Incorporation of a Agent Based approach for Integrated OLAP/OLAM Architecture

1Mr.M.Shiva Kumar, 2Dr.K.Krishnamoorthy,

ABSTRACT:
The current data mining tools is used to build knowledge based on a huge historical data. At present, businesses are facing with fast growing data that are very valuable in contributing knowledge. Knowledge should be updated regularly in order to ensure its quality and precision thus improve the decision making process. Data mining has shown great potential in extracting valuable knowledge from large databases. However, current data mining algorithms and tools are costly and several are too complex in their operations when dealing with large databases. In recent years, agents have become a popular paradigm in computing, because its autonomous, flexible and provides intelligence. Embedding agents in the current data mining processes and tools are believed to be able to solve the obstacle .Data warehouses are used for information processing, analytical processing, and data mining. An OLAP based data mining is referred to as OLAP mining, or on-line analytical mining (OLAM), which emphasizes the interactive and exploratory nature of OLAP mining. This paper focuses on an agent-based integrated OLAP/OLAM architecture that mainly focuses on data preprocessing. The aims is to provides an auto preprocessing a set of new data, which suite to data mining novice user. The proposed agent based framework consists of eight supporting agents: data cleaning agent, transformation agent, data reduction agent, data integration agent, proxy MDDB agent, input agent, GUI agent and dividing agent. Also we have three co-coordinating agents: DB API agent, MDDB agent and CUBE API agent. This paper is start by introducing the data mining process problem includes data preprocessing which agent can solve data mining problems. By applying agent in data preprocessing, a tool that intelligence yet flexible can be produced

Keywords:
Agents, Data Mining, Data Preprocessing, OLAP/OLAM architecture

1. Introduction:

Recently data mining suffers from distributed, complex, large and rapid changes of data. Most of the organizations have branches located in different places and each branch own their local database to store their business information [1]. Besides that, today's business process is mostly online and across global world. Organizations face with large updated data online. Applying data mining to acquire updated knowledge leads to complexity in data processing, managing and mining [6].

Most data mining tools suffers on this complexity that they could not handle large database efficiently. These tools support various techniques in some stages of data mining process, mainly data preprocessing, data mining and knowledge visualization, which mainly design for data miner expert's users. Consequently, the user of data mining tools are required to have a deep knowledge in data mining concept and techniques to be able to apply every task in data mining thus produce a good model for decision making. The current data mining tools also do not support real time modeling where new model or knowledge can be acquired when new data is added. In practice, a data miner will carry out remodeling to ensure new knowledge is obtained. This limitation has cause drawbacks in many applications such as in medical, stock exchange and finance. These domains are critical in having updated knowledge since it affects the human life and business performances.

The past research in stock price prediction shows that daily changes of stock market has a significant changes on the accuracy of the next day prediction [6]. Using the current data mining tools, the data mining model is predetermined where the development of knowledge is done earlier. The knowledge model is build based on the past data not from the current past data that may gives new pattern that indicates the tomorrow decision.

In order to have most recent knowledge model with the current data mining tools would be very expensive. New coming data need to be cleaned and mined over and over again. Beside that the users also need to decide when appropriate time to perform data mining which is time intense. In recent years, agents have become popular paradigm in computing because it's autonomous, flexible, and adaptive and provides intelligence [8]. Agents have been used in a wide range of problems. Agents are software entities that perform some set of tasks on behalf of users with some degree of autonomy. Assigning the tedious data miner's job to agent is very significant. The agent will handle new data and build new knowledge. Besides, integrating multi-agents in data mining will reduce complexity of data mining.

Research has shown various potential use of agents to enhance the current data mining techniques such as in

Manuscript received April 5, 2010
Manuscript revised April 20, 2010

neural network, association rules and clustering. Nevertheless, only a few research focus on data mining process. Our research is focuses on enhancing agent mining tools using agent. This paper discusses the basic information on data mining, data preprocessing and agent technology, which later presents the potential features of data mining tools using agent and finally, the design of agent-based integrated OLAP/OLAM architecture as one of the features of an agent based data mining is presented.

