AN APPLICATION TO ASSIST PREVENTION OF STIDENT DROPOUT IN DISTANCE LEARNING USING DATA WAREHOUSE DEVELOPMENT AND DATA MININGD TECHNIQUES

Sumruay Komlayut, School of Liberal Arts, Sukhothai Thammathirat Open University, Nonthaburi, 11120, Thailand, 662-504-8637,sumruaykom152@hotmail.com
Waranya Poonnawat, School of Science and Technology, Sukhothai Thammathirat Open University, Nonthaburi, 11120, Thailand, 662-504-8277, waranya2007@gmail.com
Siwaporn Rookachart, Computer Center, Sukhothai Thammathirat Open University, Nonthaburi, 11120, Thailand, 662-504-7431, riras@hotmail.com

ABSTRACT

Student dropout rates in distance learning universities are higher than those in conventional universities, therefore reducing the dropout rate is essential in a distance learning system. The application of data mining and a data warehouse to the problem of Distance Learning Students' Dropout is presented.
Applying the clustering algorithm revealed that the dropout students can be clustered into two groups using the demographic attributes. Sequence clustering revealed the pattern of students' course registration before they dropped out. Time series equations were calculated precisely to predict the amount of dropout students in each semester and each school.

Keywords: Data Mining, Data Warehouse, Dropout, Distance Learning

INTRODUCTION

The distance learning system was viewed as an appropriate way for an open university to expand higher educational opportunities. In Thailand, Sukhothai Thammathirat Open University (STOU) was formally established by Royal Charter in 1978. It was the first open university in Southeast Asia to use the distance learning system. It subscribes to the principle of life-long learning goals to improve the quality of life of the Thai people, and strives to expand educational opportunities for both those who have completed secondary education and the general public. [1]
In each year, the average number of undergraduate students enrolling is about 14,000, but the average of dropout students is more than 40%. Non-completion affects both the university and the students. The university loses time, reputation, costs of providing the study program and material to students who could not complete the program, and costs of managing examinations all over the country. Students waste their time and money. They cannot get the opportunity to get a better job, and might have a negative attitude to the distance learning system. On a national level, it means the loss of opportunity in human resource development as well. Developing a dropout data warehouse not only extracted related data from many databases into a single point of view, but also stored data in a dimensional data format, which could use the analytical applications to slice, dice, analyze and mine the data.

METHODOLOGY

In this study, the waterfall methodology [2] was used to develop a dropout students' data warehouse. There were 5 steps in this methodology. Firstly, the feasibility study was conducted to determine the

cost, time and benefits from building the dropout students' data warehouse. Secondly, the business requirements were gathered by interviewing top management; middle management; operational management and operational staffs. These requirements were analyzed to create data modeling by identifying business objects and relationships among those objects. The data warehouse structure using galaxy schema was designed as well. Thirdly, the data flow architecture, as shown in Figure 1, and system architecture were designed by exploring the source system and the organizational infrastructure.

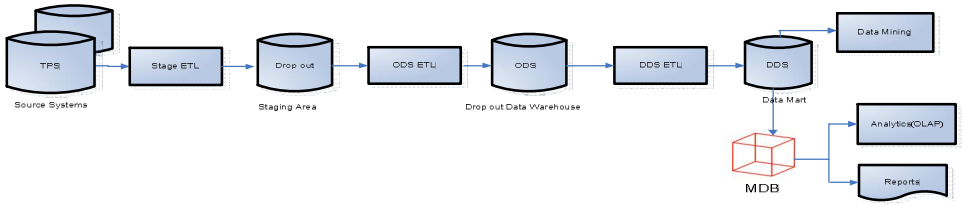

Figure 1 A data flow architecture with a stage, DW, DM and MDB

The arrows in Figure 1 show the flow of data. The data flows from the source systems to a staging area (Drop out), to a data warehouse (DW), to a data mart (DM). There were three ETL (extract, transfer and loading) packages. Stage ETL retrieved and loaded data from the source systems into a staging area, ODS (Operational Data Store) ETL retrieved and loaded data from the staging area into dropout DW. DDS (Dimensional Data Store) ETL retrieved and loaded data from DW into DM. The MDB (multi-dimensional database) was created for OLAP and reporting. Data mining techniques were used to mine the information and new knowledge from DM. The Integration Service tools of Microsoft SQL Server 2005 were used to create the ETL packages. The data cleansing process was included in these packages.
Fourthly, the dropout student data warehouse was developed using the ETL packages and the designed structure. The last step was the three front-end applications development; OLAP cube, reporting and data mining.

