An Analysis on Multi-Agent Based Distributed Data Mining System

R.HEMAMALINI *, Dr.L.JOSEPHINE MARY **

* Research scholar St.Peter's University
** Professor & HOD, MCA Dept. Sri Ram Engineering College,Veppampet

Abstract- The Distributed Data Mining **(**DDM) is a branch of the field of data mining that offers a framework to mine distributed data paying careful attention to the distributed data and computing resources. Usually, data-mining systems are designed to work on a single dataset. On the other hand with the growth of networks, data is increasingly dispersed over many machines in many different geographical locations. Also, even as most practical data-mining algorithms operate over propositional representations are known as first order learning. In existing system, the concept of knowledge is very important in data mining. In order to get the correct knowledge from the data mining system, the user must define the objective and specify the algorithms and its parameters exactly with minimum effort. If the data mining system produces large number of meaningful information by using a specialized data mining algorithm like association, clustering, decision trees etc., it will take more time for the end- users to choose the appropriate knowledge for the problem discussed. Even choosing the correct data mining algorithm involves more time for the system. Developing a data mining system that uses specialized agents with the ability to communicate with multiple information sources, as well as with other agents requires a great deal of flexibility. The main objective of this paper titled on "**An Analysis on Multi-Agent Based Distributed Data Mining System**" describes the knowledge integration, Knowledge Integration in Distributed Data-Mining and Heterogeneous vs. Homogeneous Data-Mining, a literature survey of Multi-Agent Based Distributed Data Mining System, a Model Of Multi –Agent System Based Data Mining, the improving DDM performance by combining distributed data mining and multi-agent system and Data Mining using Multiple Agents.

Index Terms- MAS – Multi agent System, DDM- Distributed Data Mining, DMA – Data Mining Agent , ILP- Identify Local Pattern, KI – Knowledge Integration.

I. INTRODUCTION

Data-mining or Knowledge Discovery is concerned with extracting knowledge from databases and/or knowledge bases using machine learning techniques. The first order learning is to enables us to explore the aspects of knowledge integration and theory refinement which do not appear in propositional systems. Software has the response to the problem of using the vast amounts of information stored on networked systems. There are many types of software agent; however, agents are typically thought of as being 'intelligent' programs which have some degree of self-sufficiency. We intend to design an open, flexible data-mining agent. A group of these agents will be able to co-operate to discover knowledge from distributed sources.

In DDM, one of the two assumptions is commonly adopted as to how data is distributed across sites: homogeneously (horizontally partitioned) and heterogeneously (vertically partitioned). Both viewpoints adopt the conceptual idea that the data tables at each site are partitions of a single global table. In the homogeneous case, the global table is horizontally partitioned . The tables at each site are subsets of the global table; they have exactly the same attributes. In the heterogeneous case the table is vertically partitioned, each site contains a collection of columns as sites do not have the same attributes. However, each tuple at each site is assumed to contain a unique identifier to facilitate matching. It is important to stress that the global table viewpoint is strictly conceptual. It is not necessarily assumed that such a table was physically realized and partitioned to form the tables at each site. The development of data mining algorithms that work well under the constraints imposed by distributed datasets has received significant attention from the data mining community in recent years. Local computation is done on Improving DDM Performance by Combining Distributed Data Mining and Multi-Agent System.

The Multi-Agent Learning is a number of co-operative distributed learning systems have been produced. Each agent has a data-source and a clustering algorithm. The agents propose rules which characterize the data seen and review other agents' proposals. Eventually consent about the knowledge extracted from the data is reached. Each agent has local knowledge and either an inductive or deductive learning algorithm. Agents attempt to solve a problem-solving task by either retrieving the knowledge required, or by using learning to acquire it. Failures result in communication with other agents which are passed sub-goals, which are then treated as tasks.

