Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7 – No. 12, January 2018 – www.caeaccess.org

A Novel Design of Sophisticated Distributed Knowledge Extraction Process on Grid Architecture

Shahina Parveen M.
Research Scholar, Department of Computer Science & Engineering
JNTU, Hyderabad, India

G. Narsimha
Professor, Department of Computer Science & Engineering
JNTUH, Karimnagar, Hyderabad, India

ABSTRACT
With the rising demands of ubiquitous applications, the complexities associated with the data are exponentially increasing. Although, such massively generated complex data doesn't pose much challenge in storage system, but it definitely strikes a challenging problem in order to perform mining. The process of discovering the valuable knowledge becomes much challenging if a distributed architecture of grid network is considered. Therefore, the proposed system introduces a novel architecture that is capable of performing error-free distributed mining over grid networks. The significant contribution of proposed system is to apply a novel and cost effective optimization technique for simplifying the data structurization problem in distributed system that is found to normalize the existing data complexity problems. The study outcome exhibits significantly low errors and minimal computational cost in presence of peak traffic condition to prove that proposed architecture offers better mining approach in contrast to existing approaches.

Keywords
Analytics, Data Mining, Distributed, Grid Computing, Knowledge Discovery, Data Complexity

1. INTRODUCTION
Grid computing plays a significant role in network technologies as it offers comprehensive utilization of networked resources to carry out certain sophisticated computing task [1]. Such network consists of highly interconnected terminals that are again connected centrally by a node controller [2]. Although, grid computing offers potential benefits toward sophisticated computational task but at the same time it is also associated with distributed data management [2]. Theoretical illustration of grid computation in literatures have claimed its supportability of heterogeneous data using the concept of data integration however such *form* of the data is more hypothetical and impractical as compared to problems of data management in existing times [3]. With the evolution of virtualized storage system over cloud, storage is absolutely not a challenging task but applying analytics on distributed storage is extremely challenging task. The prime reason behind these problems is paradigm shift in the *form* of data. At present, there is inclusion of different source point of such data where such data are more unstructured form. Hence, performing analytical operation of such forms of data itself possess a computational challenge irrespective of number of research-based approaches on mining [4]-[7]. For an effective data management in grid computing, it is required to offer better quality-of-service, higher fault tolerance, enhanced expressiveness to dynamic query, efficient resource discovery, and better transparency of data. At present some of the challenges encountered in grid computing with regards to data management are i) no standard definition of data diversification that leads to failure in identifying the data forms leading to ineffective storage / query processing of data over the grid, ii) no availability of mechanism to perform consistent tracking of level of redundancies present in data to generate computational as well as network overhead, iii) lack of mechanism to ensure higher degree of productiveness in data, and iv) absence of any robust framework for dealing with data dynamicity. All the above mentions challenges are root cause of inapplicability of existing mining approaches on data stored over grid. It has been also studied that there are various forms of open source frameworks e.g. Hadoop, Map Reduce, etc that are frequently studied in regards to the concept of big data approach [8]-[10], however, such frameworks are more inclined on cloud computing and less on grid computing. There is no dedicated framework or tool or model that identifies the complexities associated with data management n distributed grid environment. Hence, the problems become much magnified when the attempt to use *mined data as service*, which is still an open-end problem in association with grid computing. Although, the construction of grid computing has no dependency on cloud but there is a rising number of research attention where potential of cloud computing would be increased through grid computing. Hence, offering mined data as a service from grid network could offer valuable service to cloud customers and hence could significantly assists in redefining the level of services to cloud customers as well. It should be also known that utilization and coverage of cloud computing has more direct reach ability to its end customer as compared to grid which is accessed by few skilled nodes to carry out some sophisticated processing. One way to address this problem is to start investigating the better possibility of building distributed mining environment over grid architecture. Hence, the proposed manuscript introduces one such design principle that offers distributed mining with better performance over grid architecture. Section 1.1 discusses about the existing literatures where different techniques are discussed for detection schemes used in power transmission lines followed by discussion of research problems in Section 1.2 and proposed solution in 1.3. Section 2 discusses about model implementation followed by discussion of result analysis in Section 3. Finally, the conclusive remarks are provided in Section 4.

1.1 Background
This section briefs about the approaches associated with distributed mining system and data management techniques practiced in grid computing. The work carried out by Deng et

