THE ENHANCEMENT OF ODL STUDENT RECRUITING CAMPAIGN WITH DATA WAREHOUSE DEVELOPMENT AND DATA MINING TECHNIQUES

Waranya Poonnawat (waranya.poonnawat@gmail.com)
Sumruay Komlayut (sumruaykom152@hotmail.com)
Nuttaporn Henchareonlert (stashnut@stou.ac.th)
Sukhothai Thammathirat Open University, Thailand

ABSTRACT

The purpose of this research was to develop an OLAP cube data warehouse, and, using data mining techniques, to support the university's public relations, admissions, and planning divisions in the efficient recruiting of students by surveying, through interviews; the opinions of management and operational personnel, and through documents; the attributes in application forms and annual reports. User requirements, source data and systems were all examined. The data warehouse and front-end applications developed are described below. 1. Student Data Warehouse—this repository was designed to store students' historical data and to facilitate analysis and reporting following the user requirements. Students' historical data including demographic data from 2001-2005 were extracted, loaded and transformed from source systems, then they were cleaned before uploading to the data warehouse using star schema. 2. OLAP Cub—this 122 multidimensional structure enables users to analyze the students' demographic data in many dimensions such as "Number of Registered Students in each year by Semester, Major, School, Gender, Occupation, Region, etc." Predefined reports were created and published to an intranet and users were able to create ad-hoc reports through web browsers as well as XLAddin. 3. Data Mining—this technique finds hidden knowledge and patterns in ODL student data supporting decision making, using three algorithms: Naïve Bayes, Clustering and Association Rules. Occupation of students is the strongest factor influencing students' choices of Schools. Students' demographic data can be clustered into groups with similar or dissimilar characteristics such as "Single, Unemployed, Low Income (<3,000 Baht)" or "Married, Male, Studying Law, High Income", and can generate rules from frequently occurring cases such as "Occupation=Teacher-Lecturer (private sector), Marital Status=Single > School=School of Educational Studies" or "Occupation=Police, Marital Status=Single -> School=School of Law". The results from the study indicated that users were satisfied using information and applications from the data warehouse, OLAP cube and data mining techniques which enable the university to reduce costs and to reach the desired enrolment target effectively.

Keywords: data warehouse, data mining introduction, ODL Student Recruiting Campaign.

Each year Sukhothai Thammathirat Open University (STOU) funds a public relations campaign to recruit new students with traditional mass marketing which enabled its public relations division to purchase more channels and cover the mass market. The budget tended

to increase every year. The campaigns focused only on the mass market rather than on specific target groups for each school.

To overcome this problem, a data warehouse was developed to integrate the historical demographic data of students. New information was created by exploring hidden linkages within the data using OLAP and data mining. This enabled a new, more efficient and effective targeted recruiting campaign to be developed.

METHODS

Two types of data were gathered: 1) the information needs of users from interviews and 2) the record of new students who enrolled in the first semester during academic year 2001 to 2005 from the Office of Registration. Later, the information model was made and then Microsoft SQL Server 2005 was used to build a data warehouse for creating OLAP (online analytical processing) cubes, creating a data mining model and creating reports.

Three algorithms of data mining techniques were used in this research: 1) Naïve Bayes which was used for classifying students' characteristics depending on some variables as the input and forecasting or predicting some variables as the result, 2) Clustering which is used for classifying students' characteristics based on a set of variables which was similar or dissimilar, and 3) Association Rules which is used for analyzing the frequency of the recurring set of variables or data sets to make visible relationships within these variable sets and creation of rule sets variables in the relationship.

Three methods were used to create reports: 1) Report Design which created predefined or routine reports through SQL Server Business Intelligence Development Studio (BIDS), 2) Report Builder which created ad-hoc reports through Internet Explorer, and 3) XLAddin which created ad-hoc reports through Microsoft Excel 2003.

The prototype was presented to users and was evaluated by users through questionnaires.

RESULT

Data Warehouse Development

Firstly, demographic data of students obtained from users' interviews, including gender, age, major and school that students applied to study, hometown, prior certificate or education, media or channels for receiving the application news, reasons to study, and the expectation from studying at STOU were collected into the data warehouse.

Next, the information model was prepared and additional demographic data of students from application records including type of people who applied to study (i.e. normal, monk, disabled person, etc.), religion, marital status, hardware or electronic devices used for learning, income per month, money to support their study, home postal code, occupation, and so forth, was collected. Data from the Office of Registration was in .mdb file format from Microsoft Access, which was stored separately in a table for each academic year. The data were merged together in one table and one column was added for storing the academic year.