There are three ways learning can occur when data is distributed. These relate to when agents communicate with respect to the learning process:

• The first approach gathers the data in one place. The distributed database management systems is used to provide a single set of data to an algorithm is an example of this. The problem with such an approach is that it does not make efficient use of the resources usually associated with distributed computer networks.

• The second approach is for agents to exchange information at the same time as learning on local data. This is the approach taken by and no revision or integration is needed, as the agents are effectively working as a single, tightly coupled, algorithm over the entire data. This restricts the agents to using learning algorithms that have been specially modified to work in this way. Thus the main disadvantage with this approach is that it does not allow the use of future learning algorithms.
• The third approach is for the agents to learn locally, and then to share their results, which are then refine it and integrate it by other agents in their own search for data and knowledge. This model permits the use of standard algorithms, and also allows inter-operation between different algorithms. The main problem here is how to integrate the local results. We are adopting the latter approach, as it provides distributed processing together with flexibility in taking part in future algorithms.

Why Agents For DDM?

The following are in the support of using agents in DDM:
Autonomy of data sources: A DM Agent may be considered as a modular extension of data management system to handle the access to the underlying source with the given constraints on the required self-sufficiency of the system, data model.
Interactive DDM: Assisting agents may limit the amount of human user who supervises and interferences with the running data mining process.
Active Selection of sources and data Gathering: One challenge for intelligent DM agents acting in open distributed data environments in which the DM checks the data sites available and their content may change at any time. DM agents may be applied to select data sources according to the given criteria.
Scalability of DM to massive distributed data: To reduce network and DM application server load ,DM agents may be let to transfer the local data sites in a DDM system, which performs mining tasks locally.
Multi-Strategy DDM: For some complex application setting an appropriate combination of multiple data mining technique may be more beneficial than applying one particular technique.DM agents learn their action depending on the data retrieved from different sites and mining tasks to be done.
Collaborative DM: DM agents may operate independently on the data they may have gathered at local sites and then combine their respective models .DM agents may talk with each other and jointly plan a solution for the problem.
Security and Trust worthiness: Agent based DDM has to manage with the problem of ensuring data security and privacy. Agent code and data security play a big role in DDM system by which Data mining agents may get information even form partial integration to a certain extent and with some probability.

The following sections describe the relationship between theory revision, knowledge integration, Knowledge Integration in Distributed Data-Mining and Heterogeneous vs. Homogeneous Data-Mining, a literature survey of Multi-Agent Based Distributed Data Mining System, a Model Of Multi –Agent System Based Data Mining, the improving DDM performance by combining distributed data mining and multi-agent system and Data Mining using Multiple Agents.

2.1. Knowledge Integration

If this approach is adopted, then the local theories have to be integrated. It must be remembered that each agent's local results are correct for that agent's view of the data[4]. Thus the fundamental problem is to compare local theories with previously unseen data, i.e. other agent's data. This data is of course summarized by the results produced by the other agents. Both theory revision and knowledge integration as resources in helping to produce a single global result is taken care off. There appear to be three alternatives:

- The first approach involves a supervisor agent, in which one agent attempts to integrate all the local theories. However, this may lead to the transmission of large amounts of the original data, in order to test the accuracy of the new knowledge.
- The second is a democratic version of the supervisor approach, with the agents working as a team to integrate their local results.
- The third approach involve search agent taking other agent's theories and integrating them locally. It can be seen from this discussion, the knowledge integration is a fundamental one in creating a distributed data-mining architecture.

2.2. Knowledge Integration in Distributed Data-Mining

Individual agents will produce new knowledge based on their discovery goals, and the view of the distributed data. This individual knowledge will have to be integrated, so that it accounts for all views of the data. Theory Refinement and Knowledge Integration are related techniques. Theory refinement involves revising a theory with respect to new training examples. Knowledge integration involves combining two theories into a single unified theory. However, the learning techniques used for both are similar, and ILP algorithms in particular appear to make little distinction between revising clauses in response to new examples, and combining two sets of clauses and then revising them to fit the existing example.

