Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

(Special Issue for ITECE 2016)

Optimized Distributed Data Mining using Multi Agent System

Hiren Savaliya [1], Niraj Trivedi [2], Yagnik Hathaliya [3]

[1]*Computer Department, A. V. Parekh Technical Institute, Rajkot*
[2]*Computer Department, A. V. Parekh Technical Institute, Rajkot*
[3]*Computer Department, A. V. Parekh Technical Institute, Rajkot*

ABSTRACT

Data Mining is an emerging field which has a great emphasis on the concept of finding patterns and trends from large quantities of information. Data Mining is performed on Data Warehouse which is simply large amount of stored temporal information which is stored at a central site. Then it is accessed and algorithms are run on them to extract useful modules of information from that data. However there are many complex systems available in which single data mining is not feasible, for that Distributed Data Mining is necessary. Distributed Data Mining (DDM) is widely used in commercial world to get effective results faster than simple or single data mining (SDM) technique [1]. DDM simply uses divide and conquer technique in wider terms in which large amount of data are divided into small chunks and passed onto different systems where those individual systems run algorithm and generate their own results. After that all the results are combined together centrally. This solves the problem of Resource Constraint and allows better resource utilization with less expense. DDM Architecture with Multi Agent is a system having multiple agents capable of reaching goals which are very difficult for a single system to achieve.

This paper is the integration of two technologies namely multi-agent system and distributed data mining, also known as multi agent based distributed data mining, in terms of significance, system overview, existing systems, and research trends [1].

Key Terms: Data, Mining, Data, Data Mining, Distributed Data Mining, Multi Agent Data Mining, Performance Optimization

I. INTRODUCTION

Data Mining is a simple process of obtaining knowledge from database and use that information for betterment of future interaction with consumers. The need of data mining comes with some issues a multiple levels that are needed to be addressed first.

- As stated earlier it is clear that to perform data mining, data warehouse is needed and it will require a very large amount of data to be gathered on a single place which raises the issue of security that every data is not sharable. Some companies are not comfortable sharing their customers' data.
- Data can be Homogeneous or Heterogeneous, in both the cases data mining algorithm should work properly. This raises even bigger issue at implementation level. Heterogeneous data cannot be compared with other data and some information is always different from another.
- Data Mining requires the data to be stored in a single place, usually data changes very rapidly and it is to be reflected in Data Warehouse every time so tons to data is needed to be transferred to a single place this will propose an issue of the bandwidth.
- Data Warehouse can be easily considered as a single point of failure. If data stored at warehouse is destroyed somehow then the whole system fails.
- Data stored a data warehouse is accessed from data warehouse from a single system and algorithms are run on those data but even the very high end computers have a limited amount of computing power. It takes lot of time to produce some fruitful output after processing the whole data.

Security and bandwidth both are major issues with many companies face; organizations don't want to share the data itself to any third party organization, whereas organizations may be willing to release the results with others. Remaining sections of paper are organized as follows. In Section II we describe Distributed Data Mining Overview. In section III we describe Agents and their need. In Section IV we describe Agent based Distributed Data Mining. In Section V we describe Multi Agent based Distributed Data Mining. In Section VI we describe Other Issues and Future Scope.

II. DISTRIBUTED DATA MINING

Data Mining deals with greater problems stated earlier like scalability and computing resources. DDM is a branch of data mining field offering framework to distributed data over numerous locations and also pays careful attention to computing resources. Distributed data can be homogeneous where the global table can be horizontally partitioned because all the attributes are same in data whereas in heterogeneous data table must be partitioned vertically according to the attributes present in the data; each site contains collection of columns. Also each tuple in all sites are assumed to have a unique identifier to provide column matching facility useful to combine results from each site.

DDM provides an abstract architecture that allows multiple sites to compute their own results independently using own computing power and storage capability. DDM can work in two manners namely synchronous and asynchronous. In Synchronous DDM there is a Supervisor which assign task to each site and after that result at each site is combined at central supervisor and then that combined result is distributed to all the sites. In Asynchronous DDM each site works as same peer no supervisor is present there and each site runs their own work and after that partial results are passed to each other by simple message passing to one another.

