IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 41-47
www.iosrjournals.org

Study of Distributed Data Mining Algorithm and Trends

Ms. Rupali Chikhale
[*rupali.chikhale@raisoni.net, G.H. RaisoniInstitue of Information Technology, Nagpur*]

Abstract: *Data mining technology now a days is used as a mode of identifying patterns and trends from large quantities of data.The Data Mining technology use data integration method to generate Data warehouse, where all data put together on one of the site which is treated as a central site, and then data mining algorithm executed against that data to extract the useful Module Prediction and knowledge evaluation. Distributed Data Mining (DDM) is a field which deals with analyzing distributed data and proposes algorithmic solutions to perform different data analysis and mining operations in adistributed manner by considering the resource constraints. Thepaper discusses Distributed Data Mining algorithms, methods and trends to discover knowledge from distributed data in an effective and efficient way.*
Keywords:*Distributed Data Mining, Grid Computing, Multi Agent Systems*.

I. Introduction

The technologicaldevelopments in information and communication (wired and wireless)results into the appearance of distributed computing environments; which is basically originated from knowledge discovery from databases (KDD), also called as data Mining.The developments in data mining leads towards the distributed data mining (DDM) that mines data sources regardless of their physical locations, which consists of several, and different sources of large volumes of data and several computing units. The most common and prominent example of a distributed environment is the Internet, where increasingly more databases and data streams appear that deal with several diversified areas. Also the Internet use as communication media for geographically distributed information systems. Other examples of distributed mining are process monitoring using sensor networks and grids for the system where a large number of computing and storage units are interconnected over a highspeed network.

In short the objective of Distributed Data Mining (DDM) is to extract useful information, knowledge and patterns from distributed heterogeneous data bases. i.e. to compose them within a distributed knowledge base and use for the purposes of decision making. Most of the modern applications classified into the category of systems that uses DDM for distributed decision making. Applications can be of different natures and from different scopes, for example, the combination of data and information is utilized for situational awareness; data mining is also useful in scientific process in order to observe the results of diverse experiments and design a model of a phenomena, intrusion detection, analysis and handling of natural and man-caused disaster.

II. Architecture of Distributed Data Mining(DDM)

At first the paper discusses architecture of data mining system and need of data mining system for distributed system. From the diagram given below the left part of Figure:1 shows the traditional data warehouse - based architecture for data mining. This model of data miningworks by uploading critical data in data warehouse for centralized data mining. But due to long response time, lack of proper use of distributed resources and fundamental characteristics of centralized data mining algorithm,it is not suited for distributed data mining.

The solution to this is to have a distributed application calling for distributed data processing which is controlled by available resources and human factors. Consider an example of Ad hoc wireless sensor network where different sensor nodes are monitoring time critical events. Central collection of data from every node may createheavy network traffic over low bandwidth wireless network and also consume lots of power. A distributed architecture of data mining likely to reduce the communication load and power consumption by different nodes in sensor networks. Hence need Data mining Architecture that provide careful attention toward distributed data, communication and computations and resources so that it can be used optimally. As shown in Figure (Right) the objective of DDM is to perform the data mining operations based on the availability of resources and type of operations.Any site is selected for accessing the data and then performs the operations centrally. While performing this, the sites will be selected as per the storage, computation and communication capacity.

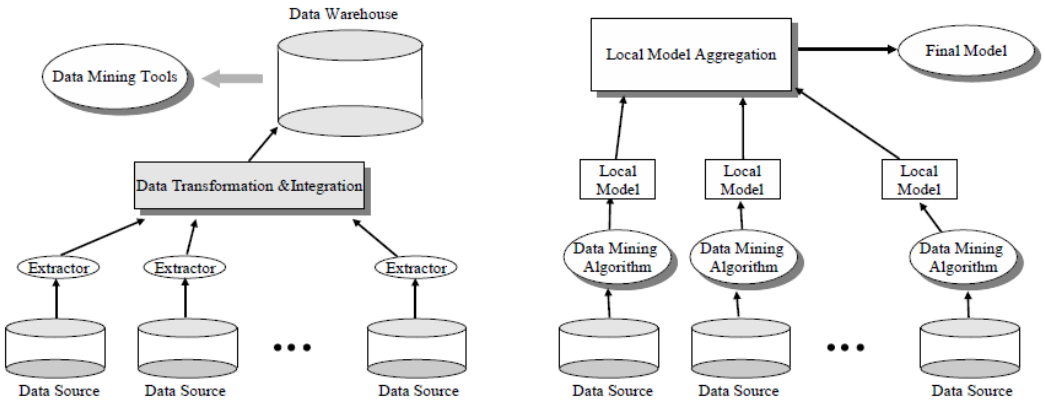

Figure 1. A data warehouse architecture (Left). Distributed Data Mining Framework (Right).

