Inf Technol Manag (2011) 12:67–79
DOI 10.1007/s10799-011-0091-8

Distributed data mining for e-business

Bin Liu · Shu Gui Cao · Wu He

Published online: 24 March 2011

Abstract In the internet-based e-business environment, most business data are distributed, heterogeneous and private. To achieve true business intelligence, mining large amounts of distributed data is necessary. Through a thorough literature review, this paper identifies four main issues in distributed data mining (DDM) systems for e-business and classifies modern DDM systems into three classes with representative samples. To address these identified issues, this paper proposes a novel DDM model named DRHPDM (Data source Relevance-based Hierarchical Parallel Distributed data mining Model). In addition, to improve the quality of the final result, the data sources are divided into a centralized mining layer and a distributed mining layer, according to their relevance. To improve the openness, cross-platform ability, and intelligence of the DDM system, web service and multi-agent technologies are adopted. The feasibility of DRHPDM was verified by building a prototype system and applying it to a web usage mining scenario.

B. Liu
State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, 100084 Beijing, China

B. Liu (✉) · S. G. Cao
College of Economics and Management, Hebei University of Science and Technology, 050018 Shijiazhuang, China
e-mail: sjzbit@gmail.com

S. G. Cao
e-mail: caoshugui@163.com

W. He
Center for Learning Technologies, Old Dominion University, Norfolk, VA 23529, USA
e-mail: whe@odu.edu

Keywords Distributed data mining · e-business · Web service · Multi-agent · Knowledge integration

1 Introduction

With the development of information technologies, such as computation, network, and database, e-business [16, 28, 43] has rapidly extended its scale, scope, and measures during the last two decades. Due to the openness, globalization, and virtualization of network economics, the e-business market has become larger and more complicated than ever. The major e-business pressures are labeled the 3Cs: competition, customers, and change [34]. To alleviate the business pressures, an effective way is to use Business Intelligence (BI) [15, 33, 42, 53], which requires enterprises to use data mining (DM) [19–21, 23, 35] tools to analyze business data. DM, as one of the promising technologies (since the 1990s), is a non-traditional data-driven method that can discover novel, useful, hidden knowledge from massive data sets. It has been considered very useful for data analysis in business, industries, and engineering [4].

A large amount of e-business data are increasingly generated at distributed locations. In many cases, it is not feasible to transfer all of the data to a central location for DM due to security issues, limited network bandwidth, or the internal policies imposed by some organizations [7]. Distributed data mining (DDM) is an extension of DM techniques in distributed data environments. DDM can also be used to effectively speed up the DM process, even if the data is not physically distributed. However, the primary purpose of DDM is to discover and combine useful knowledge from databases that are physically distributed across multiple sites [45]. Giannella et al. [14] describe two

main advantages of using DDM: (1) lower network traffic (On each data source site, it processes the data and sends the results (local model) back to the main host, instead of transferring large amount of data across the network, which can take much time [46]. As the local model is much smaller than the local data, sending only the model can substantially reduce the network traffic and require much less network bandwidth), and (2) better security (Sharing only the model, instead of the entire data, could mean better security for some organizations since it overcomes the issue of data privacy).

The rest of the paper is organized as follows: Section 0 describes the related concepts of BI and DDM. Applicable scenarios and issues of DDM for e-business are also described. Section 3 classifies modern DDM systems into three classes and provides some representative examples; the m issues with modern DDM systems are also summarized in this section. To help address these issues with DDM systems, Sect. 4 proposes a novel DDM model which divides the system into layers (based on data source relevance) to support hierarchically parallel mining. Section 5 evaluates the feasibility of the proposed DDM model by verifying a prototype system using the proposed model with a web mining experiment. Section 6 presents the conclusion and suggestions for future research.

2 Related concepts of BI and DDM

2.1 Concept of BI

BI [15, 33, 42, 53] is the concept of applying a set of technologies to convert data into meaningful information. BI tools include information retrieval, DM, statistical analysis, and data visualization. Using the tools, large amounts of data originating in different formats and from different sources can be consolidated and converted to key business knowledge. Figure 1 presents a general view on how to transfer data into BI knowledge. The process involves both business experts and technical experts. It converts a large scale of data into meaningful outcomes so as to provide decision-making support to end users [57].

2.2 DDM and its applicable scenario

DM deals with the problem of analyzing a large amount of data in a scalable manner [19]. DDM is a branch of DM that offers a framework to mine distributed data with careful attention to the distributed data and computing resources. A distributed scenario (where DDM is applicable) may have the following features [6]:

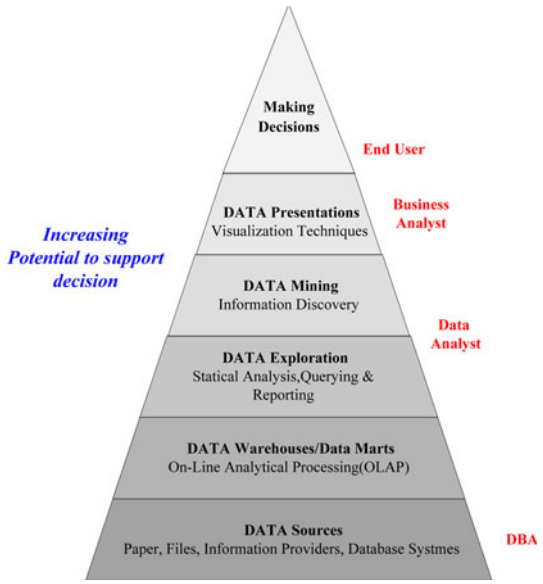

Fig. 1 BI processing [57]

1. The system consists of multiple independent sites of data and computation which communicate only through message passing.
2. Communication between the sites is expensive.
3. Sites have resource constraints.
4. Sites have privacy concerns.

2.3 Issues in DDM for e-business

In e-business, most of the daily produced data are distributed in sites (e.g., websites, departments, companies of the same business chain). Many of the sites can be connected through networks which provide the environment for DDM. Some studies [10, 64] are conducted to address the following issues in order to improve the performance of DDM systems.

- *Heterogeneous versus homogeneous DM*. In a centralized DM, most of the work is to deal with the homogeneous data, which means the data are maintained by the same DBMS and management model. If the data are heterogeneous, the local data management model is usually integrated and converted to the global model before conducting DM. Otherwise, contradictions among attributes may occur.
- *Data variety in dynamic environment*. In traditional DM, the data is regarded as static, and the mining work is executed in an environment owns enough data. In an e-business environment, the data related to business is time-varying in nature and it is a challenge to correctly transfer the time series related result.
- *Communication cost*. In centralized DM, the I/O and CPU time costs are considered when designing the

algorithm. In a distributed data environment, the communication cost, which depends on the network bandwidth and the amount of transferred information, needs to be considered.

- *Knowledge integration* [26, 49]. For DDM, the final purpose is to get the local result by analyzing local sitesand integrating the local results to produce the global result. To analyze the local data set, we can use the existing centralized DM methods. To integrate the local results, a traditional simple integration method may not work. For instance, the samples that are locally interesting may lose their value at the global level. It is necessary to collect all of the local interests and verify the global interest degree in order to produce the final result.

3 Modern DDM Systems

Modern DDM systems can be classified as follows:

3.1 DDM systems based on parallel DM agents

PArallel Data Mining Agents (PADMA) is a multi-agent [2, 44, 61] based architecture for DM. It is a system that makes use of intelligent DM agents, which are responsible for accessing, analyzing, and discovering the hidden patterns within the data warehouse. The agents work together in conjunction with each other and share the same repository or metadata [29]. The main purpose of designing PADMA is to realize the coordinated parallel mining by multi-agent technologies, which can enhance the efficiency.

