An Effective Data Preprocessing Technique for Improved Data Management in a Distributed Environment

Sharon Christa
Post Graduate Programme, RIIC
Dept of ISE,
DSCE, Bangalore,
India.

V. Suma
RIIC,
DSI,
Bangalore,
India.

Lakshmi Maduri
RIIC,
DSI,
Bangalore,
India

ABSTRACT
With the evolution of distributed computing, the databases are inherently distributed across the globe and therefore data analysis from various data sources is very essential in decision making. The core need in the current industrial environment is hence to extract information from the huge, complex and dynamic data through data mining techniques. Integrating data from multiple data sources and analysing the large, complex dynamic data is a tedious and complex work. Additionally, database consists of inconsistent and noisy data. Further, with the decrease in quality of data to be mined the quality of knowledge model obtained from it also decrease which inturn affects the decision making process. However optimization of data preprocessing can resolve the aforementioned issues. This paper provides design and development of data preprocessing software, based on intelligent agents. This software enables data preprocessing operations to be performed in an automated mode, and gives accurate results in lesser time when compared to manual data preprocessing.

General Terms
Data Mining, Data Preprocessing, Intelligent Agents.

Keywords
Discretization Agent, Transformation Agent, Data Clean Agent, Inconsistent Data, Processed Data.

1. INTRODUCTION
Data Mining is the most popular tool that is used in many fields such as bioinformatics, stock market etc [1]. Data mining is a technique to analyze large amount of data which can identify the patterns and correlations that exist in the massive amount of data in order to find the hidden knowledge in it using various algorithms. The data mining process uses various algorithms such as classification, clustering, association, prediction to realise the knowledge model for apt decision-making and predictive etc. [2].

Even though data mining play a significant role in decision making and other database related issues, inconsistencies and noise present in it results in producing unreliable outcome. It is worth to note that real world data is prone to inconsistencies, duplicate values and can use different units for same field. Mining with such data inturn give unreliable knowledge model [3]. The data mining algorithm depends on the quality of data preprocessing. Data preprocessing is the initial step that is performed prior to data mining to remove the noise and inconsistencies in the data set. It consists of four phases mainly data integration, data reduction, data transformation, and data cleaning. Further, when data is dynamically updated; one has to mine the whole data once again on order to obtain the updated knowledge model [5]. However processing and managing complex data and handling dynamic data are some of the significant features which are not currently available data mining tools [4]. To overcome the above said drawbacks it is required to develop intelligent data preprocessing software.

This paper provides an overview, design and implementation of intelligent data preprocessing software which can perform all the operations in an automated way. These intelligent agents will themselves perform all data preprocessing activities in lieu of manual setting of parameters and analysing the data which is the current day scenario in all IT industry.

The organization of the paper is as follows: Section 2. is Literature Survey which describes about the related work done by several research scholars in this domain, Section 3. Explain the Design Model, Section 4. Elucidates the architecture of the about the preprocessing model and section 5 depicts the implementation of the model section 6 provides the summary of the entire work also has some screenshots of the software.

2. LITERATURE SURVEY
Ever since the advent of technology progress of research is witnessed in effectively managing massive amount of data. According to the authors in [6], data mining is the process of extraction of interesting information or patterns from data in large databases. They state that agents can be either software or hardware entities that performs some specified set of tasks on behalf of the user having some degree of autonomy. They further states that Knowledge discovery in databases can be achieved through several steps which include data preparation, selection of data mining model, its application, and its corresponding output analysis. The authors further express that with the use of intelligent agent paradigm it is possible to automate the individual tasks. Authors in [7] state that, agents can be activated remotely over the network or can be triggered on the occurrence of certain event and starts an analysis operation. Finally, agents can help navigate and model the World-Wide Web and other areas of growing importance. By developing agent based data preprocessing software the performance of data mining tool further improves. Therefore, integration of agent specifically in data preprocessing is considered a

sensible approach in handling dynamic data [8][9]. Data preprocessing is the initial task in knowledge discovery. Agent based systems are the outstanding approach to overcome the drawbacks of data preprocessing software. In recent years, agents have become popular paradigm in computing because of its autonomous, flexible, adaptive and intelligent characteristics [10]. Data preprocessing is a set of task that include data cleaning, data integration, data transformation, and data reduction. Data cleaning activity eliminates the noise present in the data set. Data integration phase enables merging of data from multiple sources into a coherent data source. Data transformation activity transforms the data into a suitable form for data mining and data reduction activity reduces the data size [11][12]. Intelligent agents behave rationally [14]. The intelligent agents do the work with human intelligence but may not behave like human beings. The agent which is involved in processing of mining performs productive task, retrieves useful knowledge with less noise and reduced processing time when compared to the normal mining tools [13][15].