Complicating factors in Distributed Data Mining:

The major issues that affect the performance of Distributed Data Mining are as follows:

Heterogeneous data mining: If the data is heterogeneous then contradiction among the attributes will occur. Also if the data is heterogeneous; local data management model should be integrated into a global model before dealing with the data items.

*Dataconsistency:*Since data is distributed across many sites, it creates a problem of data inconsistency. The modification applied on in local data modelif notreflected to global data model or global database it may affects the final result produced after Data Mining.

*Communication cost:*Communication cost depends on network bandwidth and amount of information transferred. InDistributed Data Mining a cost model should be built.

*Knowledge integration:*Knowledge Integration deals with integrating local results to produce global results. It is the critical step in any Distributed Data Mining. During the integration process the local models should not lose its value in the global range, it must be preserved.

Data variance: In distributed environment data is not static as that of traditional data mining. Along with data the executing environment is also dynamic; hence the Distributed Data Mining algorithm should correctly transfer the time series result and time series related result.

*Privacy preserving:*The main objective of privacy preserving data mining is to develop algorithm for modifying original data in some way so that private data and private knowledge remain private even after mining process. The problem that occurs when unauthorized user derives confidential informationfrom released data which is commonly called database inference problem. Most recent efforts towards addressing the privacy issue are data distortion and cryptographic methods.

III. Distributed Data Mining Algorithms

Distributeddata mining works by analyzing data in a distributed fashion and pays careful attention towards the differences between centralized collection and distributed analysis of data. When the data sets are large scaling up the speed of the data mining task is crucial. Parallel knowledge discovery techniques address this problem by using high performance multicomputer machines. For development of data analysis algorithms that can scale up as we attempt to analyze data sets measured in terabytes and petabytes on parallel machines with hundreds or thousands of processors. This technology is particularly suitable for applications that typically deal with very large amount of data that cannot be analyzed on traditional machines in acceptable times. Most of the DDM algorithm designed upon parallelism they apply on distributed data. The same algorithm is applied on different sites producing one local model per site. All local models are then aggregated producing the final model. Each local model represent locally coherent patterns but lacks the details needed for producing globally meaningful knowledge. Hence DDM algorithm requires centralization of subset of the local data itemto compose it & Minimum data transfer needed for successful DDM algorithm.

Distributed data mining algorithm is classifiedinto three:

- DDM based on Multi Agent System
- DDM based on Meta learning
- DDM based on grid

DDM based on Multi Agent System:

Multi Agent System (MAS) offer architecture for collaborative problem solving in distributed environments. The behavior of agents depends on data from distributed sources. Agents in MAS need to be pro-active and autonomous. Agents perceive their environment, dynamically reason out actions based on conditions and interact with each other. In some applications performance of agents depends on existing domain theory.

Knowledge of MAS is complex and collective, complex in the sense that it is the outcome of data analysis and preexisting domain knowledge. Analysis of data may require advanced data mining for detecting hidden patterns, constructing predictive models and identifying outliers among others. Collective means each analysis is performed by different agents. MAS are mostly used in sensor nodes where there is a requirement for comparison of data at different nodes. Since solution for distributed problem requires collaboration, semiautonomousbehavior and reasoning, there is a perfect synergy existing between MAS and DDM. Agents are used in this system for the following purposes:

*Autonomy of data sources***:** A Distributed Mining (DM) agent may handle access to underlying data source in accordance with given constraints on required autonomy of system, data and model.

*Interactive DDM:*Interactive DDM allows a human user to supervise and interfere with running data mining process.

*Dynamic election of source and data gathering:*DM agents applied to adaptively select data sources according to given criteria such as expected amount, type and quality of data at considered source, actual network and DM server load.