1) The OLAP cube was built by using Analysis Service tools of Microsoft SQL Server 2005. Users can build the cube by defining the data source from the dropout data warehouse.

2) Reports were developed using three tools; Report Designer for designing predefined reports. Data sources from the OLAP cube need to be defined when using this tool to enable users to create multidimensional reports, Report Builder can be used to design ad-hoc reports, and Excel 2007 can be used to build a report by connecting to the OLAP cube, the pivot table tools will be displayed for the users to build reports.

3) Data mining was developed using CRISP-DM (Cross Industry Standard Process-Data Mining) methodology. CRISP-DM [3] consists of 6 phases which are; Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The Microsoft SQL Server 2005 data mining tool was used in this study.

3.1) The Business Understanding phase focuses on understanding the objectives and requirements from a business perspective, then converting this information into a data mining problem definition, and finally designing a preliminary plan to achieve the objectives.

3.2) The Data Understanding phase starts with initial data collection from the source system and proceeds with activities in order to get familiar with the data, identify data quality problems, discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. (Additional interviewing of the operational staffs at the registration office was also conducted.)

3.3) The Data Preparation phase covers all activities to construct the relevant data from the data warehouse. The data views are created by selecting the attributes that relate to each model, and then sampling the data from the data mart into the data views. This information is then fed into the modeling tools.

3.4) For the Modeling phase, the three techniques (clustering, sequence clustering, and time series) are selected and applied, and their parameters are set to optimal values.
3.5) The Evaluation phase focuses on evaluating the models and reviewing the processes executed to construct the models in order to assure that business objectives were achieved. The decisions on the use of the data mining results have been made.
3.6) The Deployment phase organizes and presents the knowledge gained from the results in a way that users can use it. In this study, the results were interpreted and presented to the users and to the academic council.

RESULTS

In this section the results are summarized as follows:

Dropout data warehouse

The dropout data warehouse of undergraduate students at STOU was developed. The 545,933 records of dropout students during the academic years 1999 to 2006 were stored into the data warehouse. The data warehouse consists of two fact tables, five dimension tables and seventeen look-up tables as in Figure 2.

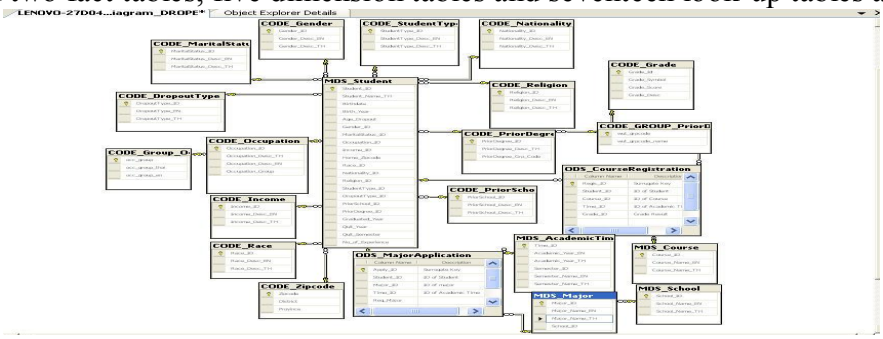

Figure 2 The structure of the dropout data warehouse

Front-end applications

The three front-end applications; OLAP cube, reporting and data mining were developed.

1) The OLAP cube was ready to drill down the amount of drop out students and the amount of registered courses with many dimensions (i.e. by quit year, by school, by gender, by occupation, by income, by dropout type, etc.). Users can use the browser tools to create the multidimensional report or to do the online analytical processing.