In various scenarios DDM is very helpful that are stated as below:

- System having multiple branches spread over different physical locations which can communicate by using message passing only.
- Transferring large amount of data between sites is expensive.
- Sites having resource constraint.
- Sites having issues regarding privacy of their data.

III. AGENT BASED DISTRIBUTED DATA MINING (ADDM)

DDM is a compound system focusing on the distribution of resources via the network and data mining processes. DDM should be scalable as the system configuration is subject to change with respect to time which proposes great details of software issues like, robustness and extensibility. So that agents are devised with the capabilities of working regardless of those issues.

Autonomous agent can come in handy that can perform multiple tasks regardless of configuration available. Upon applying, the agent studies environments and configuration of the system and generates an execution plan accordingly [2], [3], [4], [5], [6], and [7] discuss the benefits of deploying agents in DDM systems.

Agents can be considered as smaller programs that can be transferred across the network in DDM which eliminates the situation where bulk of data was needed to be transferred in the case of DM. For example, mining agent a1, located at site s1, posses algorithm alg1. Data mining task t1 at site s2 is instructed to mine the data using alg1 [1]. In this setting, transmitting a1 to s2 is a probable way rather than transfer all data from s2 to s1 where alg1 is available. Above example shows advantage of using agent in DM.

A DM Agent can be modularized so that it can be managed according to underlying data. A DM Agents are proactive which can further lessen the human interaction with the system by providing solutions along with the result. That may eliminate some intermediate results and directly provide more matured results. DM Agent can be applied to pass on with different parameters which helps agent to adapt according to the environment. DM Agent can work with users and take feedbacks that will allow DM agent to get more accurate over the time which in turn produce more accurate results over time. Some disadvantages may also arise that any DM Agent can be third party issued software, which is given unsolicited access to all underlying data. Data Security can be a major issue in Secure DDM.

ADDM can be viewed as per shown in figure 1 [1]. That is a generalized model which is a simple request response model. Data communication can be done between different set of components. Basic components of ADDM are described as follows:

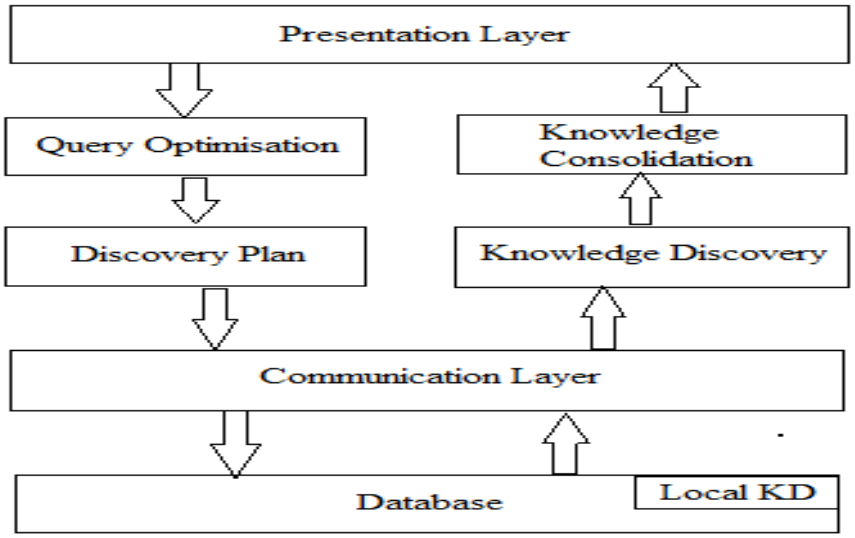

Figure 1. ADDM System Overview

Database is the base layer in which all the data is stored. ADDM works on distributed manner which can be hosted by Online Servers, Data Streams, web pages etc. It works as a knowledge base for the system from where the data is to be gathered. **Communication Layer** provides some services which represents database into a logical manner like it can be represented in terms of their data types, mining algorithms to be used, data schemas etc. **Presentation** is used to hide background implementation complexity from users and just showing them results in forms of simple messages, graphs and diagrams as per the requirement but to do that it needs to use different other layers. A **Query Optimizer** is used to optimize user requests and determine the resources needed to satisfy that query. It also determines the task of whether to perform parallel operations or not. **Discovery Plan** assigns each resource a particular task. In short it divides bigger tasks into smaller sub-tasks and allocates accordingly. **Local Knowledge Discovery** performs local basic mining tasks which defines small patterns between the data and decreases need for transferring very large information. **Knowledge Discovery** is the place where actually mining is performed and larger patters are found after executing queries on data source. **Knowledge Consolidation** gathers results from all the sources and finally gives a consolidated, compact and meaningful result.