Albashiri et al. [1] proposed an Extendible Multi-Agent Data mining System (EMADS), whose vision is that a community of DM agents (contributed by many individuals) can interact with one another under decentralized control to address DM requests. EMADS is considered both as an end-user application and a research tool. In EMADS, there is an anarchic collection of persistent, autonomous (but cooperating) DM agents operating across the Internet. Individual agents have different functionalities; the system currently comprises data agents, user agents, task agents, mining agents, and a number of ‘‘house-keeping” agents. Users of EMADS may be data providers, DM algorithm contributors, or miners of data. The current functionality of EMADS is limited to classification and meta association rule mining. Figure 2 offers a high- level view of EMADS.

Danish [8] proposed a DM architecture of “CAKE” (Classifying, Associating and Knowledge DiscovEry) using centralized metadata, which contains all the rules of classification and association,along with the data structure

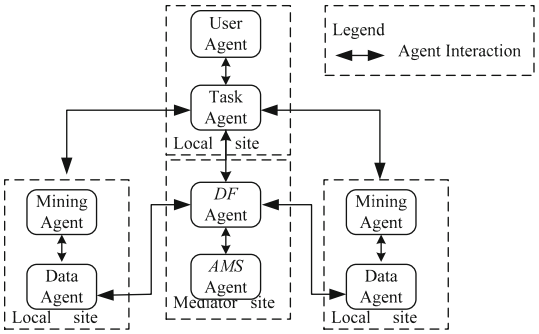

Fig. 2 High level view of EMADS conceptual framework [1]

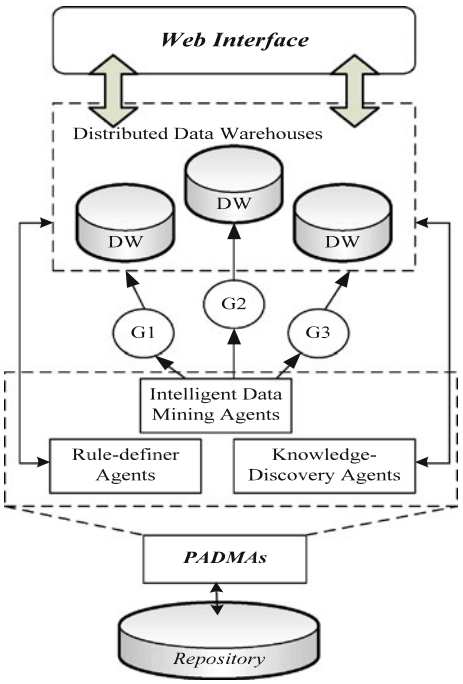

Fig. 3 CAKE (Architecture) [8]

details. The “web interface” is used to provide the users with the interface so that they can view the result. Future work on CAKE will improve its ability of dealing with heterogeneous data sources and complex mining needs. Figure 3 shows the architecture of CAKE. Rule-definer agents are used to define the metadata of the data warehouse on the basis of the rules that are going to be defined by the users. These rules are then going to be used by the “Intelligent Data Mining Agents” for DM and by “Knowledge Discovery Agents” to drive the knowledge out from the defined patterns. “Intelligent Data Mining Agents” are a group of agents which can be set up to work on a specified set of data with defined rules at any location.

Chen et al. [5] proposed a DDM system to effectively solve the problems caused by network bandwidth limit, data privacy, and system incompatibility when mining

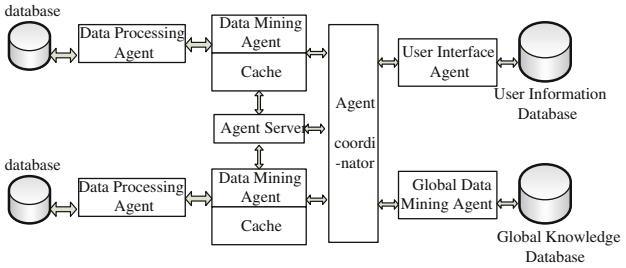

Fig. 4 DDM system architecture proposed in [5]

distributed data with a traditional centralized DM model. Figure 4 illustrates the system architecture. Taking into account the complexity of the data processing and the feasibility and reliability of system realizing, the Data Mining Agent (DMA), which plays the core role in the system, executes the mining task. The agent coordinator negotiates and synchronizes among the modules.

3.2 DDM systems based on meta-learning

Meta-learning is a recently developed technique that deals with the problem of computing a "global" classifier from large and inherently distributed databases. Meta-learning aims to compute a number of independent classifiers (concepts or models) by applying learning programs to a collection of independent and inherently distributed databases at the same time. The "base classifiers" computed are then collected and combined by another learning process. Here, meta-learning seeks to compute a "meta-classifier" that integrates in some principled fashion the separately learned classifiers to boost overall predictive accuracy [45].

The main purposes of designing this kind of system is to improve the quality of selection and the composition of DM algorithms, and to select the reasonable DM model according to the relevance of different data sources. Tozicka et al. [52] proposed a framework for agent-based distributed machine learning and DM based on (1) the exchange of meta-level descriptions of individual learning processes among agents and (2) online reasoning about learning success and learning progress by learning agents. Figure 5 illustrates a generic model of a learning step. To improve the utility of the framework, the communication

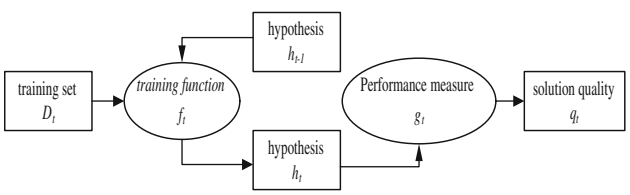

Fig. 5 A generic model of a learning step [52]

cost should be considered first; secondly, experiments with agents using completely different learning algorithms (e.g. symbolic and numerical) should be executed.

Dam et al. [7] proposed an evolutionary-based online-learning system called XCS in conjunction with the knowledge probing technique. XCS is a genetic-based machine learning algorithm that applies a reinforcement learning scheme. Luo et al. [37] considered an execution engine as the kernel of the system to provide mining strategies and services, and proposed an extensible architecture for this engine, based on a mature multi-agent environmentwhich connects different computing hosts to support intensive computing and complex process control via distribution (see Fig. 6). Reuse of existing mining algorithms is achieved by encapsulating them into agents. The algorithms also define a DM workflow as the input of the engine and detail the coordination process of various agents to process it.

Yang et al. [56] proposed a Service Oriented Architecture for Knowledge Discovery (SOA4KD) (see Fig. 7), which selects and executes the knowledge discovery algorithm by using a meta-learning and semantic web service. User requirement is divided into a content part and a quality part. An extended knowledge discovery task ontology is proposed which can acquire user requirements through a natural language interface along with domain ontology. A Knowledge Discovery Service (KDS) quality ontology which considers the unique characteristic of KDS (as well as the characteristics of general service) is proposed. In this ontology, meta-learning is used to select the most appropriate KDS, according to user requirements. However, to guarantee the reliability and integrity of the user's need being expressed in a natural language, the need is restrained in the given set.

