ISSN: 2393-8528

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 4 Issue 3; May-June-2017; Page No. 52-56

# Survey of a Multi-Agent System for Distributed Data Mining

**S. Aswanandini**

M.Sc, M.Phil., Assistant Professor, Department of Computer Science

KG College Of Arts and Science, Coimbatore, India

aswanandini@kgcas.com

Received 07 April 2017; Accepted 06 May 2017

## ABSTRACT

The main purpose of data mining is to extract information from a data set and then transform into understandable form. It extracts and analyses data to present it in a conceivable form. Distributed data mining is based on the architecture of multi-agent systems. The problem of distributed data mining is very important in networks. This paper focuses on DDM algorithms in the context of multi agents It provides the communication between the DDM and MAS. It then focuses on applications in multi-agent based system and distributed clustering algorithms. It is based on the algorithms for distributed clustering, including its privacy.

**Keywords:** *multi-agent systems, clustering, privacy, entropy, asynchronous, network, agent, cluster.*

## I. INTRODOCTION

A multi-agent system (MAS) is a computerized system of multiple interacting intelligent agents. Multi agent systems deal with problems that are difficult to solve with a single agent. Agent computing which deals with complex systems are able to compete, communicate and exercise control within the frame of their objectives. Another benefit offered by multi agent system is the distribution of agents across the network. Many different paradigms are available for distributed computing. Among them, the agent paradigm has found to be particularly suitable for supporting the construction of portable and effective frameworks for computation which is distributed. According to the agent paradigm a the distributed framework is designed based on a Multi-Agent System (MAS), i.e. a system is composed of many agents and has the capacity of reaching goals that are tedious to achieve by an individual system. In addition, MASs can manipulate self-organization and complex systems even when the individual properties of all the agents are simple.

The Distributed Data Mining (DDM) deals with the challenges in analyzing data in a distributed manner and offers many solutions algorithmically to perform different data analysis and mining operations in a fundamentally distributed manner which pays careful attention to the resource constraints. Data mining techniques have become popular which can reveal the valuable knowledge hidden in the rough data. These techniques provide high performance approaches to cope up with the enormous amount of data and the complex algorithms. Distributed computing provides an effective approach. In many applications, the single or multi agent depends on the observed data from distributed sources. In a typical distributed environment analyzing the data which is distributed is an irrelavant problem because of many constraints such as limited bandwidth (*e.g.* wireless networks), distributed compute nodes, data which is protected and similar. A network operations center is one or more locations from which control is transferred to a computer, television broadcast, or telecommunications network.

Large organizations operate on NOC to manage different networks to provide geographic redundancy in the event of one site which is not available. The increasing use of multi-database technology over the networks and various distributed, homogeneous multi-database systems, has led to the development of many database systems for real world applications. The main problems any approach to DDM is challenged issues of autonomy and privacy. In DDM, one of two assumptions is adopted commonly as to how data is distributed across

**Corresponding author: S. Aswanandini |*

sites: homogeneously (horizontally-partitioned) and heterogeneously (vertically partitioned).

Both adopt the conceptual method that the data tables at each site are partitions of a single global table. The global table is horizontally partitioned in homogeneous method. The tables at each site are subsets of the global table; they have exactly the same attributes. The table is vertically partitioned in the heterogeneous case. Each site contains a collection of columns which does not have the same attributes. Each tuple at each site is assumed to contain a unique identifier to facilitate matching. This paper provides the possible relation between MAS and DDM technology. It focuses particularly on distributed clustering, a problem finding increasing number of applications in sensor networks, information retrieval in a distributed manner and many other domains. The paper provides an overview of multi agent system in DDM including privacy-preserving ones.

## II. DISTRIBUTED DATA MINING:

Data mining deals with the problem of analyzing data in scalable manner. DDM is one of the branch of data mining that offers a framework to distributed data paying careful attention to the distributed data and computing resources. Data mining and data warehousing are inter related together. Most of the tools operate on a principal of gathering every data on to a central site and then running an algorithm against that data. A number of applications are infeasible under such a methodology leading to a need for distributed data mining. In data mining, Distributed data mining (DDM) can be considered in this broader context. The Objective of DDM is to perform the data mining operations based on the type and availability of the distributed resources. Also the data sites are downloaded to a single site to perform the data mining operations at a center location.