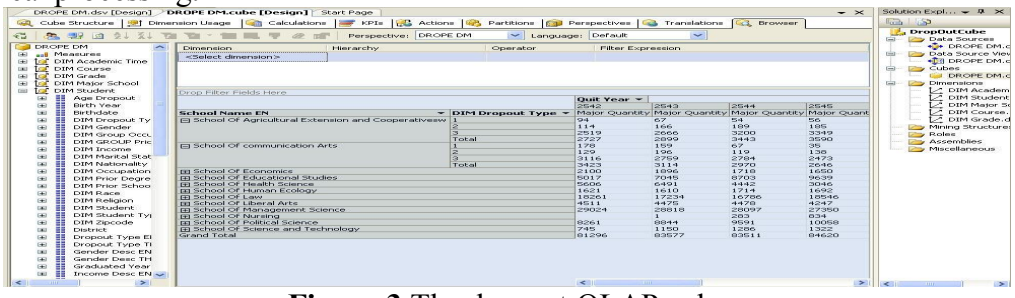

Figure 3 The dropout OLAP cube

Figure 3 is an example of the multidimensional report using the OLAP cube. It displayed the amount of dropout students with school dimension, dropout type dimension and quit year dimension. Users can drill down into each dimension and add more dimensions to the cube as required.

2) Reporting application used three tools; Reporting Builder, Reporting Designer and Excel 2007.

2.1) Reporting Builder was used to create predefined reports which corresponded to the user requirement.

2.2) But Reporting Designer is more flexible than Reporting Builder. In this study, Reporting Designer was used to create ad-hoc reports.
2.3) In order to drill down information quickly and easily, Excel 2007 was used in this study. It is another powerful tool to build reports using the cube. Users can use the PivotTable to create the multidimensional reports and professional-looking charts. Moreover, users were able to browse all of the predefined reports via web browser from the STOU intranet and can drill down for more details or drill up for summary. Users were able to print reports in a variety of file formats including XML, TIFF, PDF or Excel.
3) Three data mining models were applied to mine the information and to discover new knowledge from the data mart.

3.1) Clustering Model

The clustering model used iterative techniques to group the 64,377 records in VClusteringStudent view into twelve clusters that contained similar characteristics. The cluster diagram shows links among these clusters in Figure 4.

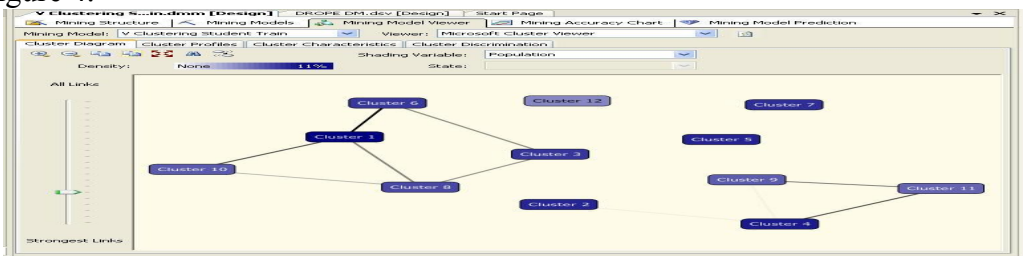

Figure 4 Cluster diagram showing strong (dark) links between clusters

Figure 4 shows the twelve clusters are similar to each other; the darker the line between two clusters, the more similar those clusters are. The color of the cluster box indicates the population size; the darker the cluster box, the more dropout students in that cluster. This study used the links to group the clusters. There are two main groups of clusters which are linked together. Group one consists of Cluster 1, Cluster 3, Cluster 6, Cluster 8 and Cluster 10, and group two consists of Cluster 4, Cluster 9 and Cluster 11. The cluster profiles can be displayed using the cluster profiles tab as in Figure 5.