3.DESIGN MODEL

In the current scenario efficient knowledge modelling is an elementary strategy and that makes data preprocessing an inevitable phase. Also data preprocessing consumes almost eighty percent of the overall time taken for knowledge discovery in data bases [17]. Methods for data preprocessing in the currently available data-mining tool is analysed. The analysis and evaluation of data indicated that the data pre-processing phase consumes most of the data mining time and effort [16]. Optimization of the aforementioned issues is the main aim of my research. Since, agents are intelligent systems that can perform operations with some level of intelligence and understanding capabilities, incorporating intelligent agents to the data pre-processing system enables to achieve efficient data preprocessing. The activity diagram of the data preprocessing architecture is given in figure 1. Dataset for data preprocessing is obtained from multiple data sources and is first analysed for the type of data, the attributes in it, the steps to be performed etc and its metadata is created. Then the data reduction process is performed in it which will reduce the size of the data and enables to select which all attributes has to be retained. After that the dataset is checked to see if all data are in the same format and is done so if it is not of same format. Then the whole data is scanned to see if any inconsistencies are there, if any tuples are left blank. For this process the metadata created is used. The whole process is done as two steps, handling missing data and smoothing the data. Thus obtained data is called the processed data which is given as the input to data mining process.

4. AGENT BASED DATA PRE - PROCESSING SOFTWARE

According to Maes, Intelligent Agent (IA) is software entity that can be used to perform the operations independently. He also states that these agents therefore can be used in lieu of user or another program [18][19][20].By integrating agent technology with data preprocessing the performance of the tool further improves. Therefore, integration of agent specifically in data preprocessing is considered a sensible approach in handling dynamic data [16] [13] [2] [5]. Each agent has specific characteristics and it vary depending on the problem domain. In a multi agent system agents communicate, co-operate and co-ordinate with the other agents. Each agent in the system acts autonomously, and co-operates with other agents. They work together for the tasks to be performed to achieve the goal of the system [18]. General characteristics of an agent are autonomy, adaptive, cooperative, interactivity, intelligent, learning, ruggedness, continuous, coordinator, and mobility [21].

The architecture of the data preprocessing software is in figure 2. Data preprocessing application acts as an interface that process the data to be mined. The dataset whit inconsistency is stored in the database with the

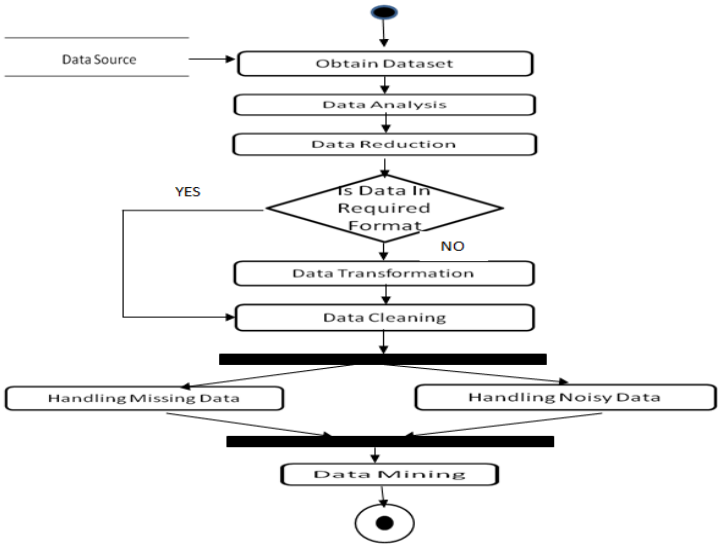

Figure 1: Activity Diagram of the Preprocessing Software

metadata of it. The whole data is analysed to find the inconsistencies, duplicate fields, multiple data formats in same attribute, missing data field. This is all done by the coordinator agent. Then according to the operations to be

performed it is given to the agents that will perform each operation, like data transformation by transformation agent, data reduction by discretization agent, data cleaning by clean miss and clean noisy agent.

The functions of the preprocessing software include data integration, data reduction, data transformation, data cleaning and data visualization. Each of the function is handled by intelligent agents.

Data integration process combines data from multiple sources like data cubes, multiple databases, and flat files. It performs schema integration and also objects matching. It makes use of the metadata for the integration. It performs correlation analysis and also chi-square test to handle redundancy.