*Scalability of DDM to massively distributed data:*A set of DM agents allow for a divide and conquer approach by performing mining task locally to each of data sites. DM agents aggregate relevant pre-selected data to their originating server for further processing and may evaluate best strategy between working remotely or migrating on data sources.

*Multi strategy DDM:*Data Mining agents uses different data mining technique to choose depending on type of data retrieved from different sites. The learning of MAS is based on multi strategy selection of DM methods.

*Security:*Agent Code and data integrity is a crucial issue in secure DDM. Subverting or hijacking a data mining agent places a trusted piece of software. If DM agents are even allowed to migrate to a remote computing environments methods to ensure authentication and confidentiality of mobile agents have to be applied. Selective agent replication may prevent malicious hostfrom blocking or destroying temporarily residing DM agents.

*Trustworthiness:*DM agents may infer sensitive information even from partial integration to a certain extent, with some probability. It is called inference problem. It enables us to integrate implicit knowledge from different source using commonly held rules of thumb.
It is based on MADM and CAKE architectures**.**

MADM (Multi Agent Data Mining)Architecture:

In distributed data mining, there is a fundamentaltrade-off between the accuracy and the cost ofthe computation.

If interest is in costfunctions which reflect both computation costsand communication costs, especially the cost of wide area communications, we can process allthe data locally obtaining local results, andcombine the local results at the root to obtain thefinal result. But if our interest is accurate result,we can ship all the data to a single node. We assume that this produces the most accurateresult. In general, this is the most expensivewhile the former approach is less expensive, butalso less accurate.

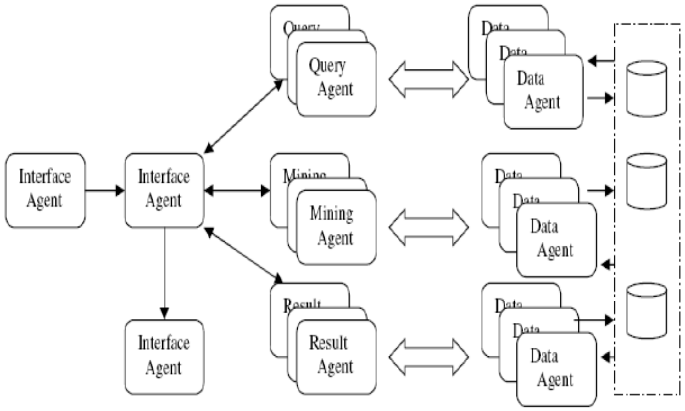

Figure -3: MADM System Architecture

Following are the components of the system:

a. *Interface Agent***:** It interacts with user (or user agent). This agent contains modules for getting input from user as well as methods for inter-agent communication. It asks the user for providing their requirements and provides the user with results after mining. The agent stores the history of user interaction, and user profiles with their specific preferences.

b. *Facilitator agent:*Facilitator agent is mainly responsible for activation and synchronization. It seeks assignments from interface agents and presents the final results to interface agent.

c. *Resource agent:*The resource agent actively maintains the Meta data information about each of the data sources. It also provides predefined and ad-hoc retrieval capabilities. It takes into account the heterogeneity of databases.

d. *Mining agent:*Data Mining agents implement some specific data mining techniques and algorithms. It carries out Data Mining Activity. It captures result of DM and communicates it to result agent or facilitator agent.

e. *Result agent:*Result Agent observes a movement of mining agents, and obtains result from mining agents. It stores details about report templates and visualization primitives that can be used to present result to user.

f. *Broker agent***:** Broker Agent serves as advisor agents that facilitate diffusion of request to agents that have expressed an ability to handle them. It keeps track of names, ontologies and capabilities of all registered agents in the system. It can reply to query of an agent with names and ontology of appropriate agent that has capabilities requested.

g. *Query agent:*Query Agent is generated at each demand of a user. The knowledge module contains Meta data information including local schemas and global schemas. The schemas are used in generating necessary queries for data retrieval.

h. *Mobile Agent:*Mobile Agents travels around their network. It processes the data and sends result back to main host. The main advantage here is low network traffic. There is also requirement for installing agent platform at each site.