2. Agent Technology

Agent technology is a popular technology in computer science and artificial intelligent area [2;3]. Its characteristics such as autonomy, dynamic, intelligent, unpredictable, modular and can be applied in larger scope of problem [3]. Generally, agent is described as a computer system that is situated in some environment that able to react autonomy in order to meets its design objective. An agent is an entity of computation solution that able to operate effectively in dynamic and open environment [4] and acts on behalf of someone else after having been authorized.

Each agent has specific characteristics [6] such as autonomy, adaptive, cooperative, interactivity, intelligent, learning, ruggedness, continuous, coordinator, and mobility. An agent should have one or combination of these characteristic. The characteristic of agent very much depends on the confront problem. Agent has been used widely in data mining area in domain engineering; business; medical and stock price prediction [6].

Thus, we can include a agent based technology in the design of integrated OLAP/OLAM architecture of data mining.

3. Data Mining

Data mining can be defined as extracting or mining knowledge from large amount of data. Data mining is the process of discovering meaningful correlations, patterns, and trends by sifting through large amounts of data and by using advanced analytical techniques such as neural network, fuzzy logic and rough set. Data mining process consists of following stages: data collection, data preprocessing, building knowledge and analysis of the knowledge. Data miner expertises need to carry out a series of experiments in order to obtain a good knowledge model. The experiment is performed systematically to ensure all data in data set are equally treated. Various techniques are selected to suit the nature of the data and therefore ensure the best knowledge model is obtained.

4. Data Preprocessing

Data preprocessing stage is an important part in data mining process. It handles various types of dirty data in large data set. Dirty data consists of noisy, incomplete, inconsistent and lost of data value. Dirty data leads to development of inaccurate knowledge model. Selection of suitable data preprocessing methods for a particular problem is very important. Nowadays this task is totally reliant on data mining expert to improve the quality of data and to increase the accuracy and the effectiveness of data mining process. Clean and quality data is important in data mining to produce a reliable and accurate model. Using dirty data will cause problems such as long learning time, model produced may not be reliable, specific and lengthy rules and poor accuracy model. All these problems need to be avoided because the objective of data mining is to extract knowledge from real world data set. The knowledge should be accurate, understandable and interesting.

There are four stages of data preprocessing that are data cleaning, data integration, data transformation, and data reduction [33]. Data cleaning is used to clean data from noise. Data integration can be applied to merge data from multiple sources into a coherent data store. Data transformation transforms data into the form that suitable for mining process while data reduction can be used to reduce the size of data.

Data require a suitable and appropriate preprocessing technique. Currently, a number of data mining tools provide preprocessing facilities while several may not have these functions. Most organizations used available data mining tools but they attended problems in making the tools fully utilized. This is due to the difficulty in selecting the suitable techniques to preprocess the data, difficulty in using and understanding the techniques for non-expert user and handling new data.

The most difficult part in data mining is data preprocessing. Most of experimental time and effort is at this stage. It is a challenge for novice users that they require expert to identify preprocessing techniques that suitable for their data. This is because non-expert users cannot determine which techniques is better and can preprocess their data suitable for the data mining task.

5. Potential Features Of An Agent Based Data Mining Tools

TABLE 1: Agent-Based Data Mining Vs the Current Data Mining

Features	Agent	Non-Agent	
		Rosetta	NConnect
Ability to propose which preprocessing technique suit the data	/	X	X
Ability to preprocess incoming new data according user profile	/	X	X
Sharing experience on mining	/	X	X
Ability to suggest possible valuable knowledge extract from the data (from sharing experience)	/	X	X
Suitable for novice user	/	X	X

Table 1 shows the potential features of using agent-based data mining versus the current data mining tools. The table shows that the non-agent data mining tools do not support coming of new data and re-model the knowledge while some significant valuable data is added.
With such basic features, the potential features shown in the table can be implemented. Using agent based data mining will reduced cost of acquiring the updated knowledge, sharing data mining experiences in various domain and suitable for novice user. The operation costs can be reduced due to the less dependent of data mining expert. The user profile of preprocessing and mining data can be used to propose to novice user appropriate preprocessing and mining techniques for particular domain.