Our initial decision to use an ILP learning algorithm [5] was based on the insight that many ILP algorithms provide support for theory revision and knowledge integration. An agent based on such an algorithm could be used to both induce and integrate knowledge. However, there is far more to consider than simply choosing an appropriate algorithm. Firstly, we must consider the nature of the data and discovery goals given to each agent. Then it must be considered when the agents should co-operate: either before, during or after learning. Finally, if it is decided that the co-operation should take place after learning, it should be decided how the agents will integrate the set of local results in order to reflect a global solution to the data-mining goal.

2.3. Heterogeneous vs. Homogeneous Data-Mining

If each agent in the system is associated with a single database, then there are two basic types of interaction to

consider. If each agent has the same discovery goal, and the same database scheme, though normally with different values, is referred to this as homogeneous data-mining. In this case the problem for the agents is to resolve partial results based on each partial view of the whole of the data.

If each agent has a different database and discovery goal, then the agents may use theories found by other agents as sub-theories. For example, if one agent learns a definition of parent, then a second agent might use this in its definition of grandparent. We refer to this as heterogeneous data mining.

3. LITERATURE SURVEY

Data mining**,** also called knowledge discovery in databases, in computer science, is the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence such as neural networks and machine learning with database management to analyze large digital collections, known as data sets[1].

Distributed data mining (DDM) is an important research area. One of the approaches suitable for the DDM is to select relevant local patterns from the distributed databases. Such patterns, often called prototypes, are subsequently merged to create a compact representation of the distributed data repositories. To assure obtaining homogenous prototypes the feature selection requires collaboration of agents [2]. The task of distributed data mining is to extract and integrate knowledge from different sources. Solving such tasks requires a special approach and tools, different from those applied to learning from data located in a single database. One of the approaches suitable for the DDM is to select relevant local patterns from the distributed databases [3]. Knowledge integration (KI) is the process by which organizations identify, acquire, and utilize knowledge from external sources. Existing studies on KI have either researched what knowledge organizations gain from which sources or what processes are needed to do that [4]. With the rapid advancement of information technology, flood of digital data collected by business, government, and scientific applications need analyzing, digesting, and understanding. Scalability has become a necessity for data mining algorithms to process large data more effectively and extract insightful information from large data [5]. In cooperative multi-agent systems, roles are used as a design concept when creating large systems, they are known to facilitate specialization of agents, and they can help to reduce interference in multi-robot domains. The types of tasks that the agents are asked to solve and the communicative capabilities of the agents significantly affect the way roles are used in cooperative multi-agent systems [6]. MAS represent a promising approach for complicated data mining tasks due to its configuration flexibility and parallel processing. This is exploited in encapsulation of computational modeling, pre-processing and post-processing methods reflecting industrial standard specifications. The ontology of computational agents' capabilities is proposed and included in the model. The agent, which implements the DL model, supports matchmaking, system integrity checking and automated creation of the system[7].Massive amounts of data that are often geographically distributed and owned by different organizations are being mined DM infrastructures to fully exploit the benefit of the knowledge contained in these very large data repositories[8]. A communication mechanism between multi-agents was established in the system, in order to achieve two approaches to deal with information processing in a distributed way. An algorithm for each approach has been modified from the original forms of decision tree building/ making algorithms so that it can work under multi-agents computing environments as well as single computing environments [9]. An ADDM system concerns three key characteristics: interoperability, dynamic system configuration, and performance aspects, discussed as follows. Interoperability concerns, not only collaboration of agents in the system, but also external interaction which allow new agents to enter the system effortlessly [10].

