Distributed data mining systems: techniques, approaches and algorithms

Ammar Alhaj Ali[1], *Pavel Varacha*[1], *Said Krayem*[1], *Petr Zacek*[1, *], and *Andrzej Urbanek*[2]

[1]Faculty of Applied Informatics, Tomas Bata University in Zlin, Czech Republic
[2]Pomorska Academia in Slupsk, Poland

Abstract. Nowadays, we are living in the midst of a data explosion and seeing a massive growth in databases so with the wide availability of huge amounts of data; necessarily we are become in need for turning this data into useful information and knowledge, where Data mining uncovers interesting patterns and relationships hidden in a large volume of raw data and big data is a new term used to identify the datasets that are of large size and have grater complexity. The knowledge gained from data can be used for applications such as market analysis, customer retention and production control. Data mining is a massive computing task that deals with huge amount of stored data in a centralized or distributed system to extract useful information or knowledge. In this paper, we will discuss Distributed Data Mining systems, approaches, Techniques and algorithms to deal with distributed data to discover knowledge from distributed data in an effective and efficient way.

1 Introduction

Data mining technology is used as a mode of identifying patterns and trends from big data, in other word the process of automatically discovering useful information in large datasets, Data mining uses artificial intelligence techniques, neural networks, and advanced statistical tools to detect trends, patterns, and relationships [1].

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [2].

Some features of distributed scenario where Distributed Data Mining are applicable are: A system consists of multiple independent sites of data and computation which communicate only through message passing, Communication between sites is expensive, and Sites have resource and privacy concerns. The primary goal of many distributed data mining method is to reduce the number of messages send. Some methods attempt to load balance across the sites to prevent performance from being dominated by time and space usage of any individual sites (Giannella C. et al., 2004).

The paper is organized as follows: Section.2 we describe data mining and architecture of a typical data mining system,Section.3 we offer knowledge discovery in database and stages of KDD process and in section.4 we offer data mining techniques such as classification clustering. Etc and section.5 and section.6 we show general view about distributed data mining systems and its architecture, in section.7 we show steps in Distributed data mining; And in section.8 offer Distributed data mining algorithm.

2 Data Mining

Data mining is a technology that blends data analysis methods with sophisticated algorithms for processing large data sets, and an active research field that aims at developing new data analysis methods for novel forms of data [3].Based on this view, the architecture of a typical data mining system may have the following major components [2] and it is illustrated by Figure 1.

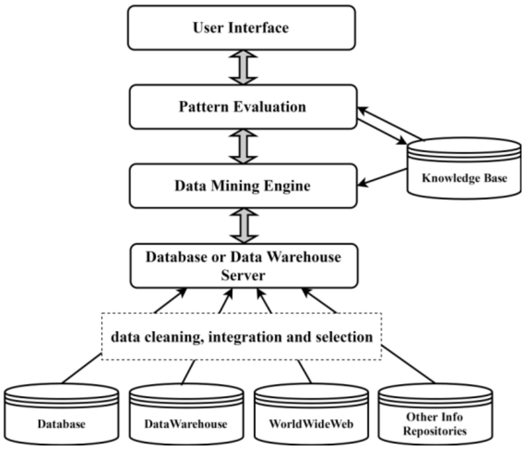

Fig. 1. Architecture of a typical data mining system (own source).

[*] Corresponding author: zacek@utb.cz

- **Database, data warehouse, World Wide Web, or other information repository**: This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.
- **Database or data warehouse server**: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.
- **Knowledge base**: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.
- **Data mining engine**: This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.
- **Pattern evaluation module**: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns.
- **User interface**: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.

3 Knowledge discovery in databases (KDD)

The terms knowledge discovery and data mining are distinct. KDD refers to the overall process of discovering useful knowledge from data [3, 4]. The process could be seen in Figure 2.

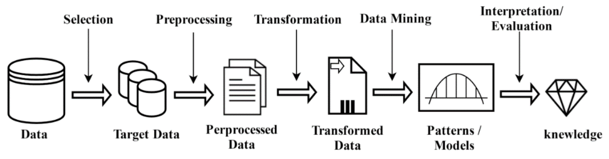

Fig. 2. Knowledge discovery process (own source).