After the initial data were reviewed, the cleansing process was started from extracting the data from source systems, transforming some data; by considering correctness, validity of

data, and data formatting in order to make it suitable to use, and then loading it into the data warehouse schema following the information model. Then the relationships between tables were created.

Front-end Applications

The front-end applications were developed and divided into three groups: OLAP cube, reports, and data mining algorithm.

The OLAP cube was created firstly by defining data sources from data warehouse schema, data source view was created by choosing the tables, then defining the cube model by determining dimensions and measure. After the deployment the OLAP cube was ready to view the measure (number of new students) with many dimensions (i.e. by year, by semester, by gender, by region, by occupation, by income, etc.).

Reports were created using three methods: Report Design, Report Builder, and XLAddin.

Report Design had several tools (i.e. label, graph, data set definition, setting parameters, etc.) for designing predefined reports (Figure 1). Data sources defined from OLAP cube enabled the creation of multidimensional reports. Users were able to browse these reports through a web browser from the STOU intranet and to view the measures which users could drill down for more details or drill up for summaries. Users were able to print reports in variety of file formats including XML, TIFF, PDF or Excel.

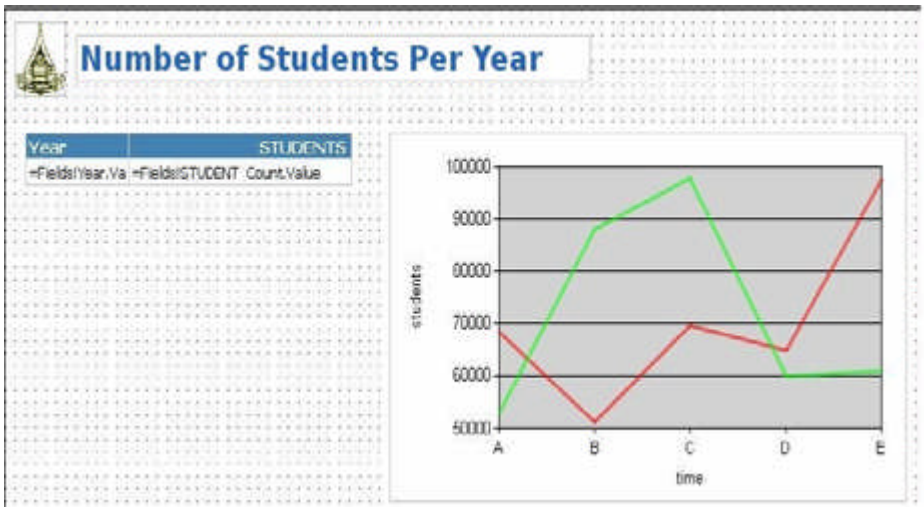

Figure 1. Creating reports with Report Design

Report Builder had fewer tools (i.e. labels, few report layouts, filters, etc.) compared to Report Design for designing ad-hoc reports (Figure 2). Data sources defined from the OLAP cube and data model needed to be created for the connection. Users were able to create multidimensional reports as well as Report Design but focused on the exception reports or specified reports. Users were able to browse these reports through a web browser on the STOU intranet and to view the measures which users could drill down through for more details or drill up for summaries. Users were able to print reports in variety of file formats including XML, TIFF, PDF or Excel.

