Distributed Data Mining in Terms of Grid Computing

K.Ravichandran,B.Gunasekar, K.R.Shankar
PG Student,PG Student,PG Student
Department of Computer application
SRM University.

Abstract—Distributed Data Mining approach deals with mining the distributed data by means of distributed resources. Data distribution and data computationfor solving larger number of problems and application execution that are distributed in nature.Distributed Data mining is a data mining where computation and data is dispersed over multiple independent sites. This paper focus on Grid, which is distributed infrastructure that enables resource sharing with coordination. This framework on distributed data mining with grid is heterogeneous and distributed in nature. Grid Computing share the different applications with computing infrastructure, thus enabling to share the computer resources that results in greater performance and efficiency. A data grid is a grid computing system that deals with the data controlled sharing and management of large amount of distributed data. This paper analyses the current distributed approaches, architecture and design concepts and provides the solution for distributed data mining in grid for the resources.

Keywords— Data mining, distributed data, grid

1. Introduction:

1.1 Distributed Database:

Distributed database stores data in multiple locations which are dispersed over interconnected network of computers. Here the storage is controlled by distributed means. Relationship between data items forms association rules[5]

1.2 Data Warehouse:

Integration of data from one or more sources which results in creation of centralized data which act as repository that is referred as Data warehouse. Data warehouse stores current and historical data. Data warehouse is a database that helps in data analysis as well as reporting. As a summary Data Warehouse is a database that extracts the source data and delivers to dimensional data that helps in querying and analysis for decision making.

1.3 Grid Computing:

Grid Computing is a computer resource collection from different locations. This also referred as peer to peer with distributed computing.

1.4 Parallel vs Distributed Data Mining:

Data Mining is extraction or mining the knowledge from data with larger volumes. In recent years, Data Mining plays a role of attraction towards industry as well as society due to available data in large volumes and making them into useful information.

Distributed Data mining deals with data analysis of distributed computing nodes. In each site it contains their own data source and algorithms. Knowledge must be derived.

Whereas in Parallel approach resources are shared across network, that results in unused capacity for parallel processing of extremely intensive computation.

2. Current distributed techniques:

Data mining is mining or extraction of data for huge volumes of data. This plays vital role in industry as it deals with huge data that helps in useful information and knowledge. Distributed data mining refers to data mining where data distributed in different sites. Each site has its own mining algorithms and data source. Grid is a infrastructure helps in coordination of sharing between the dynamic locations. Association rules forms the basis of relationship between the data and helps in

ISSN: 2278 – 7798

measurements. This analyses which method is better than other based on experiments.

Data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making[2]. Data here stored in fact and dimension table. Transaction using distributed data is one of challenging factor. Thus it maintains privacy because it has information about data source and site-specific information. Architecture as follows

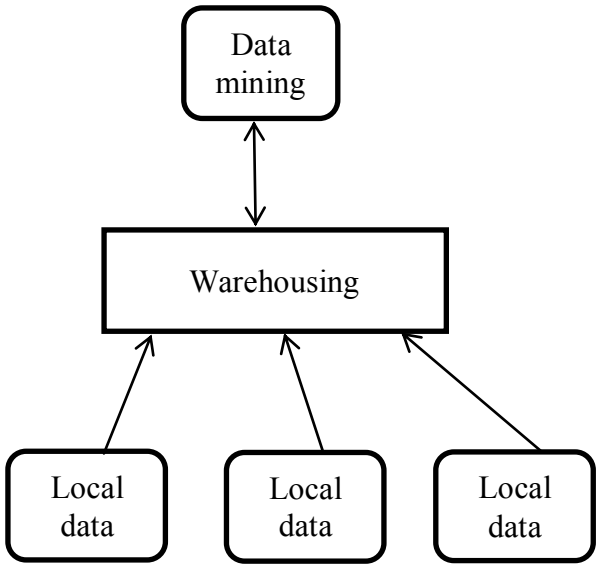

Fig 1 Data warehousing with all the data

Parallel and distributed mining deals with the problems of association rule discovery, using multiple processors in parallel. Parallel computing uses single systems with many processors work on same problem. Distributed computing uses many systems loosely coupled by a scheduler to work on related problems. It is critical to design a parallel algorithms and cost of bringing the parallel database into one site is expensive. Parallelization here refers to achieving scalability and to improve performance. Major parallelism comes under the following categories-Task and Data parallelism.