6. Agent Based OLAM/OLAP Architecture

There are many applications using agent in data cleaning, data transformation, and data reduction. In data cleaning agent has been used to handle missing data, noisy data and outliers [7]. In data transformation, agent is used to transform data to suitable format for mining [3] and normalization [7; 6]. Agent is also used for discretization, attribute reduction and dimension reduction [11].

By integrating this two popular technology, both potential can be increased [7; 3]. Agent usage in data mining especially in preprocessing is very useful especially in handling dynamic data. Even though there were a lot of applications using agent in preprocessing, there are only cover specific techniques or domain. This paper want to propose a framework that can offer a general preprocessing system that can be use for any task or domain.

7. Agent-Based Technical Requirements

In order to fulfill the features of agents stated in Table 1, therefore the agent based preprocessing should be intelligent. The agent intelligent here means that the agent is able to propose the best preprocessing technique suite to the domain. The agent will autonomously preprocess the incoming of the new data. The cleaned data is later used to update the new knowledge model when required using a specific optimization techniques.

With such functionality, the data mining tools can be used for non-expert data mining. With such complex system, the framework is developed based on a multi-agents system concept to achieve the user specific goals.

Agents-based preprocessing has the characteristic such as autonomy, adaptive, reactivity, learning, and cooperatively. The agent-based preprocessing has the following function:

- defines the structure of raw data given by the user using the preprocessing data dictionary
- propose preprocessing techniques that suitable for data
- creates user's preprocessing profile
- clean incoming new data and keeps in a temporary file

The raw data must be in unstructured table and kept as a text file. The system will identify the structure of the data, such as identify the types of the attributes such as integer or string, nominal or ordinal and perform categorization process if it is nominal. The user is able to update and validate the structure of the data. Base on the data, the agents will examine the data and do the basic cleaning process needed.
Coordinator agent will send preprocessing task to other agent depends on what preprocessing type required of the particular data. For example, if there are missing data, the agent will identify and treat missing data by replacing it with certain value. Agent will apply several techniques to fill the missing data and techniques that produce a better result will be proposed to the user as the best techniques.

8. Agent-Based Olap/Olam Architecture

There are eight supporting agents: data cleaning agent, transformation agent, data reduction agent, data integration agent, proxy MDDB agent, input agent, GUI agent and diverting agent. Also we have three co-coordinating agents: DB API agent, MDDB agent and

CUBE API agent. designed for this framework. In this framework, each agent is responsible to capture all the user preferences including the analysis of the domain data. All agents have the capabilities to communicate and cooperate with each other through coordinator agent. This framework is show in Figure 1.

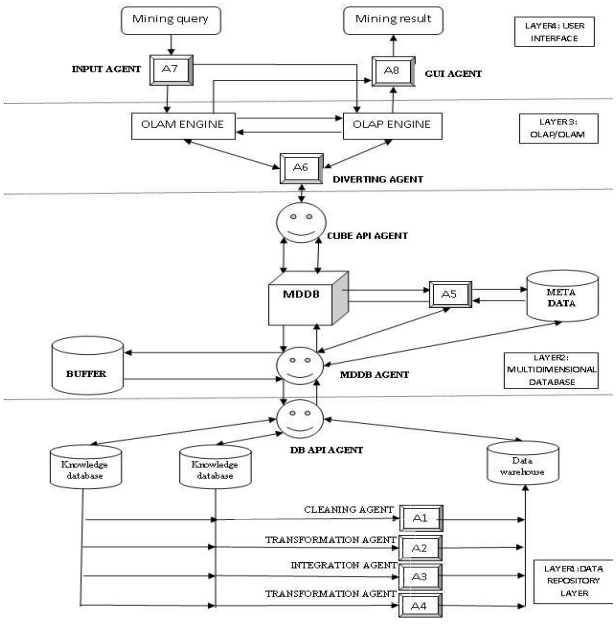

Figure 1 :Agent Architecture for OLAP/OLAM

a. User Interface (GUI & INPUT) Agents:

The User Interface Agents are used to provide the user interface with the system. The User Interface Agent proactively provides solution analysis for the data. It assists the user in formulating queries and displaying output and report as requested by user. User Interface Agent also helps users carry up activities in an autonomous way.
It gets the result from the OLAP/OLAM engines and processes it, to present the result in the user expected form.

b. Diverting Agent

This layer receives the data from cube API agent, recognizes the nature of the data and divert to either OLAP agent or OLAM agent according to the nature of data.
It is responsible to check and identify types of data and attributes in database

c. OLAP/OLAM Engines

On-line analytical processing (OLAP) can be performed in data Figure1: AGENT Based Integrated OLAM And OLAP Architecture warehouse/marts using the multidimensional data model. Typical OLAP operations roll-up, drilldown (across through), slice-and-dice, pivot (rotate), as well as statistical operations such as ranking and computing moving averages and growth rates. OLAP operations can be implemented efficiently using the data cube structure.
OLAP-based data mining is referred to as OLAP mining, or on-line analytical mining(OLAM),which emphasizes the interactive and exploratory nature of OLAP mining.

d. Cube API Agent

Cube API agent acts as a intermediatory between MDDB layer and OLAP/OLAM layer.

e. MDDB Agent

MDDB agent is responsible to build a MDDB, according to the information or data received from data repository layer.
This layer has the responsibility to update the MDDB and Meta data.
This layer also holds the entire control of MDDB and meta data. MDDB agent is like a manager, which responsible for coordinating the various tasks that needs to be performed in a cooperative problem solving between the user and other agents. It can determine the required preprocessing task based on data mining task chosen by the user. The tasks can be generated automatically based on meta-knowledge in the coordinator agent. This Agent also responsible to access the data from database and provides data for other agents in the system. New data can be added to the present database dynamically. Coordinator agent conducts a request to agents based on data condition and user preferences.

f. Proxy MDDB Agent

Proxy MDDB agent acts as a MDDB agent when the original MDDB agent gets failed. This layer reports to the MDDB agent regarding the errors in the updations of MDDB With the help of meta data.
This agent shares the job of MDDB agent when it is overloaded. It also identifies problems that occur in database such as missing and noisy data based on data dictionary. If new problem occurs, it will save it as a new knowledge. The best preprocessing techniques is stored in the user profile.

g. DB API Agent

DBI API is responsible for the exchange of data between data repository layer and MDDB layer.
This layer holds and control a buffer which contains a frequent result sets.

h. Data cleaning Agent:

Data cleaning agent handle missing and noisy data by using various types of techniques based on type of missing and noisy cases.

i. Transformation Agent

Transformation Agent is used to transform the data into appropriate forms for mining.

j. Data Reduction Agent

Reduction agent performs the reduction of redundant data's and responsible for reducing unwanted attributes by using the attributes subset selection techniques It identifies and ignores the out layers from the actual data The role of reduction agent is also to discretize the data by using discretization techniques selected.

k. Data Integration Agent

Data integration agent integrates the data from various types of repositories such as relation database, multimedia databases, flat files etc. Techniques for Cleaning Agent, Transformation Agent, Reduction Agent and Integration Agent can be added dynamically. New preprocessing techniques can be added to the system as a new knowledge. These agents also handle new data for a present database. CleanMiss Agent, CleanNoisy Agent, Transformation Agent and Discretization Agent also propose preprocessing techniques that suitable for data and save the options choose as knowledge for future use

9. Conclusions And Future Work

The proposed agent based integrated OLAP/OLAM design has a promising future in improving the data mining technology. The framework is designed scalable in such a way it allows the addition of new techniques and other data mining capabilities. By the combination of data mining and agent, data preprocessing can be simplified and yet reliable. Various techniques such as basic cleaning rules and some cleaning techniques have been implemented and tested using standard data from the UCI machine learning data repository.
However, there are spaces for improvement where more techniques need to be included in this framework. The future work of this study is to build the adaptive preprocessing agent, which handle preprocessing of the new data using the techniques keep in the user profile. The main intention of the proposed framework is to gather all preprocessing techniques available and to support the online preprocessing services which also provide an adaptive knowledge on preprocessing experience among the users within various domains of data.