al. [11] has used simulated annealing and evolutionary approach in leveraging the distributing mining system for filtering content from active network. Shah et al. [12] have developed a mechanism for query management using sliding window for improved performance over distributed data. Bu et al. [13] have carried out investigation towards mining community-based information using multi-agent based approach. Savage et al. [14] have used open-source framework for applying mining approach on contrast pattern on large database. Study toward security of extracted knowledge was presented by Tassa [15] using association rule. Angiulli et al. [16] have presented a scheme to extract outliers from the mined data. Foo et al. [16] have presented a classification scheme for distributed data that offers supportability of data replication with capability of auto-reconfiguration of classifier tree. Sun et al. [17] have presented a method for performing fragmentation of ongoing task of distributed nature for ensuring better workflow management. Tsai et al. [18] have presented a distributed mining approach that has the capability to perform clustering using derived patterns of its mobility. Foo and Schaar [19] have further optimized the classification process in order to suit the real-time analysis of distributed data with better re-configurability. However, all these above studies towards distributed mining have not been applied in grid environment. In recent time, the studies towards grid computing are carried out in different direction with respect to data management. The work carried out by Souravlas [21] has emphasized on enhancing data accessibility by implementing binary tree for identifying certain significant files. Weng et al. [22] have used clustering-based approach to address the problems of dimensional reduction in smart grid. Lois et al. [23] have presented an arithmetical model by introducing a consensus-based logic for minimizing convergence errors owing to data uncertainty. Allalouf et al. [24] have implemented similar mechanism of dimensional reduction over smart grid using network optimization principle. Garlasu et al. [25] have investigated applicability of big data-based approaches on grid computing. Study towards developing middleware for facilitating desktop grid is carried out by Saad et al. [26]. Anjos et al. [27] has highlighted the disadvantages associated to the use of existing distributed software frameworks and comments its inapplicability over heterogeneous environment of grid computing. Moretti et al. [28] have presented an abstraction-based approach in order to facilitate data computing over campus grid. Similar cadre of investigation towards data intensive work is also carried out by Zhou [29] by leveraging collaborative learning approach in web services. Rusitschka et al. [30] have presented a hypothetical framework where the potential of cloud computing is being implemented over smart grid. Hence, there are different forms of work is carried out towards distributed mining as well as in grid computing. The next section briefs up the research problems associated with existing research approaches followed by proposed solution to address the explored research issues. S.Perveen and Narsimha [31], have introduced a distributed data mining architecture on grid infrastructure. From this architecture, author identified the some potential characteristics of the Map-Reduce framework and performed appropriate enhancement over the grid architecture. By the experimental analysis, they conclude that the proposed algorithms will maximize or improve the computation processing time and enhanced the efficiency of distributed mining over grid infrastructure. In [32], author discussed about grid architecture and its important characteristics. Additionally, they reviewed on data mining approach on grid computing and its related challenges.

M.Shahina and Narsimha [33] have proposed a new clustering mechanism which extracts the data from the warehouse. In this context, author applied a statistical local modeling approach which removes the non repetitive data from the storage and showed that from this mechanism can enhance the overall processing time with increasing of traffic load as compared with other existing methods of distributed mining.

In [34], author designed a general architecture of distributed data mining and has been implemented as mining services by utilizing of grid-infrastructure. The primary goal of this prior study was to deploy different distributed data-mining algorithms over grid architecture. For the experimental analysis, they introduced two clustering algorithms and executed on that architecture and finally measured the performance parameters.

1.2 Research Problem

The significant research problems are as follows:

- Existing research work is not found to emphasize on the data replicates, which is present in exponential quantity in distributed location resulting in lower precision of mining.
- Existing knowledge extraction process in grid computing are less distributed in nature and more centralized in nature that acts as impediments to analytics on grid networks.
- At present, the complexities associated with data are predominantly high and results in massive diversification and uncertainty of data that reduces the value of knowledge after mining.
- Lack of efficient optimization technique also results in unexplored area of distributed knowledge discovery process in grid architecture.

Therefore, the problem statement of the proposed study can be stated as "*Developing a cost effective architecture in grid computing that supports distributed mining in presence of increasing complexities associated with current state of distributed data.*"

1.3 Proposed Solution

The prime purpose of the proposed system is to present a design of a sophisticated architecture that contributes to distribute mining operation exclusively over the grid architecture. For this purpose, the proposed implementation is carried out using analytical research methodology.The flow of the methodology adopted for the proposed research design is exhibited in Fig.1.