IV. MULTI AGENT BASED DISTRIBUTED DATA MINING

In DDM we need to define a balance between accuracy and cost of the computation. If the cost is the factor to be considered which includes communication cost and computation cost then processing of all data can be done at local level and local results can be obtained. Then those results are combined to generate final result but this may generate little less accurate results. If our requirement is to obtain more accurate results then all data is needed to be shipped at a single node and then DM is done at a single site. The assumption is that it will produce the most accurate results but shipping all the data to a single location can be very cost effective and takes more time than the former discussed approach.

Most MADM adapts similar architecture as in figure 2 [20] with the most common structural components. They use KQML or FIPA-ALC, which are standard agent communication models facilitating communication between agents. There are many agents that can be used in mining but most common agents used while DM is defined as below:

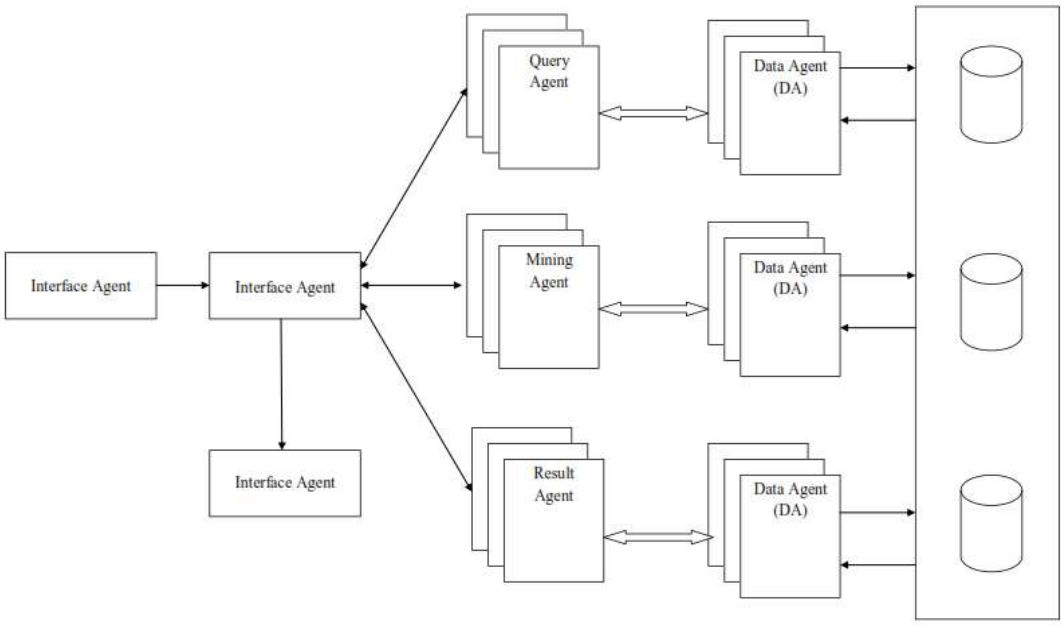

Figure 2. General Architecture of MADM

4. 1 User Agent (Interface): This layer interacts with user and interacts with them to gather the requirements and required factors to be considered while performing DM operation. Also it provides user with the result of the mining operation performed in the background. It also converts the result in form of some user friendly messages graphs, etc. It also contains modules for inter agent communication and taking inputs from user. Some interface agents even stores user profiles and their past search history to provide their user with specific preferences.

