Integration and Interaction of Distributed Data Mining with Agent Technology

Meenu Gupta[1], Rajeev Kumar[2]
Assistant Professor, Dronacharya College of Engineering[1]
Assistant Professor, GCEW College of Engineering[2]

ABSTRACT

In recent years, more and more researchers have been involved in research on both agent technology and distributed data mining. A clear disciplinary effort has been activated toward removing the boundary between them, that is the interaction and integration between agent technology and distributed data mining. We refer this to *agent mining* as a new area. The marriage of agents and distributed data mining is driven by challenges faced by both communities, and the need of developing more advanced intelligence, information processing and systems. In this paper presents an overall picture of agent mining from the perspective of positioning it as an emerging area. We summarize the main distributed data mining, driving forces, disciplinary framework, applications, and trends and directions, data mining-driven agents, and mutual issues in agent mining. Arguably, we draw the following conclusions: (1) agent mining emerges as a new area in the scientific family, (2) both agent technology and distributed data mining can greatly benefit from agent mining, (3) it is very promising to result in additional advancement in intelligent information processing and systems. However, as a new open area, there are many issues waiting for research and development from theoretical, technological and practical perspectives.

1. INTRODUCTION

Traditional warehouse-based architectures for data mining suppose to have centralized data repository. Such a centralized approach is fundamentally inappropriate for most of the distributed and ubiquitous data mining applications. In fact, the long response time, lack of proper use of distributed resource, and the Fundamental characteristic of centralized data mining algorithms do not work well in distributed environments. A scalable solution for distributed applications calls for distributed processing of data, controlled by the available resources and human factors. For example, let us suppose an ad hoc wireless sensor network where the different sensor nodes are monitoring some time-critical events. Central collection of data from every sensor node may create traffic over the limited bandwidth wireless channels and this may also drain a lot of power from the devices. A distributed architecture for data mining is likely aimed to reduce the communication load and also to reduce the battery power more evenly across the different nodes in the sensor network. One can easily imagine similar needs for distributed computation of data mining primitives in ad hoc wireless networks of mobile devices like PDAs, cell phones, and wearable computers. The wireless domain is not the only example. In fact, most of the applications that deal with time-critical distributed data are likely to benefit by paying careful attention to the distributed resources for computation, storage, and the cost of communication. As an other example, let us consider the World Wide Web: it contains distributed data and computing resources. An increasing number of databases (e.g., weather databases, oceanographic data, etc.) and data streams (e.g., financial data, emerging disease information, etc.) are currently made on line, and many of them change frequently. It is easy to think of many applications that require regular monitoring of these diverse and distributed sources of data. A distributed approach to analyze this data is likely to be more scalable and practical particularly when the application involves a large number of data sites. Hence, in this case we need data mining architectures that pay careful attention to the distribution of data, computing and communication, in order to access and use them in a near optimal fashion. Distributed Data Mining (sometimes referred by the acronym DDM) considers data mining in this broader context. DDM may also be useful in environments with multiple compute nodes connected over high speed networks. Even if the data can be quickly centralized using the relatively fast network, proper balancing of computational load among a cluster of nodes may require a distributed approach. The privacy issue is playing an increasingly important role in the emerging data mining applications. For example, let us suppose a consortium of different banks collaborating for detecting frauds. If a centralized solution was adopted, all the data from every bank should be collected in a single location, to be processed by a data mining system. Nevertheless, in such a case a distributed data mining system should be the natural technological choice: both it is able to learn models from distributed data without exchanging the raw data between different repository, and it allows detection of fraud by preserving the privacy of every bank's customer transaction data. For what concerns techniques and architecture, it is worth noticing that many several other fields influence Distributed Data Mining systems concepts. First, many DDM systems adopt the Multi-Agent System (MAS) architecture, which finds its root in the Distributed Artificial Intelligence (DAI). Second, although Parallel Data Mining often assumes the presence of high speed met work connections among the computing nodes, the development of DDM has also been influenced by the PDM literature. Most DDM algorithms are designed upon the potential parallelism they can apply over the given distributed data.. In figure 1 a

general Distributed Data Mining framework is presented. In essence, the success of DDM algorithms lies in the aggregation. Each local model represents locally coherent patterns, but lacks details that may be required to induce globally meaningful knowledge. For this reason, many DDM algorithms require a centralization of a subset of local data to compensate it. The ensemble approach has been applied in various domains to increase the accuracy of the predictive model to be learnt. It produces multiple models and combines them to enhance accuracy.