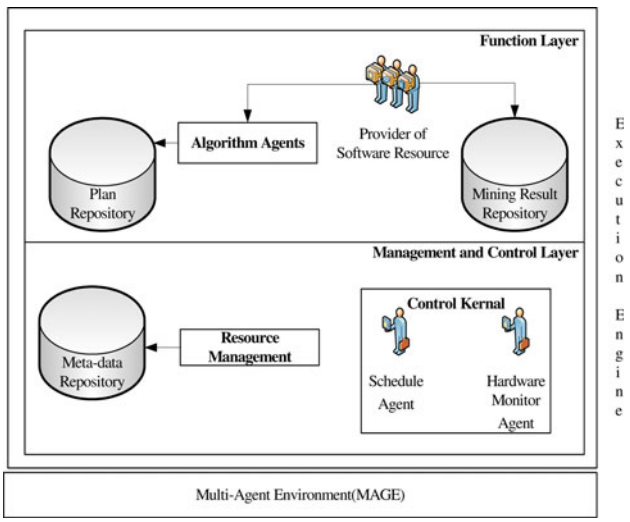

Fig. 6 System architecture of execution engine[37]

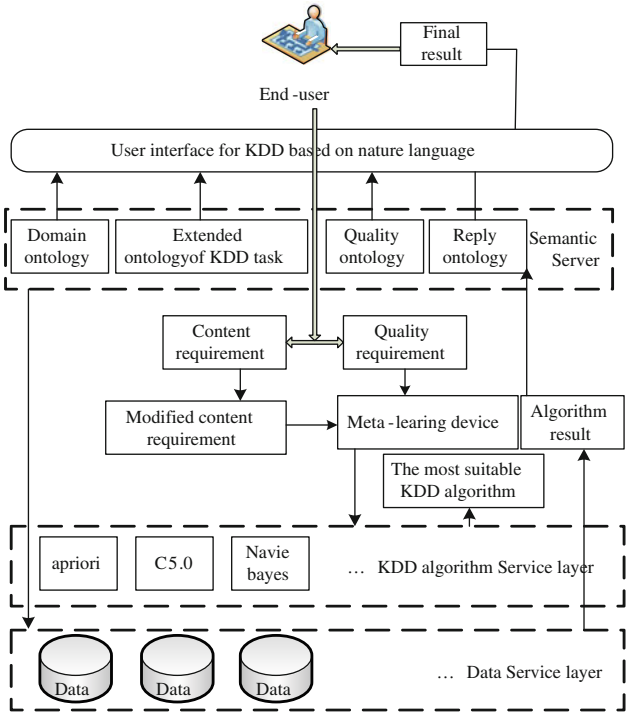

Fig. 7 Architecture of SOA4KD[56]

3.3 DDM systems based on Grid

DM in grid [13, 30] computing environment represents a specific incarnation of DDM motivated by resource sharing via local and wide area networks [48]. Increased performance, scalability, access, and resource exploitation are the key drivers behind such endeavors. However, several factors hamper large-scale DM applications on a grid. To begin with, Grid computing [50, 58, 62] itself is relatively new, and relevant standards and technologies are still evolving. A plethora of DM technologies and a staggering number of largely varying DM application scenarios further complicate matters. Finally, DM clients range from highly domain-oriented end users to technology-aware specialists. For highly domain-oriented end users, user transparency and ease-of-use is paramount. Technology-aware specialistsneed to control certain detailed aspects of DM and grid technology [48].

Today, new DDM projects aim to mine data in a geographically distributed environment which is based on grid standards and platforms, in order to hide the complexity of heterogeneous data and lower level details. So their architectures are becoming more sophisticated to articulate with grid platforms as well as to supply a user-friendly interface for transparently executing DM tasks. When running computationally intensive processes such as DM operations in a dynamic grid environment, it is advantageous to have an accurate representation of the available resources and their current status. A grid-enabled environment has the potential to solve this problem by providing core processing capabilities with secure, reliable, and saleable high bandwidth access to various distributed data sources and formats across various administrative domains [59].

Based on the principle of SOA [17, 24, 31, 32, 54, 60], standardization and open source, Stankovski et al. [48] proposed a DDM system based on a Data Mining Grid (DMG). Figure 8 depicts the DMG system architecture in four layers. Generally, components in higher layers make use of components organized in lower layers. The layer at the bottom represents software and hardware resources; The Globus Toolkit 4 layer depicts some of the system's core grid middleware components; the high-level services layer shows components providing central DMG services; and the client components layer depicts the DMG applications' client side components.

Cesario et al. [3] proposed a general DM architectural model (see Fig. 9) that can be exploited for different DM algorithms deployed as Grid services for the analysis of dispersed data sources. In the future, they intend to deploy more DM algorithms (clustering, frequent item sets, and association rule). Since the proposed architecture does not have the splitting functionality of datasets or a data transfer utility, they will build a framework that can give this functionality to the users.

Luo et al. [36] systematically analyzed the issues of agent Grid and implemented an Agent Grid Intelligent Platform (AGrIP) which provides an infrastructure for agent-based DDM in a Grid environment. They proposed a

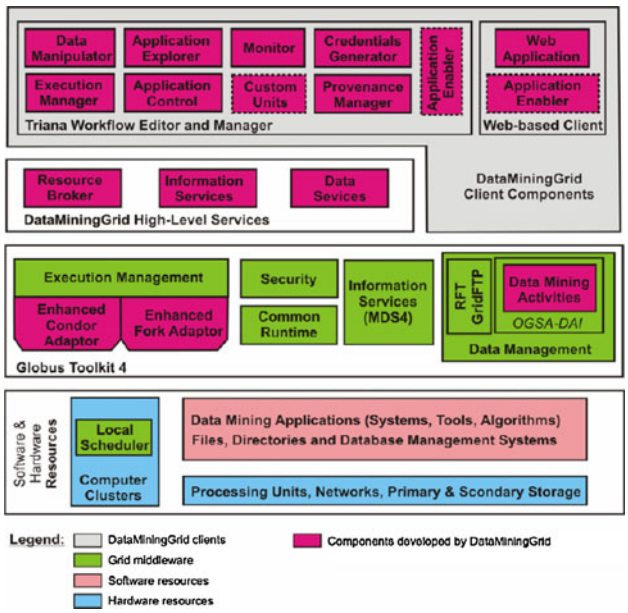

Fig. 8 The DMG system architecture [48]

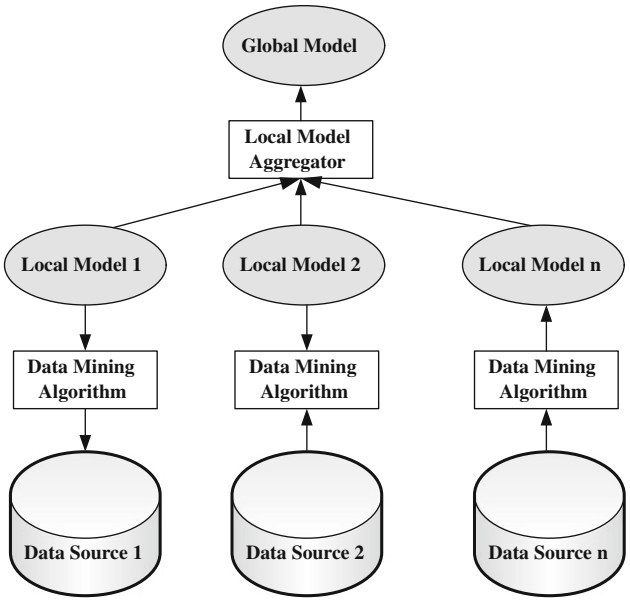

Fig. 9 Typical architecture of a DDM algorithm [3]

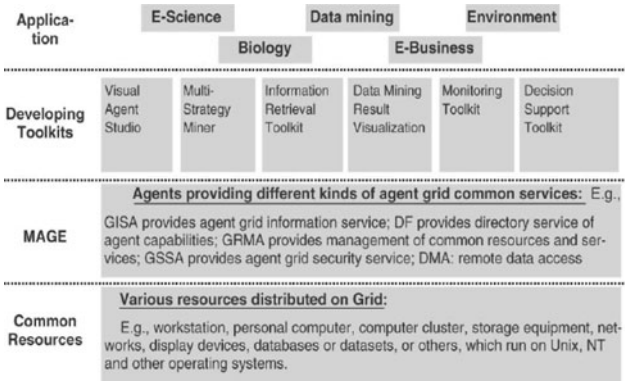

Fig. 10 Architecture of AGrIP [36]

four-layer model for AGrIP platform from an implementation point of view, as illustrated in Fig. 10:

- Common resources: various resources distributed in Grid environment, such as workstations, personal computers, computer clusters, storage equipment, databases, data sets, or others, which run on Unix, NT, and other operating systems.
- Agent environment: the kernel of Grid computing which is responsible for resources location and allocation, authentication, unified information access, communication, task assignment, and agent library.
- Developing toolkit: the development environment, containing agent creation, information retrieval, and distributed DM, to let users effectively use Grid resources.
- Application service: certain agents organized automatically for specific application purposes, such as e-science, e-business, decision support, and bio-information.