4. 2 Manager Agent (Facilitator): As per the name this agent provides management functions between the agents. It works as a facilitator between different agents for communication and synchronization. This agent manages overall process and assigns task to each agents after receiving task from user agent. After that it seeks services of group of agents and prepares the final result which is to be passed on to user agent. We can say that this agent is responsible for inter-agent communication [20]. The sequence of tasks to be executed is created from specific "ontologies" stored in the knowledge module using a rule-based approach. The agent task includes getting relevant data sources, requesting services from agents, generating queries, etc. The knowledge module also contains meta-knowledge about capabilities of other agents in the system.

4. 3 Data Agent (Resource Agent): This maintains meta-data about the information available about each data source. It fetches necessary data sets from the data available according to need of mining agent for specific mining operation. It also manages issues regarding vertical and horizontal partitioning of data tables in cases of heterogeneous and homogeneous data respectively. It then devices queries to fetch the data from database and after the process is done, the results of process are passed back to the manager agent.

4. 4 Mining Agent: Actual data mining techniques, algorithms and tasks to be performed to fetch the knowledge from multi factored data is done by mining agent. Interface module is responsible for inter agent communication. Process module contains methods to initiate and execute data mining task, getting the result of mining and communicating the result back to result agent or manager agent.

4. 5 Result Agent: Result agent works with mining agents and by observing work of mining agent it produces knowledge as a result. There can be more than one result agent working with further more mining agent. Each gets their own result and then all those partial results are combined as a final result and passed to facilitator. The knowledge module provides with the templates in which results can be presented to users.

4. 6 Broker Agent: Broker agent a.k.a. matchmaking agent works as an advisory module that has data regarding capabilities of each agent in the system that are expressed by the agent itself. So broker agent may give the task to each agent as per their capabilities or recommend them to the facilitator about the capabilities of

particular agent. For that a new agent must register themselves with their capabilities to the broker so that broker agent can advertise them. So that a new agent can be a part of the multi agent system already available.

4. 7 Query Agent: These agents are generated according to the demand of user. It handles each user query. To generate a Query agent we need a Knowledge module which contains meta-data about information available which includes local schemas and global schemas.

4. 8 Ontology Agent: As in MADDM database used is often more complex and distributed, this agent is needed to store overall knowledge about ontology of the data available like schemas present, their structure, and relationship between schemas. This will help Query agents to generate proper queries as per the task required to be done.

4. 9 Mobile Agent: Some tasks require mobility of agents. This agent is the type of agent which travels around the network processing data and sends back the result to a central host machine. This is a cost effective approach which requires just transferring the agent instead of transferring lots and lots of data to another system.

4. 10 Local Task Agent: In most cases data agent is located in each machine locally, so this works as a local task agent. It performs local tasks and then submits the information to facilitator (MA). A local agent only work on local database and retrieve information from local database only and then returns the result to the system.

4. 11 KDD System agents: Some complex systems contains other agent to help main agents like Knowledge Discovery from Database (KDD) System Agents, This includes data preparation and data evolution. These Agents are:

4.11. 1 Pre Processing Agent: It prepares the data for data mining and performs cleansing of data before applying it to the data set for data mining task.

4.11. 2 Post Data Mining Agent: This evaluates the task performed by Data Mining agent and generates post information about quality of task done. Which would be helpful to refine the agents performance for further mining tasks

V. OTHER ISSUES AND FUTURE SCOPE

The interface and combination of two technologies have new challenges to explore. Considering various ingredients for the combination is becomes a key to hastily improve the systems development process and usability, There are different perspectives to examine which is depicted below.

5. 1 Research prospective: It mainly deals with applications of the system that require a data mining technology to pay careful attention to the distributed computing, communication, and storage. The future scope regarding research prospective is an approach to develop MADDM is using Swarm intelligence which is very much associated to intelligent agents. Hence researchers pay concentration to the possibility to implement DDM systems with swarm intelligence. Swarm intelligence (SI) is the collective behavior of decentralized, self-organized systems, natural or artificial. The concept is employed in work on artificial intelligence [5] [19].

5. 2 User prospective: It mainly deals with the human-computer interaction and these issues in DDM have some distinctive challenges. The system level support requirement to do the group interaction, Introduce the alternate interface such as mobile devices, and The main challenges is regarding the security while you concern about the user perspective.

