International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(10), Jan, 2014, pp. 53-66, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

A distributed data mining architecture based on nonlinear model

Maedeh Tashakkorian

Mazandaran University of Science and Technology

*Corresponding Author's Email: *m.tashakkorian@ustmb.ac.ir*

Abstract

Development of computer networks and distributed database technologies has encouraged distributed data storage and a new technical generation of distributed data mining. This field has a wide range of applications. The present paper deals with applying multi-agent technology in data mining. At first we talk about agent and data mining technologies development briefly. Then, the benefits of combining multi-agent with distributed data mining would be explained. And at last a distributed data mining model based on multi-agent technology will be offered and the mining methods of this model on association rules in the distributed databases will be analyzed.

Keywords: Distributed data mining, Multi-agent system, Association rules

1. Introduction

Many novel application programs and information technologies use data obtained from distributed resources. For instance, the transaction data of banks, hospitals, manufacturing plants, government agencies, insurance companies, census data in a particular year, and many others are among data acquired from this kind of resources [1-3]. Even the temporal and spatial relationships, can provide different views for a single database [4]. Knowledge discovery and data mining technology is an effective tool for mining large volume of data and extracting efficient new patterns and relationships between data. This field has a wide range of applications. Databases that create or receive these data usually belong to natural or legal persons who are pursuing their own goals and benefits, and are not willing to offer their

Vol. 4(10), Jan, 2014, pp. 53-66, ISSN: 2305-0543
Available online at: http://www.aeuso.org

knowledge for free. The topic that has attracted lots of attention in the field of data mining these days is the distributed state of data. Development of computer networks and distributed database technologies has encouraged distributed data storage and a new technical generation of distributed data mining. Distributed data mining is consisted of finding the semi-automatic of hidden patterns in the data, when data or deduction mechanisms are distributed. The distributed state of deduction mechanisms means that it is necessary to consider the communication costs between different knowledge extracting mechanisms. In previous studies mostly centralized techniques were taken into consideration, and also the target data have a flat and homogenous structure.

Problems like knowledge extraction without having access to all of the existing data, setting up an effective and optimum communication link with other knowledge extracting mechanisms, and also exchange of knowledge or intermediate instead of raw information are only a handful of difficulties facing distributed data mining. Hence, although distributed data mining is considered as a key solution to the main problems of data mining, it is the cause of many challenges and difficulties itself. Effective resolution of these problems will result in greater and better use of data mining and applying current potentials in domains that in spite of dire need for data mining its use is quite limited. Examples of data mining techniques include but not limited to classification models, clustering and association rules.

In the area of association rules, a significant effort on the discovery of a variety of patterns, including frequent item-sets and spatial or temporal association rules has been carried out. The common challenges in this area fall into two groups [3]; 1. Identifying patterns from a large database, and 2. Discovering new patterns after integration of multiple databases. The main objective of distributed databases is the latter one. Normally, distributed data mining algorithms first analyze the data at the local level, and then by applying knowledge integration methods it combines the local mining results at the general level to

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(10), Jan, 2014, pp. 53-66, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

achieve overall findings [5]. A sample of distributed data mining framework has been shown in figure 1.