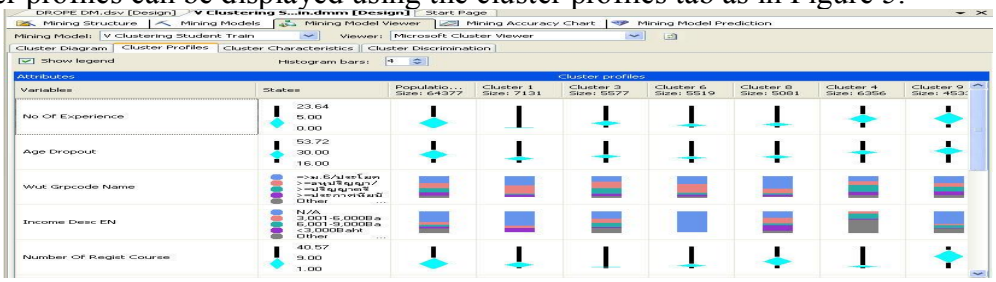

Figure 5 The profile of each cluster

The cluster profile was used to find out the characteristics of the dropout students. Cluster 1 and Cluster 3, which have a strong link, identify that the characteristics of students in Cluster 1 and Cluster 3 are similar. Most of them are young females (16-24 years old) with 0-5 years of work experience, low income (less than 3,000 baht) and studying in the School of Management Science. The number of registered courses was 3-9 courses before dropout. The characteristics of students in another group, Cluster 4 and Cluster 11 are dissimilar to the characteristics of students in Cluster 1 and Cluster 3. Most of the students in Cluster 4 and Cluster 11 are middle aged males (36-53 years old) with10-23 years of work experience, moderate to high income (more than 12,000 baht) and studying in the School of Law. They registered for 3-9 courses before dropout as well.

3.2) Sequence clustering model

The sequence clustering model was a hybrid of sequence and clustering techniques [4]. It was based on the Markov chain theory. The data view, VSeqClustStudentRegist contained 51,864 records of the course registration data of dropout students which were processed using this model. The sequence clustering model grouped the data into 8 clusters. Similar clusters (clusters with similar probability distributions such as clusters 1 and 6) are closer to each other as in Figure 6.

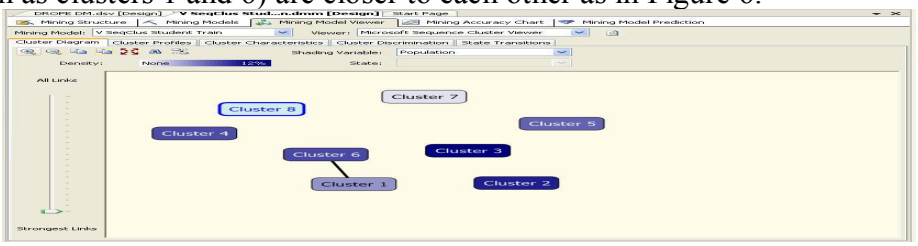

Figure 6 Cluster diagram showing a strong link between Clusters 1 and 6

In each cluster, the patterns of course registration can be navigated using the cluster transition pane. Figure 7 displays the sequence of course registration of students in Cluster 1 before dropping out. Each node is a sequence state, and each edge is the transition between these two states. Each edge has a direction and weight. The weight is the transition probability. From the figure, there is a strong link of transition from Course 20001 toward Course 20101 registration. Among those dropout students who registered for Course 20001, 85% also registered for Course 20101. About 66% of dropout students recurrently register for Course 22222. The results indicated that the content of these courses might be considered for revision. The schools responsible for these courses should find ways to help students before they dropout.

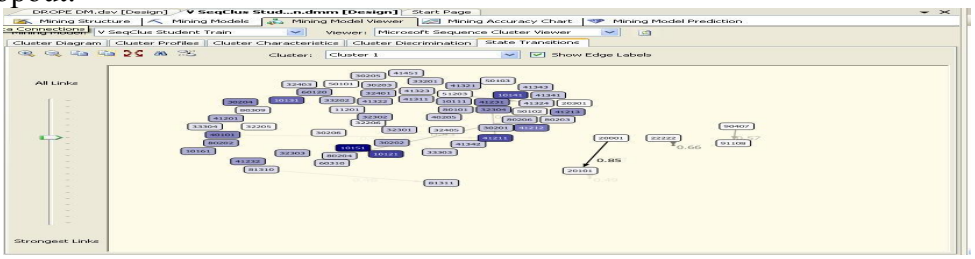

Figure 7 Cluster transitions of the registered courses of Cluster 1

3.3) Time series model

The time series model used in this study was a Microsoft Time Series, which is a hybrid of auto regression and decision tree techniques. The amount of dropout students in each school can be forecast by viewing the chart tab as in Figure 8.