Typically, voting (weighted or un weighted) schema are employed to aggregate base model for obtaining a global model. As we have discussed above, minimum data transfer is another key attribute of the successful DDM algorithm.

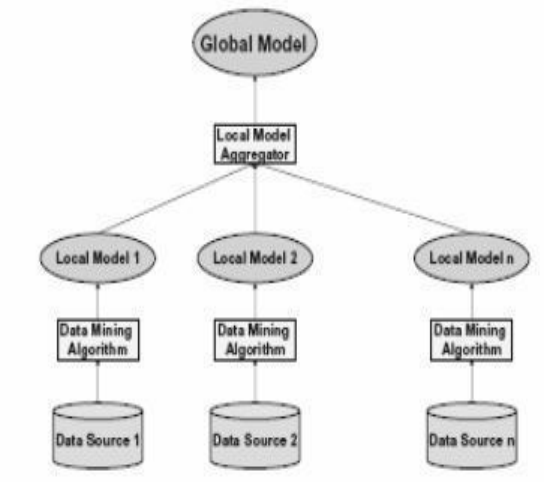

Fig 1: General Distributed Data Mining Frame Work

2. CHALLENGES OF DISTRIBUTED DATA MINING

Data mining and machine learning currently forms a mature field of artificial intelligence supported by many various approaches, algorithms and software tools. However, modern requirements in data mining and machine learning inspired by emerging applications and information technologies and the peculiarities of data sources are becoming increasingly tough. The critical features of data sources determining such requirements are as follows: In enterprise applications, data is distributed over many heterogeneous sources coupling in either a tight or loose manner. Distributed data sources associated with a business line are often complex, for instance, some is of high frequency or density, mixing static and dynamic data, mixing multiple structures of data; Data integration and data matching are difficult to conduct; it is not possible to store them in centralized storage and it is not feasible to process them in a centralized manner; In some cases, multiple sources of data are stored in parallel storage systems; Local data sources can be of restricted availability due to privacy, their commercial value, etc., which in many cases also prevents its centralized processing, even in a collaborative mode; In many cases, distributed data spread across global storage systems is often associated with time difference; Availability of data sources in a mobile environment depends on time; The infrastructure and architecture weaknesses of existing distributed data mining systems requires more flexible, intelligent and scalable support. These and some other peculiarities require the development of new approaches and technologies of data mining to identify patterns in distributed data. Distributed data mining (DDM), in particular, Peer-to-Peer (P2P) data mining, and multi-agent technology are two responses to the above challenges.

3. AGENT MINING INTERACTION AND INTEGRATION

The emergence of agent mining results from the following driving forces: The critical challenges in agents and data mining respectively, the critical common challenges troubling agents and data mining the complementary essence of agents and data mining in dealing with their challenges, and the great add-on potential resulting from the interaction and integration of agents and data mining. Agents and data mining are facing critical challenges from respective areas. Many of these challenges can be tackled by involving advances in other areas. Fig.2. illustrates these challenges. In this section, we specify both individual and mutual challenges in agent and mining disciplines that may be complemented by the interaction with the other disciplines.