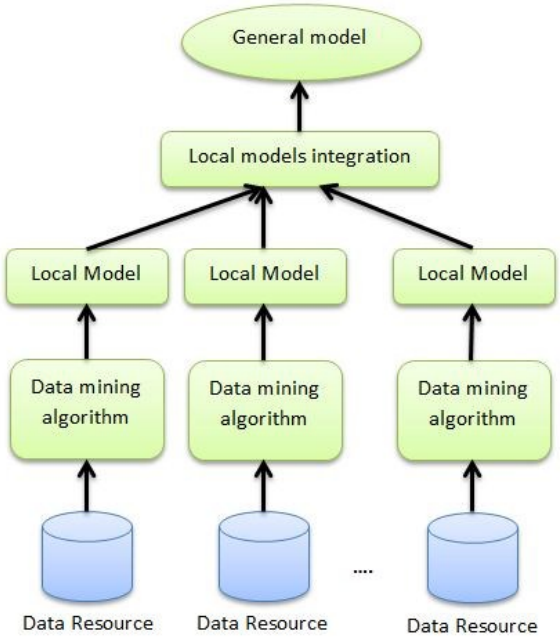

Figure 1: distributed data mining framework

An agent is a type of software which can act autonomously and provide the desired goals [6]. Smart agents, are advanced agents that can react to the environmental changes in relation with other agents and perform smart computations to reach their goals [7]. Agents are active, flexible, social, goal directed, autonomous, and capable of reasoning. Based on these facts, there is an urge to use multi-agent technology for distributed data mining in a way that agents will carry out request analysis, pretreatment operations and data mining. Although it may look simple to solve these problems through mining a database and then investigating and merging the patterns, it is still a very challenging issue and needs more work to improve data privacy preservation and time consumption of pattern recognition in databases. This entails

Vol. 4(10), Jan, 2014, pp. 53-66, ISSN: 2305-0543
Available online at: http://www.aeuso.org

the design of an architecture to handle the challenges of this domain. There are three fold of these challenges:

- How to design a mining architecture with ability to recognize distributed patterns?
- What kinds of information need to be swapped among distributed sites?
- What are the advantages of using multi-agent systems for distributed data mining? And how can we utilize the ability of intelligence, learning and reasoning of agents in this field?

This paper displays our recent findings in resolving the mentioned concerns. In the following sections we offer a distributed data mining architecture on the basis of multi-agent systems. In section 2 recent activities in the field of agents and multi-agent systems in data mining are discussed. At section 3 we offer a glance on distributed data mining and in section 4 the aforementioned architecture will be elaborated in details. Finally, we come to conclusion in section 5.

2. Related works

Data mining is semi-automatic extraction of models, patterns, changes, abnormalities from large data sets [8]. The studies performed in the field of distributed databases mining have caused impressive improvements in classification [9-11], clustering [12, 13], association rules mining [14-17], and similarity evaluation of databases [18, 19]. Generally, two major groups are evident in the studies concerning distributed data mining. First group includes works which try to solve the problem of data distribution by accumulation of data in a central point using the enhanced network algorithms and protocols. Also, assuming that in distributed data mining accuracy and costs (of data transmission) are correlated, they try to lower the costs to achieve a satisfactory solution through adopting an accuracy improving strategy. This is what the Papyrus system is based on [20]. Second group of studies consider different ways of data weighting and present methods for data mining without transmission of raw data to a central

Vol. 4(10), Jan, 2014, pp. 53-66, ISSN: 2305-0543
Available online at: http://www.aeuso.org

point and usually by transfer of intermediate data between sites. Normally in all of these methods it is tried to minimize the communications. In study [21] a procedure is suggested for extracting association rules from distributed data with minimum communication between processes extracting them. Moreover, as a sample of the research done in the subject of privacy preservation in the process of data mining, research [22] can be mentioned.

Well-known systems that implement distributed data mining using agents consist JAM[1] [23] and PADMA [24] systems. JAM, is the name of a distributed system which performs data mining by making use of agents. This system is composed of a number of sites all of which have their own database and a few learners. The PADMA system includes some agents, user interface and an organizer. Main role of the organizer in the system is receiving user's requests and referring them to the agents and then collecting the results from agents and combining them.

3. Agent based data mining

When data mining technology enters the real world of problem solving especially in the field of data management and complicated programs, it will face several challenges such as data redundancy, dynamic changes of the environment, communications overload, privacy of resource data, organizational limitations in the distributed data resources, data pretreatment, compatible learning, interactive mining and utilizing human intelligence. Multi-agent technology performs pretty well in the domains of user interactions, autonomous computing, self-organizing ability, cooperation, communication, negotiation, peer to peer computing, and mobile computing. These are advantages of this technology and can highly improve the data mining process and solve its complicated problems in the areas of data treatment, information treatment, pattern mining, modeling, user interactions, substructures and services.