Stages of KDD process [5]:

- **Selection**: This stage consists of creating a target data set from your available data sources.
- **Preprocessing**: This stage consists of cleaning and preprocessing the data set to be used for data mining.
- **Transformation**: This stage consists of identifying and implementing any data transformations that are required.
- **Data Mining**: This stage consists of selecting the appropriate data mining algorithms and selected data set.
- **Interpretation/Evaluation**: This stage consists of reviewing the results and the outputs.

4 Data mining techniques

There are several major data mining techniques have been developing and using in data mining projects recently including:

- **Classification**: It maps the data into predefined groups. It is used to develop a model that can classify the population of records at large level [7].
- **Association**: is usually to find frequent item set findings among large data sets [6].
- **Clustering**: Clustering as the name suggests is the process of grouping data into classes [6].
- **Prediction**: is a data mining technique that is used to identify the relationship between independent variables and relationship between dependent and independent variables [6].
- **Sequential Patterns**: it is one of the data mining techniques that seek to discover similar patterns in data transaction over a business period [6].
- **A decision tree**: is a flow chart like tree structure, where each node denotes test on an attribute value, each branch represents the result of the test, and tree leaves represent classes [6].

5 Distributed Data Mining

Distributed Data Mining (DDM) is a field which deals with analysing distributed data and proposes algorithmic solutions to perform different data analysis and mining operations in distributed manner by considering the resource constraints [8]. Where distributed computing plays an important role in the data Mining process for several reasons.

First: Data Mining often requires huge amounts of resources in storage space and computation time. To make systems scalable, it is important to develop mechanisms that distribute the work load among several sites in a flexible way.

Second: Data is often inherently distributed into several databases, making a centralized processing of this data very inefficient and prone to security risks [9].

6 Architecture of Distributed Data Mining (DDM)

In Figure 3, section A (left), that shows the traditional data warehouse - based architecture for data mining. This model of data mining works by uploading critical data in data warehouse for centralized data mining and then perform data mining [10]. But these algorithms create many issues regarding bandwidth and privacy protection, since a sufficient bandwidth is required to transfer huge size of data and unable to preserve privacy of sensitive data, so this is not suited model for distributed data mining [10].

DDM architecture includes multiple sites each having independent computing power and storage capability. Each site performs local computation on its own; the architecture for DDM is as shown in Figure 3, section B (Right), From the Figure it is clear that, in distributed local models are generated on different nodes and finally aggregate to form a global model which represents the mining result of the entire dataset [10].

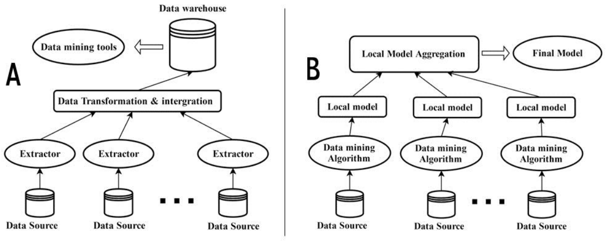

Fig. 3. Data warehouse architecture (A) & Distributed Data Mining framework (B) (own source).

7 Steps in distributed data mining

Today's real-world databases are highly susceptible to noise, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low quality data will lead to low-quality mining results [11].

Major tasks in data preprocessing involves:

- **Data Cleaning**: It involves fill in missing values, identify outliers and smooth out noisy data, correct inconsistent data, resolve redundancy caused by data integration [11, 12].
- **Data Clustering**: is an unsupervised classification. Data Clustering is used to place data elements into related groups without advanced knowledge of group definitions [12].
- **Data Integration**: integration can help reduce and avoid redundancies and inconsistencies in the resulting data set [11].
- **Data Transformation**: The data are transformed or consolidated into forms appropriate for mining [12].
- **Data Reduction**: representation of data set will be smaller in volume but yet produces same result [12].
- Data Discretization is essential before preprocessing for further analysis. In data Discretization, we use some classification algorithm to reduce the data size by Discretization [13].