3.4 Issues of the modern DDM systems

Based on the above analysis of modern DDM systems, three main issues are summarized as follows:

- Most DDM systems' architecture is closed and that lack openness and platform independence can make it difficult to dynamically manage the DM algorithms. However, in e-business, the decision making support cases (such as customer segmentation, personal service, and cross selling) are complex and need to be solved by dynamically combining multiple DM algorithms.
- The data source relevance has not been given enough consideration, and the single DDM approach cannot guarantee the quality of the final global result.
- There is a lack of effective methods to integrate the result of local DM.

4 A novel DDM model for e-business

Based on our years of practical experience in the data mining area, in order to improve the efficiency of e-business DDM system and explore a solution to address these common issues (see Sect. 3.4), we propose a Data source Relevance based Hierarchical Parallel Distributed data mining Model (DRHPDM) (see Fig. 11) which adopts web service [11, 25, 39] and multi-agent [41] technologies.

4.1 Features of DRHPDM

The main features of **DRHPDM** are listed as follows:

- To improve the openness, cross-platform ability, and intelligence of the DDM system, web service encapsulating the DM algorithm and multi-agent will be adopted. Web service is a new distributed computing model which has the features of platform dependency, unified data representation and supporting component reuse [11, 25, 39]. It can effectively realize the publishing, discovering, and calling of function bodies [47]. Multi-agent has the features of agent (e.g. autonomy, initiative and adaptability) and can realize the cooperation of agents. It can effectively support the execution of DDM from local to global use [41].
- The data sources which have strong relevance to one another will construct the Local Centralized Data Mining Layer (LCDML), which can improve the quality of the final global DDM results.
- When mining the local data source, other resources can help realize parallel mining. In summary, the data source and the other resources of the local site can construct the Local Parallel Data Mining Layer (LPDML).

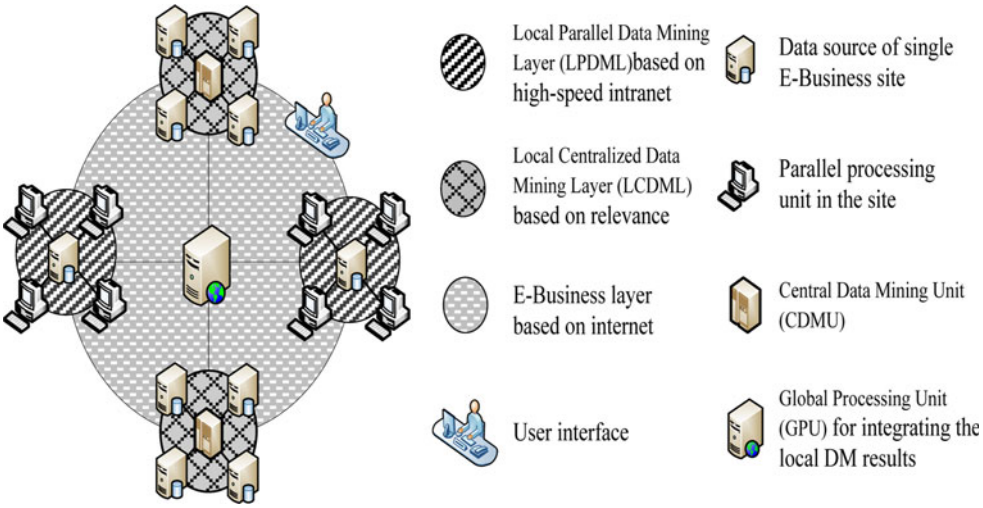

Fig. 11 Hierarchical parallel DDM model in E-Business

- The local mining results will be transferred to the Global Processing Unit (GPU) that is responsible for integrating the local results in order to produce the final global results for DDM.

4.2 Workflow of DRHPDM

For an e-business enterprise whose distributed sites (holding different data sources) are connected by Internet, the following workflow will be activated when the user (decision maker, market analyzer, data analyzer, etc.) wants to extract knowledge from a large amount of data (see Figs. 12, 13):

Step 1: The user logs into the system and submits the DM requirement to the GPU.
Step 2: On the GPU, the DM requirement is analyzed and divided into a series of DM tasks.
Step 3: Consequently, the tasks are broadcast to all the distributed sites which hold data sources belonging to the user's corporation.
Step 4: As a single site can deal with different businesses, a site may have different data sources corresponding to different businesses. When a site holding different task-related local data sources receives the DM task, the site will register the information (including the IP, the data file, the data type, etc.) in the GPU.
Step 5: On the GPU, according to the registered information, the data sources are divided into sets corresponding to different DM tasks. The relevance among the data sources of the same set will be appraised by a special program (see Sect. 4.3). After the appraisal, the data sources that have high relevance to one another are grouped together to produce centralized DM layers (or LCDMLs—see Sect. 4.1); the others will be discretely mined to produce the parallel DM layers (or LPDMLs—see Sect. 4.1). Finally, the GPU notifies the related data sources about the grouping results.
Step 6: The data sources, which are notified to be discretely mined, will implement locally parallel DM in their sites. The data sources, which are notified to be centrally mined, will transfer necessary data to the Central Data Mining Unit (CDMU). The CDMU can be one of the sites holding the data sources, or another site with enough computing and storage resources can be used.
Step 7: The DM results (models) are transferred to Local Managing Agent (LMA) (see Sect. 4.4) for local integration. The integrated models are then consequently transferred to GPU for global integration.
Step 8: To improve the readability, the final global model is transformed and submitted to the user.