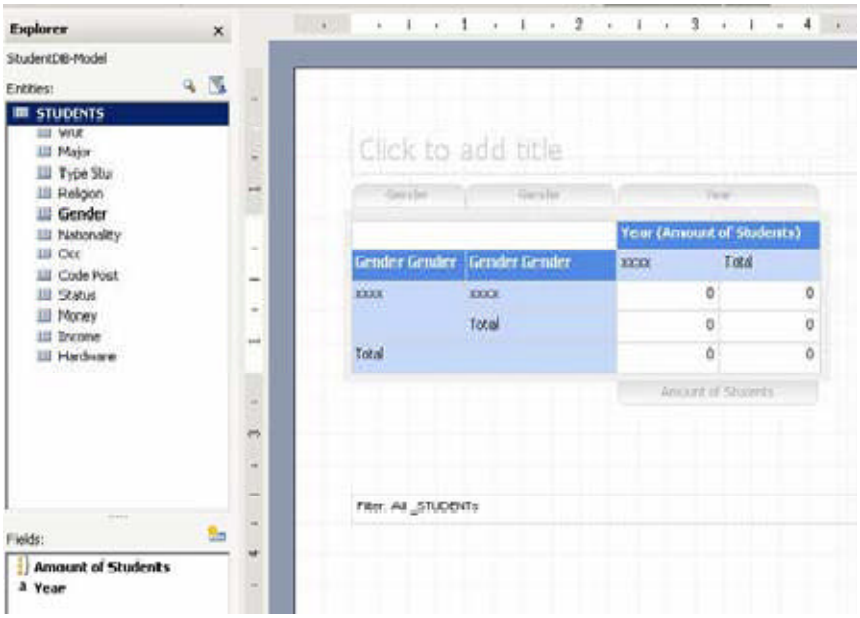

Figure 2. Creating reports with Report Builder

XLAddin has a number of tools (i.e. formula, graphic, pivot table, etc.) for designing both predefined reports and ad-hoc reports (Figure 3). The data sources defined from OLAP cube and the connection needed to be created first. Users were able to create reports with Excel's features with which most were familiar. Users were able to deploy these reports to STOU intranet, and browse them through a web browser.

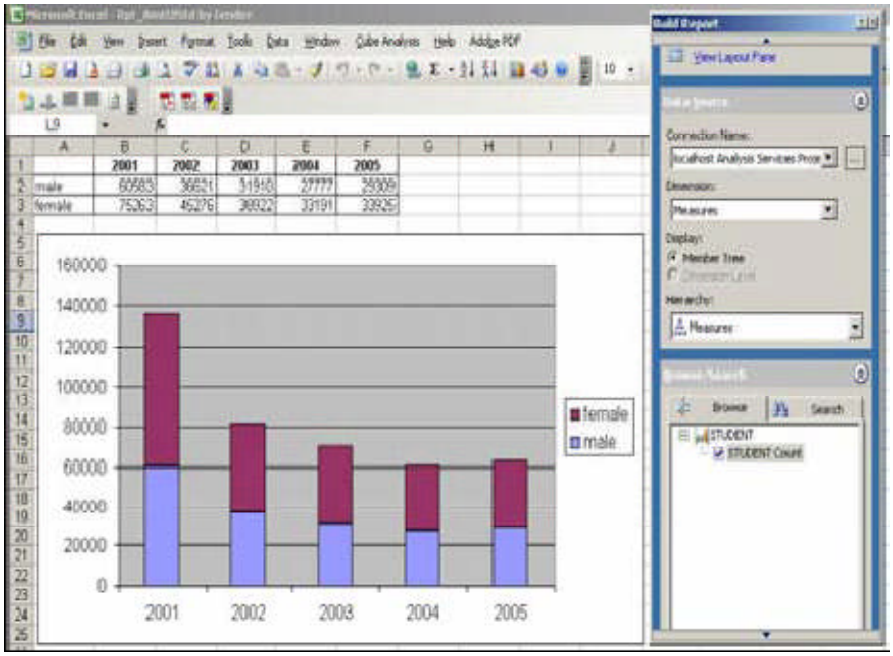

Figure 3 Creating reports with XLAddin

Reports which were generated from three methods can be displayed on web site through the STOU intranet (Figure 4).

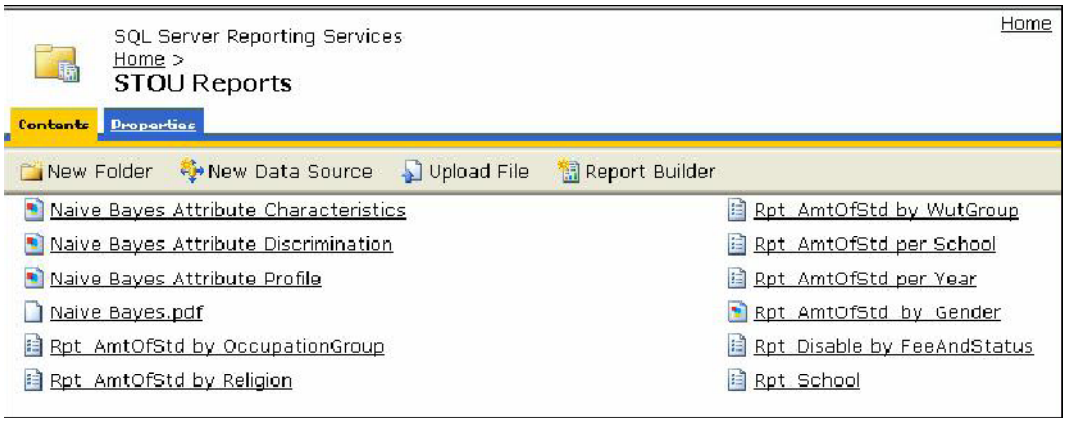

Figure 4. Reports on STOU's Intranet

Data Mining Models were created by using three algorithms: Naïve Bayes, Clustering, and Association Rules.