8 Distributed Data Mining Algorithms

Distributed data mining algorithm is classified into three classes [8]:

1. DDM based on Multi Agent System.
2. DDM based on Meta learning.
3. DDM based on grid.

8.1 DDM based on Multi Agent System (MADM and CAKE architectures)

DDM application can be further enhanced with agents. ADDM (Agent-based Distributed Data Mining) takes data mining as a basis foundation and is enhanced with agents; therefore, this novel data mining technique inherits all powerful properties of agents and, as a result, yields desirable characteristics. In general, constructing an ADDM system concerns three key characteristics: interoperability, dynamic system configuration, and performance aspects [12] [14, 15]:

8.1.1 MADM (Multi Agent Data Mining) Architecture

In distributed data mining, there is a fundamental trade-off between the accuracy and the cost of the computation. If interest is in cost functions which reflect both computation costs and communication costs, especially the cost of wide area communications, we can process all the data locally to obtain local results and combine the local results at the root to obtain the final result, but if our interest is accurate result, we can ship all the data to a single node. We assume that this produces the most accurate result. In general, this is the most expensive while the former approach is less expensive, but also less accurate [8] [14, 15].

Following are the components of the system [8] [12] [14, 15]:

- **Interface Agent**: It interacts with user (or user agent).
- **Facilitator agent**: Facilitator agent is mainly responsible for activation and synchronization.
- **Resource agent**: The resource agent actively maintains the Meta data information about each of the data sources.
- **Mining agent**: Data Mining agents implement some specific data mining techniques and algorithms.
- **Result agent**: Result Agent observes a movement of mining agents, and obtains result from mining agents.
- **Broker agent**: Broker Agent serves as advisor agents that facilitate diffusion of request of agents.
- **Query agent**: Query Agent is generated at each demand of a user.
- **Mobile Agent**: Mobile Agents travels around their network. It processes the data and sends result back to main host.

8.1.2 CAKE (Classifying Associating and Knowledge Discovery) architecture

The CAKE architecture is based on centralized Parallel Data Mining Agents (PADMAs). CAKE is a 4-tier architecture where the Distributed Data Mining is implemented using parallel Data Mining Agents (PADMAs) using centralized metadata which contains all the rules of Classification and Association along with its data structure details and web interface used to provide the users with the interface to view the result [8] [12] [16]. Whole architecture could be seen in Figure 4.

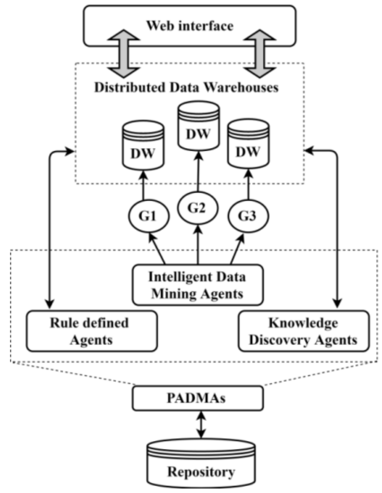
Web interface
Distributed Data Warehouses
DW
DW
DW
G1
G2
G3
Intelligent Data Mining Agents
Rule defined Agents
Knowledge Discovery Agents
PADMAs
Repository

Fig. 4. CAKE Architecture (own source).

8.2 DDM based on Meta learning

Meta learning system is implemented by JAM (Java Agents for Meta-learning) system. JAM is a distributed agent-based data mining system. It provides a set of the learning agents which are used to compute classifier agents at each site. The launching and exchanging of each classifier agents take place at place at all sites distributed data mining system by providing a set of Meta learning agents which combine the computed models at different sites. JAM is a first system that employs Meta learning as a means to mine distributed databases [8] [12] [17].

The Figure 5 is an example of architecture of Meta learning system.

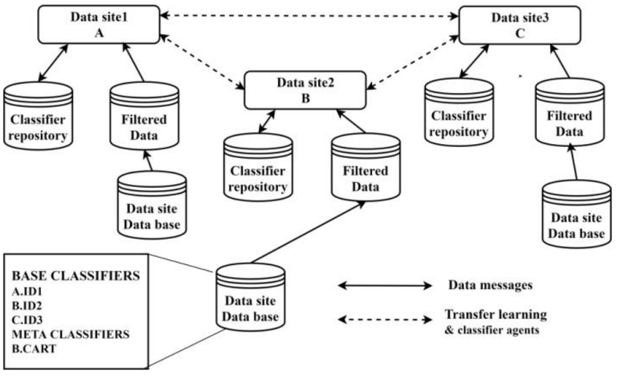

Fig. 5. JAM architecture (own source).

8.3 DDM based on Grid

Grid provides distributed computing environment that enables coordinated resource sharing within dynamic organizations consisting of individuals, institutions, and resources. The main objective of grid computing is to allow organizations and application developers to establish distributed computing environments which use computing resources on demand [8].