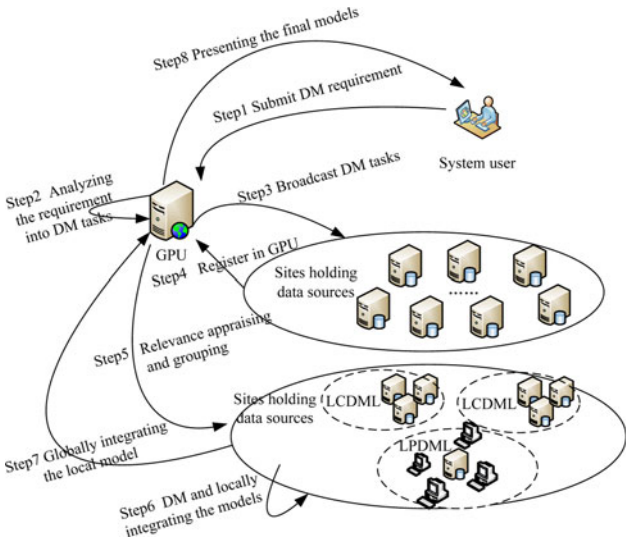

Fig. 12 Workflow of DRHPDM

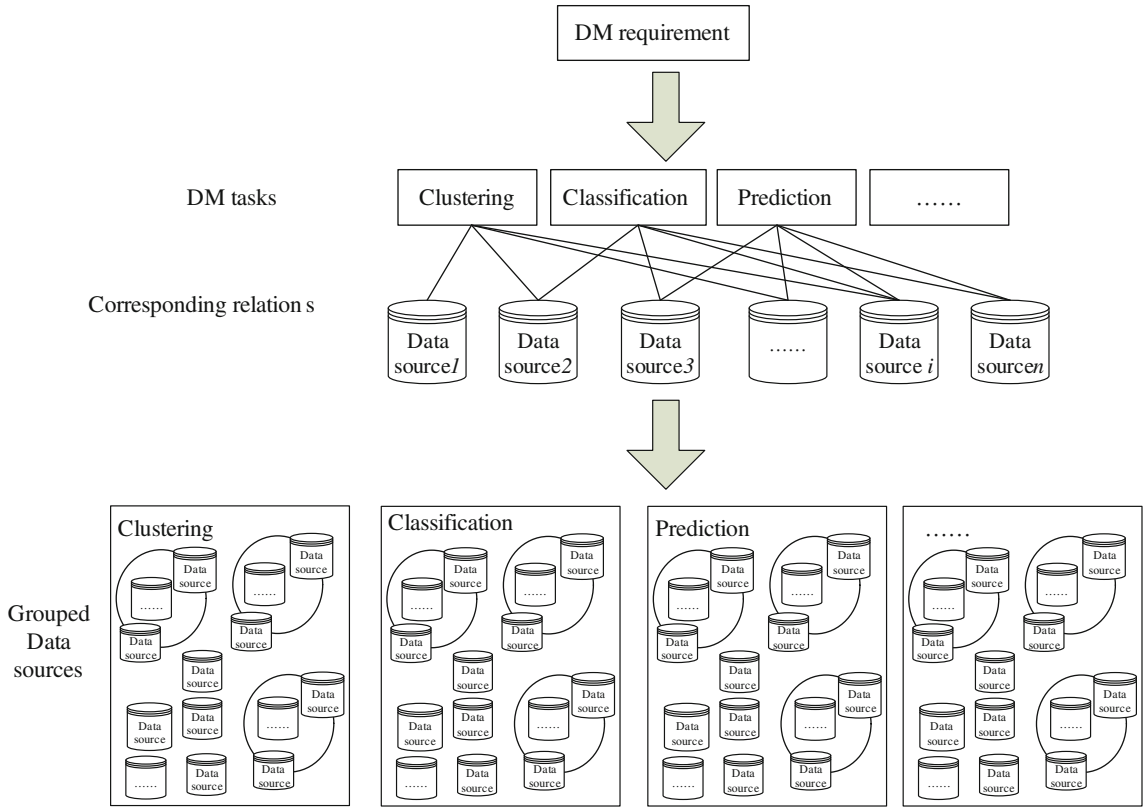

Fig. 13 Dataflow of DRHPDM from step1 to step 5

4.3 Measuring the data source relevance based on ontology and semantics

Ontology [18, 55] serves as the metadata schemas, providing a controlled vocabulary of concepts, each with explicitly defined and machine-processable semantics. By defining shared and common domain theories, ontology helps people and machines to communicate concisely by supporting semantics exchange, rather than just syntax [38]. Semantic models based on ontology can accurately and completely depict the concepts and the relevance among concepts [40], and the interaction among data sources in the ontology semantic model layer offers such features as completeness, accurateness, and efficiency. As a result, using ontology and semantics technology, we can build a data source relevance measuring model which will measure the relevance of the DM task-related databases existing in the corresponding data sources. In addition, the databases with strong relevance should be centralized mined in order to guarantee the quality of final DM result.

As shown in Fig. 14, the DM task-related database should be reverse-engineered to produce its E-R (Entity Relation) model first. To improve the quality of the E-R model, it should also learn from instance data. Sequentially, the E-R model is translated to the data source ontology and is expressed by Web Ontology Language (OWL).

The data source semantics model is defined as follows:

$$DSSM = \{G(V, E), \Gamma, \Lambda, N, T, A^o\} \quad (1)$$

Where G denotes a graph based on Unified Modeling Language (UML) class graph, the note set V corresponds to the entity set, and edge set E denotes the relations among entity concepts; Γ denotes the set of entity concept and $\Gamma = \{c_1, c_2, \ldots, c_n\}$; Λ denotes the set of entity relation and $\Lambda = \{r_1, r_2, \ldots, r_n\}$; the mapping relation between V and Γ is built by function set N; the mapping relation between E and Λ is built by function set T; and A^o is the restrictive axiom on Λ.

Sequentially, the relevance of the entity concepts (vocabulary) belonging to the data source semantic model is measured and the semantic relevance can be computed. The executive solution can be chosen from one of the following two measuring methods (according to the situation):

- Measuring semantic relevance based on concept lattice: The form context (G ∪ {g}, M, J) should be built according to domain knowledge related to concrete e-business. The concept lattice can be set up according to the form context; the semantic relevance between objects (terms) can be computed according to the hierarchy of the concept lattice.
- Measuring semantic relevance based on HowNet: Firstly, the mechanism for computing relevance

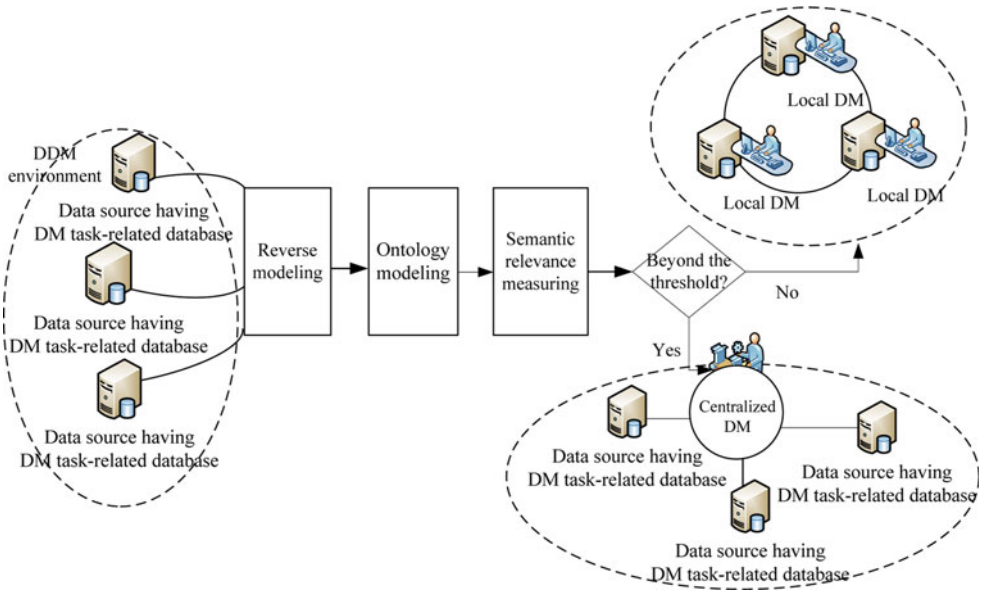

Fig. 14 Measuring data source relevance

between the sememes (atomic or indivisible units) should be set. Then the relevance between the terms can be computed according to the computing results between the sememes. Finally, the relevance between data sources can be produced according to the relevance between the sememes.