REFERENCES

[1] Rao, Vuda Sreenivasa. "Multi agent-based distributed data mining: An overview." International Journal of Reviews in Computing 3 (2009): 83-92.
[2] Baik, Sung Wook, Jerzy Bala, and Ju Sang Cho. "Agent based distributed data mining." Parallel and Distributed Computing: Applications and Technologies. Springer Berlin Heidelberg, 2004. 42-45.
[3] Giannella, Chris, Ruchita Bhargava, and Hillol Kargupta. "Multi-agent systems and distributed data mining." Cooperative Information Agents VIII. Springer Berlin Heidelberg, 2004. 1-15.

[4] Gorodetsky, Vladimir, Oleg Karsaeyv, and Vladimir Samoilov. "Multi-agent technology for distributed data mining and classification." Intelligent Agent Technology, 2003. IAT 2003. IEEE/WIC International Conference on. IEEE, 2003

[5] Parunak, H. Van Dyke, and Sven A. Brueckner. "Engineering swarming systems." Methodologies and Software Engineering for Agent Systems. Springer US, 2004. 341-376.

[6] Klusch, Matthias, Stefano Lodi, and M. Gianluca. "The role of agents in distributed data mining: Issues and benefits." Intelligent Agent Technology, 2003. IAT 2003. IEEE/WIC International Conference on. IEEE, 2003.

[7] Klusch, Matthias, Stefano Lodi, and Gianluca Moro. "Issues of agent-based distributed data mining." Proceedings of the second international joint conference on Autonomous agents and multiagent systems. ACM, 2003.

[8] SBailey, Stuart, et al. "Papyrus: a system for data mining over local and wide area clusters and super-clusters." Proceedings of the 1999 ACM/IEEE conference on Supercomputing. ACM, 1999.

[9] V. Gorodetskiy. Interaction of agents and data mining in ubiquitous environment. In Proceedings of the 2008 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'08), 2008.

[10] Panait, Liviu, and Sean Luke. "Cooperative multi-agent learning: The state of the art." Autonomous Agents and Multi-Agent Systems 11.3 (2005): 387-434.

[11] Raymond S. T. Lee and James N. K. Liu. ijade web-miner: An intelligent agent framework for internet shopping. IEEE Transactions on Knowledge and Data Engineering, 16(4):461–473, 2004.

[12] Wooldridge, Michael, and Nicholas R. Jennings. "Agent theories, architectures, and languages: a survey." Intelligent agents. Springer Berlin Heidelberg, 1994. 1-39.

[13] Hassan Zaid, Syed Zahid, et al. "ADMI: A Multi-Agent Architecture To Autonomously Generate Data Mining Servrces." (2004).

[14] Gorodetskiy, V., et al. "Agent-based service-oriented intelligent P2P networks for distributed classification." Hybrid Information Technology, 2006. ICHIT'06. International Conference on. Vol. 2. IEEE, 2006.

[15] Kargupta, Hillol, Ilker Hamzaoglu, and Brian Stafford. "Scalable, Distributed Data Mining-An Agent Architecture." KDD. 1997.

[16] Hershberger, Daryl E., and Hillol Kargupta. "Distributed multivariate regression using wavelet-based collective data mining." Journal of Parallel and Distributed Computing 61.3 (2001): 372-400.

[17] L. Cao, C. Luo, and C. Zhang. Agent-Mining Interaction: An Emerging Area. Lecture Notes in Computer Science, 4476:60, 2007.

[18] F. Bergenti, M. P. Gleizes, and F. Zambonelli. Methodologies And Software Engineering For Agent Systems: The Agentoriented Software Engineering Handbook. Kluwer Academic Publishers, 2004.

[19] Bansal, Jagdish Chand, et al. "Balanced artificial bee colony algorithm." International Journal of Artificial Intelligence and Soft Computing 3.3 (2013): 222-243.

[20] Pandey, Trilok Nath, Niranjan Panda, and Pravat Kumar Sahu. "Improving performance of distributed data mining (DDM) with multi-agent system." IJCSI International Journal of Computer Science Issues 9.2 (2012): 73-82.