4. AN ANALYSIS ON MULTI-AGENT BASED DISTRIBUTED DATA MINING SYSTEM

One or more agents per network node are responsible for examining and analyzing a local data source. In addition, an agent may query a knowledge source for existing knowledge. The agents communicate with each other during the discovery process [6]. This allows the agents to integrate the new knowledge they produce into a globally coherent theory. In addition, a supervisory agent, responsible for coordinating the discovery agents may exist. A graphical interface allows the user to assign agents to data sources, and to allocate high level discovery goals. It allows the user to critique new knowledge discovered by the agents, and to direct the agents to new discovery goals, including ones that might make use of the newly discovered knowledge. The following Agents are employed in Data Mining:

Categorization Agent:-Categorization agent classified terms in-lieu of format. Such subject have separate category in various form; which category is valuable for any instance and it follow that event summarization.

Objective:-three level of category maintain (high, middle, low) it perform under based.

Data mining agent*:* - A data mining agent [7] is a software program built for Pre-purpose of finding information efficiently. It is a type of intelligent agent that operates valuable information to find the relationship between different pieces of information. It is a type of agent to detect major trend changes

Objective: - DMA finding suitable new pertaining information efficiently

***Filtering Agent*:**-A Filtering agent filter required information; it will check contents and attachment task both include in On End of Data Filters (smart screen, intelligent content filtering, file filtering, Multiple AV Scan)

Objective: - Searching message and check filtering reinforcement.

Information Agent:-An individual or company that is charged with explaining the various transaction of another party to anyone who need to know

Objective: - Delivery of commodity to give or keep Information.

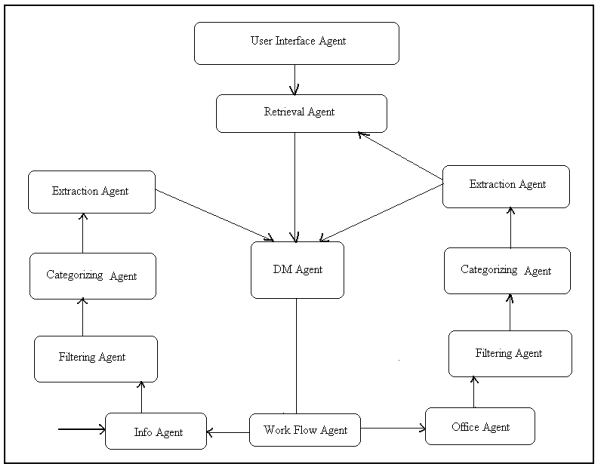

Figure : 1 A model of Multi –Agent system Based data Mining

User Interface Agent: - The user interface is the space where interaction between humans and machines occur effectively and control of the machine feedback from machine, operate making decisions.
Objective: - Interaction between human and machines operate all levels.
Interface Agent: - Interface agent to be a program that can also affect the objects in a direct manipulation interface, but without exploit instruction from user. The interface agent reads input that the user presents to the interface and it can make changes to the objects the user sees on the screen.
Objective: - Defining of objects it depends on input and output.
Extraction Agent:-Extraction agent extracts set of information regarding object and is used its dilute for any such information for further needs. Any information fetch it explain every criteria of objects. Collect complete information about the concept. It shows object detailed in well-mannered.
Objective:-Launch every bit of information in detailed manner.
Retrieval Agent:-In retrieval agent retrieve information which one has been extracting. Whether this agent executes information using data sets and visualization effects etc. it will display exactly induced information as well. Such functional value used in this format for retrieving procedure.
Objective: - Execute information according to demand specification.
Office Agent: - According to portal, office agent will choose information where it finds suitable. Different types of office agent occur; they work and uses are totally different. It depends on functioning values and does their own business. It is very purposeful ordinal nature of point.
Objective:-According to the work through this agent is altered as per work nature.
Workflow agent:-the workflow agent can be configured for polling on demand processing. When configured for polling, the workflow agent periodically polls (quires) in the database work queue for batches of task to be processed. Workflow depends on polling.
Objective: - Priority techniques use to analyze workflow.