4.1.2 CAKE (Classifying Associating and KnowledgeDiscovery) architecture:

The CAKE architecture is based on centralizedParallel Data Mining Agents (PADMAs). CAKE is a 4-tierarchitecture where the Distributed Data Mining isimplemented using parallel Data Mining Agents(PADMAs) using centralized metadata which contains allthe rules of Classification and Association along with itsdata structure details and web interface used to providethe users with the interface to view the result.

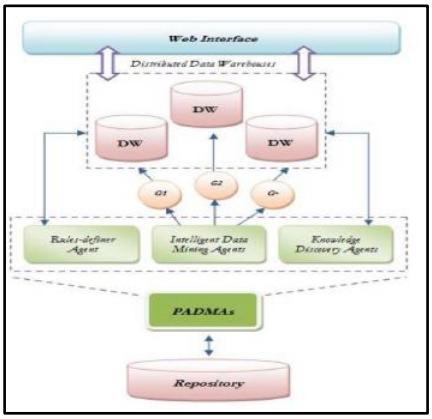

Figure-4 : CAKE Architecture

The PADMAs are executed on thesites where Data Warehouse exists.For improving theperformance and privacy factor; agents are required onlyonce to retrieve their respective configuration.PADMAs are combination of Multi-Purposeagents, which are of three major categories according totheir respective roles at each stage of Data Mining process:

a. *Rule defined agents:*They used to define the Meta- data of Data Warehouse on the basis of rules that are going to be defined by the users. These rules are combinations of conditions and weighted values, defined to perform the operation for evaluating the data and identify the dependency and relationship between attributes to ascertain the hidden knowledge.
b. *Intelligent data mining agents:*The Data Mining Agents are group of agents, which can setup to work on specified set of data on any location with defined rules. These groups of agents will work together to mine the data and compute the desiredresult.
c. *Knowledge discovery and agents:*The Knowledge Discovery Agents are used to determine the final computed results in success or failure along with the explanation on computed data. These decisions are taken on the basis of defined requirements in the repository.
d. The Metadata or agent repository is a database that itself used by PADMAs to perform their respectivejobs. It contains he data warehouse metadata that is necessary or required by PADMAs. All the agents are toaccess the Repository once they are being initialized.

DDM based on Meta learning:

Meta learning system is implemented by JAM system. JAM is a distributed agent based data mining system. It provides a set of learning agents which are used to compute classifier agents at each site. The launchingandexchanging of each classifier agents take place at place at all sites distributed data mining system by providing a set of Meta learning agents which combined the computed models at different sites. JAM is a first system that employs Meta learning as a means to mine distributed databases.

The Meta learning executes the learning process in parallel on subsets of training data sets which improves efficiency. Executing the same serial program in parallel improves time complexity. In Meta learning is in small subsets of data which can easily accommodate in main memory instead of huge amount of data. Meta learning combines different learning system each having different inductive bias, as a result predictive performance is increased. Meta learning constitutes a scalable machine learning method because it generalizes to hierarchical multilevel Meta learning. Most of these algorithmsgenerate classifiers by applying the same algorithms on different database.

First all local classifiers are computed on each local data site by executing learning agents. Then these local computed classifiers are combined with each local classifiers through Meta learning agents. Each local data sites are administered by local configuration file which is used to perform learning and Meta learning task. After computing base and Meta classifiers JAM system executes the modules for classification of desired data sets. The Configuration File Manger (CFM) is used as server which is responsible for keeping the state of system up to date.

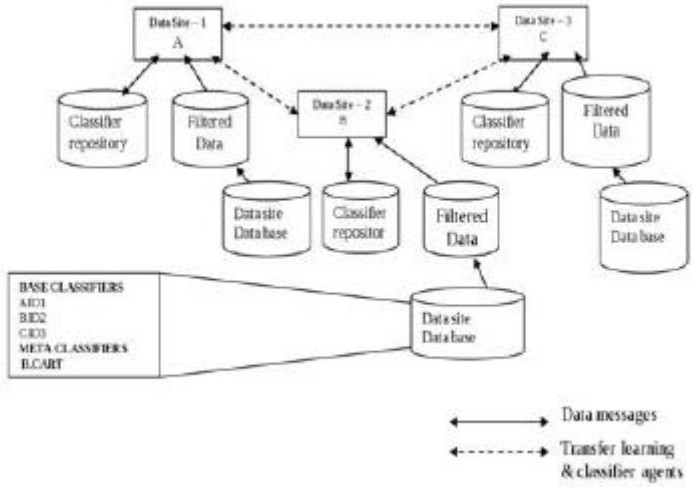

Figure -4 : JAM Architecture

DDM based on Grid:

Grid computing represents evolution of distributed computing and parallel processing technologies. Grid provides distributed computing environment that enables coordinated resource sharing within dynamic organizations consisting of individuals, institutions, and resources. The main objective of grid computing is to allow organizations and application developers to establish distributed computing environments which use computing resources on demand.