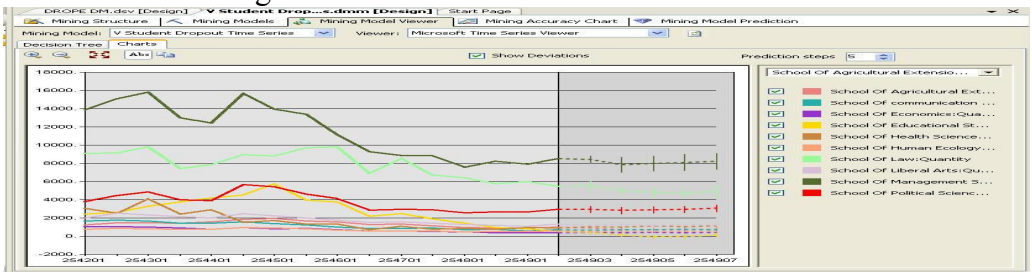

Figure 8 The forecasting values of the dropout student amount of each school in each semester from 1999 to 2006.

Figure 9 is divided by a vertical line. The left side of the line represents the historical series values, and the right side of the line represents the future forecast. The forecasted values are indicated using a dotted line. The amount of dropout students in the School of Management Science has the highest values. This information is useful to the schools for planning prevention of dropout students and can be used as the KPI to measure the success of the plan.

DISCUSSION

The clustering technique grouped students into clusters by using their characteristics. These findings help STOU know the target groups. Most of the students who are young, female, and have a little bit of experience with low income may illustrate the essential factors that affect their dropout; they lack experience and discipline to study in a distance learning system. Also, the students who are middle aged, male, and have more than 10 years experience, and also study in School of Law may not have a genuine reason for getting the qualification. The Sequence clustering technique revealed the course registration pattern of dropout students. The number of courses registered was 3-9 courses, which since STOU allow students to register for 1-3 courses per semester, implied that the dropout students will stay in the system only 1-2 semesters (each course is 6 credits). If STOU could assist these students to pass 2 semesters, they may remain in the system.

CONCLUSION

The data warehouse of dropout students was of benefit to STOU as a source of information. The proposed methods have been applied to a data warehouse concerning the dropout problem of STOU. The OLAP and reporting applications can provide detail and summary in multidimensional reports to management via intranet. Data mining models revealed the characteristics of the dropout students. The hidden pattern of course registration was discovered. Such patterns can be described through Markov chains as a series of transitions characterized by course. Data mining is the process of discovering new knowledge. It can help users to uncover patterns of behavior, but it may not identify the causes of business problems. Humans need to identify the causes [5]. Application of data mining techniques can avail administrators the opportunity to advise revision of course materials, and course tutors to advise students at appropriate times. Therefore this new knowledge can be used in order to identify policies, planning, and arranging activities aimed at lowering the dropout level.

FURTHER STUDY

The results were based on the quantitative data in the data warehouse. Additional interviewing of the dropout students should be conducted to collect the qualitative data about the reasons for dropout, and the assistance needed from the university. These data could be analyzed to find factors that affect the dropout by using other data mining techniques.

REFERENCES

[1] Thammathirat, S. (2008), "Mission and Objectives", retrieved on September 6, 2009 from the WorldwideWeb:http://www.stou.ac.th/eng/AboutSTOU/mission.asp http://www.stou.ac.th/eng/AboutSTOU/mission.asp

[2] Rainardi, V. (2008). "Building a Data Warehouse: With Examples in SQL Server", Springer-Verlag, New York Inc.

[3] Olson, D. and Delen, D. (2008), "Advanced Data Mining Techniques", Springer-Verlag Berlin Heidelberg.

[4] Tang, Z. and MacLennan , Jamie (2005), "Data Mining with SQL Server 2005", The United States of America, Indiana; Wiley Publishing Inc.

[5] Larose, D. (2004), "Discovering Knowledge in Data: An Introduction to Data Mining", Wiley Publishing, Inc.