Distributed Data Mining (DDM)[3] is a branch of the data mining that provides a framework to mine distributed data paying careful attention to distributed data and computing resources.

Grid computing has emerged for coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations in industry and business .Grid computing is a novel computational model, distinguished from conventional distributed computing by its focus on large-scale resource sharing, innovative applications and high-performance orientation. Today grids can be used as effective infrastructures for distributed high-performance computing and data processing. A grid is a geographically distributed computation infrastructure composed of a set of heterogeneous machines that users can access via a single interface. Grids therefore, provide common resource-access technology and operational services across widely distributed virtual organizations composed of institutions or individuals that share resources.

3. Distributed Data mining:

Distributed data mining uses mining algorithms by using the flat files. This focus on the complex and data types which are advanced. The Trend moves from stand- alone applications like centralized towards the distribution[4]. Organizations operate using global markets need to perform distributed data mining integrated with the knowledge from data. Distributed data mining addresses the impact of distribution of users, computational resource and software in data mining process.

Integration of data is expensive and also poor performance and scalability. So these factors lead to the emergence of distributed data mining.

Large amounts of data are collected and warehoused. Data are generated and stored at high speed in local databases, from remote sources. Simulations generation of data are performed. E-commerce and e-business applications store and manage huge databases about products, clients and transactions[7]. We are focusing on storage of data rather than extracting knowledge from the data. Traditional techniques are not feasible for raw data. Data mining helps scientists in hypothesis formation in biology, medicine, physics, and engineering. Companies use data mining techniques to provide customized services and support decision making. Datasets must be shared by large communities of users that pool their resources different sites belonging to a single company, or from a large number of laboratories, plants, or public organizations.

ISSN: 2278 – 7798

Data is distributed using homogenous and heterogeneous across sites. Both data tables are partitions from a single global table. Tables in each site is subsets from global table, they have same attributes. In homogenous it is horizontally partitioned. In heterogeneous it is vertically partitioned, with collections of columns. Each row contains unique identifier for matching purposes.

Some application scenarios are [1]

- Intrusion detection
- Credit card fraudulent detection
- Analysis and prediction in business
- Finance
- Astrology
- Anomoly detection

Data mining uses sophisticated data analysis tools for pattern discovery that is valid, discover the unknown informations and relationship of large data. It is technique for collection and management of data with analysis and prediction for future outcomes.

However it may lead to situations in which the information is located in different physical locations. So here we can mine effectively using the heterogeneous sites. Example astrology information located in different databases and the data comes from two different sources, so combing this will be time consuming and expensive. Information overload on very large data take longer time for obtaining the results. One approach to solve this problem is by parallel algorithms. Parallel and distributed knowledge discovery is based on the use of networks for the mining of data in a distributed and parallel fashion. It manages and analyze data, which is geographically distributed in different data warehouses. This is represented in vertical fashion where instances are called as value.

4. Grid Environment:

Grid is a distributed system with non-interactive workloads that involve large number of files. It distinguish from the conventional computing is that grid is more loosely coupled, geographically dispersed and heterogeneous in nature. Grid helps in co-ordination of resources with sharing in different organizations.

A Data Grid can include and provide transparent access to semantically related data resources that are different managed by different software systems and are accessible through different protocols and interfaces.

Grid applications are [6]

- Simulation on remote supercomputers.
- Visualization of large data.
- Distributed processing of demand data analysis.
- Scientific instruments with remote computers and data archives.

In this section we are going to focus on knowledge grid. This is integration of both data mining and grid computing technique. Data mining tools are integrated with grid services. This performs data mining in larger volumes over grid helps in discoveries and to develop models and business information and services. This offers distributed knowledge discovery as well as parallel on the top of grid middleware.