Data transformation involves smoothing, aggregation, generalization, normalization in which data is scaled to a specific range and also attributes construction. Smoothing is done by various regression methods, binning etc whereas attribute construction will construct new attributes from the existing one. These methods help to make data more appropriate for mining.

Strategies in data reduction are data cube aggregation; attribute subset selection, dimensionality reduction, numerosity reduction, and discretization and concept hierarchy generation. This is performed so that the time taken for complex data analysis and also mining huge amounts of data can be avoided.

Data cleaning is a two step process which includes handling missing values and handling noisy data. It is done with discrepancy detection that makes use of the metadata which give the knowledge about domain and data type. While scanning the dataset, if the tuples have no recorded value then various strategies are used like ignore the tuple, fill each missing values manually, use the most probable value, use the attribute mean to fill in case of numeric data, use a constant like unknown or infinity. Binning, regression and clustering are done to remove the errors and smoothing the data.

The software has mainly five agents: coordinator agent, discretization agent, transformation agent, clean miss agent, clean noisy agent and the responsibilities of each agents are: Coordinator 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 can be generated automatically based on meta-knowledge in the coordinator agent. CleanMiss Agent and clean noisy agent handle missing and noisy data by using various types of techniques based on type of missing and noisy cases. Transformation Agent is used to transform the data into appropriate forms for mining. The role of reduction agent is to discretize the data by using discretization techniques selected.

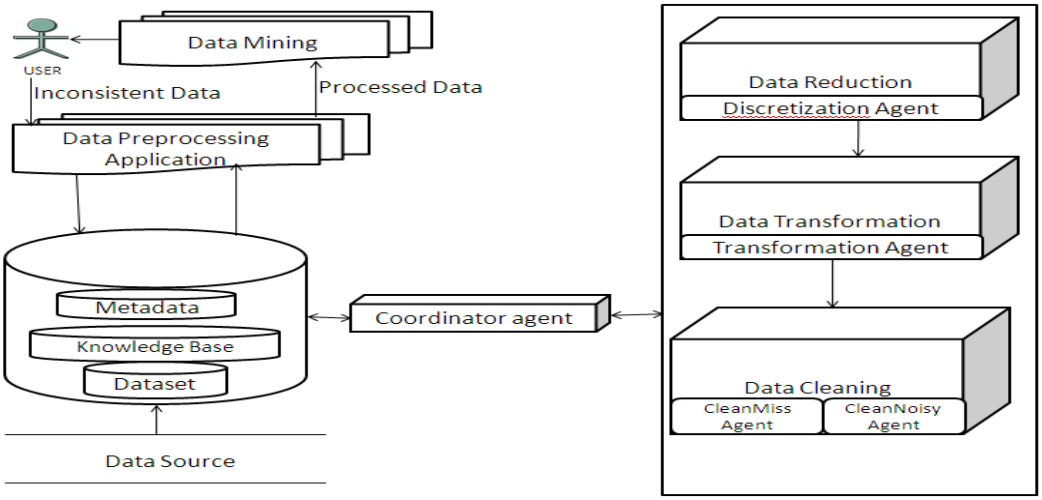

Figure 2: Architecture of the Preprocessing System.

5.IMPLEMENTATION OF THE DATA PRE - PROCESSING ARCHITECTURE

Each phase of the preprocessing activity is assigned to an agent who has to report ultimately to the coordinator agent. The db further has the repository of each of these processing activities. The architecture of the preprocessing system is implemented using NetBeans IDE in which each agent is developed. Figure 3, Figure 4, Figure 5 provides the snapshots of the implementation.

It is worth to note at this point that data analysis is performed by the coordinator agent and hence the need for manual interpretation of the analyzed data in the tabular format is also overcome. Figure 3 depicts the user authentication and user interface implementation. When the user opts to upload a data set to be analysed the file uploading page opens up. Figure 4 depicts the file uploading page. Once dataset is uploaded its each attribute is identified and checked for existence of any missing values, the range of values that it takes and for the existence of duplicate values which can be removed if desired. Figure 5 depicts the attributes that are having blank tuples and also provides options for handling them. Further data reduction if chosen provides various options to reduce the data in terms of value, range etc.

