. Distributed data mining in grid computing environments. Future Generation Computer Systems, vol. 23, no. 1, pp. 84-91

[22] K. Cardona., J. Secretan., M. Georgiopoulos., and G. Anagnostopoulos.2007. A grid based system for data mining using MapReduce. Technical Report TR, AMALTHEA

[23] L. Huo., Y. Fang., and H. Hu.2008. Dynamic service replica on distributed data mining grid. In Computer Science and Software Engineering, International Conference, vol. 3, pp. 390-393

[24] D. Talia.2009. Distributed data mining tasks and patterns as services. In Euro-Par Workshops-Parallel Processing, Springer Berlin Heidelberg, pp. 415-422

[25] D. Talia., P. Trunfio., and O. Verta.2008. The Weka4WS framework for distributed data mining in service-oriented Grids. Concurrency and Computation: Practice and Experience, vol. 20, no. 16, pp. 1933-1951

[26] F. Huang., Z. Li., and X. Sun.2008.A data mining model in knowledge grid", In Wireless Communications, Networking and Mobile Computing, WiCOM'08. 4th International Conference, pp. 1-4

[27] M.P. Atkinson., J.I. V.Hemert., L. Han., A.Hume., and C.S. Liew.2009. A distributed architecture for data mining and integration. In Proceedings of the second international workshop on Data-aware distributed computing, pp. 11-20

[28] M. Lackovic., D. Talia., and P. Trunfio.2009. A Service-Oriented Framework for Executing Data Mining Workflows on Grids. In Grid and Pervasive Computing Conference,. GPC'09. Workshops, pp. 72-79

[29] M. Brescia., S. Cavuoti., R.D.Abrusco., O.Laurino., and G.Longo.2012. DAME: A Distributed Data Mining and Exploration Framework within the Virtual Observatory. In Remote Instrumentation for e-Science and Related Aspects, Springer, pp.267-284

[30] M.B.H. Hmida., and Y. Slimani.2010. Meta-learning in grid-based data mining systems. International journal of communication networks and distributed systems, vol. 5, no. 3, pp. 214-228

[31] M. Kantarcioglu., and R. Nix.2010. Incentive compatible distributed data mining. In Social Computing (SocialCom), IEEE Second International Conference, pp. 735-742

[32] T.J. Oyana.2010.A new-fangled FES-k-means clustering algorithm for disease discovery and visual analytics. EURASIP Journal on Bioinformatics and Systems Biology, vol. 746021, no. 1

[33] V.S. Rao., and S. Vidyavathi.2010. Distributed data mining and mining multi-agent data. (IJCSE) International Journal on Computer Science and Engineering, vol. 2, no. 04, pp. 1237-1244

[34] R. Tlili., and Y.Slimani.2011. Executing association rule mining algorithms under a Grid computing environment. In Proceedings of the Workshop on Parallel and Distributed Systems: Testing, Analysis, and Debugging, pp. 53-61

[35] A. Prusiewicz, and M. Zieba.2011. The proposal of Service Oriented Data Mining System for solving real-life classification and regression problems. In Technological Innovation for Sustainability, Springer Berlin Heidelberg, pp. 83-90

[36] M.F. Santos., W.Mathew., and C.F.Portela.2011. Grid Data Mining for Outcome Prediction in Intensive Care Medicine. Enterprise Information Systems, pp. 244-253

[37] R. Mallik., N. Sarda., H. Kargupta., and S. Bandyopadhyay.2011. Distributed data mining for sustainable smart grids. Proc. of ACM SustKDD, vol. 11, pp.1-6

[38] T.N. Pandey., N. Panda., and P. K. Sahu. 2012. Improving performance of distributed data mining (DDM) with multi-agent system. IJCSI International Journal of Computer Science, issues. 9, no. 2

[39] R.B. Prajapati., and S. Menaria.2012. Multi Agent-Based Distributed Data Mining. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 1, no. 10, pp. 76.

[40] Y. Zhang., D.Sow., D.Turaga., and M. V.D. Schaar.2014. A fast online learning algorithm for distributed mining of big data. ACM SIGMETRICS Performance Evaluation, vol.41, no. 4, pp. 90-93

[41] K. Belbachir., and H. Belbachir.2013. Parallel Mining Association Rules in Calculation Grids

[42] P.Vishvapath., S.Ramachandram., A.Govardhan.2014. GWSVM Algorithm for a Grid System. International Journal of Computer Science and Information Technologies, vol. 5 (5), pp. 6871-6876

[43] A. Amini., H.Saboohi., T.Y.Wah., and T. Herawan. 2014. A fast density-based clustering algorithm for real-time internet of things stream. The Scientific World Journal

[44] H. Maaß., H.K. Cakmak., F. Bach., R. Mikut., A. Harrabi., W. Süß., W. Jakob., Kl-U.Stucky., U.G. Kühnapfel., and V. Hagenmeyer.2015. Data processing of high-rate low-voltage distribution grid recordings for smart grid monitoring and analysis. EURASIP Journal on Advances in Signal Processing, no. 1, pp. 1-21

[45] A.O. Ogunde., O. Folorunso., and A.S. Sodiya.2015. The Design of an Adaptive Incremental Association Rule Mining System. In Proceedings of the World Congress on Engineering, vol. 1

[46] M. Rebbah., M.E.A. Yemres., M. Khaldi., and M. Debakla. 2015. Hybrid Distribution for Association Rules Extraction on Grid Computing. Retrieved, 24th Nov, 2015

[47] T. Srinivasan., and B. Palanisamy.2015. Scalable Clustering of High Dimensional Data Technique using SPCM with ANT Colony Optimization Intelligence. Hindawi, the Scientific World Journal, pp. 5

[48] G. Zhou.2015. Cloud Platform Based on Mobile Internet Service Opportunistic Drive and Application Aware Data Mining. Journal of Electrical and Computer Engineering, vol. 50, pp. 357-378

10. AUTHOR PROFILE

Shahina Parveen has worked as Assistant Professor , Department of ISE, Bhageerathi Bai Narayan Rao Manay Institute of Technology, Bangalore. She has got 9 years of teaching experience. She has obtained Bachelor of Engineering from JNT University in the year 2005. She studied Masters of Technology from ANU, Guntur, AP and was awarded in the year 2010. Now she is a Ph.D. student 4th year of CSE at JNT University, Hyderabad, India. She has published papers in both national and international conferences.

Dr. G. Narshimha is working as asssociate professor at JNTUH, Karim Nagar, Telangana, India. He has completed his B.E in ECE at Osmaniya University, Hyderabad and obtained Master degree in CS&E in 1999 at Osmaniya University. He has awarded doctrate in CS&E Osmaniya University Hyderbad, India in July 2009. He has about 17years of teaching experience. He has published 70 papers in both national & international conferences followed by 38 international and nation journals. 3 PhD's are awarded and 13 research schcolars are working under him. He is life member of indian society for technical education (MISTE), MIEEE.