A Proposed DDM algorithm and framework for EDM Of Gujarat Technological University

KEYWORDS Educational Data Mining (EDM), Distributed Data Mining (DDM), KDD, Middle Ware, Query Processing, Data Warehouse

Dineshkumar B. Vaghela	Dr. Priyanka Sharma
PhD Scholar, GTU, Chandkheda.	PhD Guide, GTU, Chandkheda.

ABSTRACT Educational Data Mining (EDM) is an emerging interdisciplinary research area and with the essence of data mining concepts, it deals with the processing and knowledge acquisition from the data originating in educational context. EDM uses computational approaches to process and analyze educational data to study educational questions. Distributed Data Mining (DDM) plays important role because of; first, mining requires huge amounts of resources in storage space and computation time. To make systems scalable, it is important to develop mechanisms which distribute the work load among several sites. Second, data is often inherently distributed among several sites, making centralized processing of this data may prone to security risks and inefficient. Distributed Data Mining explores techniques of how to apply Data mining in a non-centralized way. DDM can be effectively used for EDM as because of many universities having many colleges of different disciplines. In this research paper, we have proposed an algorithm and the frame work for Gujarat Technological University (GTU) for Educational Data Mining using the concept of Distributed Data Mining.

INTRODUCTION

The extensive usage of information technology in educational institutes leads the generation and collection of large volumes of data storage in different formats like records, files, documents, images, sound, videos and many new data formats. For better decision making, the data collected from large repositories of different applications require proper method of extracting knowledge. There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational environments [3]. The Educational Data Mining (EDM) aims at the discovery of useful information from large collections of data [1].

Educational Data Mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor, and many others for classification, prediction, association rule generation, discover and extract patterns of stored data [2]. The discovered knowledge can be used for prediction regarding enrolment of students in a particular course, faculty performance analysis, placement prediction, prediction about students' performance and so on.

The University has many colleges and offers different courses. The admission process is centralized, but the college and course allocation to the students are different based on their merit marks. So at the end of the admission process, each college has its separate list of the students and the institute will only be responsible for their data for the entire course duration. Each institute maintains students' personal, academic results, co-curriculum activities, placement related data, faculty performance analysis and so on.

As the data to be processed are generated, stored, maintained and processed at different geographical sites i.e. institutes. We have proposed the framework and the algorithm with distributed data mining to process the data of different sites as a whole for decision making.

In this paper first we have discussed the related work for Educational Data Mining; in the later sections we have discussed the proposed algorithm and the framework for EDM using DDM. At last we have discussed the conclusion and the future work.

RELATED WORK

Educational Data Mining (EDM) is an immerging trend and attractive discipline which does various predictions in educational institutes or universities. There have been various works in this area involving different techniques of data mining.

A. Evaluation of Students academic performance Shreenath Acharya et al. [4] presented an overview on Knowledge Discovery on Databases to predict the student's academic trends. Lots of work has been done in this using data mining. Using the historical information and the current semester data students' performance can be predicted. Many of association rule mining algorithms can be used to generate the frequent item sets which can be useful to predict the students' performance. Kalpesh Adhatrao et al.[5] have applied different decision tree algorithms like ID3(Iterative Dichotomiser) and C4.5 to predict students academic performances and checked their accuracy to choose the appropriate algorithm that could be effectively applied. Mohammed M. Abu Tair and Alaa M. El-Halees [6] have given a case study on how to use different data mining techniques like Association Rule Mining, Clustering, Classification and Outlier Detection at various phases to improve the academic performance of graduate students.

B. Predicting School dropouts N. Quadri1 et al [7] presented a technique to determine the list of students who are likely to drop out of the college. Various attributes like their gender, family background, attendance etc. were considered while mining. Decision tree algorithm has been suggested as to predict the dropout. Gerben W. Dekker [8] gave a case study which included attributes that could act as a strong predictor for success. The famous Decision tree algorithms like C4.5 and CART were found suitable to predict dropouts.

C. Students behavioral prediction M.Sindhuja et al. [9] have proposed a methodology which explored Behavior Attitude Relationship Clustering (BARC) Algorithm to show the improvement in students performance in terms of predicting their behavioral, attitude and relationship with faculty members and Tutors. Hierarchical clustering was used to group students based on their similarity measures and the study was experimented using Weka Tool. Ryan S.J.D. Baker [10] presented a machine-learning technique that could instantaneously find when a student using an intelligent tutoring system is off-task, i.e., engaged in things which does not include the system or a learning task. Only the system log files were considered while mining and Latent Response Models were used as the statistical basis for all of the detectors of off task behavior.