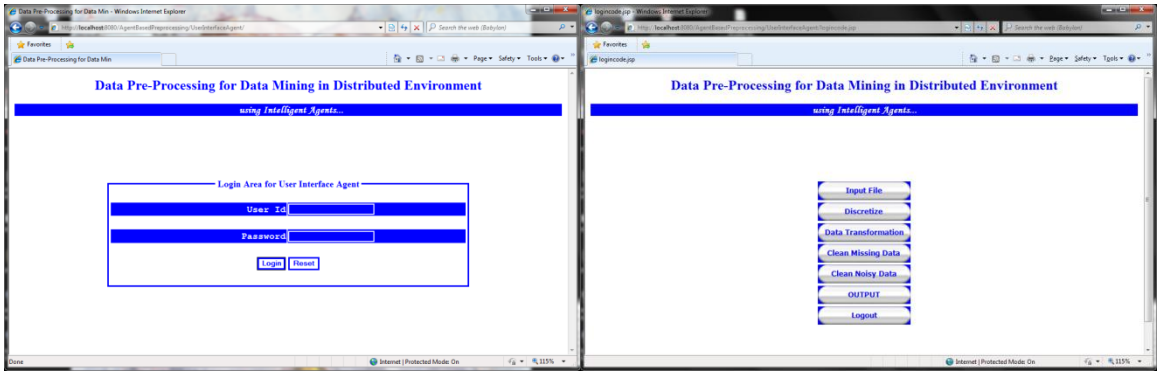

Fig 3: User Authentication and User Interface Design

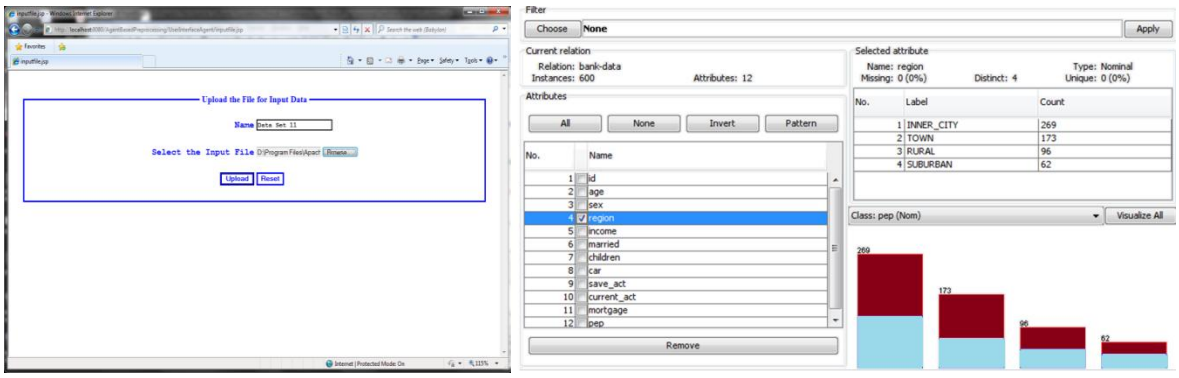

Fig 4: File Uploading and Output data in Weka Tool

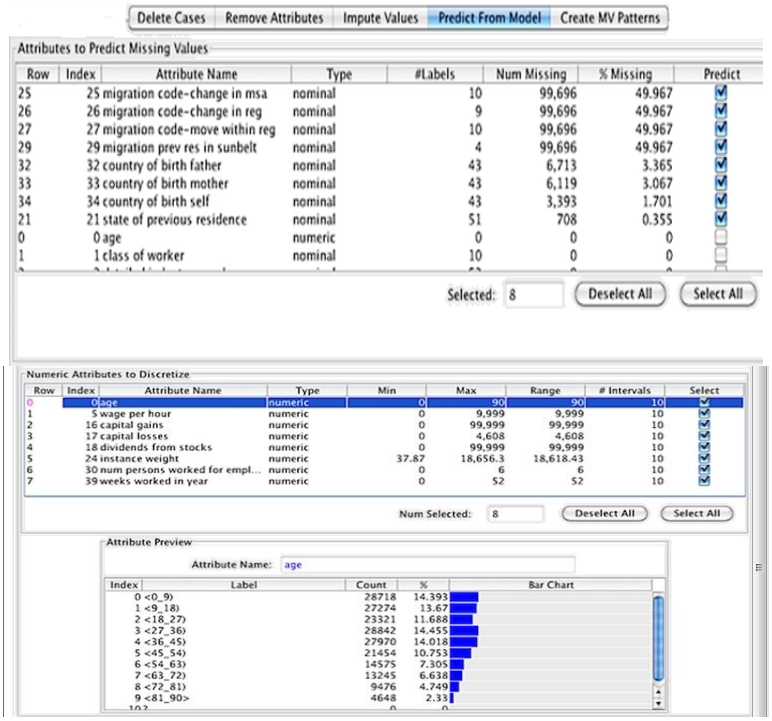

Figure 5: Data Miss and Discretize Agent

6. CONCLUSION

The data preprocessing software ensure to improve the quality of data mining as it is designed by considering scalability characteristics. However our forthcoming papers will provide the implementation results of data integration and elimination of data inconsistency using data cleaning agent. With the evolution of distributed computing, the databases were inherently distributed across the globe. The core need in the current industrial environment is to extract information from the huge, complex and dynamic data through data mining techniques. However, existing data mining tools are not effective in efficiently processing the dynamic and inconsistent data. Therefore, it is imperative to develop intelligent data preprocessing agents which can themselves perform all data reprocessing activities in lieu of manual setting of parameters and analysing the data which is the current day scenario in all IT industry. This paper provides an overview design and implementation of intelligent data preprocessing software. The implementation of intelligent data preprocessing technique using intelligent agent ensures to improve the data quality in terms of time and human intervention.

7. REFERENCE

[1] Peng Jin, Yun-Long Zhu And Kun-Yuan Hu. August , 2007 A Clustering Algorithm For Data Mining Based On Swarm Intelligence Proceedings Of Sixth International Conference On Machine Learning Cybernetics, Hong Kong, 19-22

[2] Pyle, D. 1999 Data Preparation for Data Mining. Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA

[3] C., Lavrac, N., Moyle, S., Kavsek, B. 2001Integrating Aspects of Data Mining, Decision Support and Meta-Learning: Internal SolEuNet Session, ECML/PKDD'01 workshop notes 43–52

[4] B. Liu and A. Tuzhilin 2008: Managing and Analyzing Large Collections of Data Mining Models, Communications of ACM, Vol. 51, No. 2.

[5] Zulaiha Ali Othman, Azuraliza Abu Bakar, Abdul Razak Hamdan, Khairuddin Omar and Nor Liyana Mohd Shui, 2007 "Agent based preprocessing," International Conference on Intelligent and Advanced Systems.