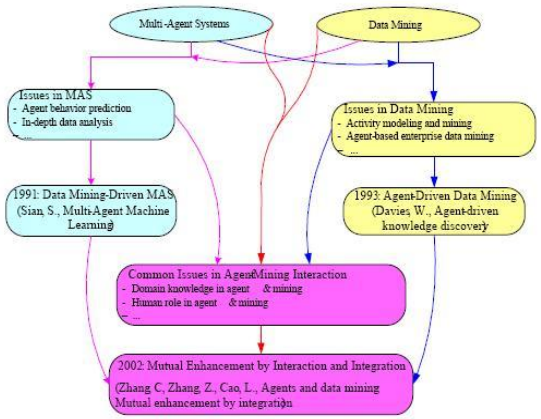

Fig. 2: Challenges in Agents and Data Mining.

4. MUTUAL CHALLENGES IN AGENT AND DISTRIBUTED DATA MINING

As addressed in [5, 6, 7], agents can enhance data mining through involving agent intelligence in data mining systems, while an agent system can benefit from data mining via extending agents' knowledge discovery capability. Nevertheless, the agent mining interaction symbiosis cannot be established if mutual issues are not solved. These mutual issues involve fundamental challenges hidden on both sides and particularly within the interaction and integration. Fig. 2. presents a view of issues in agent mining interaction highlighting the existence of mutual issues. Mutual issues constraining agent-mining interaction and integration consist of many aspects such as architecture and infrastructure, constraint and environment, domain intelligence, human intelligence, knowledge engineering and management, and nonfunctional requirements.

Architecture and infrastructure Data mining always faces a problem in how to implement a system that can support those brilliant functions and algorithms studied in academia. The design of the system architecture conducting enterprise mining applications and emerging research challenges needs to provide (1) functional support such as crossing source data management and preparation, interactive mining and the involvement of domain and human intelligence, distributed, parallel and adaptive learning, and plug-and-play of algorithms and system components, as well as (2) nonfunctional support for instance adaptability, being user and business friendly and flexibility. On the other hand, middle to large scales of agent systems are not easily built due to the essence of distribution, interaction, human and domain involvement, and openness. In fact, many challenging factors in agent and mining systems are similar or complementary.

Constraint and environment Both agent and mining systems need to interact with the environment, and tackle the constraints surrounding a system. In agent communities, environment could present characters such as openness, accessibility, uncertainty, diversity, temporality, spatiality, and/or evolutionary and dynamic processes. These factors form varying constraints on agents and agent systems. Similar issues can also be found from real-world data mining, for instance, temporal and spatial data mining. The dynamic business process and logics surrounding data mining make the mining very domain-specific and sensitive to its environment.

Domain intelligence Domain intelligence widely surrounds agent and mining systems. Both areas need to understand, define, represent, and involve the roles and components of domain intelligence. In particular, it is essential in agent mining interaction to model domain and prior knowledge, and to involve it to enhance agent-mining intelligence and actionable capability.

Human intelligence Both agent and mining need to consider the roles and components of human intelligence. Many roles may be better played by humans in agent-mining interaction. To this end, it is necessary to study the definition and major components of human intelligence, and how to involve them in agent mining systems. For instance, mechanisms should be researched on user modeling, user and business friendly interaction interfaces, and communication languages for agent-mining system dialogue.

Knowledge engineering and management To support the involvement of domain and human intelligence, proper

mechanisms of knowledge engineering and management are substantially important. Tasks such as the management, representation, semantic relationships, transformation and mapping between multiple domains, and meta data and meta-knowledge are essential for involving roles and data/knowledge intelligence in building up agent-mining

Nonfunctional requirements Nonfunctional requests are essential in real-world mining and agent systems. The agent-mining simians may more or less address nonfunctional requirements such as efficiency, effectiveness, action ability, user and business friendliness.

5. FRAMEWORK OF AGENT AND MINING INTERACTION AND INTEGRATION

This section aims to draw a concept map of agent mining as a scientific field. We observe this from the following perspectives: evolution process and characteristics, agent-mining interaction framework.

5.1 Evolution Process and Characteristics

As an emerging research area, agent mining experiences the following evolution process, and presents the following unprecedented characteristics.

From one-way interaction to Two-way interaction: The area was originally initiated by incorporating data mining into agent to enhance agent learning [20].Recently, issues in two-way interaction and integration have been broadly studied in different groups.

