RESEARCH PLAN PROPOSAL

Efficient data mining techniques for application in elementary educational system with special reference to out of school children

For registration to
Doctor of Philosophy

IN THE FACULTY OF COMPUTER SCIENCE

To

THE IIS UNIVERSITY, JAIPUR

Submitted by:
Astha Pareek
ICG/2010/11460

Under the Supervision of:

Dr.Manish Gupta
Dy.Director (IT), RCEE,
Deptt. Of Education, GoR

Dr.Amita Sharma
Asst.Professor (CS & IT)
ICG-THE IIS
UNIVERSITY, Jaipur

Department of CS & IT
December, 2011

1. Topic

Efficient data mining techniques for application in elementary educational system with special reference to out of school children.

2. Introduction

Data Mining techniques are used to extract meaningful information and to develop significant relationships among variables stored in large data set/ data warehouse. According to Alaa el-Haees(2009)Data mining, a branch of computer science is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. Data mining techniques are currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. The related terms *data dredging*, *data fishing* and *data snooping* refer to the use of data mining techniques to sample portions of the larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These techniques can, however, be used in the creation of new hypotheses to test against the larger data populations.

Data Mining is a process of extracting previously unknown, valid, potentially useful and hidden patterns from large data sets (Connolly, 1999).The amount of data stored in educational databases is increasing rapidly.Efficent data mining techniques are required in order to get required benefits from such a large data and to find out hidden relationships between variables (Han and Kamber, 2006). Clustering and decision tree are most widely used techniques for future prediction from educational databases.. The main goal of clustering is to partition children into homogeneous groups according to their characteristics and abilities (Kifaya, 2009). These applications can help teachers/ educational administrators and society to enhance the quality of education at different levels.

Data mining, sometimes also called Knowledge Discovery in Databases (KDD), can find relationships and patterns that exist but are hidden among the vast amount of educational data. It combines machine learning, statistical and visualization techniques to discover and extract knowledge in such a way that humans can easily comprehend. For universities, the knowledge discovered by data mining techniques would provide a personalized education that satisfies the demands of students and employers. In order to deliver meaningful analysis, data mining techniques can be applied to provide further knowledge beyond the data explicitly stored. Compared to traditional analytical studies, data mining is forward looking and is oriented to individual students. For example, the clustering aspect of data mining can offer comprehensive characteristics analysis of students, while the predicting function from data mining can help the university to act before a student drops out or to plan for resources based on the knowledge of how many students will transfer or take a particular course.

Data mining in general consists of five major elements:

- Extract, transform and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.

- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Data mining can be applied to a number of different applications, such as data summarization, learning classification rules, finding associations, analyzing
changes, and detecting anomalies (Han et al., 2006,Westphal et al., 1998). Sometimes, data mining has to deal with unstructured or semi-structured data, such as text. Text mining is defined as, "the automatic discovery of previously unknown information by extracting information from text"(Spasic et al., 2005). Data mining is widely applied in many areas such as retail, financial, communication, and marketing organizations

Data Mining Techniques:

There are various techniques of data mining to find interesting patterns in the data set. Some of the important data mining techniques are:

- Artificial neural network: - Non-linear predictive models that resemble biological neural network in structure.
- Decision trees: -A decision tree consists of nodes and branches, starting from a single root node. Each node represents a test or decision. Depending on the outcome of the decision, one chooses a certain branch and when a terminal node (leaf) is reached, a decision on a class assignment is made.
- Genetic Algorithm: - Genetic algorithm are techniques that act like bacteria growing in a Petri dish. We set up a data set and then give the Genetic Algorithm ability to do different things, such as to decide whether a direction or outcome is favourable. The Genetic algorithm will move in a direction that will hopefully optimize the final result. GAs are used mostly for process optimization, such as scheduling, workflow, batching and process re-engineering. Optimization techniques use processes such as genetic combination, mutation and natural selection in a design based on the concept of natural evolution.
- Clustering:-Clustering is the task of organizing data into groups(known as clusters) such that the data objects that are similar to (or close to) each other are put in the same cluster. A cluster is therefore a collection of objects which are “similar” amongst them serves and are “dissimilar” to the objects belonging to other clusters.
- Association Rules: - Association rule mining finds interesting association relationship among large set of data items. A typical example of association analysis is market basket analysis. This process analyses customers buying patterns by finding association between different items that the customer places in their shopping baskets.
- K-means algorithm: The K-means algorithm works only for data sets that consist of numeric attributes. Each cluster is associated with centroid (centre point) and each point is assigned to the cluster with the closest centroid.
- Visualization: - It is a class of graphical techniques used to visualize the contents of the database. Visualization techniques are very useful tools for discovering patterns in data sets and this method is used to get rough idea to find patterns in the data set.
- Nearest Neighbour: A class of learning algorithm that uses a simple form of tabular look up to classify. It defines that in order to predict what a predication value is in one record it looks for other records with similar predictor values in nearest to the unclassified records in the data set.
- Neural Network: - Neural network is an approach of computing that involves developing mathematical structures with the ability to learn. The methods are the result of academic investigations to model nervous system learning. Neural network have the remarkable ability to derive meaning from complicated or imprecise data and can be used

to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

- Prediction:-It is used to develop a model which can derive a single aspect of the data (predicted variable) from some combination of other aspects of the data (predicted variable). It is used for detecting student behaviour, predicting and understanding student's educational outcomes etc.
- Discovering with models: - A model phenomenon developed with prediction, clustering or knowledge engineering, is used as a component for further prediction of relationships. Such as discovering of relationships between students behaviour, and students characteristics or contextual variables and analysis of research in question across wide variety of contexts.

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- no operational data, such as industry sales, forecast data, and macro economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

The patterns, associations, or relationships among all this *data* can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in the light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

The present study aims at analyzing various techniques such as k-means and decision trees in determining how different reasons affect children dropout during an academic year using in an educational institute. Decision tree analysis is a popular data mining technique that can be used to explain the interdependencies among different variables(reasons of dropout).Other helpful tools in the analysis are:PTR(People Teacher Ratio), use of ICT and Infrastructure facilities at School etc.

3. REVIEW OF LITERATURE

Data Mining

Data mining is often set in the broader context of knowledge discovery in database or KDD. The KDD process involves several stages: selecting the target, processing data, transforming them if necessary, performing data mining to extract pattern and relationship and then interpreting and assessing the discovered structures (David Hand, heikki Mannila, Padhraic smyth).

Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making (Connolly, 1999). Data mining software's allow the users to analyze data from different dimensions, categorize it and summarize the relationships identified during the mining process (Han and Kamber, 2006). Different data mining techniques are used in various fields of life such as medicine, statistical analysis, engineering, education, banking, marketing, sales, etc. (Zhao and Maclennan., 2005). Cluster analysis is used to segment a large set of data into subsets called clusters. Each cluster is a collection of data objects that are similar to one another and are placed within the same cluster but are dissimilar to objects in other clusters. (Behrouz.et.al., 2003, Dongsong, 2004).

Data Mining in Education

Data mining is an emerging methodology being used in educational system to enhance our understanding of various patterns in School education such as to find out key reasons of Out of School Children(OOSC) from the age group of 6-14 years and focus on identifying, extracting and evaluating variables responsible for the OOSC.

(Ayesha et.al 2010) conducted a study on student learning behaviour. Under this study, the data mining technique named k-means clustering is applied to analyze student's learning behaviour and to discover knowledge that comes from educational environment.

K-means clustering (Alaa el-Halees 2009) is a widely used method that is easy and quite simple to understand. Cluster analysis describes the similarity between different cases by calculating the distance. These cases are divided into different clusters on the basis of similarity. K-means is also a well known clustering algorithm (Han and kamber 2006) tends to uncover relations among variables present in the dataset. Erdogan and Timor (2005) used educational data mining to identify and enhance educational process which can improve their decision making process. Henrik (2001) concluded that clustering was effective in finding hidden relationships and associations between different categories of students.

