IMPROVE STUDENT RETENTION IN HIGHER EDUCATION TO USE DATA MINING

NAMITA AWASTHI[1], VIMAL KUMAR AWASTHI[2]

[1]P.G, Student, Department of Computer Science & Engineering, KIT Kanpur, AKTU, U.P India
[2]Assistant Prof., Department of Computer Science & Engineering, KIT Kanpur, AKTU, U.P India

***

Abstract: - Data mining combines machine learning, statistics, and visualization techniques to discover and extract insights. One of the biggest challenges facing higher education is improving student retention (National Audition Office, 2007). Our project uses data mining and natural language processing technologies to monitor students, analyze their academic behavior, and provide a foundation for effective intervention strategies. Our goal is to identify potential problems as early as possible and follow up on intervention options to improve student retention. In this article, we discuss how data mining can help identify 'at risk' students, assess course or module suitability, and tailor interventions to increase student retention.

Key Words: Data Mining, Higher Education, Student Retention, Student Intervention.

INTRODUCTION

As the cost of storage and processing power declines, data storage has become easier and cheaper. Universities are faced with the immense and rapid growth of the volume of educational data (Schönbrunn and Hilbert, 2006). Data mining, sometimes also known as Database Knowledge Discovery (KDD), can find relationships and patterns that exist but are hidden among the vast amount of educational data. It combines machine learning, statistics, and visualization techniques to discover and extract insights in ways that humans can easily understand. For universities, the knowledge discovered by data mining techniques would provide a personalized education that meets the demands of students and employers.

To provide meaningful analysis, data mining techniques can be applied to provide additional insights beyond explicitly stored data. Compared to traditional analytical studies, data mining is prospective and targeted at individual students. For example, the aggregation aspect of data mining can offer a comprehensive analysis of student characteristics, while the predictive function of data mining can help the university act before a student. drop out or plan resources based on knowledge of number of transfer students. or take a private lesson.
Student retention is an indicator of the university's academic performance and enrollment management. Poor student retention could negatively affect the university and lead to serious financial hardship. In this article, we use our project as a case study to discuss how to apply data mining to improve student retention. The rest of the paper is structured as follows. Section 2 provides background related to the project. An overview of student retention can be found in section 3. Section 4 discusses the data source and project methodology. The results of the experiment are discussed in section 5. Finally, section 6 summarizes this article.

BACKGROUND AND RELATED WORK

Data mining can be applied to several different applications, such as data synthesis, training, classification rules, association search, change analysis, and anomaly detection (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 the text" (Spasic et al., 2005). Data mining is widely applied in many fields, such as retail, financial, communications, and marketing organizations.

For universities, data mining techniques could help provide a more personalized education, maximize the efficiency of the education system, and reduce the cost of educational processes. You can guide us to increase student retention rate, increase education improvement rate, and increase student learning outcomes.

Gabrison uses the data mining prediction technique to identify the most effective factor in determining a student's test score and then adjust those factors to improve the student's test score performance the following year (Gabrison, 2003). Luan uses data mining to group students and determine which students can easily accumulate their courses and which take longer courses (Luan, 2002). These clusters help universities to identify the needs of each group and make better decisions about how to offer courses and programs, how much time is required for teaching, etc. In (Minaei-Bidgoli et al., 2004), the authors use a data mining classification

technique to predict the final grades of students based on their functionality of using the web. This can identify students at risk early and allow the tutor to provide appropriate advice in a timely manner.

To understand the factors that influence college student retention, questionnaires are often used to collect data, including student personal history, student behavior implications, student perceptions, for example in (Superby et al., 2006) the authors applied different approaches such as the forest decision tree, neural networks and linear discriminant analysis to their questionnaires. However, perhaps due to the small sample size, the accuracy of the prediction is not very good. Herzog (2006) collected data from the institution's student information system, the American College Test student profile, the National Student Clearinghouse, and the SPSS software chosen to estimate student retention and time to graduation. Nearly 50 characteristics including demographics, campus experience, college experience, and financial aid are applied to predict student retention. Research shows that the decision tree and neural networks work better when larger data sets are available.

The MCMS (Mining Course Management Systems) project at Thames Valley University (TVU) proposes to build a knowledge management system based on data mining. The different data sources of today's university systems (such as the library system, student administration, e-learning) are integrated as a data warehouse based on designed data models. Data mining technologies are applied to predict individual student performance, as well as the relevance of the course or module. Meanwhile, text mining and natural language processing (NLP) technologies are used to generate user-friendly results for better understanding (Oussena, 2008).