From single need-driven to mutual needs-driven: Original research work started on the single need to integrate one into the other, whereas it it is now driven by both needs from both parties. As discussed in [12, 8], people have found many issues in each of the related communities. These issues cannot be tackled by simply developing internal techniques. Rather, techniques from other disciplines can greatly complement the problem-solving when they are combined with existing techniques and approaches. This greatly drives the development of agent-driven data mining and data mining-driven agents.

Intrinsic associations and utilities: The interaction and integration between agents and data mining is also driven and connected by intrinsic overlap, associations, complementation and utilities of both parties, as discussed in [5, 6]. This drives the research on mutual issues, and the synergetic research and systems coupling both technologies, into a more advanced form.

Application drives: Application request is one of the key driving forces of this new trend. we present some major application domains and problems that may be better handled by both agent and mining techniques. Major research groups and researchers [6] in respective communities tend to undertake both sides of research. Some of them are trying to link them together to solve problems that cannot be tackled by one of them alone, for

instance, agent-based distributed learning [30, 31, 32, 25, 26], agent-based data mining infrastructure [4, 5, 26], or data mining driven agent intelligence enhancement [4].

5.2 Agent-mining interaction framework

The interaction and integration between agents and data mining are comprehensive, multiple dimensional, and inter disciplinary. As an emerging scientific field, agent mining studies the methodologies, principles, techniques and applications of the integration and interaction between agents and data mining, as well as the community that focuses on the study of agent mining. On the basis of complementation between agents and data mining, agent mining fosters a synergy between them from different dimensions, for instance, resource, infrastructure, learning, knowledge, interaction, interface, social, application and performance. As shown in Fig. 3, we briefly discuss these dimensions.

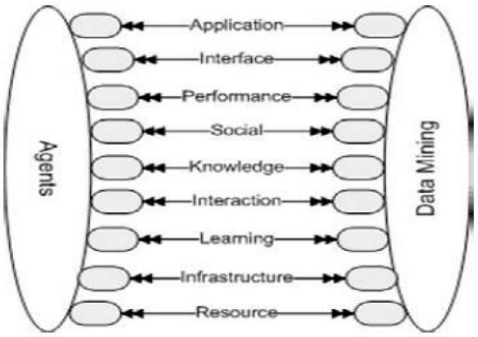

Figure 3: Multidimensional Agent Mining

Resource layer – interaction and integration may happen on data and information levels; Infrastructure layer – interaction and integration may be on infrastructure, architecture and process sides

Knowledge layer – interaction and integration may be based on knowledge, including domain knowledge, human expert knowledge, meta-knowledge, and knowledge retrieved, extracted or discovered in resources

Learning layer – interaction and integration may be on learning methods, learning capabilities and performance perspectives

Interaction layer – interaction and integration may be on coordination, cooperation, negotiation, communication perspectives

Interface layer – interaction and integration may be on human-system interface, user modeling and interface design;

Social layer – interaction and integration may be on social and organizational factors, for instance, human roles;

Application layer – interaction and integration may be on applications and domain problems;

Performance layer – interaction and integration may be on the performance enhancement of one side of the technologies or the coupling system.

From these dimensions, many fundamental research issues/problems in agent mining emerge. Correspondingly, we can generate a high-level research map of agent mining as a disciplinary area. Figure .4 shows such a framework, which consists of the following research components: agent mining foundations agent-driven data processing, agent-driven knowledge discovery, mining-driven multi-agent systems, agent-driven information processing, mutual issues in agent mining, agent mining systems, agent mining applications, agent mining knowledge management, and agent mining performance evaluation.