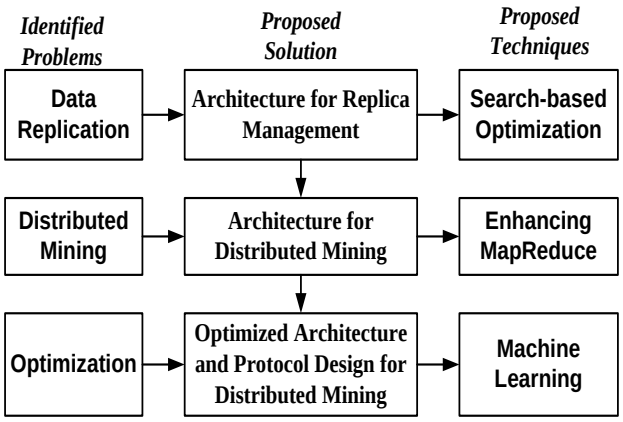


Fig.1 Flow of Proposed Methodology of Distributed Mining

In order to obtain solution towards designing an effective distributed mining approach, the proposed system identifies three significant research problems i.e. data replication, distributed mining, and optimization. These are also the significant research gap that the proposed system targets to bridge up. The first part of the proposed system emphasize on addressing the data replication problem where search-based optimization has been deployed for obtaining better structurization form of data. The outcome of unique data is obtained from the study to minimize response time. The second part of the solution deals with introducing a novel distributed mining process by improving the operation of existing MapReduce. The mechanism offers more applicability and found to offer better computational cost and reduced time. The proposed system is further enhanced by introducing an optimization of the proposed algorithm by considering a novel machine learning approach to control the computational performance of distributed mining. Presence of any forms of errors is typically impossible to model with potential impact on outcome and hence the apply machine learning approach to solve this problems. A convergence test is performed by considering hypothetical error applied to machine learning in order to obtain the best possible knowledge out of it. The system performs re-check of error values of the newly obtained data and compares with the updated threshold value to obtain newly formed data with reduced errors. The next section illustrates about the implementation process involved in proposed design methodology.

2. SYSTEM DESIGN & IMPLEMENTATION

In order to design a better form of distributing mining technique, it is essential to address some of the significant factors associated with the complexity of the data over grid architecture. As grid computing is more characterized for their potential capability to harness the maximum computational capability to execute sophisticated task; therefore, the execution of this set of process will not actually pose a bigger challenge. However, the challenge lies in precision of this concept design and implementation which is directly meant to be executed for distributed grid architecture. For better study effectiveness, divide and conquer approach is applied to accomplish this research objective. This section will discuss sequentially the architecture developed and their operation along with their relationship in order to leverage distributed mining.

2.1 Architectural Discussion

The core design basis of the architecture is to optimize the clustering operation with an aid of local models that can directly contribute to enhance the quality of mined data. There are three architectures designed viz. i) architecture for replica management, ii) architecture for distributed mining, and iii) optimized architecture and protocol design for distributed mining. The brief discussions of the architecture are as follows:

2.1.1 Architecture for Replica Management

This architecture addresses the 1st research problems where it is stated that presence of redundant data in grid nodes will adversely affect the distributed mining process and therefore this architecture offers a scheme where the replicates present in the grid nodes can be identified in order to perform effective data storage over the grid nodes. The first essential component of this architecture is *Processing Input* which is the responsible for performing pre-processing operations in order to minimize the level of noise present. Once the noises are reduced, the next component i.e. *Replica Clustering using Local Mode* performs clustering operation where the replicates are arbitrarily chosen, centroid point of the plane of clustering is obtained, followed by applying *Statistical approach* for comparing it with local models. This process will essentially result in generation of an elite local model to flag the best source of replicates within the grid nodes. The final step is to initiate a component called as *Replica Selection Process* that mainly applies *evolutionary search-based optimization technique* in order to further enhance the clustering process. The architecture for replica management is shown in Fig.2.

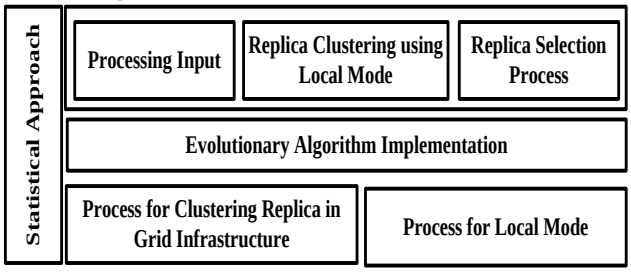


Fig.2 Architecture for Replica Management (Image)

The first *Process for Clustering Replica in Grid Infrastructure* takes the input of pre-processed data from multiple warehouses and implies local models after it. In this process, it applies a search-based optimization scheme where a fitness function is constructed in order to identify the state where the elite outcome of replica is obtained. The second *Process for Local Mode* than takes the input of replicates and initiates three levels of local modes using statistical approach. The main contribution of this architecture is to leverage the quality of data by emphasizing on obtaining the complete information of the incoming data with its respective source of origination and then perform a clustering process that leads to an outcome of unique and redundant data. In this algorithm, any redundant data are discretized but just prioritize them using search-based optimization technique in order to fasten up the query processing system. This will eventually mean that there is no much change in original storage of the distributed data; however, the proposed clustering mechanism extracts the relevant information about the data, applies pre-processing operation followed by evolutionary technique which selects only the best replicates for further processing. It is to be noted

that any replicates are not to be removed from the original storage in order not to affect data availability but only ensures that extra replicates doesn't effect in further analytical process or induce computational burden.

2.1.2 Architecture for Distributed Mining

This architecture addresses the 2nd research problems that states insufficient availability of distributed mining techniques in grid architecture. Motivated from the design principle of distributed open source framework MapReduce, the proposed system performs enhancement of its conventional designs in order to evolve up with novel distributed mining that works on grid networks. It is also endowed with the capability to mine massive volumes of data. This architecture presents a method where distributed data is explored for its significance so that mining operation leads to superior knowledge discovery. Fig.3 exhibits the proposed architecture for distributed mining.

