AN OVERLAY-BASED DATA MINING ARCHITECTURE TOLERANT TO PHYSICAL NETWORK DISRUPTIONS

S.Sumathi*1, C.Senthilraja*2

*1 PG Scholar (M.Phil-CS), Siri Psg Arts & Science College for Women, Salem, Tamilnadu, India
*2 M.C.A.,M.Phil, Head of the Department (CS), Siri Psg Arts & Science College for Women, Salem, Tamilnadu, India

Abstract—Big data generated from various aspects like online transactions, social websites, logs and search queries is increasing rapidly and thus the demand for data mining has risen as a noteworthy zone. An overlay-based parallel information mining executes completely dispersed information administration and handles processing by utilizing the overlay system, which can achieve high flexibility. The talk incorporates a survey of best in class systems and stages for preparing and overseeing huge information and also the endeavours expected on enormous information mining. Nonetheless, the overlay-based parallel mining structural planning is not fit for achieving data mining administrations if there is an occurrence of the physical system disturbance that is created due to switch/correspondence line breakdowns on the grounds that various hubs are expelled from the overlay system. To get the estimated arrangement and better results, the proposed framework utilizes K-medoids algorithm for cluster formation and overlay based system. Proposed work gives enhancement in terms Energy Consumption in data gathering, reduced delay and Node Coverage

I. INTRODUCTION

Data Mining and Knowledge Discovery are usually defined as the extraction of patterns or models from observed data, usually the ability to explore much richer and more expressive models, providing new and interesting domains for the application of learning algorithms. The wide availability of large-scale data sets from different domains has created a demand to automate the process of extracting valuable information from them. For example, consider Facebook application, we upload various types of information such as text, images and video. The process of effective mining of such data is known as big data mining [1]. Today is the era of Google, the thing which is unknown is searched in Google and within fractions of seconds; we get the number of links as a result. This would be the best example for the processing of Big Data. This Big Data is not any different thing than out regular term data. The Big Data is nothing but a data in an extreme large amount available of heterogeneous, autonomous sources, which get updated in fractions of seconds. Conventional parallel data mining architectures with centralized management schemes lack scalability, which causes bottleneck in the entire system and this leads to decrease in performance of the system as the number of nodes increases [4]. As a remedy for improving scalability, an overlay-based parallel data mining architecture has been proposed. Since all the nodes execute both management and processing functions by using overlay network, this architecture can balance the management load. Furthermore, this architecture also achieves higher service availability against the breakdown of master node because it keeps providing the data mining until the overlay network is disrupted. However, distributed networks are intolerant of (i.e., not resilient to) network failures due to the lack of a centralized infrastructure to manage the frequent joining and departure of an enormous number of nodes. There has been a great discussion on this critical issue. While the distributed network management has been studied to solve this issue in general. In order to construct distributed networks, robust enough against network-failures without a management system the present study was undertaken. Categorically, we fixate on the degree distribution of the distributed networks and propose a method to construct networks predicated on the degree distribution robust against network failures. This survey is limited to the Big Data Mining Applications and Overlay Based Parallel data mining, the rest of this paper is organized as follows: In Section 2, we introduce the basic terminology on data mining, big data and Overlay Based Parallel data mining, the rest of this paper is organized as follows: In Section 2, we introduce the basic terminology on data mining, big data and Overlay Based Data Mining. In Section 3, we describe previous work on parallelization of most well-known data mining and Big Data algorithms in the research community. With each algorithm we present our envisioned overlay-based parallel data mining architecture are discussed, as well as the experimental works.

II. BASIC CONCEPTS AND TERMINOLOGY

In this competitive world, top level management needs to take right decisions at the right time for giving better service to customers and to provide a better organizational image. Decisions based on better analysis results in
increasing profit and decreasing loss. Management is dependent on better analytical and data mining services for this purpose. Fig.1 describes the relationship between each step in the process, the technologies that can be used to complete each step in Microsoft SQL Server. The Data mining offers a wide range of algorithms used for analysis, pattern discovery and prediction. This includes techniques such as association rule mining, decision trees, regression, support vector machines and much more. In the last twenty years, a lot of research has been done on improvising performance of data mining techniques.