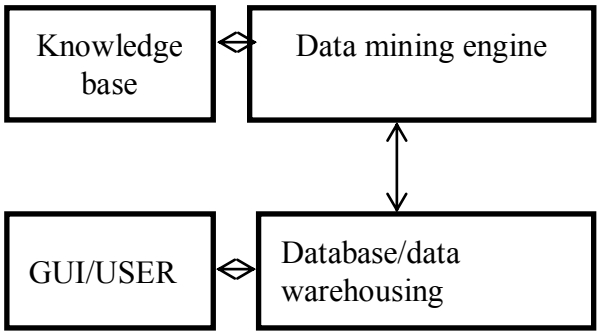

Fig 2 Data Mining Architecture

Knowledge grid Directory services used in monitoring and discovery and as a tool used in knowledge grid. This architecture includes repository of data sources, algorithms and tools knowledge from mining process. Knowledge Grid uses basic grid mechanisms for knowledge discovery based on tools and services. Based on Globus Toolkit, services developed in different ways. Services offered in Globus Toolkit[2] are

Grid Security Infrastructure: Security by authentication and connection to the open network.

Monitoring and Discovery Service: Infrastructure in publishing and information access for grid.

Globus resource allocation manager: Useful for process creation, management, monitoring and allocation of resources.

ISSN: 2278 – 7798

Heartbeat monitors: Helps in detection and reporting for the failure of the process.

Grid File Transfer Protocol: Secure transfer mechanism that allows parallel data transfers with high authentication.

Replica Catalog and Replica Management: Replica Catalog maps between the logical names to one or more files in physical storage ensures to keep track of replicated files. Replica Management is a combination of Replica Catalog and Grid File Transfer Protocol.

Advancement in Grid Applications include Knowledge grid(Knowledge Discovery and Data Mining),Semantic Grid(Ontology and metadata), Grid Services(Information Grid, Data grid, Computation Grid),Grid fabric.

Data mining service have many components tied to Grid Service-service data access, service data element, and service implementation. Service registry, service creation, authorization, concurrency, notification and management come under the associated rules in discovery service. Two types of Grid services are Predictive and Apriori. Apriori Algorithm is used in the grid environment uses similar set of rules which is fast in retrieving data when compared with Predictive.

Growing Grid complexity leads in Knowledge management functions for system as well as user needs and help in knowledge discovery. Also helps in pervasive computing for context awareness and adaption techniques. Useful in self configuring and automation, Discovery and fault tolerance of dynamic resources.

5. Conclusion:

Today many organizations, companies, and scientific research centers manage and produce large amount of complex data and information.Climate data, astrology data and transaction data are just some examples of massive amounts of digital data repositories that today must be stored and analyzed to find useful knowledge in them. This data and information can be effectively exploited if it is used as a source to produce knowledge necessary to support decision making. With a rapid growth in technologies, distributed data mining plays most important role in the process of automation, analysis and knowledge from the large volume of data in the field of engineering, science and medicine.

6. References:

[1] Jiawei Han, MichelineKamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2002

[2] Zakim J, Pan Y, "Introduction: recent developments in parallel and distributed data mining," Journal of Distributed Parallel Database, 2002,pages 123-127 vol-11.

[3] M.Cannataro and D. Talia, "KNOWLEDGE GRID Architecture for Distributed Knowledge Discovery", CACM,2003,Pages 89-93,Vol 46

[4] Albert Y. Zomaya, Tarek El-Ghazawi, OphirFrieder, "Parallel and Distributed Computing for Data Mining", 1999.

[5] M. Z. Ashra, D. Taniar, and K. A. Smith, "A Data Mining Architecture for Distributed Environments", IICS 2002,pages 27-38,Vol 2

[6] M. Cannataro, D. Talia, Semantics and Knowledge Grids:Building the Next Generation Grid, IEEE Intelligent Systems,2004, page 56–63.

[7] H. Kargupta and C. Kamath and P. Chan, Distributed and Parallel Data Mining: Emergence, Growth, and Future Directions, In: Advances in Distributed and Parallel Knowledge Discovery, AAAI/MIT Press,200, page.409–416.

ISSN: 2278 – 7798