Data Mining in Elementary Education

Education is an essential element for the betterment and progress of a country. It enables the people of a country to be civilized and well mannered. Mining in educational environment, called Educational Data Mining, concerns with developing new methods to discover knowledge from educational databases in order to analyze students trends and behaviours toward education (Alaa el Halees 2009,Erdogan and Timor 2005,Galit 2005)Lack of deep and enough knowledge in elementary educational system may prevent system management to achieve quality objectives. Data mining methodology can help bridging this knowledge gaps in education system.

From a number of field studies and literature survey so far prevalent, it can be drawn that a multiplicity of factors are responsible in one way or the other for the vast proportion of the children being not in school as well as gender disparity in access of schooling. The major problem of children neither working nor attending school encompasses the demand site constraint and supply constraint. The demand site constraints include household activities; economic reasons like poverty, domestic chores, lack of interest of children themselves and that of parental motivation. The supply side constraint is mainly related to the factors like dilapidated building and educational infrastructures that are responsible for substantial unmet demand for education.

National Family Health Survey (NHFS-2), 1998-99 gives reasons of children never attending the school or not currently attending school. The reasons are specified by this survey include school being too far away, transport not available, education not consider necessary, children required for household work farm work, cost of education being not affordable much, no proper school facilities available for girls, children and parents not interested in studies etc.

As per CTS survey 2010 in Rajasthan, some of the reasons for Out of School Children in the age group 6-14 years are summarised below:

1	Agriculture work
2	Care of siblings
3	Grazing cattle
4	Poor Economic Condition
5	Lack of Educational Facility
6	Ignorance of Guardian
7	Long Illness
8	Non Friendly Environment of School
9	Migration of family
10	Child marriage
11	Regular Absence from school
12	Homeless
13	Without adult protection

4. Justification and likely benefits

1. In order to get required benefits of such a large volume of data and to find hidden relationships between variables, different data mining techniques will be used.
2. For this we will be using clustering techniques to predict reasons for OOSC in the age group of 6 to 14 years at elementary education level in Rajasthan.
3. Partitioning of children into homogeneous groups (clustering) will be done according to their characteristics and abilities.
4. We hope that the information generated after the implementation of data mining technique may be helpful for teachers/community as well as for educational administrators.
5. This work may help in improving children enrolment, reduce dropout ratio by taking appropriate steps at right time to improve the quality of education.

5. Objectives

The objectives of the present work are:

1. Identification of the main reasons of being OOSC at elementary education level by using new data mining techniques, to be proposed by us .The

information acquired may help us to ascertain what measures can be taken to make OOSC in school.

2. Applying data mining techniques such as K-means clustering and decision tree techniques on collected data to obtain desired information.
3. To Compare and analysis the results from these techniques and suggest for necessary planning in view of dominant OOSC reasons.
4. Analysing the factual state of different reasons that cause children to dropout during an academic year.
5. It is planned to take OOSC reasons as variables in the present work to determine main reasons responsible for OOSC and the existence of interdependence among the variables, if any.

6. Plan of Work and Methodology

6.1. For accomplishment of the objectives of research the steps includes:

Phase I: Data Collection

The first phase of our research is collection of data for out of school children in elementary education in Rajasthan.

Phase II: Analysis

In the next phase, an analysis of data will be done for the purpose of making data to be useful for further phases. In this phase we will arrange the data according to objectives and categorize the data for further processing.

Phase III: Tools and Techniques

In this phase we propose to use data mining tools and techniques on the data collected and arranged.

Phase IV: Results and Discussions

In this phase results will be drawn, analyse and discussed.

Phase V: Publication of findings.

In the last, the results will be published in the relevant Journals of repute.

In elementary education system, the OOSC is determined by dropout as well as by never enrolled.

The proposed model may make prediction about the main reasons of OOS based on class dropouts, main reasons of never enrolled children etc. Identified reasons as well as system inform the OOSC social/demographic category-wise. The proposed model also deals with entry rate of students in a particular class and dropout rate with specific reasons. Model may be developed using DMX queries available in visual studio 2005.