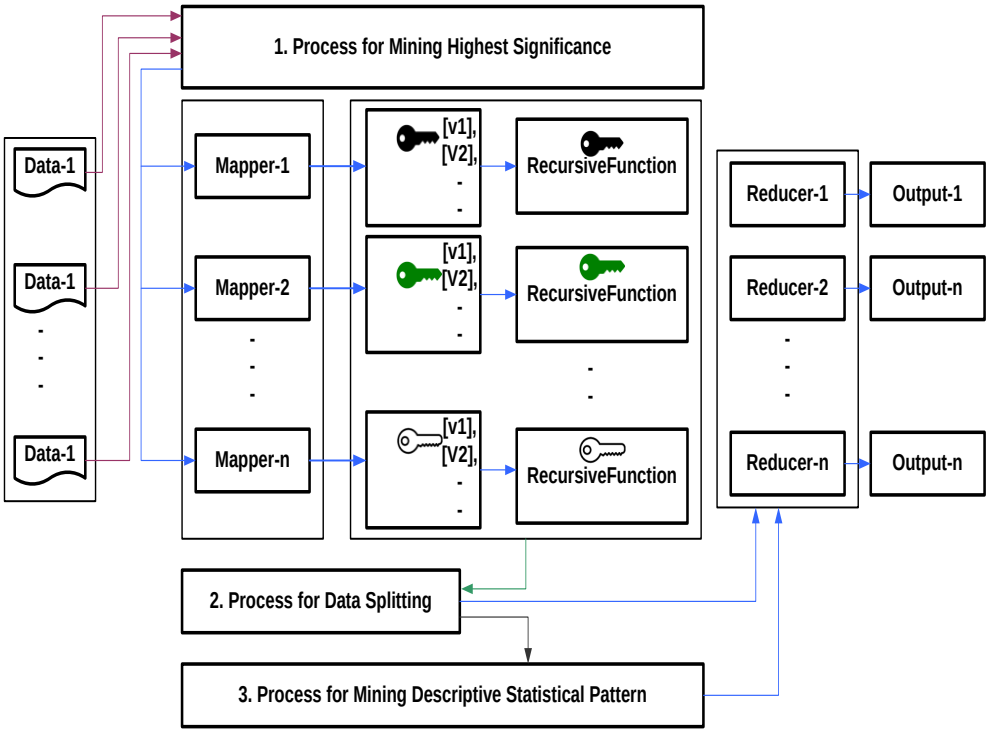


Fig.3 Architecture for Distributed Mining (Image)

The complete execution of this architecture is dependent on 3 significant processes. The *first process* applies MapReduce and performs second level of pre-processing followed by indexing of the received distributed data over grid network. This data is subjected to registration process to further explore the demands for performing further mining. The *second process* addresses the problems of higher dimensionality in data followed by clustering the data over different number of mappers followed by aggregation task by reducers. The system also estimates *data correlation* between two columns of distributed data that results in better search outcomes. This operationally is an enhancement on existing MapReducer that allows the data to be prioritized on the basis of the numerical value of data correlation. The *third process* relates to apply mining technique on the unique data using statistical approach. This process takes the input of different columns and computes mean of them in order to explore unique statistical pattern. Owing to simplistic approach of this process, there is significantly lower computational burden as well as significant advantage in the form of deterministic pattern of search as well as discrete information about the dimensionality of aggregated data. Apart from this, the architecture also offers clear visualization to any forms of variations in any form of statistical pattern for an effective knowledge extraction.

2.1.3 Optimized Architecture and Protocol Design for Distributed Mining

This architecture address the 3rd research problem that states the performing optimization to distributive mining will require broader coverage of complexities associated with data. This is the last part of proposed architectural design which also implements a design of protocols for performing an effective distributed data mining over grid networks. This architecture acts as compliments to enhancing the previous architecture of data mining. For a given set of data, the proposed system chooses to initially address the problem associated with i) *diversity of data*, ii) *ambiguity in data*, and iii) *data dynamicity*. In order to formulate this concept, a case study of distributed enterprise system is considered where dynamic data is generated across the grid and statistical-based approach is considered for evolving up with a distributed and perceive clustering model. The complete algorithm discussed in proposed system is meant to be executed in the controller component of grid architecture with an assumption that all the communication links are well established and secured. The optimization of the proposed architecture is carried out considering a novel machine learning approach to control the computational performance of distributed mining.