4.4 Knowledge integration model

Based on the sharing of findings in the integration model, agents can share the knowledge with each other. As shown in Fig. 15, the Local Managing Agent (LMA) is responsible for collecting, recording, and publishing the knowledge mined by the Data Mining Agent (DMA). When a DMA in a local site needs knowledge, it will first verify whether other sites can provide such knowledge. If so, the knowledge on other sites will be transferred to the local site. Otherwise, the DMA should locally mine the knowledge. If the LMA has too much knowledge to hold, DMAs will be divided into blocks, with DMAs with strong relevance being divided into the same block. The knowledge mined in one block can be interchanged with other blocks by the Global Integrating Agent (GIA) which exists

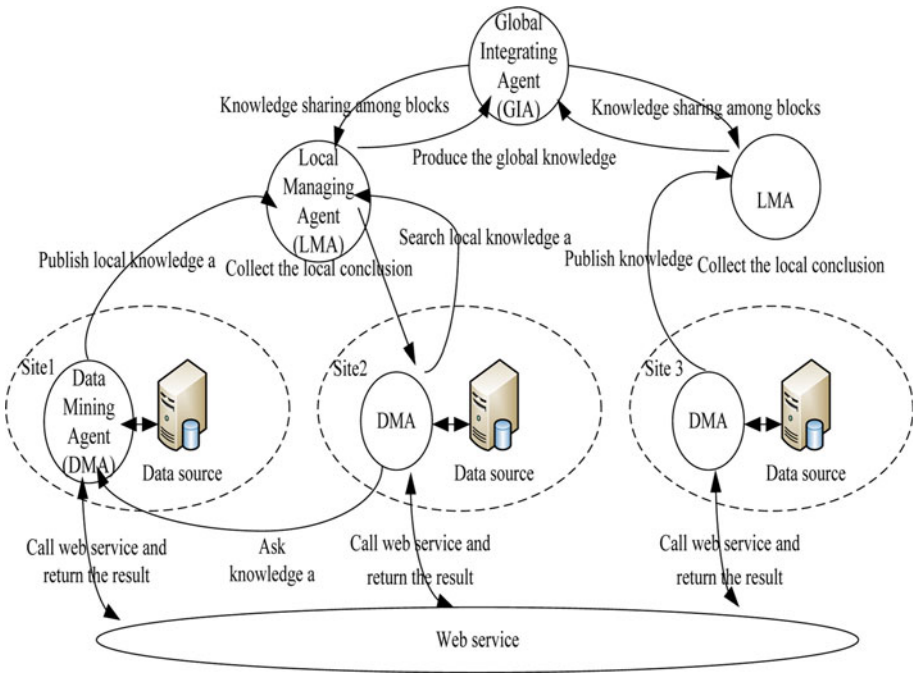

Fig. 15 Knowledge-integrating model

in the GPU and is responsible for producing the global knowledge.

4.5 Web service composition

A complicated e-business mining task often needs multiple DM algorithms to cooperate. In DRHPDM, the algorithms are encapsulated into web services; it is necessary to study the web service composition method to realize flexible cooperation among DM algorithms.

Figure 16 depicts the whole procedure of web composition:

Step 1: A service applicant (e.g. DMA) submits its service requirement including purpose and restraints (e.g., Quality of Service (QoS)) to the composition system;
Step 2: A compositing agent analyzes the requirement and extracts the core content to the Planning and Designing Subsystem (PDS). Sequentially, PDS produces several functional flows and processes the reasoning and verifies the work for these flows. PDS accesses the web service library and chooses the most suitable services; finally, the execution flow is made by PDS according to the functional flows.
Step 3: According to the execution flow and QoS, the Appraisal and Optimizing Subsystem (AOS) produces the optimized suitable execution flow.
Step 4: The Execution and Monitoring Subsystem (EMS) receives the execution flow from the AOS and register it so that it can be used by users.

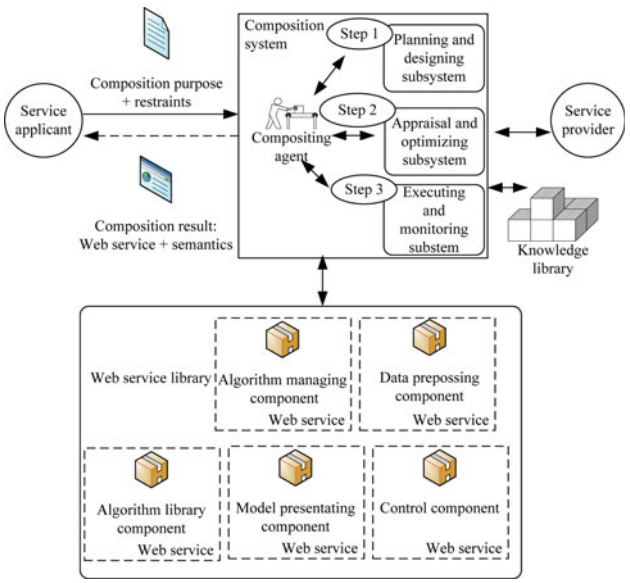

Fig. 16 Web service compositing model

Step 5: The composition results (including web service and the related semantics) are provided to the service applicant.

5 Experimental evaluation

Nowadays, Web Usage Mining (WUM) [9, 12] has become a powerful way for realizing BI. Accurate Web usage information could help attract new customers, retain current customers, improve cross marketing or sales, increase effectiveness of promotional campaigns, track leaving customers, and serve as the most effective logical structure for a user's Web space [22, 27]. In this section, we verify the feasibility of DRHPDM with a WUM case. We have performed the experiments using the scenario that an e-business manager wants to improve the website topology by using WUM, with a condition that log files be distributed in different local websites.

To establish the experimental environment, we adopted real server log files which record visitors' accessing behavior (such as time, page visits, and IP addresses). The log files were provided by a Chinese e-business website located in Hebei Province, which consists of a LAN with 10 heterogeneous nodes. To simulate the real world, the server log files were divided into six parts and were randomly distributed to six nodes; the DM algorithms were encapsulated into web services and were deployed into different nodes. We also built a prototype system of DRHPDM using block-based design methods to guarantee its scalability. The system was developed with Java to guarantee its cross-platform ability.

The workflow of the prototype system was designed according to the description in Sect. 4.2. The main experimental steps are listed as follows:

Step 1: After a series of preprocessing operations including data clearing, user identifying, session identifying and path identifying, we identified 320 sessions from 5,236 users and the total number of URLs decreased from 2,120 to 210;
Step 2: As Fig. 17a shows, the LAN nodes connecting to the local machine were displayed;
Step 3: As Fig. 17b shows, the listed nodes are those that have task related data (log files) and have registered themselves in the local machine which is working as the GPU and running the system;
Step 4: As Fig. 17c shows, the nodes holding the task related data were divided into thirds, according to the relevance among themselves;
Step 5: When changing the node of the "DM Web Service" listbox, the DM algorithms provided by the node in the Web service form was consequently