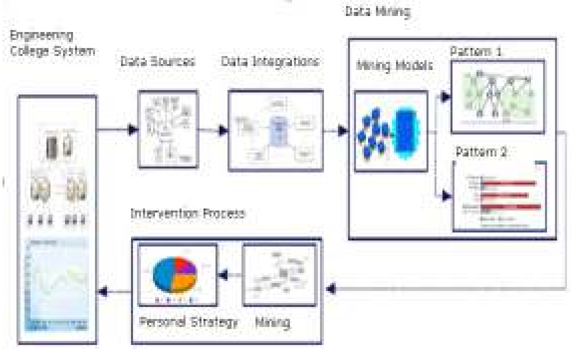
D. Distributed data mining: algorithms, systems and applications, Byung-Hoon Park and Hillol Kargupta[11] presented a research paper on different distributed data mining algorithms and also discussed about the systems and applications of the distributed data mining.
In the next section, we have discussed the proposed algorithm, system and framework for distributed data
mining of educational data mining of Gujarat Technological University which has different zones in entire Gujarat. Each zone has several numbers of colleges for ease of administrative work at regional level.

PROPOSED ALGORITHM
We have considered the engineering colleges zone wise under taken by Gujarat Technological University. The admission process is carried out by the ACPC (Admission Committee for Professional Courses) for all the engineering colleges under GTU. We have proposed one common algorithm which will be followed by every educational data mining task like students' admission prediction, students' performance predictions, faculty performance prediction, student drop out, predicting college failure using historical/past data, grouping students based on their results … etc. Here below we have proposed an algorithm.

Step 1: Data Integration at each zonal server

Step 2: For very first data mining (For each category of operation), do process on entire data set stored at each zonal data warehouse. Say the result as RZi where i=1,2,3,….N (Number of zones) Incremental Approach: For other subsequent data mining operations, fetch only the newly arrived data set NDZi of ith zone and process it, say the result NRZi. Then consolidate it with previous result and let say the newly consolidated result as FRZi.

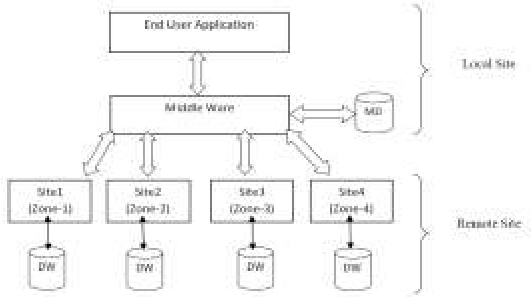
Step 3: Using proposed message passing model, collect the results FRZi from different zonal servers to the client machine and display the final consolidated result. First of all kind of data from different engineering colleges are integrated at respective zonal server, some of the data pre-processing techniques can be applied. For each category of task performed the EDM, which later consolidated with the results generated of only newly arrived data set (i.e. incremental approach). This approach will be helpful to reduce the overall processing time of data mining task as we need not to apply the algorithm on the same data which have been processed earlier. The mentioned EDM zonal wise which can later be used with the results of other zonal results in order to perform the Data Mining operation for the entire GTU. In the below figure we have mentioned the proposed system at each zonal server.



Engineering College System: Each engineering college has its educational system, maintaining the data related with the student admission, students results, student performance, faculty performance … etc.
These data are stored in different databases.

Data sources & Data integration: The data generated from the educational system will be the data source, which are integrated to the zonal server using the data integration.
Data Mining: Each zonal server has the data warehouse of all the engineering colleges under it. It also has data mining algorithms to process the user's query which will generate different results.
Intervention system: The results generated with the data mining will be applied to the personal strategy to take the decisions.

PROPOSED FRAMEWORK
For the Educational Data Mining using the concept of Distributed Data Mining, below is the proposed framework which has end user application and the middle ware at the local site, and the remote site (zone) contains its own data warehouse and the process for different EDM tasks. The Meta Directory (MD) contains the information regarding each zonal server for EDM. The Middle Ware works as the interface between end user application and the remote site. Her Middle Ware does the marshaling and other operations such as locating the zonal server using MD.



REFERENCE [1] Heikki, Mannila, Data mining: machine learning, statistics, and databases, IEEE, 1996. | [2] U. Fayadd, Piatesky, G. Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6, 1996. | [3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000. | [4] Shreenath Acharya, Madhu N, "Discovery of students' academic patterns using data mining techniques" IJCSE,Vol. 4 No. 06 June 2012. | [5] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao "PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS" IJDKP, Vol.3, No.5, September 2013. | [6] Mohammed M. Abu Tair and Alaa M. El-Halees "Mining Educational Data to improve Students' performance "JICT, Volume 2 No. 2, February 2012. | [7] Mr. M. N. Quadri1,Dr. N.V. Kalyankar "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques"GJCST, Vol. 10 Issue 2 (Ver 1.0), April 2010. [8] Gerben W. Dekker "Predicting students drop out: a case study" 2nd Int. Conf. on Educational Data Mining, April 10, 2009. | [9] M.Sindhuja et al. "Prediction and Analysis of students Behaviour using BARC Algorithm",IJCSE, Vol. 5 No. 06 Jun 2013. | [10] Baker.R., "Modeling and understanding students' off-task behaviour in intelligent tutoring systems. In Conference on Human Factors in Computing Systems", San Jose, 2007 California, 1059-1068 | [11] Byung-Hoon Park and Hillol Kargupta,"Distributed data mining: algorithms, systems and applications"