[1] Java agents for Meta-learning

Vol. 4(10), Jan, 2014, pp. 53-66, ISSN: 2305-0543
Available online at: http://www.aeuso.org

What is meant by agent-based data mining is some versions of multi-agent systems which have been created for the enhancement of data mining. If multi-agent technology is applied not only the aforementioned problems are reduced to a minimum, but also it will bring numerous advantages [25, 26]. Some of these advantages include:

- **Decentralized control:** Decentralized control is almost the most important characteristic of multi-agent systems which distinguishes them from parallel or distributed methods. Decentralized control means that each agent in a multi-agent system works autonomously.
- **Destruction resistance:** This feature is one of the characteristics of decentralized control, meaning that even if some of the agents encounter difficulties and stop working the system as a whole will continue to operate.
- **Reduces the complication of system development:** Multi-agent system is distributed not only in structure but also in logic, therefore it is very suitable for distributed mining in parallel computations and can lower the complication of system development. This characteristic is also due to the decentralized control of these systems because system usages can be enhanced through increasing the number of agents.
- **Increases system intelligence:** Compared to other software systems, agent has a high capacity and capability and can present a proper and precise service.
- **Increases system visibility and freedom:** In other words, in implementation methods agent is a kind of capsulation model.
- **Reduces system traffic:** An agent located in a local site is able to send the results of its mining operations to associated sites for analysis which can somewhat decrease data traffic on the network.

Vol. 4(10), Jan, 2014, pp. 53-66, ISSN: 2305-0543
Available online at: http://www.aeuso.org

- **Increases system stability:** In the face of outer disorders and disturbances agent is capable of adapting to the new environment to adjust the parameters by means of interactive learning and guarantee efficiency and stability of the system.

- **Distributed data mining with multiple strategies:** For some complex programs a suitable combination of several data mining techniques performs better and more usefully than a single specific technique. Data mining agents can learn, they choose between their activities on the basis of the type of data of different sites and mining operations that have to be carried out.

- **Cooperative distributed data mining:** Data mining agents are able to perform mining operations independently on data from different sites and in the end combine the resulting models. They can also share the knowledge they have extracted to make use of other agents capabilities.

4. An architecture for distributed data mining based on multi-agent systems

Considering the tasks a distributed data mining system has to perform, its functions can be classified into three categories; user interaction functions, management functions and data treatment functions. Each function is related to an agent. For the purpose of mining the data from various sites, data pre-treatment and data mining agents are considered as a pair. The data pretreatment agent prepares the data for the data mining agent. The data mining agent finishes the data mining task and sends the results to the to the central data mining site for integration and extraction of the right results. Fig. 2 displays a distributed data mining system based on the above-noted concepts. This system is generally composed of four layers; user layer, management layer, processing layer and resource layer. All of these four layers and their fundamental elements will be accounted for in the following sections.

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(10), Jan, 2014, pp. 53-66, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