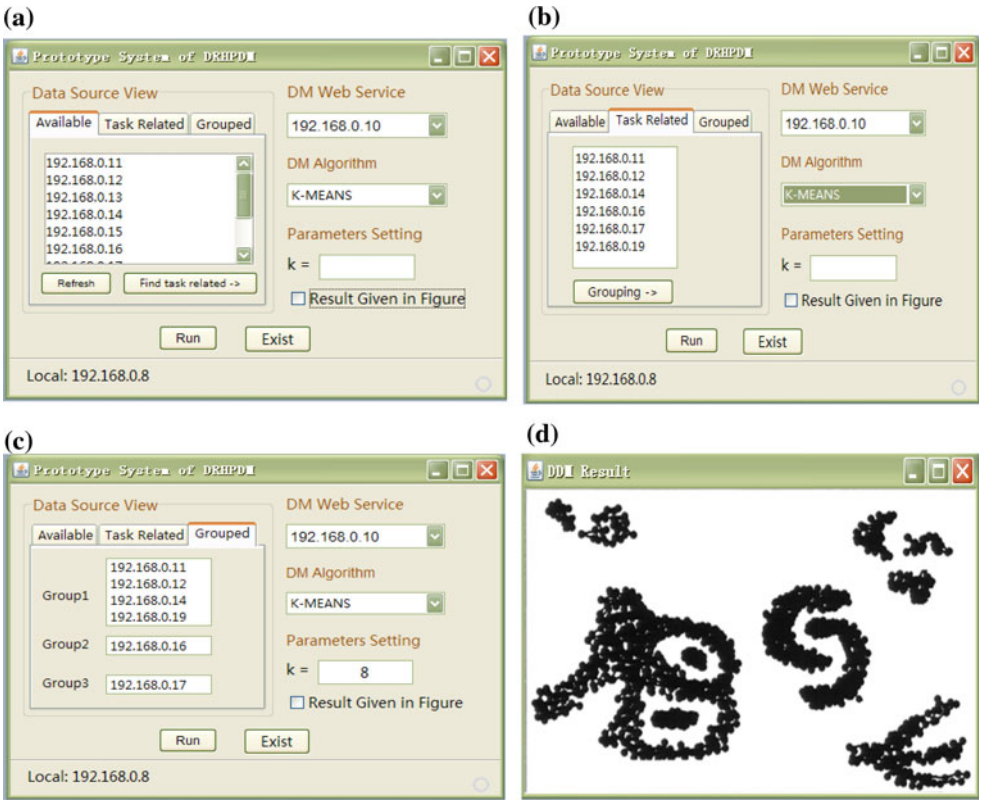

Fig. 17 Prototype system of DRHPDM

presented in the "DM Algorithm" listbox. As Fig. 17c shows, we selected the "K-Means" algorithm provided by the LAN node with the IP "192.168.0.10" to mine the patterns from log files. The parameters that should be given were listed in the "Parameters Setting" area;

Step 6: As Fig. 17d shows, after the "Run" button was clicked, the URLs were able to be clustered into five main parts with distributed K-Means algorithm in about 20 s.

According to the clusters shown in Fig. 17d, the web page URLs numbered from 82 to 112 are clustered together. When checking their relations using the website topology, we found that they all belong to the "online payment" module. In other words, the clustering results are consistent with the real business module of the website.

On the other hand, the URL numbered 37 was clustered together with the URLs numbered from 140 to 156, but according to the website topology, the web page 37 does not have direct hyperlinks to the web pages from 140 to 156 (which belong to "personal information management" module). To obtain in-depth analysis of the phenomenon, we checked the content of page 37 and found that it provided the entrance to realize "self-assistant website construction." That is to say that, according to the cluster, we found a regular pattern that showed that most of the website's visitors were apt to access the "personal information management" module and the "self-assistant website construction" module in one session. There are over 20 modules (such as user registration, personal information management, self-assistant website construction, games, job advertising, bargains, etc.) on this e-business website. As a result, we can add the hyperlinks on page 37 to the homepage of the "personal information management" module to help users find information more quickly and thus to enhance the user experience and to improve the information seeking efficiency. In conclusion, the WUM, using our prototype system, does provide the capability to help improve the website topology.

6 Conclusion and future research

Nowadays, with the support of information technology, e-business has been rapidly growing. How to make use of the e-business data to support decision making has become a main focus of the BI area. In this paper, we reported on an in-depth investigation on the issues of DDM in the e-business data environment and modern DDM systems. Based on the literature review and on our in-depth analysis, we proposed the "Data source Relevance based Hierarchical Parallel Distributed data mining Model (DRHPDM)." Giving enough consideration to the data relevance and to improvingthe final mining result, the data sources are divided into two kinds of layers: a centralized

mining layer and a distributed layer. By adopting web service and multi-agent, DRHPDM shows the capability to realize flexible, cross-platform mining.

In the future, we will conduct more research on the realization of local centralized data mining as well as on the threshold of the relevance among data sources. We will also involve a variety of users to systematically evaluate the prototype system, in order to further improve the usability, efficiency, and effectiveness of our prototype system and the DRHPDM model. Our proposed DRHPDM model is transformative and can be applied to other fields such as e-learning, e-government, etc. Finally, our proposed DRHPDM model can be further extended and integrated with recent advances in distributed text mining [51, 63] to help discover more hidden knowledge, patterns, and insights.

Acknowledgments The research is supported by Natural Science Foundation of Hebei Province (No. G2010000903), and the Doctor Start-up Research Fund of Hebei University of Science and Technology (No. QD200945).

References

1. Albashiri KA, Coenen F, Leng P (2009) EMADS: an extendible multi-agent data miner. Knowl Based Syst 22(7):523–528
2. Brintrup A (2010) Behaviour adaptation in the multi-agent, multi-objective and multi-role supply chain. Comput Ind 61(7): 636–645
3. Cesario E, Talia D (2008) Distributed data mining models as services on the grid. In: IEEE International Conference on Data Mining Workshops, Pisa, TBD, Italy, pp 486-495
4. Chen GQ, Wei Q, Liu D, Wets G (2002) Simple association rules (SAR) and the SAR-based rule discovery. Comput Ind Eng 43(4):721–733
5. Chen ZY, Liu S F, Liu G (2008) The multi-agent knowledge management system model for pervasive computing. In: 3rd international conference on pervasive computing and applications, Alexandria, Egypt, pp 70–73
6. Da Silva JC, Giannella C, Bhargava R, Kargupta H, Klusch M (2005) Distributed data mining and agents. Eng Appl Artif Intell 18(7):791–807
7. Dam HH, Abbass HA, Lokan C (2005) DXCS: an XCS system for distributed data mining. In: Proceeding of the 2005 conference on genetic and evolutionary computation, Washington, DC, USA, pp 1883–1890
8. Danish K (2008) CAKE-classifying, associating and knowledge discovery-an approach for distributed data mining (DDM) using parallel data mining agents (PADMAs). In: IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, Sydney, Australia, pp 596–601
9. Das R, Turkoglu I (2009) Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. Expert Syst Appl 36(3):6635–6644
10. Davies WHE, Edwards P (1995) Agent-based knowledge discovery. Working Notes of the AAAI Spring Symposium. Information Gathering from Heterogeneous, Distributed Environments, Palo Alto, California, pp 34–37
11. David M, Massimo P, Matthias W (2007) Bringing Semantics to Web Services with OWL-S. World Wide Web 10(3):243–277
12. Eirinaki M, Vazirgiannis M (2003) Web mining for web personalization. ACM Trans Internet Technol 3(1):1–27
13. Foster I, Kesselman C, Tuecke S (2001) The anatomy of the grid: enabling scalable virtual organizations. Int J High Perform Comput Appl 15(3):200–222
14. Giannella C, Bhargava R, and Kargupta H (2004) Multi-agent systems and distributed data mining. In: Cooperative information agents VIII: 8th international workshop, CIA 2004, Erfurt, Germany, pp 1–15
15. Gong ZG, Muyeba M, Guo JZ (2010) Business information query expansion through semantic network. Enterp Inf Syst 4(1):1–22
16. Gordijn J, Akkermans H (2001) Designing and evaluating e-business models. IEEE Intell Syst 16(4):11–17
17. Graml T, Bracht R, Spies M (2008) Patterns of business rules to enable agile business processes. Enterp Inf Syst 2(4):385–402
18. Gruber TR (2002) Toward principles for the design of ontologies used for knowledge sharing? Technical report KSL-93-04, Knowledge Systems Laboratory, Stanford University
19. Han JW, Kamber M (2001) Data mining: concepts and techniques. Morgan Kaufman Publishers, San Francisco
20. Hand D, Mannila H, Smyth P (2001) Principals of data mining. MIT press, Cambridge
21. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin
22. Heer J, Chi EH (2001) Identification of Web user traffic composition using multi-modal clustering and information scent. In: Proceedings of the workshop on web mining, SIAM conference on data mining, Chicago, USA, pp 51–58
23. Hsu C, Wallace WA (2007) An industrial network flow information integration model for supply chain management and intelligent transportation. Enterp Inf Syst 1(3):327–351
24. Izza S (2009) Integration of industrial information systems: from syntactic to semantic integration approaches. Enterp Inf Syst 3(1):1–57
25. Jaamour R (2005) Securing web services. Inf Sec J Global Persp 14(4):36–44
26. Jasjit S (2008) Distributed R&D, cross-regional knowledge integration and quality of innovative output. Res Policy 37(1): 77–96
27. Jespersen SE, Thorhauge J, Pedersen TB (2002) A hybrid approach to web usage mining. In Proc. of 4th International Conference Data Warehousing and Knowledge Discovery (DaWaK'02) Aix-en-Province, France, pp 73–82
28. Kakousis K, Paspallis N, Papadopoulos GA (2010) A survey of software adaptation in mobile and ubiquitous computing. Enterp Inf Syst 4(4):355–389
29. Kargupta H, Hamzaoglu I, Stafford B (1997) Scalable, distributed data mining using an agent based architecture. In: Proceedings the third international conference on the knowledge discovery and data mining. AAAI Press, Menlo Park, California
30. Kumar A, Kantardzic MM, Madden S (2006) Guest editors' introduction: distributed data mining-framework and implementations. IEEE Internet Comput 10(4):15–17
31. Lee SM, Olson DL, Lee SH (2009) Open process and open-source enterprise systems. Enterp Inf Syst 3(2):201–209
32. Liu D, Deters R, Zhang WJ (2010) Architectural design for resilience. Enterp Inf Syst 4(2):137–152
33. Luhn HP (1958) A business intelligence system. IBM J Res Dev 2(4):314–319
34. Luo HY, Gao JL, Ji WL (2008) Research on data mining in e-business websites. In: 2008 International conference on computer science and software engineering
35. Luo J, Xu L, Jamont JP, Zeng L, Shi Z (2007) Flood decision support system on agent grid: method and implementation. Enterp Inf Syst 1(1):49–68

