CAKE – Classifying, Associating & Knowledge DiscovEry An Approach for Distributed Data Mining (DDM) Using PArallel Data Mining Agents (PADMAs)

Danish Khan
[1] *Habib Bank Ltd. (HBL), ASD- IT&SG, 6th Floor, HBL Plaza, I.I Chundrigar, Karachi, Pakistan.*
[2] *PAF – KIET (City Campus) 28-D, Block-6 P.E.C.H.S, Karachi, Pakistan.*
cdsdanish@hotmail.com

Abstract

This paper accentuate an approach of implementing Distributed Data Mining (DDM) using Multi-Agent System (MAS) technology, and proposes a data mining technique of "CAKE" (Classifying, Associating & Knowledge DiscovEry). The architecture is based on centralized PArallel Data Mining Agents (PADMAs).

Data Mining is part of a word, which has been recently introduced known as BI or Business Intelligence. The need is to derive knowledge out of the abstract data. The process is difficult, complex, time consuming and resource starving. These highlighted problems addressed in the proposed model.

The model architecture is distributed, uses knowledge-driven mining technique and flexible enough to work on any data warehouse, which will help to overcome these problems. Good knowledge of data, meta-data and business domain is required for defining rules for data mining. Taking into consideration that the data and data warehouse has already gone through the necessary processes and ready for data mining.

1. Introduction

In today's world, the increasing processing power and sophisticated technologies has increased the business need, and now people expect more from systems. These days, the computer systems are not only used for storing data but also for providing information and forecasting. This way it plays a vital role and its need has bought up the concept of Data Warehouse, which is:

"A subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management's decision" [1]

Accompanying these requirements Data Warehouses has to facilitate with outsized amount of data to be stored for analysis, which has increased the functionality and responsiveness to access data which was not possible while using transaction oriented system, and it is the only way to get the answer of desired questions without affecting production system.

Extract, Transformation & Loading also known as ETL is a Data Warehouse pre-process comprises of combination of activities, which are executed to transform the subject-oriented summarized data from Operational System(s) to Data Warehouse(s). The below mentioned pre-processes, are required to be performed in-order to remove the outliers, inconsistency and incompleteness of data to maintain the quality of data and its result.

- *Data Cleaning:* Filling missing data, removing outliers and resolve inconsistency.
- *Data Integration:* Data from multiple sources or systems combine in a single standardized format.
- *Data Transformation:* Making data from multiple sources to a single aggregation form.
- *Data Reduction:* Is a selection of required data to be available in data warehouse.
- *Data Discretization:* It deals with reduction of data, but especially numerical data.

Data Warehouse deals with the storing of subjected data. Now what we require is the information in a manner that can help in accomplishing our objectives. The process of extracting and viewing of data in a Multidimensional Model is part of OLAP (Online Analytical Processing), which are set of tools used by knowledge workers or business domain experts to dig out what they exactly required. The below diagram shows the complete course of activities.