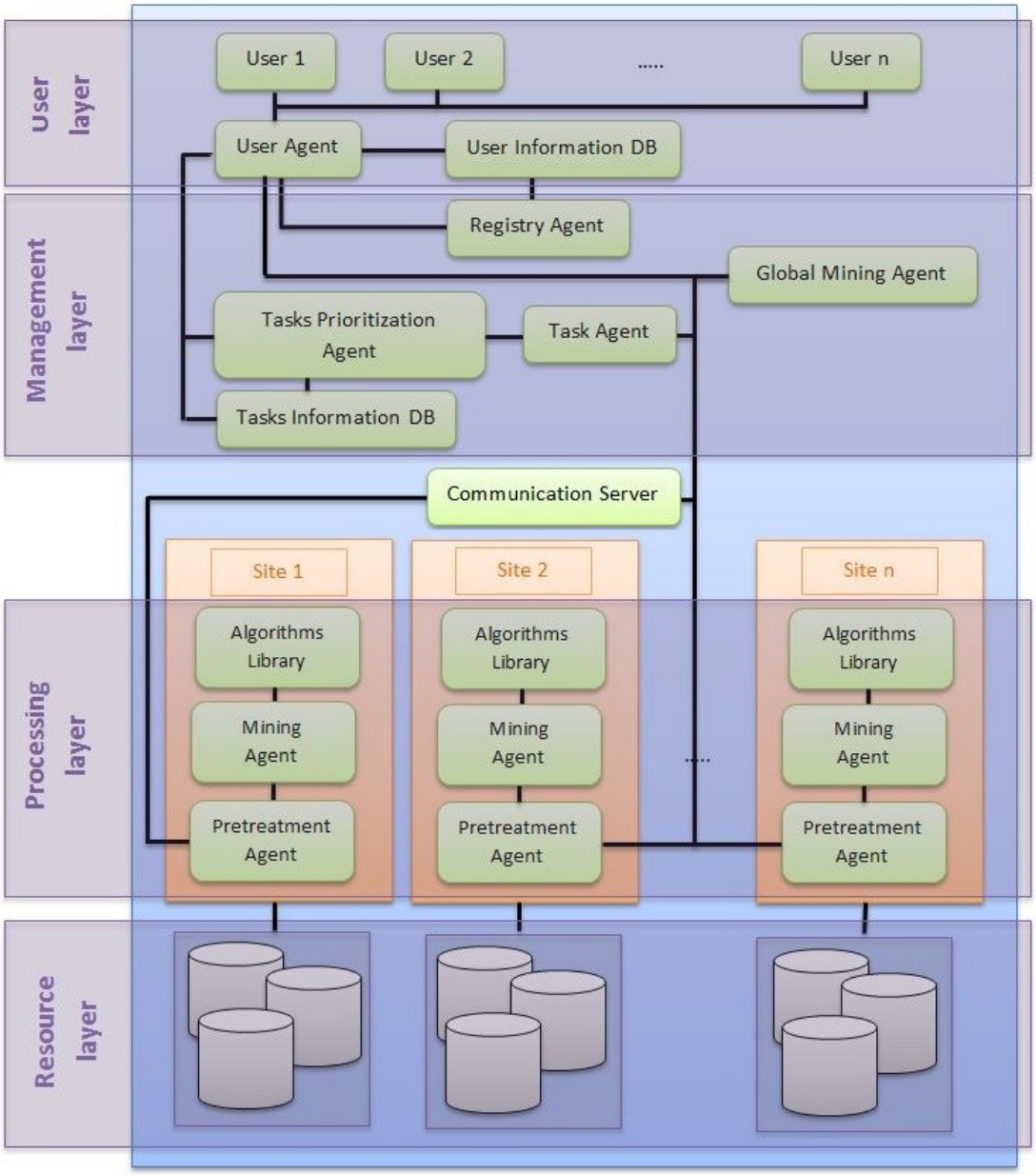

Figure 2: Distributed data mining architecture based on multi-agent systems

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(10), Jan, 2014, pp. 53-66, ISSN: 2305-0543
Available online at: http://www.aeuso.org

4.1. First layer, the user layer

The first layer of this architecture includes the users, user agent and user information database.

- **User agent:** The main task of this agent is finalizing the interactions between the system and the user. For this purpose the user agent presents an interface through which the user submits its requests and the results are presented back to the user. Intelligence of the user agent is exhibited in the long term, meaning that in the data mining process the user's mining requests information is saved in the tasks information database which can demonstrate the user's interests. Also, this information can highly improve quality of the system services. This agent is in connection with the tasks prioritization agent for the treatment of user's requests.

- **User information database:** In this database user's records and information are saved.

4.2. Second layer, management layer

Components of this layer include registry agent, tasks prioritization agent, task agent, global data mining agent and tasks information database.

- **Tasks prioritization agent:** This agent saves the number of requests in a given time range and the requests information. Then, it designates a weight for every request based on the saved information which is actually the aim of this agent. When the system and network are in idle status this agent can find the saved information the weight of which is greater than the critical value. Based on these weights the information needed for mining requests is sent to the task agent.