### Architecture of a data warehouse

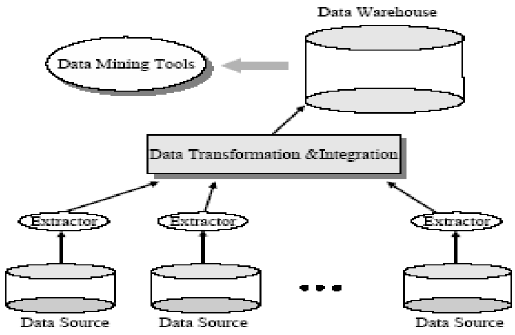

Figure 1:

The objective of DDM is to perform the data mining operations based on the type and availability of the distributed resources. The data sets are being downloaded to a single site and perform the data mining operations at a central location. The field of DDM has emerged as an active area of study. The bulk of DDM methods operate over an abstract architecture which includes multiple sites having independent computing power and storage capacity.

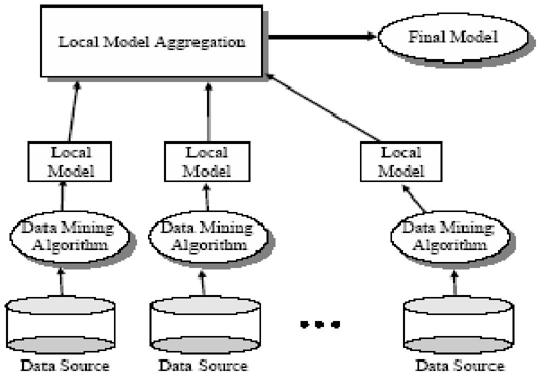

Figure 2:

## III. Distributed data mining framework

Data mining which is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. Data mining uses the sophisticated data analysis tools to discover previously unknown, valid patterns and relation-ships in large data set. These tools can include many models like statistical, mathematical and machine learning methods. It provides information within the data that queries and reports can't effectively reveal.

## IV. AGENTS FOR DDM

### Autonomy of data sources

A DM agent is considered as a modular extension of a data management system to handle the access to the underlying data source in accordance with the given constraints on the required autonomy of the system model and data. This is in full compliance with the paradigm of cooperative information systems.

### Interactive DDM

Pro-actively the assisting agents may limit the amount a human user has to interfere and supervise with the running data mining process, e.g., DM agents may anticipate the individual limits of the potentially large search space and proper intermediate results

### Multi-strategy DDM

DM agents may learn in due course of their deliberative actions in which combination of multiple data mining techniques can be choosen depending on the type of data retrieved from different sites and data mining tasks to be pursued. The learning methods of this category is similar to the adaptive selection of coordination strategies in amulti-agent system as proposed.

## V. SECURITY

If there is any failure to implement least privilege at a data source which could give any mining agent unsolicited access to sensitive data. For security, Agent code and data integrity is a crucial issue.

DDM: Hijacking a DM agent places as a trusted piece of (mobile) software—thus any sensitive data carried or transmitted by the agent—under the control of an intruder. If the DM agents are allowed to migrate to remote computing environment methods to ensure authentication and confidentiality of a mobile agent which has to be applied.

## VI. Distributed Clustering

Data clustering is the task of partitioning a multivariate data set into groups maximizing intra-group similarity and inter-group dissimilarity. In a distributed environment, it is required that the data objects are not transmitted for efficiency and security reasons between sites. An approach to clustering exploits the local maxima of a density estimate (i.e.) to search for connected regions which are populated by similar data objects. Instead of looking at the broad spectrum of DDM algorithms, we restrict to distribute clustering methods and their applicability in MAS.

There are two types in it. Efficiency focused and privacy focused. Algorithms focusing on efficiency strive to increase communication and computational efficiency. But in some cases they can offer nice privacy preservation, which is not their primary goal. Algorithms focusing on privacy hold privacy maintenance as their primary goal and they also try to maximize communication and computational efficiency, they first preserve privacy.