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.236

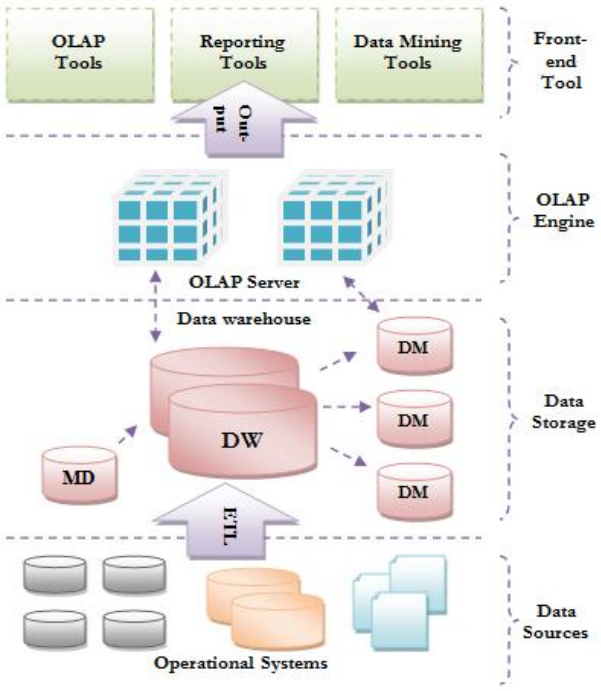

Figure 1.1 *– Data warehouse Architecture*

Data Mining (DM), can simply be explained as an automated process of discovering unanticipated knowledge from massive amount of data. Data Mining involves complex Data Structures, Algorithms, Statistics and Artificial Intelligence. It also includes learning from previous knowledge and recognizing hidden data pattern and providing the realistic results along with rationalization. Knowledge Discovery in Database also known as KDD a synonym of Data Mining, which comprises of three stages:

- The understanding of business and data.
- Performing the pre-processes tasks.
- Data Mining and Reporting.

Data Mining techniques includes Classification, Association & Clustering, which are used to mine data on the basis of different defined rules and patterns.

A term in Data Mining has been introduced known as "Distributed Data Mining (DDM)", it's an approach of performing Data Mining on Distributed Data Warehouses over different remote locations, which either contains the same data distributed over different locations or different data related to the same subject. There are several DDM approaches, which are developed using MAS that includes BODHI, PADMA and JAM, all these approaches deals with the centralized architecture. While another approach of DDM known as Papyrus deals with Peer-to-Peer (P2P) working style. [2]

PArallel Data Mining Agents (PADMAs) is Multi-agent based architecture for Data Mining. It is a system that makes use of Intelligent Data Mining Agents, which are responsible for accessing, analyzing and discovering the hidden patterns within the Data Warehouse. They all work together in conjunction with each other and share same repository or meta-data. [4]

2. Related Work

Classification and prediction algorithms for machine learning typically require all training data to be resident in memory during decision tree construction. The data mining system does not utilize any functions of the database/ data warehouse system. It fetches data from an external source (e.g. a flat file), processes the data using some data mining algorithm and then stores the mining results in another file. Such a system fails to take advantage of the fact that in a database / data warehouse data tend to be well organized, indexed, cleaned, integrated or consolidated. Classification is a two-step process. In the first step (the learning step), a model is built to describe a predetermined set of data classes or concepts and learning algorithm (decision tree). The attribute selection algorithm, using average class entropy function (ACE) is used to determine results. [10]

A possible number of data mining methods for an accurate prediction for the lifetime of metallic components are explained. When the user queries the framework, the knowledge base is first consulted to search for matching between existing rules and user inputs. If user inputs are matched to a rule, we produce the result directly from the rule. If we can not find a matching rule, new data should be input into predictors to produce a result. [6]

Data-mining algorithms are required which can interrogate metadata and meta-knowledge attached to data points. The proposed model of SMART algorithms requires two key components. (1) A range of tools and techniques and (2) Expert knowledge. It is suggested that the data-mining profession should go the next step and develop multi-modal algorithms that contain a range of mining and modeling tools and techniques. 'Intelligent' agents can be used for this purpose. An alternative strategy would be to use swarm technology. The swarms would consist of armies of agents each containing one or more modeling techniques. In addition there are procedures such as active learning techniques. [8]

All the Data Mining above proposed approaches are providing different methods but neither of them is proposing a complete framework or model that is full filling the needs and addressing towards the problems that could be faced in a complex Data Warehouse(s).

In this paper we have tried to cater a kind of complete framework that can work on any Data Warehouse(s).

3. Architecture

Our proposed architecture is based on 4-tier, which are identified to be the major components in the complete process. Distributed Data Mining (DDM) is implemented using PArallel Data Mining Agents (PADMAs) using centralized meta-data, which contains all the rules of Classification and Association along with the data structure details and web interface is used to provide the users with the interface to view the result.