8.3.1 Visual environment for grid applications (VEGA) architecture

The design includes many services used to compose, validate, and execute a parallel and distributed knowledge discovery computation and other services to store and analyze the discovered knowledge. The main services are the following:

- **Data access service (DAS)**: The data access service is responsible for searching, selecting, extracting, transforming, and delivering data to be mined.
- **Tools and algorithms access service (TAAS)**: This service is responsible for searching, selecting, and downloading data mining tools and algorithms. As before, the metadata regarding their availability, location, and configuration are stored in the KMR (knowledge metadata repository) and managed by the KDS (Knowledge Directory Service), whereas the tools and algorithms are stored in the local storage facility of each K-Grid (Knowledge-grid) node.
- **Execution plan management service (EPMS)**: An execution plan is represented by a graph describing the interaction and data flows among resources. In simple cases, a user can directly design the execution plan by using a visual composition tool where programs are connected to data sources.
- **Results Presentation Service (RPS)**: Result visualization is important in the knowledge discovery process to help users in the interpretation of the discovered patterns. This service specifies how to generate, present and visualize the knowledge models extracted (e.g., association rules, clustering models, classification models), after storing them in the Knowledge Base Repository (KMR). The result metadata are stored in the KMR to be managed by the KDS. Where metadata about the resources selected for the computation are then stored into the Task Metadata Repository (TMR) [18]. The design process could be seen in Figure 6.

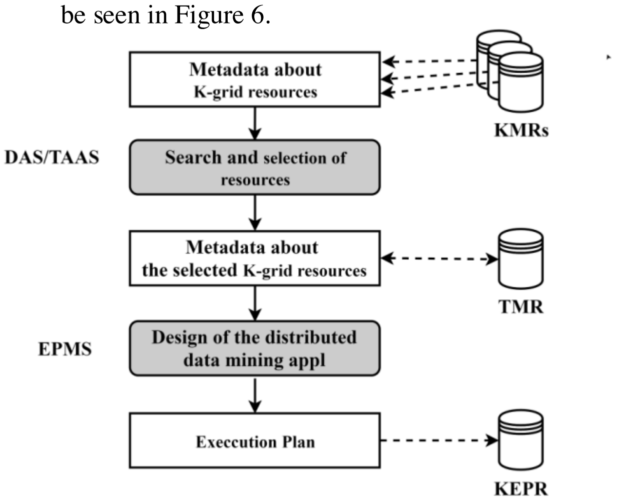
Metadata about K-grid resources
KMRs
DAS/TAAS
Search and selection of resources
Metadata about the selected K-grid resources
TMR
EPMS
Design of the distributed data mining appl
Exeecution Plan
KEPR

Fig. 6. Design process of data mining computation. Mining System (own source).

VEGA provides following Execution plan management service (EPMS) operations:

- Task Composition
- Task Consistency Checking
- Execution Plan Generation.

Task composition is performed by means of graphical interface which provides a user with a set of graphical objects representing grid nodes and resources.

Task consistency checking: is to obtain a correct result and consistent model of computation. The validation process is performed by means of two components → the model pre-processor and model post-processor.

In Execution plan generation: the computation model is translated into an execution plan represented by an XML document. The task is performed by execution plan generator [8] [18]. The necessary visual environment could be seen in Figure 7.

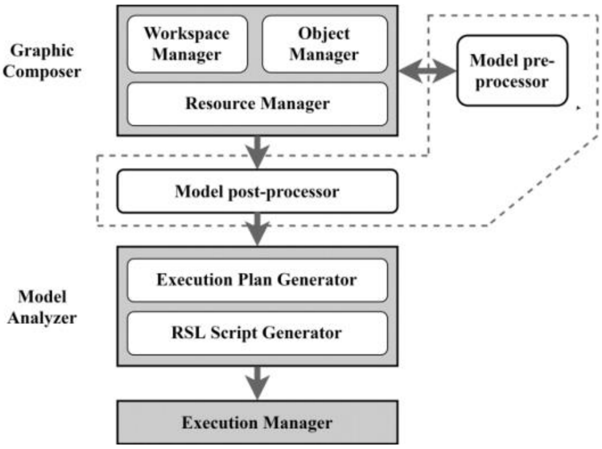

Fig. 7. Visual Environments for Grid Applications (own source).