Reference

[1] C. Aflori and F. Leon, "Efficient distributed data mining using intelligent agents," 2004.
[2] M.H. Dunham, Data Mining: Introductory and Advanced Topics, Upper Saddle River: Prentice Hall, 2003.
[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," in AI Magazine 17: pp. 37-54, 1996.
[4] A. Mohtar. Multiagent Approach to Stock Price Prediction. University Kebangsaan Malaysia. 2006.
[5] P. Nurmi, M. Przybilski, G. Linden, and P. Floreen, "An architecture for distributed agent-based data preprocessing," in AIS-ADM 2005, pp. 123-133, 2005.
[6] Agent Working Group, "Agent technology," OMG Document ec/2000-08-01, Version 1.0, 2000.
[7] Y. Bo, W. Ya-Dong, and S. Xiao-Hong, "Research and design of distributed training algorithm for neural network," Proceedings of the International Conference on Machine Learning and Cybernetics, 2005.
[8] C. Li, and Y Gao , "Agent-based pattern mining of discredited activities in public services," in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2006.
[9] Abd. Manan Ahmad, AG. Noorajis Ag. Nordin, Emrul Hamide Md. Saaim, Fairol Samaon and Mohd Danial Ibrahim, "An architecture design of the intelligent agent for speech recognition and translation," in IEEE, 2004.
[10] Weiser Mark, "The Computer for the Twenty-first Century", *Scientific American*, 1991, 265(3): 94-100.
[11] MAGNUS BLOMSTROM, KOKKO, "Human Capital and Inward FDI", *NBER working paper*, 2003, 1
[12] Krishma K, Murty M. N, "Genetic k-means algorithm", *IEEE Tram on System, Man, and Cybernetics: Part B*, 1999, 29(3):433-439.
[13] Jin Hong, Zhou Yuan-hua, "Video Disposing Technology Basing on Content Searching", *Journal of Image and Graphics*, 2000, 5(4):276-282.
[14] H Kargupta, B Park, et al, "Mob Mine: Monitoring the Stock Market from a PDA", *ACM SIGKDD Explorations*, 2002, 3(2): 37-46.
[15] D Terry, D Goldberg, D Nichols, et al, "Continuous Queries over Append-only Databases", *Proceedings of the ACM SIGMOD Int'l Con.f on Management of Data*, Madison: ACM Press, 2002. 1-26.
[16] JS Vitter, "Random Sampling with a Reservoir", *ACM Trans. on mathematical software*, 1985, 11(1): 37-57.
[17] Weiser Mark, "The Computer for the Twenty-first Century", *Scientific American*, 1991, 265(3): 94-100.

[18] MAGNUS BLOMSTROM, KOKKO, "Human Capital and Inward FDI", *NBER working paper*, 2003, 1
[19] Krishma K, Murty M. N, "Genetic k-means algorithm", *IEEE Tram on System, Man, and Cybernetics: Partb*1999, 29(3):433-439.
[20] Jin Hong, Zhou Yuan-hua, "Video Disposing Technology Basing on Content Searching", *Journal of Image and Graphics*, 2000, 5(4):276-282.

Biography

Mr.M.ShivaKumar, Research Scholar, Department of CSE, Karpagam Universty, Coimbatore, T.N,India.he has published papers in various conferences (National & international). He has good academic line of experience, Presently working has Associate Professor, in the Department of CSE,Kuppam Engineering College, Kuppam,A.P ,india.

Dr. K. Krishna Moorthy,Professor, Department of CSE,Sona College of Technology,Salem,T.N,India,has vast Experience and published papers in various conferences (National & international)