From past to present, three different trends have been observed in the data mining process. Most of the sequential algorithms were part of the centralized approach where all data is needed to be stored on a central node, which was the first trend. The second trend was observed in terms of parallelizing centralized algorithms. For parallelization, the two main approaches were used: Task parallelism and Data parallelism. Parallel computing techniques took a boost with the advent of multi core CPUs and cheaper GPUs. A combination of both GPU and CPU resulted in multifold performance benefit.

Due to largeness in size, decentralized control and different data sources with different types, the Big Data becomes much complex and harder to handle. They cannot be managed by the local tools those we use for managing the regular data in real time. For major Big Data-related applications, such as Google, Flicker, Facebook, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. And when Big Data is divided into a number of subsets, and apply the mining algorithms on them, the results of these mining algorithms will not always point us to the actual result as we want when we collect the results together.

Overlay-based parallel data mining is one of the architectures that improve the service availability against server break-downs. In this architecture, all the servers execute both management and processing functions. The overlay network is constructed by all servers and utilized to and processing nodes, which are similar to the master nodes in the conventional architecture. This overlay architecture can keep providing the service even if some nodes are removed from the overlay network. Fig. 2 shows an example of mapping and reduction processes in the overlay-based parallel data mining architecture. When a data processing request is injected, a node that received the request (node A in the Fig. 2) executes a reception function using the overlay network [10]. In other words, the mappers are found by the node using flooding message, where mappers are randomly selected (nodes B, C, and D in the Fig. 1). Then, a mapper that initially finished the mapping process (node D in the Fig 2) becomes a reducer, and it requests to other mappers to transmit the processed data to itself, where the request message can be forwarded by using flooding scheme. When the processed data are received from mappers, the reducer executes the reduction process and outputs the analysed result [24].

In this architecture, since the connectivity of overlay network dramatically affects the service availability of data mining, there are numerous works, which tackled the connectivity issue from the various points of view, example, context- cognizant, graph theory predicated, and intricate network theory predicated overlay network construction schemes [11]. These works make overlay networks, tolerant to minute-scale server breakdowns, but do not consider the sizable voluminous-scale server breakdowns, i.e., a physical network disruption. Therefore, this paper develops an overlay-based parallel data mining architecture that is tolerant to physical network disruption so that data mining is available at all times and at any place.

III. CHALLENGING ISSUES IN DATA MINING WITH BIG DATA.

There are three sectors at which the challenges of Big Data arrive. They are:

• Mining platform.
• Privacy.
• Security
• Design of mining algorithms.

Basically, the Big Data is stored in different places and the data volumes may get increased as the data keeps on increasing continuously. So, to collect all the data stored in different places is that much expensive. Suppose, if we use these typical data mining methods (those methods which are used for mining the small scale data in our personal computer systems) for mining of Big Data, and then it would become an obstacle for it. Though we have super large main memory, the typical methods are required to load the data in main memory [12]. Variety, Volume, Velocity and Accuracy are essential characteristics of big data. Variety, data from multiple
sources inherently possesses many types and different forms like structural, semi structured and unstructured data. Scalability, large volume of big data requires high scalability of its data management and mining tools. Speed of data mining depends on the data access time and efficiency. To maintain the privacy is one of the main aim of data mining algorithms. Presently, to mine information from Big Data, parallel computing based algorithms such as Map Reduce are used. The large data sets are divided into a number of subsets and then, mining algorithms are applied to those subsets, in such algorithms. Finally, summation algorithms are applied to the results of mining algorithms to meet the goal of the Big Data mining. During this whole procedure, the privacy statements obviously break as we divide the single Big Data into a number of smaller datasets. 

An emerging topic in data mining is privacy preserving data mining, the basic idea of privacy preserving data mining is performing data mining algorithms effectively without compromising the security of sensitive information contained in the data. Fuzzy fingerprint is one of the data mining technique that enhances data privacy during data leak detection operation which is based on sensitive data. The main goal of privacy preservation is protecting private data while processing or releasing sensitive information. S. Moncrieff et.al proposes a solution which is based on environmental context to dynamically alter the privacy levels in the smart house. Fabio Borges proposes a privacy preserving protocol for smart metering systems to ensure customers’ privacy and security in the network data. The security concerns have become a major barrier to the widespread growth of cloud computing. Distributed architecture is used to eliminate the risks during data mining based attacks [20].