Naïve Bayes algorithm was used to find the relationship between the forecast variable field (School) and other variables which effected students' school choices. Seven variables were used to forecast (Figure 5). The most effective variable was Occupation. For instance, students in the field of education, most teachers were teachers in private school (64%).

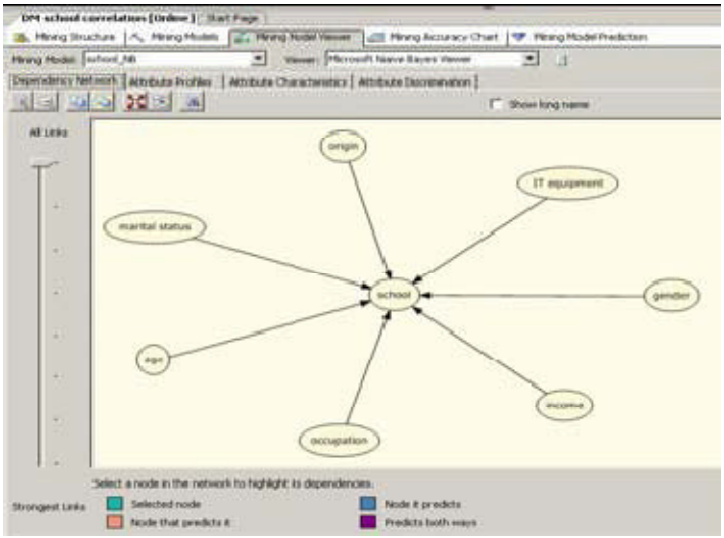

Figure 5. Dependency Network's algorithm Naïve Bayes

The Clustering algorithm was used to cluster or group similar or dissimilar data about Students (Figure 6). By default setting, the data was organized in 10 clusters/groups. The more similar of two clusters/groups were shown by a dark line and the least similar of two clusters/groups were shown by a light gray line. For instance, two interesting clusters/groups were Cluster 2 and Cluster 7. The similar characteristics were most students from both clusters learned in the School of Management Science, lived in Bangkok, and were female. The different characteristics were most students in Cluster 2 were unmarried, earned 6,001-9,000 baht per month, were between 24 and 31 years old, but most students in Cluster

7 were married, earned more than 18,000 baht per month, and were between 38 and 44 years old.

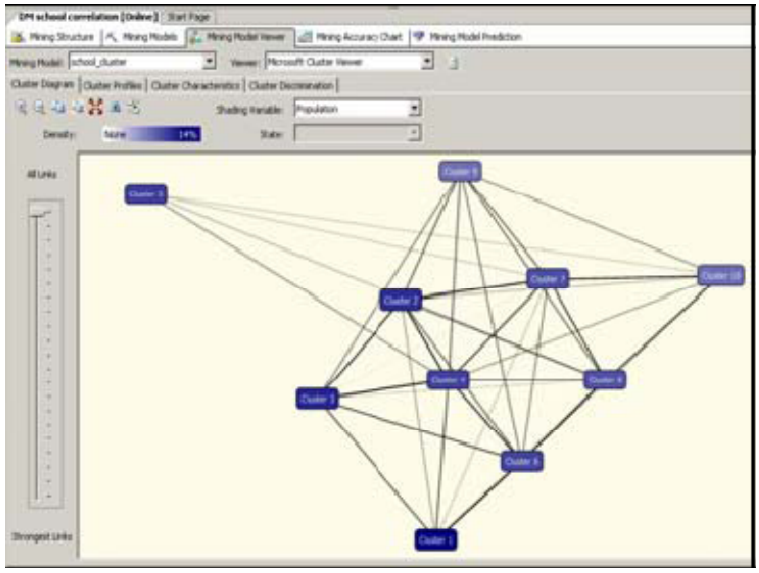

Figure 6 Cluster Diagram of Clustering Algorithm

The Association Rules algorithm was used to analyze the probability of the occurrence of repeated events. The data was analyzed to find duplicate or frequent relationships between variables, subsequently the rules were created and the important values and the probability of the relationship were calculated for each rule (Figure 7). For instance, 1) the rule about most students who were teachers in private schools, they usually applied to the School of Education; the important value was 1.304 and probability was 0.928 and 2) the rule about most students who were teachers in privates school with marital status as single, they usually applied to the School of Education; the important value was 1.180 and probability was 0.938.