36. Luo JW, Wang MG, Hu J, Shi ZZ (2007) Distributed data mining on agent grid: issues, platform and development toolkit. Future Gener Comput Syst 23(1):61–68
37. Luo P, He Q, Huang R, Lin F, Shi ZZ (2005) Execution engine of meta-learning system for KDD in multi-agent environment. In: Proceedings of the international workshop on autonomous intelligent systems: agents and data mining, St. Petersburg, Russia, pp 149–160
38. Maedche A, Staab S (2001) Ontology learning for the Semantic Web. IEEE Intell Syst 16(2):72–79
39. Marijn J, Jeffrey G, René WW (2006) Web service orchestration in public administration: challenges, roles, and growth stages. Inf Syst Manag 23(2):44–55
40. Pan JZ (2007) A flexible ontology reasoning architecture for the semantic web. IEEE Trans Knowl Data Eng 19(2):246–260
41. Pechoucek M, Marik V (2008) Industrial deployment of multi-agent technologies: review and selected case studies. Auton Agents Multi-Agent Syst 17(3):397–431
42. Petrini M, Pozzebon M (2009) Managing sustainability with the support of business intelligence: integrating socio-environmental indicators and organisational context. J Strateg Inf Syst 18(4):178–191
43. Piao CH, Hanc XF, Wu H (2010) Research on e-commerce transaction networks using multi-agent modeling and open application programming interface. Enterp Inf Syst 4(3):329–353
44. Pipattanasomporn M, Feroze H, Rahman S (2009) Multi-agent systems in a distributed smart grid: design and implementation. In: Power systems conference and exposition, Washington pp 1–8
45. Prodromidis AL, Chan PK, Stolfo SJ (2000) Meta-learning in distributed data mining systems: issues and approaches. In: Advances in distributed and parallel knowledge discovery, The MIT Press, pp 81–114
46. Rao, V.S.(2009) Multi agent-based distributed data mining: an over view. Int J Rev Comput 83–92. http://www.ijric.org/volumes/Vol3/11Vol3.pdf
47. Ryu SH, Casati F, Skogsrud H, Benatallah B, Saint-Paul R (2008) Supporting the dynamic evolution of web service protocols in service-oriented architectures. ACM Trans Web 2(2):1–46
48. Stankovski V, Swain M, Kravtsov V et al (2008) Digging deep into the data mine with Data Mining Grid. IEEE Internet Comput 12(6):69–76
49. Sumner M (2009) How alignment strategies influence ERP project success. Enterp Inf Syst 3(4):425–448
50. Tan WN, Xu YC, Xu W, Xu LD, Zhao XH, Wang L, Fu LL (2010) A methodology toward manufacturing grid-based virtual enterprise operating platform. Enterp Inf Syst 4(3):283–309
51. Theussl S, Feinerer I, Hornik K (2009) Distributed Text Mining with tm. The R User Conference 2009. Retrieved on Feburary 4th, 2011 at http://www.r-project.org/conferences/useR-2009/slides/Theussl+Feinerer+Hornik.pdf
52. Tozicka J, Rovatsos M, Pechoucek M (2007) A framework for agent-based distributed machine learning and data mining. In: International conference on autonomous agents and multi-agent systems, Hawai'i, USA, pp 1–8
53. Trkman P, McCormack K, Oliveira MPV, Ladeira MB (2010) The impact of business analytics on supply chain performance. Decis Support Syst 49(3):318–327
54. Wang C, Ghenniwa H, Shen WM (2008) Real time distributed shop floor scheduling using an agent-based service-oriented architecture. Int J Prod Res 46(9):2433–2452
55. Wang K, Bai XY, Li J, Ding C (2010) A service-based framework for pharmacogenomics data integration. Enterp Inf Syst 4(3):225–245
56. Yang L, Zuo C, Wang YG (2005) Research and implementation of service oriented architecture for knowledge discovery. Chin J Comput 28(4):445–457
57. Yang H, Simon F (2009) A framework of business intelligence-driven data mining for e-business. 2009 Fifth International Joint Conference on INC, IMS and IDC
58. Yu PJ, Buyya PR (2005) A taxonomy of scientific workflow systems for grid computing. ACM SIGMOD Record 34(3):44–49
59. Zhang N, Bao H (2009) Research on distributed data mining technology based on Grid. In: First international workshop on database technology and applications, Wuhan, pp 440–443
60. Zhang T, Ying S, Cao S, Zhang JK, Wei J (2008) A modeling approach to service-oriented architecture. Enterp Inf Syst 2(3):239–257
61. Zhang Y, Bhattacharyya S (2007) Effectiveness of Q-learning as a tool for calibrating agent-based supply chain network models. Enterp Inf Syst 1(2):217–233
62. Zhao Y, Raicu I, Lu S (2008) Cloud computing and grid computing 360-degree compared. In:Grid Computing Environments Workshop, 2008, pp 1–10
63. Zhou B, Jia Y, Liu CY, Zhang X (2010) A Distributed Text Mining System for Online Web Textual Data Analysis. In: Proceedings of 2010 international conference on cyber-enabled distributed computing and knowledge discovery (CyberC), 1–4, Oct. 2010
64. Zhuang Y, Chen JM, Xu D, Pan JG (2007) Distributed data mining based on multi-agent system. Computer Sci 34(12):163–167