- **Task agent:** These are temporary agents which are created automatically by tasks prioritization agent to examine the data mining requests and will exist until those requests are completely fulfilled. The role of connecting to the agents and if necessary

Vol. 4(10), Jan, 2014, pp. 53-66, ISSN: 2305-0543
Available online at: http://www.aeuso.org

activating and organizing them is appointed to the task agent. Usually after creation of a task agent, it requests the intended service from a group of pretreatment and data mining agents and reports the mining results to the user agent.

- **Tasks information database:** This database is used to save information of the requests received by tasks prioritization agent from the user agent.

- **Registry agent:** This agent is made up of functions which control all accesses, modifications, registry, etc. for different users such as the administrator, ordinary users and developers.

4.3. Third Layer, Processing Layer

This layer is composed of mining agent, pretreatment agent and algorithms library.

- **Pretreatment agent:** This agent is located at the local site and holds some information associated with the site's data resources. Its main mission is to carry out pretreatment operations and present normal data. For this purpose, it extracts the intended data from data resources and performs pretreatment operations like data normalization and finally sends the data to the data mining agent. This agent is comprised of four main functions of data extraction, reduction, format conversion and data simplification. The data appointed for a mining operation can be a data collection or part of one or several collections.

- **Mining agent:** Data mining agents include methods for initialization, performing mining operations and presenting the results to the intended task agent. Mining functions are performed by two types of mining agents; global data mining agent and local data mining agent. Local data mining agents carry out the mining operations in local sites and utilize the data received from data pretreatment agent. The global data mining agent on

Vol. 4(10), Jan, 2014, pp. 53-66, ISSN: 2305-0543
Available online at: http://www.aeuso.org

the other hand is located at the central site and works on data obtained from data mining operations of local agents at different sites. Furthermore, when the local data mining agent is in idle status, it can examine the existing algorithms and choose the most effective one for mining operations.

- **Algorithms library:** At this library, mining algorithms and results of utilizing them in a certain mining operation are saved. These algorithms can be deleted, updated or developed.

4.4. Forth Layer, Resources Layer

This layer includes the data resources. As mentioned above, every layer of this architecture is made up of smaller parts. To implement system connections between different agents a communication server is used, and each agent transmits the information to the next stage and next agent via this server. In this article, a four layered architecture for distributed data mining based on agent technology has been suggested.

This architecture uses the well-known PADMA model, with extra features. These features are user agent, global data mining agent, user information database, tasks information database, tasks prioritization agent and task agent. These capabilities increase extensibility, intelligence and system security and better use of resources. In addition, they decrease data redundancy as well. By adding the user agent, system intelligence and extensibility of systematic functions will be improved. Furthermore, the tasks information database reduces the time a user spends over finding the desired information without redundancy. Tasks prioritization agent organizes the tasks by investigating information stored in the database and by taking into account the number of received requests. These features make a better use of recourses and reduce data redundancy. In addition, data mining agent searches the algorithms of a library and chooses the best one for the mining operation while the system is idle.

Vol. 4(10), Jan, 2014, pp. 53-66, ISSN: 2305-0543
Available online at: http://www.aeuso.org

Conclusion

Integration of data mining and agents or in short agent mining has opened a new research field which promises development and enhancement in intelligent information processing. This topic can be analyzed from two viewpoints: data mining for agents, in other words agents act on the basis of mining results; and agents for data mining which means data mining based on agents and is referred to as multi-agent data mining. In this article, a brief summary of distributed data mining has been initially given. As mentioned before, this technology faces complicated challenges such as huge data resources, data transmission problems and combination of data. Afterwards, agents and multi-agent systems have been proposed as solutions to these problems and to improve this technology and the benefits of using this technology in distributed data mining have been described. Subsequently, a four layered architecture for data mining in distributed environments based on agent technology has been suggested and each layer and its components have been separately accounted for. Finally, this architecture was compared to the well-known PADMA model and its advantages have been recounted.

References

[1] J. Wang, Data mining: opportunities and challenges: Irm Press, 2003.

[2] V. S. Rao and S. Vidyavathi, "Distributed Data Mining and Mining Multi-Agent Data," IJCSE) International Journal on Computer Science and Engineering, vol. 2, pp. 1237-1244, 2010.