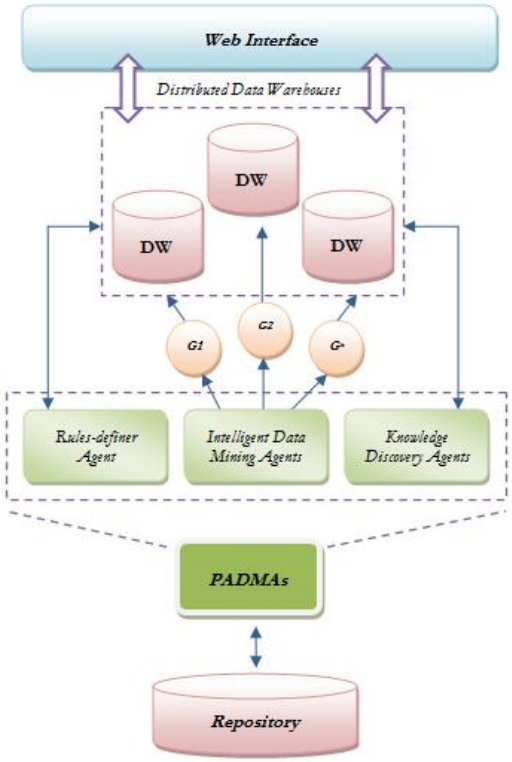

Figure 3.1 *– CAKE (Architecture)*

3.1. Distributed Data Warehouses

The Distributed Data Warehouses are either physically or logically distributed on different sites. The sites might be containing the same data distributed on multiple locations or single subjected data distributed on different locations.

The PADMAs are going to be executed on the sites where Data Warehouse exists to improve the performance and privacy factor, only once the agents are required to retrieve their respective configuration.

3.2. PADMAs

The PArallel Data Mining Agents (PADMAs) are the combination of Multi-Purpose Agents, which are of three major categories according to their respective roles at each stage of Data Mining process, and described as follows:

3.2.1. Rule-definer Agents. These are used to define the Meta-data of Data Warehouse on the basis of 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 Data Mining and by Knowledge Discovery Agents for driving the knowledge out from the defined patterns.

The rules are combination of conditions and weighted values, defined to perform the operation for evaluating the data and identify the dependency and relationship between attributes to ascertain the hidden knowledge.

Table 3.1 – Sample scoring of field items (Decision Tree)

Age	Weight	Marital Status	Weight
<20	0	Single	0
>20 <30	2	Married	2
>30	1	Separated	1

Table 3.1: shows the age and marital status column contains the possible values and assigned weighted value. When the respective column satisfies a certain condition a value is being assigned. E.g. A person of age 25 and is married, its score will be 4.

Table 3.2 – Sample scoring of field items (Association)

Column1	+	Column2	Sum Score	Weight
Age	+	Marital Status	0	0
Age	+	Marital Status	>1 <=2	1
Age	+	Marital Status	>=3 <=4	2

Now, considering the same situation here we need to identify the association between data that we can easily identify it with the computed column and assign it with high priority score to uniquely identify our required results.

3.2.2. Intelligent Data mining Agents. The Data Mining Agents are group of agents, which can be setup to work on specified set of data on any location with defined rules. These groups of agents works together to mine the data and compute the desire result. All the calculation that has been shown in the above tables is the job of Data Mining Agents to determine the weight score.

3.2.3. Knowledge-Discovery Agents. The Knowledge Discovery Agents are used to determine the final computed results in success or failure along with the explanation on the computed data. These decisions are taken on the basis of defined requirement in the repository.

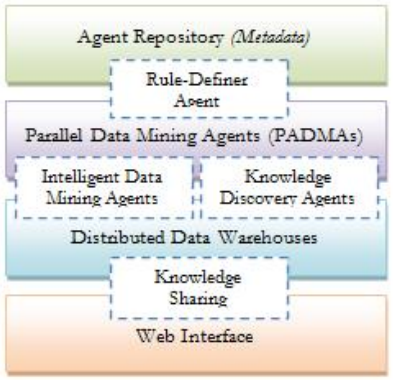

Figure 3.2 *– CAKE (Component Diagram)*

3.3. Agent Repository (Metadata)

The Repository is a database that itself used by the PADMAs to perform their respective jobs. It contains the Data Warehouse meta-data, rules defined and any data that is necessary or required by the PADMAs. All the Agents are to access the Repository once they are being initialized, so that they themselves can update their respective tasks assigned in the form of groups created by using "Rule Definer Agents".