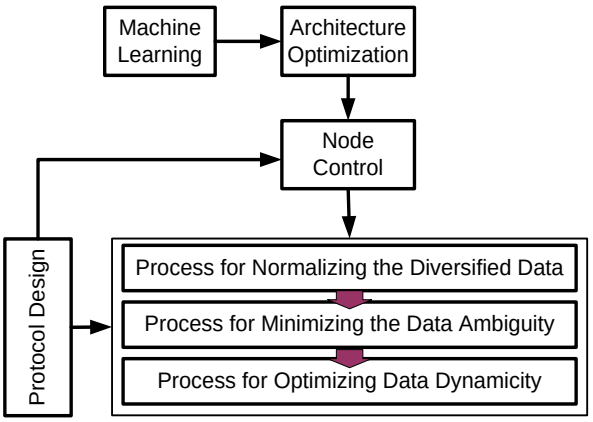


Fig. 4 Optimized Architecture and Protocol Design for Distributed Mining

The prime role of *first process* of this architecture is to construct a distributed software framework that can effectively maintain the diversified data of various forms in terms of distributed archiving over grid architecture. This process considers the input of raw data, cluster, threshold that after processing leads to generated of normalized data. The proposed study uses a threshold-based context that offers precise substitution of any form of ambiguous data prior to perform any form of mining operation. A data structure over grid could be rendered less ambiguous if all the data in the sub-matrix are very much well defined and there is a dedicated location to keep a track of it. The second process of proposed architecture takes the input of normalized data and data index that after processing yields finally allocated data on the cell with either missing value or error-prone value. The output obtained from the second process could have possibilities of estimation errors as well as channel-based errors. Presence of such errors is typically impossible to model with potential impact on outcome and hence machine learning approach is applied to solve these problems. Here, the convergence test is performed by considering hypothetical error applied to machine learning in order to obtain the best possible knowledge out of it.

2.2 Framework Connectivity

It is essential to understand about the complete utilization of proposed concept. With reference to Fig.5, it can be seen that *first architectural block* initiates with input processing for distributed data that finally results in uniquely generated and sorted data with a specific storage model. However, it only assists in minimizing the structural complexity of data so that the data could be make ready for subjecting it to a novel mining technique. This sorted data now acts as input for *second architectural block* resulting in extraction of discrete knowledge from the uniquely sorted data. The outcome of the second architectural block is capable of performing distributed mining with significant contribution of dimensionality reduction too. The study contribution is further enhanced by *third architectural block* that is capable for performing two task viz.optimization and protocol design. The task of optimization is results in performing distributed mining by identifying further more complexities associated with data diversity, data uncertainty, data dynamicity, and data productiveness while the task of protocol design presents a mechanism of performing extraction of knowledge discovery in grid environment. The final outcome of the last block also results in computation of mined data quality that quantifies the success rate of final mining approach. All the blocks of the proposed architecture was designed using statistical approach and probability theory. The proposed architecture is basically meant to be implemented within the data warehouse as the primary target. However, owing to lesser computational burden, it can be also executed in form of analytical software over grid networks as well as in cloud networks too. The proposed system also assists in cost effective distributed data modeling that significantly contributes to speed up the process of dynamic data storage over grid. This phenomenon also facilitates faster mining performance with better and error-free knowledge discovery.

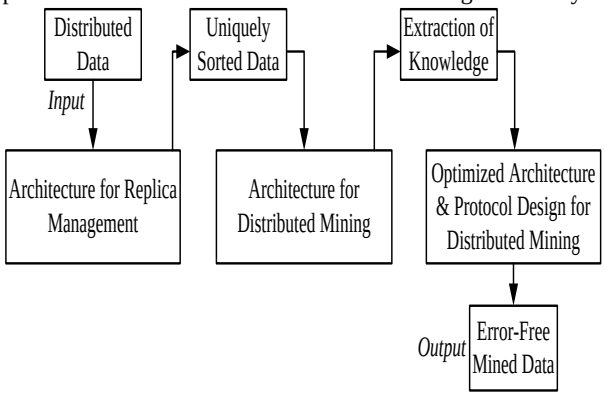


Fig. 5 Linkage among the Proposed Frameworks

3. RESULT ANALYSIS

As proposed system mainly targets to offer distributed mining operation hence emphasized on time as the performance factor. Using MATLAB, the proposed system is scripted and testified using synthetic data. This section presents the comparative analysis of all the architectures presented with respect to various existing techniques.

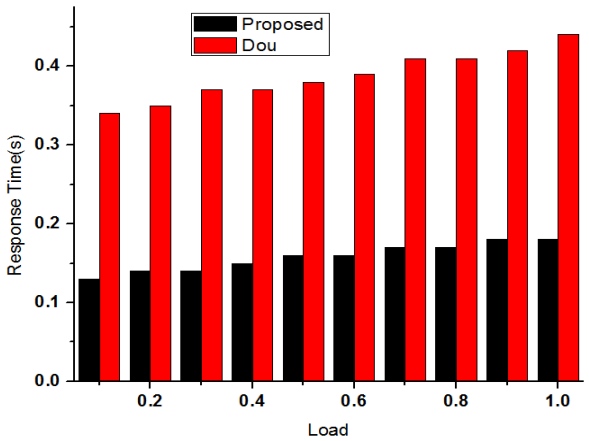
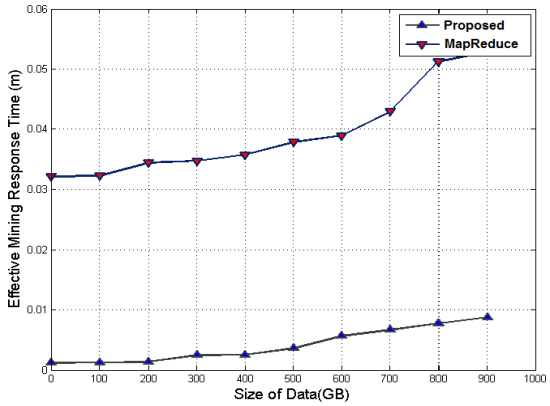