From the above study, we found that, Data mining process is not so easy and the algorithm used for mining is very complicated. The data needs to be integrated from the various heterogeneous data sources as is not available at one place. Data mining derives its name from the similarities between searching for valuable business information in a large database. For example, finding the linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable one. Both processes require either shifting through an immense amount of material, or reasonably probing it to find exactly where the value resides. The databases of sufficient size, quality and the data mining technology can generate new business opportunities by providing some capabilities.

IV. TECHNIQUES IN DATA MINING PROCESS
A. Decision trees
Tree- shaped structures that represent sets of decisions. These decisions generate the rules for classification of a dataset. Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) are included under specific decision tree methods.

B. Genetic algorithms
Genetic combination, mutation, and natural selection are the process used in Optimization techniques for design based on the concepts of evolution.

C. Nearest neighbor method
The technique that classifies each record in a dataset based on a combination of the classes of the records(s) most similar to it in a historical Sometimes called the k-nearest neighbour technique, where k is the number of neighbours.

D. Artificial neural networks
Non-linear predictive models, which are based on biological neural systems.

At present, on the level of the mining platform sector, parallel programming models like Map Reduce are being used for the purpose of analysis and mining of data. Map Reduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. Improving the performance of Map Reduce and enhancing the real-time nature of large scale data processing have received a significant amount of attention with Map Reduce parallel programming being applied to many machine learning and data mining algorithms. These data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize the model.

V. ASSESSMENT OF PHYSICAL NETWORK DISRUPTION
In this section, we aim to investigate the effect of physical network disruption on the service availability of data mining. Additionally, we confirm the effectiveness of our proposed architecture in comparison with existing architecture that are designed without considering physical network, i.e., neighboring nodes are randomly selected and data blocks are distributed in a random manner. In this evaluation, we show the number of available nodes and number of tasks that are successfully processed in order to verify the effectiveness of the proposed neighbor selection and task allocation schemes, respectively.

Mathematical expressions in previous section are used for our performance evaluation. We suppose that the physical network follows power-law degree distribution, which is a well known fact, and its topology is a tree structure, where the number of nodes including servers and routers is set to 104. The overlay network is constructed by all nodes and follows the bimodal degree
distribution, where the degree of leaf nodes is set to 3. We suppose that a processing task is partitioned into 5 data blocks and the total number of processing tasks is 103. We evaluate the performance of data processing after different types of physical network disruptions occur.

VI. EXPERIMENTAL SETUP

6.1 Result Analysis

We conclude that our results will give the better performance in terms of available nodes, delay and energy consumption.

Fig 3. Number of available nodes in different physical network disruptions.

While the number of available nodes in the existing network represents the lower value, the proposed network achieves maximum number of available nodes regardless of the physical network disruption scenarios.

VII. CONCLUSION

In an Overlay-based data mining architecture, which completely distributes management and processing functions by using overlay based network technologies that can potentially provide scalable data mining in large-scale network. However, but physical network disruption is one of the major issues and that goes on decreases the service availability of data mining and hence increased energy consumption and delay. To overcome with these issues it is important to propose new schemes, hence we proposed Overlay network based on neighbor selection and task allocation schemes. Along with that in order to reduce energy consumption, in proposed system K-Medoids algorithm was used to create optimal number of clusters and based on minimal distance algorithm the service ability of every network was enhanced efficiently for better performance. The results obtained from our work demonstrated the effectiveness of our proposed system in terms of energy consumption, end to end delay and number of alive nodes. Thus, our proposed schemes can be Wireless Sensor Networks are important for monitoring the changes in the environmental conditions so that preliminary cautions can be taken to deal with the problem. Moreover Cluster based Wireless Sensor Network are hierarchical networks where the data extraction from super nodes can do effectively and overlay usage also gives better results in case of decreased service ability from neighbor nodes.

REFERENCES