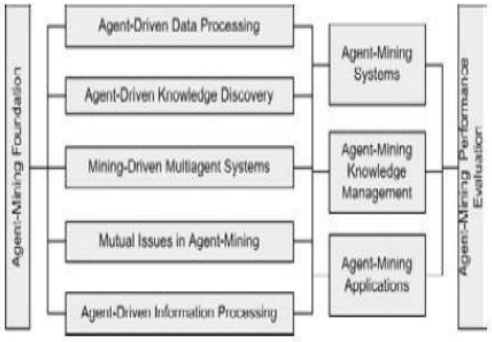

Figure 4: Agent Mining Frame work

6. APPLICATIONS

As we can see from many references, the proposal of agent mining is actually driven by broad and increasing applications. Many researchers are developing agent mining systems and applications dealing with specific business problems and for intelligent information processing. For instance, we summarize the following application domains.

- Artificial immune systems
- Artificial and electronic markets
- Auction
- Business intelligence
- Customer relationship management
- Distributed data extraction and preparation
- E-commerce
- Finance data mining
- Grid computing
- Healthcare
- Internet and network services, e.g., recommendation, personal assistant, searching retrieval, extraction services
- Knowledge management Marketing
- Network intrusion detection
- Parallel computing, e.g., parallel genetic algorithm

- Peer-to-peer computing and service
- Semantic web
- Text mining
- Web mining.

7. CONCLUSIONS

Agent and distributed data mining interaction and integration has emerged as a prominent and promising area in recent years. The dialogue between agent technology and data mining can not only handle issues that are hardly coped with in each of the interacted parties, but can also result in innovative and super-intelligent techniques and symbionts much beyond the individual communities. This chapter presents a high-level overview of the development and major directions in the area. The investigation highlights the following findings: (1) agent mining interaction is emerging as a new area in the scientific family, (2) the interaction is increasingly promoting the progress of agent and mining communities, (3) it results in ever increasing development of innovative and significant techniques and systems towards super-intelligent symbionts. As a new and emerging area, it has many open issues waiting for the significant involvement of research resources, in particular practical and research projects from both communities. We believe the research and development on agent mining is very promising and worthy of substantial efforts by both established and new researchers.

8. REFERENCES

[1] [1] Aciar, S., Zhang, D., Simoff, S., and Debenham, J.: Informed Recommender Agent: Utilizing Consumer Product Reviews through Text Mining. Proceedings of IADM2006. IEEE Computer Society (2006)

[2] [2] Batik's., Cho, J., and Bala, J.: Performance Evaluation of an Agent Based Distributed Data Mining System. Advances in Artificial Intelligence, Volume 3501/2005 (2005)

[3] [3] Cory, J., Butz, Nguyen, N., Takama, Y., Cheung, W., and Cheung, Y.: Proceedings of IADM2006 (Chaired by Longbing Cao, Zili Zhang, Vladimir Samoilov) in WI-IAT2006 Workshop Proceedings. IEEE Computer Society (2006)

[4] [4] Cao, L., Wang, J., Lin, l., and Zhang, C.: Agent Services-Based Infrastructure for Online Assessment of Trading Strategies. Proceedings of IAT'04, 345-349 (2004).

[5] [5] Cao, L.:

[6] http://wwwstaff.it.uts.edu.au/ lbcao/publication/publications.htm.

[7] [6] Cao, L., Luo, C. and Zhang, C.: Agent-Mining Interaction: An Emerging Area. AIS-ADM, 60-73 (2007).

[8] [7] Cao, L., Luo, D., Xiao, Y. and Zheng, Z. Agent Collaboration for Multiple Trading Strategy Integration. KES-AMSTA, 361-370 (2008).

[9] [8] Cao, L.:Agent-Mining Interaction and Integration Topics of Research and Development.http://www.agentmining.org/

[10] [9] Cao, L.: Data Mining and Multiagent Integration. Springer (2009).

[11] [10] Cao, L. and Zhang, C. F-trade: An Agent-Mining Symbiont for Financial Services. AAMAS 262 (2007).

[12] [11] Cao, L., Yu, P., Zhang, C. and Zhao, Y. Domain Driven Data Mining. Springer (2009).

[13] [12] Cao, L., Gorodetsky, V. and Mitkas, P. Agent Mining: The Synergy of Agents and Data Mining. IEEE Intelligent Systems (2009).