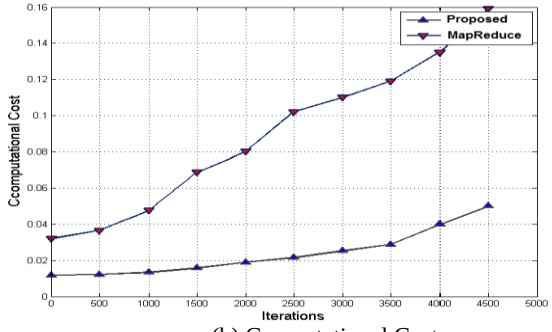
Fig. 6 Comparative Analysis of Response Time For First Architecture

The study outcome shown in Fig.6 shows the comparative analysis of proposed system by implementing the first architecture with one of the existing literature of Dou et al. [31] as they too have used the similar evolutionary-based approaching performing distributed mining. The response time of the proposed system is found to be 40% reduced as compared to existing system irrespective of adoption of similar evolution mining approach. The prime reason behind this trend is that existing approach is found using predominant

amount of Rule set in order to performing knowledge discovery. This process of executing mining approach using Rule set involves increasing amount of computational resources and thereby it involves more processing time with increasing incoming traffic load. However, proposed system applies the local model in order to perform clustering of the replicates. This process is further followed by search-based optimization using fitness function desired resulting in faster response time. Further, the second architecture is compared with the standard distributed software framework MapReduce. It is because the proposed technique implemented in second architectural design is basically an improvement over conventional design of MapReduce. A study the response time required to perform knowledge discovery (Fig.7(a)) as well as computational cost (Fig.7(b)). The graphical outcome shows that proposed system offers reduced computational cost and reduced response time in comparison to MapReduce showing that a minor change in existing standards of distributed software framework eventually enhancing the mining performance. The study outcome also is in agreement with good scalability demand as there is a least increment in response time exhibiting that proposed system could suitably fit the demands of dynamic and distributed mining over grid networks.



(a) Mining Response Time



(b) Computational Cost

Fig.7 Comparative Analysis of Response time & Computational Cost for Second Architecture

It is found that an effective mining algorithm will need to be *i*) scalable, *ii*) have the potential of identifying the data diversity and assists in structured storing mechanism over distributed system, *iii*) should be able to overcome any form of errors or challenges related to data while performing mining, and *iv*) should have higher accuracy of the mined data. Finally, investigated the outcome of the third architecture. In this part, the study outcome of the proposed system was assessed using error in y-axis and incoming traffic load in x-axis considering 500-1000 nodes. . For this purpose the proposed machine learning scheme is modified by adding up non-linear processing units in order to extract significant information about the mined outcome i.e. knowledge while performing clustering. It also has the capability to perform clustering using supervised and unsupervised learning algorithms using gradient descent approach meant for error minimization. Apart from this, the study outcome was compared with the most frequently used mining approach i.e. Linear Regression based approach (LR) and Multi-Layered Perception (MLP). The proposed study computes error as the degree of difference between the experimental outcome of data significance and hypothetical outcome of it using machine learning. Fig.8 shows that proposed study offers good performance in reducing errors as compared to the existing approaches of LR and MLP. The LR based technique offers an impressive statistical modeling with better supportability until the data are homogeneous in nature. However, real-time data are less homogeneous and more diversified causing the LR technique to fail in establishing any significant relationship between its variables. This also results in significant consumption of computational resources and processing time. With an aid of activation function and iterative learning process, the MLP technique also offers significant reduction of error. After initializing the statistical error value more than 0.7, the MLP technique is found to reduce down the error to 0.5 successfully. However, in order to do so, MLP technique has higher inclusion of iterative steps as compared to LR technique which offers slight increase in trials in order to reduce the error. The proposed system OCSLM offers a simple and non-iterative data structuring principle that uses threshold-based approach not only to identify the type of incoming data but also allocates the incoming data in a novel matrix form. Structuration of the data in the matrix form not only offers better data management but also allows any mining approach to be applied on distributed data.