3.4. Web Interface

The Web Interface is provided to the users for viewing the mined results. There could be any Reporting or Analytical tool that can be used to satisfy the needs of the end-user.

4. Example

There are 3 major steps, which are required to be executed in-order to complete the Data Mining process, which is proposed in the paper, the diagram is as follows:

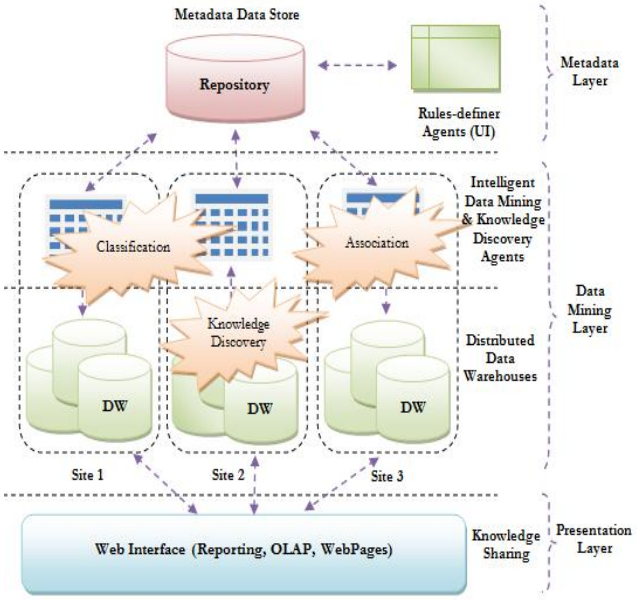

Figure 4.1 *– The 4-tier Architecture*

4.1. Metadata Layer

The process of defining the data about data, this includes data warehouse and its data structure description along with the rules for performing data mining. For every attribute on which decision is required to be taken and is a part of mining operation must be defined and then decision table is required to be made with weight for each value and the weight assignment column. The weight assigned column is later on used for calculating the results, and for computing the association between attributes same approach will be followed, but with the dependency of values between those attributes which can be defined easily as they are combination of each defined weight assignment.

There is also a details of remote Data Warehouse and configure the groups description of Data Mining and Knowledge Agents. Once the agent is executed which task is required to be performed by which agent on which source and what are their objectives which are required to be achieved.

The Repository resides on a centralized location from where all the information is being fetched by the executed agent group according to its role.

4.2. Data Mining Layer

The objective of Parallel Data Mining is achieved by configuration setup of agents and their groups. An

agent can be executed from any location with a specified profile on initializing, the agent starts executing its job. In a group there can be multiple agents that execute from multiple locations to achieve their configured objective. These objectives are defined while configuring the Metadata. One type of Mining Agent performs the operations on the data while the other one discovers the required result for the computed data in the warehouse.

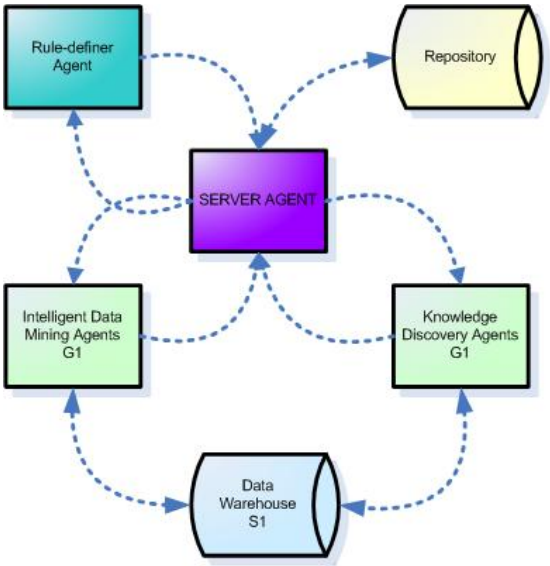

Figure 4.2 *– Agents Diagram for Single Site & Group*

4.3. Presentation Layer

The WebPages and Reports can be developed to provide user friendly results to the users. The users can access the mining results from all the sources. Furthermore we can also provide any other OLAP tool which supports the feature of simulating the results as a whole.