DISCUSSION

1. Most executives wanted to know basic information about students' characteristics but some were unable to clearly identify their needs. In addition, numerous errors were found, such as incomplete and incorrect data in the records, some foreign keys were not found in the other table, as primary keys, and table structures were different from year to year, etc.
2. The evaluation result of the benefits of using front-end applications of the data warehouse was agreed. Most users would like to use the application to make decisions for designing a recruiting campaign for their students.
3. The data mining techniques allowed users to identify relationships and find patterns of data, in addition to exploring new knowledge which had never available before. According to Marakas (2003) a data warehouse is a strategic information repository which collects historical data and focuses on making available information needed by users. Data mining searches patterns and relationships of new hidden information. Information in the data warehouse analyzed by data mining was able to answer questions that users had never thought of before. And according to Hari Mailvaganam (2005) data

mining techniques could find patterns and useful information which was in the data. Moreover, they could be used to analyze or predict and to classify.

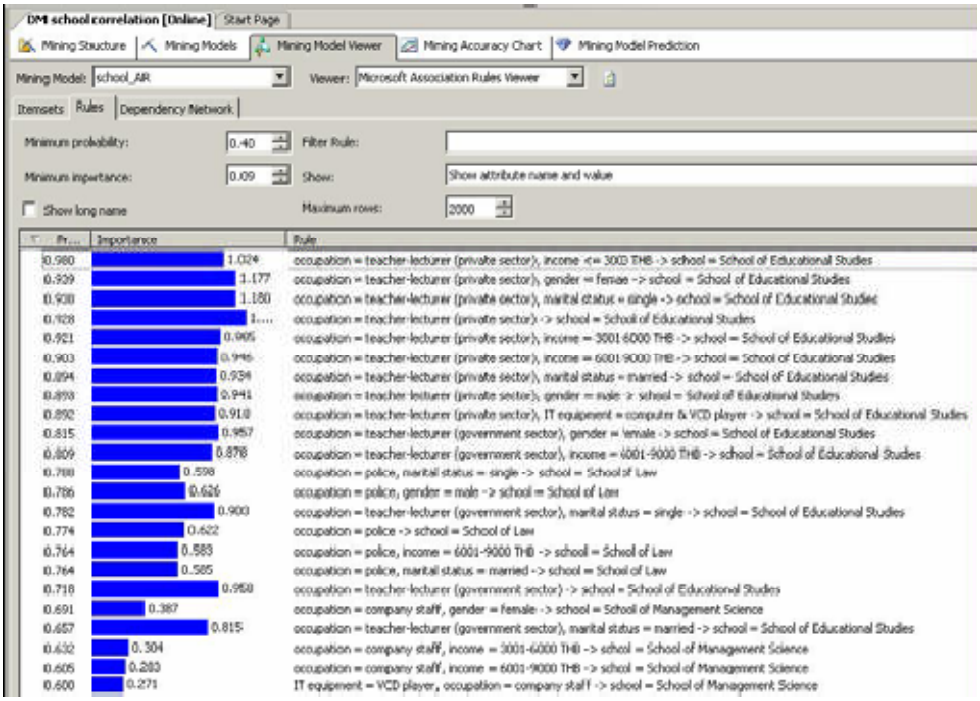

Figure 7 Rules of algorithms Association Rules

BIBLIOGRAPHY

Chitchanok Songsiri et al. (2002). Using data mining technique to search the most suitable major for students. *The 39th Academic Conference Proceedings*. Faculty of Engineering, Kasetsart University.

Berson et al. Customer acquisition and data mining". Retrieved on August 10, 2005 from the World Wide Web: *http://www.thearling.com.*

Chapman, Pete et al. CRISP-DM 1.0: Step-by-step data mining guide. Retrieved on April 16, 2006 from the World Wide Web: *http://www.crisp-dm.org.*

Chenoweth et al. Seven key interventions for data warehouse success. *Communication of the ACM, January 2006/Vol.49.*

Hari Mailvaganam. *Data warehousing overview: From metadata to data analysis.* Retrieved on August 6, 2005.

Marakas M., George. (2003). *Modern data warehousing, mining, and visualization: Core concepts*. New Jersey: Prentice Hall.

Tang, Zhao Hui, & MacLennan, Jamie. (2005). *Data mining with SQL Server 2005.* The United State of America, Indiana: Wiley Publishing Inc.

Kwak, Dukhoon. (2008). *A study on evaluation regional campuses of KNOU, Seoul.*