This approach can be implemented to a society of agents. For example, all participating agents in a real scenario belong to different competing organizations; agree to cooperate to achieve the common goal, without disclosing the contents of their data banks to each other. Each agent will neglect other agents to evaluate the risks and advantages which derive from participating to the distributed mining task.

In particular, some amount of security risks arise from the potential ability of other agents to carry out inference attacks on density estimates .The sensitive information of resulting disclosure could be exploited as a competitive advantage by the organizations which own the malicious agents. An agent has to evaluate in other aspects in order to decide autonomously whether it should investigate a probabilistic model of trustworthiness of participating agents, whether it should participate or not, the relation between trustworthiness and the topology of participating agents, and the probability of coalition attacks.

## VII. DISTRIBUTED CLUSTERING AND MULTI-AGENT SYSTEMS

The power of multi-agent-systems can be further enhanced by integrating efficient data mining capabilities and DDM algorithms also offer a nice choice for multi-agent systems which are designed to deal with distributed systems.
Clustering algorithms may play an important role in many sensor-network-based applications. Clustering algorithms are required for segmentation of data detected by the sensor nodes for situation awareness, detection of outliers for event detection which are only a few examples.

The distributed and resource-constrained nature of the sensor networks which demands a fundamentally distributed algorithmic solution to the clustering problem. Therefore, for analyzing sensor network data or data streams distributed clustering algorithms may come in handy. Clustering of sensor networks offers many challenges, including

a. Limited communication bandwidth.
b. Constraints on computing resources.
c. Limited power supply.
d. Need for fault-tolerance.
e. Asynchronous nature of the network

The Distribution clustering algorithms for this domain must address these challenges. There exist several exact distributed clustering algorithms, particularly for homogeneous data. For heterogeneous data, the number of choices for distributed clustering algorithms is relatively limited. However, there exist several different techniques for the latter scenario.

Most of the distributed clustering algorithms are still in the domain of academic research with a few exceptions. Therefore, for moderately large number of nodes the scalability properties of these algorithms are mostly studied.

Although the communication-efficient aspects of the distributed clustering algorithms help in addressing the concerns regarding the restricted bandwidth and power supply, the need for fault-tolerance and P2P communication-based algorithmic approach are yet to be adequately addressed in the literature.

## VIII.AGENT-BASED DISTRIBUTED DATA MINING

ADDM takes data mining as a basis foundation and is enhanced with agents; therefore, this novel data mining technique inherits all powerful properties of agents which yields desirable characteristics. Constructing an ADDM system has three key characteristics: interoperability, performance aspects and dynamic system configuration. Interoperability considerations, not solely collaboration of agents within the system, however additionally external interaction which permit new agents to enter the system seamlessly.

The architecture of the system should be open and flexible so that it can support the interaction including communication protocol, integration policy, and service directory. Communication protocol encloses encryption, message encoding, and transportation between agents, nevertheless, these are standardized by the Foundation of Intelligent Physical Agents (FIPA).

Integration policy specifies how a system such as an agent or a data site behaves when an external component an agent requests to enter or leave. Several agents and data sources are involved in a mining task, in which agents are configured to equip with an algorithm and deal with given data sets. Change in data affects the mining task as an agent may be still executing the algorithm.

Lastly, performance can be improved because the distribution of data is a major constraint. In distributed environment, tasks can be executed in parallel, in exchange, concurrency issues arise. Quality of service control in system perspectives and performance of data mining is desired; however it can be derived from both data mining and agent's fields.

An ADDM system can be categorized into a set of components. We may generalize activities of the system into request and response, each of which involves a different set of components. The basic components of an ADDM system are given as:

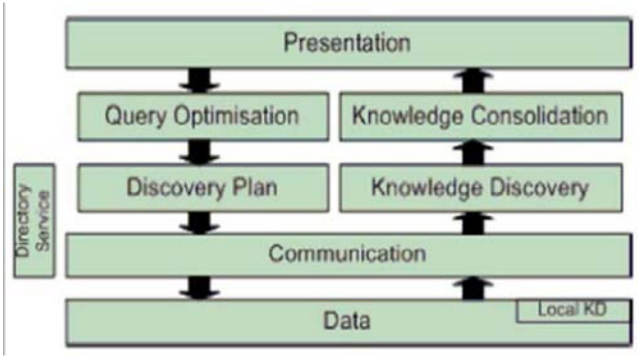

**Figure 3:**

### 1) Data:

Data is the foundation layer of our interest. In distributed environment, data can be hosted in various forms, such as online relational databases, data stream, web pages, etc., in which purpose of the data is varied.