9 The Analysis and Discussion

In Table 1, we have compared various advantages and disadvantages of each framework [8] [12].

Table 1. Advantages and disadvantages of DDM Architectures (own source)

Type of DDM	DDM Frameworks	Advantages	Disadvantages
DDM based on parallel data mining	MADM	Easy to build architecture	Agent constrained processing, agent-action ability.
	CAKE	Clear distinction of functionality between agents	Local data sources have Restricted availability due to privacy.
DDM based on Meta learning	JAM	Adaptive learning, Interactive Mining	Learning capability agent needs to be fed up with learning and reasoning algorithm
DDM based on grid	VEGA	Improved speed of execution compared to any other data mining algorithm.	Data fusion and preparation are difficult

And in Table 2 we have analyzed practical application where each framework can be most appropriate [12].

10

Table 2. DDM frameworks and practical applications (own source)

Type of DDM	DDM Frameworks	Application
DDM based on Parallel Data Mining	MADM	Web mining, text mining
	CAKE	Intrusion Detection, Customer Relationship management, Parallel genetic Algorithm, Financial Data Management
DDM based on Meta learning	JAM	Business Intelligence, Artificial Immune System, Knowledge management and Marketing, Semantic Web
DDM based on Grid	VEGA	Peer to Peer Computing and Service, Ecommerce, Internet and Network Services

10 Conclusion

Last years, we have seen the staggering numbers of data sizes, the volume of data produced is doubling every two years and unstructured data alone makes up 90 percent of the digital universe. But more data does not necessarily mean more knowledge. So data mining services has become important and integral process of every company or organization to gain competitive edge in the business. where data mining has benefited most of the companies with products need to sell or not; medical researchers use the facts that are helpful with vaccines required to develop by analyzing recent disease patterns; assist engineers with highways need to be build & much more.

As a new and emerging domain, it has many open issues waiting for the significant involvement of research resources. We believe the research and development on data mining is very encouraging and worthy of much more efforts by new researchers.

In this paper we have given a brief overview of techniques and architectures for DDM and classified data distributed mining algorithms into three classes: DDM based on Multi Agent System, DDM based on Meta learning and DDM based on grid. And for each classification we have shown the major features and pros and cons. Finally, we have offered advantages and disadvantages for each framework and proposed the most suitable application where each framework can be most appropriate.

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014) and by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089. Also supported by grant No. IGA/CebiaTech/2017/007 from IGA (Internal Grant Agency) of Tomas Bata University in Zlin.

References

1. https://en.wikipedia.org/wiki/Data_mining
2. J. Han, M. Kamber. Data Mining: Concepts and Techniques, (2006)
3. V. Grossi, D. Pedreschi, F. Turini, Data Mining and Constraints: An Overview (2016)

4. C. Priyadharsini, A. S. Thanamani, An Overview of Knowledge Discovery Database and Data mining Techniques (2016)
5. https://www.linkedin.com/pulse/knowledge-discovery-data-kdd-process-mohammad-valadkhani
6. S. Garg, A. K. Sharma, Comp. Analysis of Data Mining Techniques on Education Dataset (2016)
7. R. Tamilselvi, S. Kalaiselvi, An Overview of Data Mining Techniques and Applications (2013)
8. R. Chikhale, Study of Distributed Data Mining Algorithm and Trends (2012)
9. http://www-ai.cs.uni-dortmund.de/auto?self=$ejr31cyc.
10. R. T. Sunny, S. M. Thampi, Survey on Distributed Data Mining in P2P Networks (2012)
11. J. Han, M. Kamber, J. Pei. Data Mining: Concepts and Techniques (2012)
12. C. Srimathi, M. Subaji, A. S. Baby, D. Raveendran, A study on distributed data mining frameworks (2016)
13. https://www.tutorialspoint.com/data_mining/dm_issues.htm.
14. V. S. Rao, multi agent-based distributed data mining: an overview (2010)
15. A. V. Rao, Agent Based Approach to Knowledge Discovery in Datamining (2014)
16. B. Liu, S. G. Cao, W. He, Distributed data mining for e-business (2011)
17. S. K. Sen, S. Dash, S. P. Pattanayak, agent based meta learning in distributed data mining system (2012)
18. M. Cannataro, A. Congiusta, A. Pugliese, D. Talia, P. Trunfio, Distributed Data Mining on Grids: Services, Tools, and Applications (2004)