4.1. Improving DDM Performance By Combining Distributed Data Mining And Multi-Agent System

Typically communication is a bottleneck. Since communication is assumed to be carried out exclusively by message passing, a primary goal of many DDM methods in the literature is to minimize the number of messages sent. Some methods also attempt to load-balance across sites to prevent performance from being dominated by the time and space usage of any individual site. As pointed out in , Building a monolithic database, in order to perform non-distributed data mining, may be infeasible or simply impossible in many applications. The cost of transferring large blocks of data may be unaffordable and result in very inefficient implementations. Surveys provide a broad, up-to-date overview of DDM touching on issues such as: clustering, association rule mining, basic statistics computation, Bayesian network learning, classification, and the historical roots of DDM. The collection describes a variety of DDM algorithms (association rule mining, clustering, classification, preprocessing, etc.), systems issues in DDM (security, architecture, etc.), and some topics in parallel data mining. Survey discusses parallel and distributed association rule mining in DDM. Survey discusses a broad spectrum of issues in DDM and parallel data mining and provides a survey of distributed and parallel association rule mining and clustering. Many of the DDM applications deal with continuous data streams. Therefore, developing DDM algorithms that can handle such stream scenarios is becoming increasingly important.

4.2. Data Mining using Multiple Agents

Data-mining systems differ in certain ways from the machine learning algorithms which they are typically derived from. Firstly, they have to cope with large amounts of data. For example, learning over a census database containing information on millions of families is very different from looking at a few hand-crafted examples of 'model' families. The second problem is that real world data has a tendency to contain errors and missing information [8].

Finally, a data-mining system aims to discover knowledge that is novel, useful, and understandable, which typically requires a human to focus the search and to provide feedback on the knowledge discovered. Our high-level model is shown in Figure 2. One or more agents per network node are responsible for examining and analyzing a local data source. In addition, an agent may query a knowledge source for existing knowledge such as rules or predicate definitions [9] .The agents communicate with each other during the discovery process. This allows the agents to integrate the new, individual knowledge they produce into a globally coherent theory. A user communicates with the agents via a user-interface. In addition, a supervisory agent, responsible for coordinating the discovery agents may exist. Figure 2 shows this agent next to the user-interface. The interface allows the user to assign agents to data sources, and to allocate high level discovery goals. It allows the user to critique new knowledge discovered by the agents, and to direct the agents

to new discovery goals, including ones that might make use of the new knowledge.

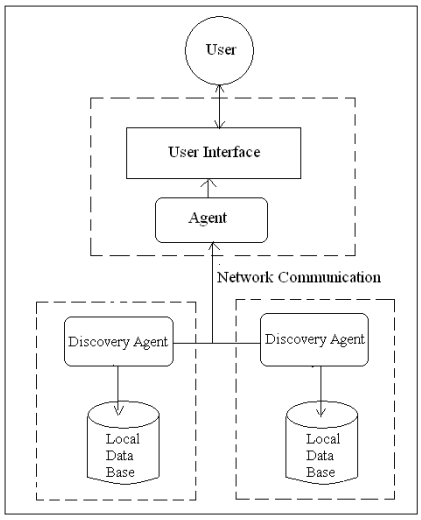

Figure 2: Data-Mining Using Multiple Agents

As far as possible, intention is to base the work on the integration of existing technologies in the field of software agents and first order learning. This is in order to concentrate on the core issues of distributed data-mining. We intend to use agents based on Agent Oriented Programming, and the techniques developed as part of the Knowledge Sharing Effort.

Intelligent agents represent an important opportunity to optimize knowledge management. Agents and data mining can work together to achieve required target. Data mining agents perform various functions of data mining. It is increasingly significant to develop better methods and techniques to organize the data for better decision making processes. In a competitive world, modern organizations focus on locating, storing, transferring and efficiently using their own information in order to better manage their intellectual capital. Main focus is to find methods and techniques to organize huge data provided by transactions or other activities and to extract useful patterns, relations, associations from data etc. In different applications, it is necessary to know what to do, when and how to do it, in order to complete the pre-established tasks for the proposed objectives, by means of self-decision systems. These systems are known in literature as agents. Intelligent agents act robustly in a flexible, open environment. Knowledge discovery process can be assisted by agents in order to increase the quality of knowledge and to simplify the main processes of identifying patterns from huge data volumes. Intelligent agents and data mining share the same objectives in order to assist decision making process.