### 2) Communication:

The system chooses the related resources from the directory service, which maintains a list of data sources, mining algorithms, data schemas , data types, etc. The communication protocols vary depending on the implementation of a system such as client-server, peer-to-peer, etc.

### 3) Presentation:

The user interface (UI) gets interacted with the user which receives and responds to user. The interface also simplifies complex distributed systems into user-friendly message such as network diagrams, visual reporting tools, etc. When a user requests for data in the data mining through the UI many components are involved.

### 4) Query optimization:

A query optimizer analyses the request that confirm sort of mining tasks and chooses correct resources for the request. It also determines whether it is possible to parallelize the tasks, since the data is distributed and can be mined in parallel.

### 5) Discovery Plan:

A planner allocates the sub-tasks with related resources. At this stage, mediating agents play vital roles to coordinate multiple computing units since mining sub-tasks performed asynchronously in addition as results from those tasks. On the other hand, when a mining task is done, the following components are taken place.

### 6) Local Knowledge Discovery (KD):

In order to transform data in to patterns which adequately represent the data and reasonable to be transferred over the network at each data site, mining process may take place locally depending on the individual implementation.

### 7) Knowledge Discovery:

Also known as mining, it executes the algorithm as required by the task to obtain knowledge from the specified data source.

### 8) Knowledge Consolidation:

In order to present to the user with a compact and meaningful mining result, it is necessary to normalize the knowledge obtained from various sources. The component involves a complex methodology to combine knowledge and patterns from distributed sites. Consolidating homogeneous knowledge/patterns is promising and yet difficult for heterogeneous case.

## IX. CONCLUSION AND FUTURE ENHANCEMENT

Multi-agent systems are designed for collaborative problem solving in distributed environments. Many of these application environments deal with empirical analysis and mining of data. The paper focused on distributed clustering algorithms. This paper suggests that previous centralized data processing techniques might not work well in several distributed environments wherever knowledge centralization is also troublesome due to restricted information measure because of limited bandwidth. Distributed data mining algorithms may offer a better solution as they are designed to work in an environment which is distributed by paying careful attention to the computing and communication resources.

## References

1. V.Gorodetsky,O.Karsaev,and V. Samoilov. Infrastructural Issues forAgent-Based Distributed Learning. In Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society Washington, DC, USA, 2006
2. D. Jensen, M. Rattigan, and H. Blau. Information awareness: A prospective technical assessment. In Proceedings of the Ninth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, pages 378{387, Washington, DC, 2003. ACM Press.
3. L. Cao, C. Luo, and C. Zhang. Agent-Mining Interaction: An Emerging Area. Lecture Notes in Computer Science, 4476:60, 2007.
4. W. Davies and P. Edwards. Distributed Learning: An Agent-BasedApproach to Data-Mining. In Proceedings of Machine Learning 95Workshop on Agents that Learn from Other Agents, 1995.
5. F. Bergenti, M. P. Gleizes, and F. Zambonelli. Methodologies And Software Engineering For Agent Systems: The Agentoriented Software Engineering Handbook. Kluwer Academic Publishers, 2004.
6. Sung W. Baik, Jerzy W. Bala, and Ju S. Cho Agent based distributed data mining. Lecture Notes in Computer Science, 3320:42–45, 2004.
7. ”Principles of data mining”-D.J.Hand, Heiki Mannila, Padhraic smith, 2001.
8. ”Principles of Data Mining”-Max A.Bramer, 2007.
9. ”Data Mining methods and models” Daniel T.Larose, 2006.
10. ”Data Mining-Oppurtunities and challenges”-John wang, 2003