Grid computing can control the computing power of a large numbers of servers, desktop computers, clusters and other kind of hardware. Grid-based data mining would allow corporate companies to distribute compute-intensive data analysis among a large number of remote resources. At the same time, it develops new algorithms and techniques that would allow organizations to mine data where it isstored. This is in contrast to the practice of having to select data and transfer it into a centralized site for mining. As centralized analysis is difficult to perform because data is becoming increasingly larger, geographically dispersedand for security and privacy considerations. Here the VEGA architecture for the same is discussed as follows:

VEGA architecture

The design process starts by searching and selecting the resources needed to compose the application. The step is accomplished by means of Data Access Service - DAS and Task Access Service - TAAS tools that analyze the XML Meta data documents which represent the available resources of the participant K-Grid nodes stored into their KMRs. Meta data about resources are stored in Task Management Repositories- TMRs, a local storage space that contains information about resources selected to perform a computation.

VEGA provides following EPMS operations:

a) Task Composition.

b) Task Consistency Checking

c) Execution Plan Generation

Task composition is performed by means of graphical interface which provides a user with a set of graphical objects representing grid nodes and resources. Task composition is implemented by software components such as resource manager, object manager, workspace manager.

Task consistency checking is to obtain a correct result and consistent model of computation. The validation process is performed by means of two components: the model preprocessor and model postprocessor.

In *Execution plan generation* the computation model is translated into an execution plan represented by an XML document. The task is performed by execution plan generator

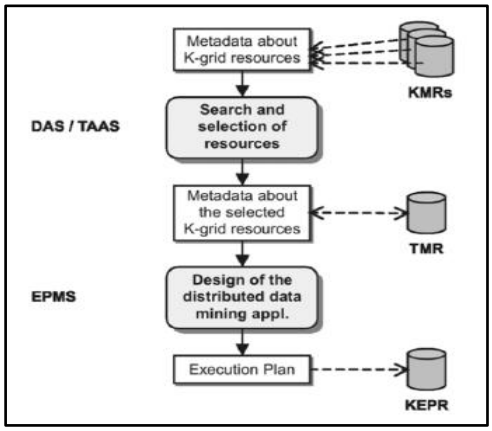

Figure-5: Design process of data miningcomputation

Comparison of DDM Algorithm with Advantages and Disadvantages:

Type of DDM	DDM Frameworks	Advantages	Disadvantages
DDM based on parallel data mining	**MADM**	**Easy to build architecture**	**Agent constrained processing, agent-action ability.**
	CAKE	Clear distinction of functionality between agents	Local data sources have restricted availability due to privacy.
DDM based on Meta learning	JAM	Adaptive learning, Interactive Mining	Learning capability agent needs to be fed up with learning and reasoning algorithm
DDM based on Grid	VEGA	Improved speed of execution compared to any other data mining algorithm.	Data fusion and preparation are difficult

IV. Conclusion

In this paper, data mining algorithms are classified data mining algorithms into three types: DDM based on parallel data mining, DDM based on Meta learning, DDM based on grid. For each classification we have analyzed the prominent data mining framework, advantages and disadvantages of framework and proposed the most suitable application where each framework can be most appropriate. Finally, we analyzed the paper using important parameters required for a good distributed data mining algorithm.

References:

[1]. http://www.csee.umbc.edu/~hillol/PUBS/review.pdf
[2]. http://www.distributeddatamining.org/
[3]. http://www.dfki.de/~klusch/papers/eaai.pdf
[4]. http://link.springer.com/article/10.1007%2Fs10799-012-0124-y#page-1
[5]. http://neuron-ai.tuke.sk/babik/19310438.pdf
[6]. http://dml.cs.byu.edu/~cgc/docs/atdm/Readings/DistributedDM.pdf