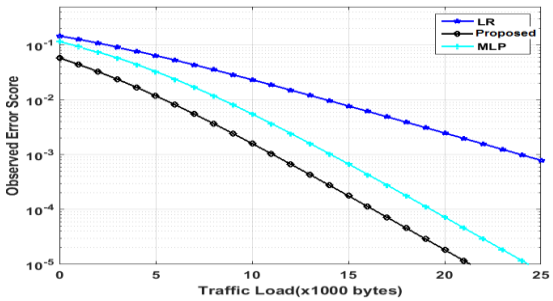


Fig.8 Comparative Analysis of Error for Third Architecture

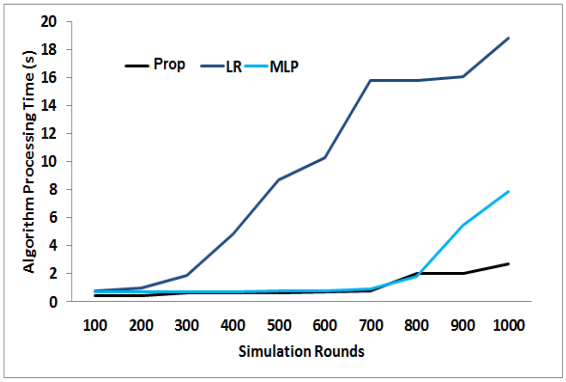


Fig.9 Comparative Analysis of Algorithm Processing Time

The outcome shown in Fig.9 exhibits that proposed system offers significantly lower algorithm processing time in comparison to existing LR and MLP scheme. The prime reason behind this is – unlike existing approaches of MLP, proposed system performs less iterative. Moreover, the approach uses dual step of minimization of errors for which reason convergence and accuracy of proposed system is much better as compared to the existing approaches. This outcome also proves that it is suitable for performing mining on any complex forms of distributed data over grid.

4. CONCLUSION

There is a significant research gap explored between existing mining algorithms and grid architecture. It is noticed that majority of the conventional analytical algorithms are designed without considering various complexities associated with present state of data. This phenomenon results in inapplicability of existing analytical approaches on distributed grids. Hence, after investigation found that existing data is associated with problems e.g. increasing diversity of data, data ambiguity, increasing dynamicity of data, etc. The proposed system also introduces multiple forms of solution e.g. search-based optimization was used for addressing replication problems, enhancement of existing Map Reduce was used for improving knowledge discovery process, and machine learning was used for addressing optimization problem in proposed system. The study outcome exhibits better response time and reduced computational cost with highly reduced errors in contrast to existing system.

The proposed work can be considered for further improvement in the computation cost as well as still more focus on error mitigation on distributed grid. Also, the security, reliability, scalability factor can be considered to bring more robustness in the proposed system.

5. REFERENCES

[1] P. Kacsuk, Dieter Kranzlmüller, ZsoltNémeth, Jens Volkert, Distributed and Parallel Systems: Cluster and Grid Computing, Springer Science & Business Media, 2012

[2] N. P. Preve, Grid Computing: Towards a Global Interconnected Infrastructure, Springer Science & Business Media, 2011

[3] A. Poduval, Do More with Soa Integration: Best of Packt, Packt Publishing Ltd, 2011

[4] C. W. Tsai, C. F. Lai, M. C. Chiang and L. T. Yang, "Data Mining for Internet of Things: A Survey," in *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 77-97, First Quarter 2014.

[5] M. De Sanctis, I. Bisio and G. Araniti, "Data mining algorithms for communication networks control: concepts, survey and guidelines," in *IEEE Network*, vol. 30, no. 1, pp. 24-29, January-February 2016

[6] L. F. C. Rezende*et al*., "Survey and prediction of the ionospheric scintillation using data mining techniques," in *Space Weather*, vol. 8, no. 6, pp. 1-10, June 2010.

[7] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," in *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153-1176, Secondquarter 2016.

[8] A. Cuzzocrea and R. Moussa, "Multidimensional database modeling: Literature survey and research agenda in the big data era," *2017 International Symposium on Networks, Computers and Communications (ISNCC)*, Marrakech, 2017, pp. 1-6.

[9] O. B. Sezer, E. Dogdu and A. M. Ozbayoglu, "Context Aware Computing, Learning and Big Data in Internet of Things: A Survey," in *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1-1.

[10] Y. Cao *et al*., "Binary Hashing for Approximate Nearest Neighbor Search on Big Data: A Survey," in *IEEE Access*, vol. PP, no. 99, pp. 1-1.

[11] S. Deng, C. Yuan, J. Yang and A. Zhou, "Distributed Mining for Content Filtering Function Based on Simulated Annealing and Gene Expression Programming in Active Distribution Network," in *IEEE Access*, vol. 5, pp. 2319-2328, 2017.

[12] Z. Shah, A. N. Mahmood, Z. Tari and A. Y. Zomaya, "A Technique for Efficient Query Estimation over Distributed Data Streams," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 10, pp. 2770-2783, Oct. 1 2017.

[13] Z. Bu, Z. Wu, J. Cao and Y. Jiang, "Local Community Mining on Distributed and Dynamic Networks From a Multiagent Perspective," in *IEEE Transactions on Cybernetics*, vol. 46, no. 4, pp. 986-999, April 2016.

[14] D. Savage, X. Zhang, P. Chou, X. Yu and Q. Wang, "Distributed Mining of Contrast Patterns," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 7, pp. 1881-1890, July 1 2017.

[15] T. Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 4, pp. 970-983, April 2014.