[14] [13]. Cao, L. Integrating Agent, Service and Organizational Computing. International Journal of Software Engineering and Knowledge Engineering, 18(5): 573-596 (2008)

[15] [14]. Cao, L. and He, T. Developing Actionable Trading Agents. Knowledge and Information Systems: An International Journal, 18(2): 183-198 (2009).

[16] [15]. Cao, L. Developing Actionable Trading Strategies, Knowledge Processing and Decision Making in Agent Based Systems, 193-215, Springer (2008).

[17] [16]. Cao, L., Zhang, Z., Gorodetsky, V. and Zhang, C.. Editor's Introduction: Interaction between Agents and Data Mining, International Journal of Intelligent Information and Database Systems, Inderscience, 2(1): 1-5 (2008).

[18] [17]. Cao, L., Gorodetsky, V. and Mitkas, P. Editorial: Agents and Data Mining. IEEE Intelligent Systems (2009).

[19] [18]. Cao, L. Agent & Data Mining Interaction, Tutorial for 2007 IEEE/WIC/ACM Joint Conferences on Web Intelligence and Intelligent Agent Technology (2007).

[20] [19]. Cao, L., Zhang, C. and Zhang, Z. Agents and Data Mining: Interaction and Integration, Taylor & Francis (2010).

[21] [20]. Brazdil, P., and Muggleton, S.: Learning to Relate Terms in a Multiple Agent Environment.EWSL91 (1991)

Integration of Agents and Data Mining. Technical

[22] [21]. Davies, W.: ANIMALS: A Distributed, Heterogeneous Multi-Agent Learning System. MSc Thesis, University of Aberdeen (1993)

[23] [22]. Davies, W.: Agent-Based Data-Mining (1994)

[24] [23].Edwards, P., and Davies, W.: A Heterogeneous Multi-Agent Learning System. In Deen, S.M. (ed)

[25] Proceedings of the Special Interest Group on Cooperating Knowledge Based Systems. University of Keele

[26] (1993) 163-184.

[27] [24].Gorodetsky, V., Liu, J., Skormin, V. A.: Autonomous Intelligent Systems: Agents and Data Mining book. Lecture Notes in Computer Science Volume 3505 (2005)

[28] [25]. Gorodetsky, V.; Karsaev, O.and Samoilov, V.: Multi-Agent Technology for Distributed Data Mining and

[29] Classification. IAT 2003. (2003) 438 - 441

[30] [26]. Gorodetsky, V., Karsaev, O. and Samoilov, V.: Infrastructural Issues for Agent-Based Distributed Learning. Proceedings of IADM2006, IEEE Computer Society Press

[31] [27]. Han, J., and Kamber, M.: Data Mining: Concepts and Techniques (2nd version). Morgan Kaufmann (2006)

[32] [28]. Kaya, M. and Alhajj, R.: A Novel Approach to Multi-Agent Reinforcement Learning: Utilizing OLAP

[33] Mining in the Learning Process. IEEE Transactions on Systems, Man and Cybernetics, Part C, Volume 35,

[34] Issue 4 (2005) 582 - 590

[35] [29].Kaya, M.and Alhajj, R.: Fuzzy OLAP Association Rules Mining-Based Modular Reinforcement Learning Approach for Multi-Agent Systems. IEEE Transactions on Systems, Man and Cybernetics, Part B, Volume 35, Issue 2, (2005) 326 - 338

[36] [30]. Klusch, M., Lodi, S.and Gianluca, M.: The Role of Agents in Distributed Data Mining: Issues and Benefits.

[37] Intelligent Agent Technology (2003): 211 - 217

[38] [31]. Klusch, M., Lodi, S. and Moro,G.: Agent-Based Distributed Data Mining: The KDEC Scheme. Intelligent

[39] Information Agents: The AgentLink Perspective Volume 2586 (2003) Lecture Notes in Computer Science

[40] [32]. Klusch, M., Lodi, S.and Moro,G.: Issues of Agent-Based Distributed Data Mining. Proceedings of AAMAS, ACM Press (2003)