Agents, are special types of software applications, have become increasingly popular in computing world in recent years. Some of the reasons for this popularity are their flexibility, modularity and general applicability to a wide range of problems like data filtering and analysis, information brokering, condition monitoring and alarm generation, workflow management, personal assistance simulation and gaming. Intelligent agents can help making computer systems easier to use, enable finding and filtering information, customizing views of information and automating work.

5. Conclusions and Future Enhancement

This paper describes some of our work about an agent-based approach to distributed knowledge discovery. Our goal is that agent-based knowledge discovery will allow us to maximize the usage of distributed computing resources, and minimize the preface and propositional results which do not reflect on the more complex aspects, as well as facilitate the easy integration and use of multiple agents . As far as possible, our intention is to base our work on the integration of existing technologies. This is in order to concentrate on the core issues of how agents can resolve different views of the world. However, the main area of investigation is how the agents should interact, and how the knowledge should be integrated. We are currently investigating how agents should be selected to work on a given discovery goal. A solution for this problem could be an intelligent system based on agents. Data mining and intelligent agents can make a common front to help people in the decision making process, to elaborate decisional

REFERENCES

[1]. Clifton, Christopher. "Encyclopedia Britannica: Definition Of Data Mining". Retrieved 2012.
[2]. Ireneusz Czarnowsk,"Agent And Multi-Agent Systems: Technologies And Applications" - Kesamsta , Pp. 724-733, 2009.
[3]. Ireneusz Czarnowski, Piotr Jedrzejowicz International Conference On Adaptive And Natural Computing Algorithms - Icannga , Pp. 609-618.2009.
[4.] Naeemkhalid Janjua, Farookhkhadeer Hussain, Omarkhadeer Hussain "Semantic Information And Knowledge Integration Through Argumentative Reasoning To Support Intelligent Decision Making ": Information Systems Frontiers - Isf , Vol. 15, Pp. 167-192, 2013
[5]. Daxin Tian, Kuifeng Ma "A Scaling-Up Machine Learning Algorithm" Vol. 7, No. 1, Pp. 39-55 , 2010.
[6]. Adam Campbell, Annie S. Wu "Multi-Agent Role Allocation: Issues, Approaches, and Multiple Perspectives "Vol. 22, No. 2, Pp. 317-355, 2011.
[7]. Roman Neruda, Ondřej Kazík "Modeling Data Mining Processes In Computational Multi-Agent Systems" 10.1145/2077489.2077500, 2011.
[8]. M-Tahar Kechadi" The Data Wave: Data Management And Mining", Workshop On Enabling Technologies: Infrastructure For Collaborative Enterprises - Wetice , Pp. 7-11, 2010
[9]. Sung Gook Kim, Kyeong Deok Woo, Jerzy W. Bala, Sung Wook Baik "Distributed Data Mining System Based On Multi-Agent Communication Mechanism",Pp. 100-109, 2010.
[10]. Priyanka Makkar Et. Al. / (Ijcse) International Journal On Computer Science And Engineering Vol. 02, No. 04, 1237-1244 "Distributed Data Mining And Mining "2010.

AUTHORS

First Author – R.HEMAMALINI, Research scholar St.Peter's University ,Asst Professor, Jaya College of Arts and Science-Thiruninravur., hemasureshbabu@yahoo.co.in
Second Author –. Dr.L.JOSEPHINE MARY, Professor & HOD, MCA Dept. Sri Ram Engineering College,Veppampet, josimgr@yahoo.co.in