The proposed model identifies the reasons for OOSC so that teachers/communities/Govt. can take appropriate steps at right time to improve the enrollment of student in schools. It deals with the measures to increase enrollments in order to predict students whose OOS reasons may be removed. This model may evolve the parameters, which may assist the policy makers for retention of children at different levels and take appropriate steps for minimizing

6.2 Data set

In this study, data gathered in Child Tracking Survey 2010(CTS2010) conducted by RCEE, Govt. of Rajasthan for the age group 6-14 children will be collected and analyzed using a data mining technique namely k-means clustering and decision tree technique. In order to apply these techniques following steps will be performed in sequence:

About 12 Lacks OOS Children in the age of 6-14 will be used, it is compatible and efficient to use with the database management system i.e., the relational database and the other reason is that the data is being maintained in this database.

6.3 Application software

The programming environment used for application is Visual studio 2005/higher for building data mining model. It is suitable for development of mining model and is compatible with SQL Server 2005, in which data is being maintained/ stored.

6.4 Data mining Process

Data mining process consists of following steps:

6.4.1 Preparations

In this step data stored in different tables is joined in a single table, after joining process errors are removed.

6.4.2. Data Selection and Transformation

In this step we determine the fields of study used for analysis. Data is already in the form of SQL2005 data structure.

6.4.3. Implementation of Mining Model

In this step, k-means clustering algorithm is proposed to be applied to the processed data to get valuable information. K-means is an old and most widely used clustering algorithm developed by MacQueen in 1967(Erdogan and Timor 2005). The basic K-means Algorithm may be understood in the following steps:

I. Select K points as the initial centroids
II. **repeat**
III. Form K clusters by assigning all points to the closest centroid
IV. Recompute the centroid of each cluster
V. **Repeat until** centroids don’t change

7 References, Bibliography, Webliography

[1]. Alaa el-Halees (2009)”Mining Students Data to analyze e-Learning Behaviour: ACaseStudy”.
http://uqu.edu.sa/files2/tiny_mce/plugins/filemanager/files/30/papers/f158.pdf

[2]. Behrouz.et.al., (2003) “Predicting Student Performance: An Application of Data Mining Methods With The Educational Web-Based System”. Frontiers in Education, Vol-1 2003, T2A - 13-18. FIE 2003. 33rd

[3]. Connolly T., C. Begg and A. Strachan (1999) Database Systems: A Practical Approach to Design, Implementation, and Management (3rd Ed.). Harlow: Addison-Wesley.687

[4]. Erdogan and Timor (2005) A data mining application in a student database. Journal of Aeronautic and Space Technologies July 2005 Volume 2 Number 2 (53-57)

[5]. Galit.et.al (2007) Examining online learning processes based on log files analysis: a case study. Research, Reflection and Innovations in Integrating ICT in Education.

[6]. Henrik (2001) Clustering as a Data Mining Method in a Web-based System for Thoracic Surgery: © 2001

[7]. Han,J. and Kamber, M., (2006) "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.

[8]. Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining". http://www.britannica.com/EBchecked/topic/1056150/data-mining. Retrieved 2010-12-09

[9]. ZhaoHui. Maclennan.J, (2005). Data Mining with SQL Server 2005 Wihely Publishing, Inc.

[10]. Shaeela Ayesha, Tasleem Mustafa (2010) Data Mining Model for Higher Education System.

[11]. Cesar Vialardi,Javier Bravo,Lelia Shfti(2009) Recommendation in Higher education using Data mining techniques.

[12]. Kifaya (2009) Mining students evaluation using associative classification and clustering communication of the IBIMA.

[13]. Alaa el-Halees (2009) Mining Students Data to analyze e-Learning Behaviour: A Case Study.

[14]. Erdogan and Timor (2005) A data mining application in a student database. Journal of Aeronautic and Space Technologies July 2005 Volume 2 Number 2 (53-57)

[15] David Hand, heikki Mannila, Padhraic smyth, “Principles of Data Mining”, PHI, ISBN 978-81-203-2457-2.

[16] Margret H Dunham, “Data Mining Introductory and Advanced Topics”, Pearson education ISBN 978-81-7758-785-2.