5. Future Work & Open Issues

The architecture that is presented in the paper is based on centralized repository data and supporting distributed data mining approach, many concerns were tried, identified and addressed, but still there are many situations to be catered; some of them are as follows:

- *Heterogeneous:* Using of different RDBMS at different distributed remote sites where Data warehouse resides.
- *Complexity:* More architecture redesigning will be required to cater complex mining needs.

In, the next phase the focus will be on supporting the heterogeneous architecture in our approach by adding mediators to support different RDBMS and to increase the privacy throughout the complete lifecycle, and also using Agent Ontology for sharing knowledge.

6. Conclusion

In summary, this study is a preliminary work. The major focus of this paper is on providing architecture with a technique that provides a baseline for Distributed Data Mining using various modified data mining methods. Secondly, the proposed model tried to cover up certain highlighted problems and using of weight assignment procedure to avoid complex conditioning, which improves the performance.

The fact is data is of no use until and unless the knowledge is extracted out from it and data mining is the only way to discover hidden knowledge from massive amount of data.

7. Acknowledgement

Thanks to my teacher Mr. Khalid Khan and my friends and colleagues who helped, guided and encouraged me in writing this paper. I would also like to thank Huma Ahmed for her valuable comments and corrections.

8. References

[1] Building the Data Warehouse, W. H. Inmon, 4th Edition

[2] M. Klusch, S. Lodi, G. Moro. “Agent-based Distributed Data Mining: The KDEC Scheme. Intelligent Information Agents - The AgentLink Perspective.” Lecture Notes in Computer Science 2586 Springer 2003.

[3] Chris Clifton “Privacy Preserving Distributed Data Mining” Department of Computer Sciences. November 9, 2001

[4] PArallel Data Mining Agents (PADMA) – http://www-fp.mcs.anl.gov/ccst/research/reports_pre1998/algorithm_development/padma/kargupta.html

[5] Giuseppe Di Fatta, Giancarlo Fortino “A Customizable Multi-Agent System For Distributed Data Mining”

[6] Esther Ge1 Richi Nayak1 Yue Xu2 Yuefeng Li2. “Data Mining for Lifetime Prediction of Metallic Components” Australian Computer Society, Inc.

Australasian Data Mining Conference (AusDM 2006), Sydney, December 2006.

[7] Josenildo C. da Silva2, Chris Giannella1, Ruchita Bhargava3, Hillol Kargupta1;4, and Matthias Klusch2 "Distributed Data Mining and Agents"

[8] Warwick Graco, Tatiana Semenova, Eugene Dubossarsky "Toward Knowledge-Driven Data Mining" 2007 ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM2007), August 12, 2007, San Jose, California, USA.

[9] Dan Li, Sherri Harms, Steve Goddard, William Waltman, Jitender Deogun "Time-Series Data Mining in a Geospatial Decision Support System"
NSF Digital Government Grant No. EIA-0091530 and NSF EPSCOR, Grant No. EPS-0091900
[10] PATRICIA E.N. LUTU "An Integrated Approach for Scaling up Classification and Prediction Algorithms for Data Mining" Proceedings of SAICSIT 2002, Pages 110 – 117

[11] Yen-Ting Kuo, Andrew Lonie, Liz Sonenberg, Kathy Paizis "Domain Ontology Driven Data Mining - A Medical Case Study" 2007 ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM2007), August 12, 2007, San Jose, California, USA.