[16] F. Angiulli, S. Basta, S. Lodi and C. Sartori, "Distributed Strategies for Mining Outliers in Large Data Sets," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1520-1532, July 2013.

[17] B. Foo, D. S. Turaga, O. Verscheure, M. van der Schaar and L. Amini, "Configuring Trees of Classifiers in Distributed Multimedia Stream Mining Systems," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 245-258, March 2011

[18] S. X. Sun, Q. Zeng and H. Wang, "Process-Mining-Based Workflow Model Fragmentation for Distributed Execution," in *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, no. 2, pp. 294-310, March 2011.

[19] H. P. Tsai, D. N. Yang and M. S. Chen, "Mining Group Movement Patterns for Tracking Moving Objects Efficiently," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 2, pp. 266-281, Feb. 2011.

[20] B. Foo and M. van der Schaar, "A Distributed Approach for Optimizing Cascaded Classifier Topologies in Real-Time Stream Mining Systems," in *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 3035-3048, Nov. 2010.

[21] S. Souravlas and A. Sifaleras, "Binary-Tree Based Estimation of File Requests for Efficient Data Replication," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 7, pp. 1839-1852, July 1 2017.

[22] Y. Weng, R. Negi, C. Faloutsos and M. D. Ilić, "Robust Data-Driven State Estimation for Smart Grid," in *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1956-1967, July 2017.

[23] V. Loia, V. Terzija, A. Vaccaro and P. Wall, "An Affine-Arithmetic-Based Consensus Protocol for Smart-Grid Computing in the Presence of Data Uncertainties," in *IEEE Transactions on Industrial Electronics*, vol. 62, no. 5, pp. 2973-2982, May 2015.

[24] M. Allalouf, G. Gershinsky, L. Lewin-Eytan and J. Naor, "Smart Grid Network Optimization: Data-Quality-Aware Volume Reduction," in *IEEE Systems Journal*, vol. 8, no. 2, pp. 450-460, June 2014.

[25] D. Garlasu*et al*., "A big data implementation based on Grid computing," *2013 11th RoEduNet International Conference*, Sinaia, 2013, pp. 1-4

[26] W. Saad, H. Abbes, C. Cérin and M. Jemni, "Toward a data desktop grid computing based on BonjourGrid meta-middleware," *2013 International Conference on Electrical Engineering and Software Applications*, Hammamet, 2013, pp. 1-5.

[27] J. C. S. Anjos, W. Kolber, C. R. Geyer and L. B. Arantes, "Addressing Data-Intensive Computing Problems with the Use of MapReduce on Heterogeneous Environments as Desktop Grid on Slow Links," *2012 13th Symposium on Computer Systems*, Petropolis, 2012, pp. 148-155.

[28] C. Moretti, H. Bui, K. Hollingsworth, B. Rich, P. Flynn and D. Thain, "All-Pairs: An Abstraction for Data-Intensive Computing on Campus Grids," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 1, pp. 33-46, Jan. 2010.

[29] R. Zhou, "Data-Intensive Scientific Workflows for Grid Computing with CSCL," *2010 Fourth International Conference on Genetic and Evolutionary Computing*, Shenzhen, 2010, pp. 845-848.

[30] S. Rusitschka, K. Eger and C. Gerdes, "Smart Grid Data Cloud: A Model for Utilizing Cloud Computing in the Smart Grid Domain," *2010 First IEEE International Conference on Smart Grid Communications*, Gaithersburg, MD, 2010, pp. 483-488.

[31] Wenxiang Dou, Jinglu Hu, Kotaro Hirasawa and Gengfeng Wu, "Distributed multi-relational data mining based on genetic algorithm," *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, Hong Kong, 2008, pp. 744-750.

[32] Shahina Praveen M, and G. Narsimha, "SADM: Sophisticated Architecture of Distributed Mining over Grid Infrastructure", International Journal of Computer Science and Electronics Engineering (IJCSEE), vol. 4, Issue. 3, 2016

[33] Shahina Praveen M, and G. Narsimha, "Scaling Effectivity of Research Contributions in Distributed Data mining over Grid Infrastructures", Communications on Applied Electronics (CAE), vol. 3, no. 8, 2015

[34] Shahina Praveen M, and G. Narsimha, "Optimized Clustering with Statistical-Based Local Model for Replica Management in DDM over Grid", Springer, pp. 23-33, 2016

[35] Shahina Praveen M, and G. Narsimha, "Distributed Data Mining Approaches as Services on the Grid Infrastructure", National Conference on Soft Computing and Knowledge Discovery, 2012

6. AUTHOR'S DETAIL

Shahina Parveen Mhas worked as Assistant Professor , Department of ISE, BhageerathiBai 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 in the dept of CSE at JNT University, Hyderabad, India. She has published many papers in both national and international conferences.

Dr. G. Narsimha is working as 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 interanational and nation journals. 7 PhD are awarded and 11 research schcolar are working under him. He is life member of indian society for technical education (MISTE), MIEEE.