[3] X. Zhu, B. Li, X. Wu, D. He, and C. Zhang, "CLAP: Collaborative pattern mining for distributed information systems," Decision support systems, 2011.

[4] B. C. Chen, L. Chen, Y. Lin, and R. Ramakrishnan, "Prediction cubes," in 31st VLDB Conference, Norway, 2005, pp. 982-993.

[5] S. Paul, "An optimized distributed association rule mining algorithm in parallel and distributed data mining with xml data for improved response time," International Journal of Computer Science and Information Technology, vol. 2, 2010.

[6] M. J. Wooldridge, An introduction to multiagent systems: Wiley, 2009.

Vol. 4(10), Jan, 2014, pp. 53-66, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

[7] I. Rudowsky, "Intelligent agents," Communications of the Association for Information Systems, vol. 14, p. 275, 2004.

[8] R. Grossman, "A top-ten list for data mining," SIAM News, vol. 34, 2001.

[9] S. Stolfo, A. L. Prodromidis, S. Tselepis, W. Lee, D. W. Fan, and P. K. Chan, "JAM: Java agents for meta-learning over distributed databases," 1997, pp. 74-81.

[10] M. Aoun-Allah and G. Mineau, "Distributed data mining: why do more than aggregating models," 2007, pp. 2645-2650.

[11] P. Luo, H. Xiong, K. Lü, and Z. Shi, "Distributed classification in peer-to-peer networks," 2007, pp. 968-976.

[12] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," in 21st ICDE Conference, 2005, pp. 341-352.

[13] S. Datta, C. Giannella, and H. Kargupta, "K-Means Clustering Over a Large, Dynamic Network," in 2006 SIAM Conference on Data Mining, 2006.

[14] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in VLDB Conferences, Santiago, Chile, 1994, pp. 487-499.

[15] D. W. Cheung, V. T. Ng, A. W. Fu, and Y. Fu, "Efficient mining of association rules in distributed databases," IEEE Transactions on Knowledge and Data Engineering, vol. 8, pp. 911-922, 1996.

[16] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in ACM SIGMOD, 2000, pp. 1-12.

[17] M. Z. Ashrafi, D. Taniar, and K. Smith, "ODAM: An optimized distributed association rule mining algorithm," Distributed Systems Online, IEEE, vol. 5, 2004.

[18] T. Li, M. Ogihara, and S. Zhu, "Association-based similarity testing and its applications," Intelligent Data Analysis, vol. 7, pp. 209-232, 2003.

[19] G. I. Webb, S. Butler, and D. Newlands, "On detecting differences between groups," in 9th ACM SIGKDD Conference, 2003, pp. 256-265.

[20] S. Bailey, R. Grossman, H. Sivakumar, and A. Turinsky, "Papyrus: a system for data mining over local and wide area clusters and super-clusters," 1999, p. 63.

[21] A. Schuster and R. Wolff, "Communication-efficient distributed mining of association rules," Data Mining and Knowledge Discovery, vol. 8, pp. 171-196, 2004.

[22] A. Schuster, R. Wolff, and B. Gilburd, "Privacy-preserving association rule mining in large-scale distributed systems," 2004, pp. 411-418.

[23] A. L. Prodromidis, S. J. Stolfo, S. Tselepis, T. Truta, J. Sherwin, and D. Kalina, "Distributed data mining: the JAM system architecture," 2001.

Vol. 4(10), Jan, 2014, pp. 53-66, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

[24] H. Kargupta, I. Hamzaoglu, and B. Stafford, "PADMA: Parallel data mining agents for scalable text classification," in High Performance Computing, 1997.

[25] K. A. Albashiri, F. Coenen, and P. Leng, "EMADS: An extendible multi-agent data miner," Knowledge-Based Systems, vol. 22, pp. 523-528, 2009.

[26] Y. Jie, "Research on Association Rules in Distributed Data Mining," Energy Procedia, vol. 13, pp. 8575-8580, 2011.