[6] Cristian Aflori and Florin Leon 2008: "Efficient Distributed Data Mining using Intelligent Agents Authors"

[7] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth 2005: "From Data Mining to Knowledge Discovery in Databases"

[8] I. A. Mohtar, 2006 "Multiagent Approach to Stock Price Prediction," University Kebangsaan, Malaysia.

[9] P. Nurmi, M. Przybilski, G. Linden, and P. Floreen, 2005 "An architecture for distributed agent based data pre-processing." Pp. 122-132.

[10] C. Li, and Y Gao, 2006"Agent-based pattern mining of discredited activities in public services," Proceedings of the 2006 IEEE/WI C/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[11] Dr. T.R. Gopalakrishnan Nair, Lakshmi Madhuri, Sharon Christa, Dr. V. Suma, 2012 "Data Preprocessing Model Using Intelligent Agents" International Conference on Information Systems Design and Intelligent Applications.

[12] Agent Working Group, 2000 "Agent technology," OMG Document ec/2000-08-01, Version 1.0.

[13] Stuart Russell and Peter Norvig 1995 "Artificial Intelligence: A Modern Approach", c Prentice-Hall, Inc.

[14] Sharon Christa, K. Lakshmi Madhuri and V. Suma 2012, "A Comparative Analysis of Data Mining Tools in Agent Based Systems", International Conference on Systemics, Cybernetics and Informatics

[15] Ranjit Bose, Vijayan Sugumaran 1998, "IDM: An Intelligent Software Agent Based Data Mining Environment," IEEE.

[16] K. Sycara, et.al 1996 "Distributed Intelligent Agents," IEEE Intelligent Systems, pp- 35-46.

[17] P. Maes July 1994, "Agents that Reduce the Work and Information Overload," Com. of ACM, Vol. 36,No. 7, pp. 29-39.

[18] S. Masina, K. Y. Lee, and R. Garduno-Ramirez 2004, "An Architecture of Multi-Agent System Applied to Fossil-Fuel Power Unit," IEEE Power Engineering Society General Meeting, pp. 1982-1988.

[19] Abd. Manan Ahmad, AG. Noorajis Ag.Nordin, Emrul Hamide Md. Saaim, Fairol Samaon and Mohd Danial Ibrahim 2004, "An architecture design of the intelligent agent for speech recognition and translation" IEEE.

[20] D. Kehagias, K. C. Chatzidimitriou, A. L. Symeonidis, and P.A. Mitkas 2004 "Information agents cooperating with heterogeneous data sources for customer-order management," ACM Symposium on Applied Computing, pp. 52-57.

[21] Dr. Joseph P. Bigus, Jennifer Bigus "Constructing Intelligent Agents with JAVA"