In MCMS, model-based data integration is applied to extract data from multiple systems into a single data warehouse for reporting and analysis (Kim, et al., 2009). As data in the data warehouse is cleansed, pre-processed, and transformed, it can greatly improve the effectiveness and efficiency of data mining processes. The knowledge uncovered by data mining techniques will allow the university to take a more advanced approach to teaching students, predicting individual student behaviors and course performance. Language Processing and Text Mining Technology

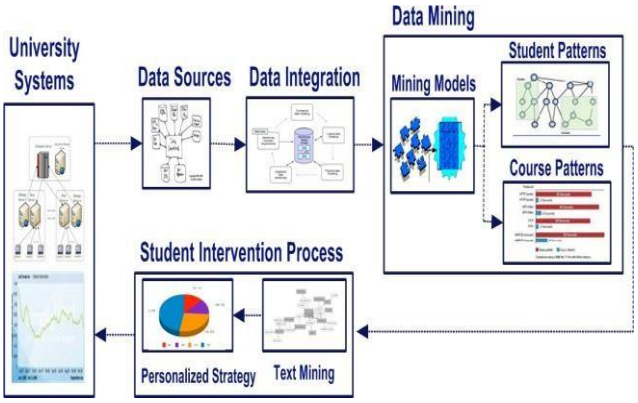

Figure 1: General Process of MCMS.

STUDENT RETENTION

One of the biggest challenges facing higher education is improving student retention. In general, more students staying in college means better academic programs and higher earnings. A report from the UK Parliament's Select Committee on Public Accounts (Public Accounts Committee, 2001) showed that while university participation is around 43 per cent, around 28,000 full-time students and 87,000 part-time students They have started undergraduate courses in 2004-2005. no longer in higher education a year later. 91.6 percent of full-time students were entering their second year, while only 78.1 percent were required to complete it (Public Accounts Committee, 2001). The Higher Education Funding Council for England (HEFCE), the body that distributes public funds to universities in England, ties its annual scholarships to the number of students who remain at the university and do not necessarily pass exams each year. The sum at stake is around £ 2,500 per full-time student per year. The loss of government grants to UK institutions due to student dropouts is around £ 105 million each year (Yorke et al., 2004).

Tuition revenue is also closely related to student retention. For a medium-sized university that receives around 2000 new students each year. If 5% of freshmen drop out of school, the missed fees will increase, if there are international students among them, the missed fees will be much higher. Additionally, dropped out students have an initial recruitment cost and new students must be recruited to keep the number of college students stable.

The most accepted model in the literature on student retention is that of Tinto (Tinto, 1995). Examines the factors that contribute to a student's decision to pursue higher education. He states that the decision to persevere or quit smoking is strongly predicted by his degree of school integration and

social integration. Tinto argues that academically, performance, personal development, academic self-esteem, enjoyment of subjects, identification with academic standards, and one's role as a student contribute to a

general feeling of integration of the student in the university (Tinto, 1995). Students who are highly academically integrated are more likely to persevere and graduate. It is the same from a social point of view. Students who have more friends at their university, have more personal contact with academics, enjoy being in university, are likely to make the decision to persist. Poor retention is typically caused by unclear career goals, course uncertainty, lack of academic challenge, transition or adjustment issues, limited or unrealistic expectations, lack of commitment, and poor performance. According to Tinto, students are more likely to stay in the course if there are connections between their own academic goals and the academic and social characteristics of the university. If students find that the particular course can combine their chosen subject and education, and greatly help them achieve their goals, their chances of completing it would be greatly increased.

There are also other models of student retention. For example, Thomas develops his model of "institutional habitus" (Thomas, 2002) based on Tinto's theory, which can be divided by academic and social experience. Academic experience covers staff attitudes, teaching and learning, and assessment. Different learning styles are supported and diversity of backgrounds is appreciated. The tutors are friendly, helpful, and approachable. The assessment gives students the opportunity to be successful and the staff is available to help. The social experience is all about friendship, mutual support, and social media. Thomas noted that a factor in his students' persistence was making them feel more comfortable with their friends.

Seidman developed a formula of student retention (Seidman, 1996) in which:

Retention = Early Identification + (Early +Intensive + Continuous) Intervention

Sideman formula and shows that early identification of at-risk students, as well as maintaining intensive ongoing intervention, is key to increasing student retention. He also explains how universities can prepare their programs and courses so that students have the greatest chance of success, both personally and academically. It is important to collect familiar information from students, as this information could help to better understand each student. He believes that we could make a difference in helping students achieve their academic goals and their institutions by increasing their retention rates.

For MCMS, we collected as much data as possible from different data sources to cover Tinto's model, including academic integration and the social integration aspect of students, which are discussed in the next section. Seidman's formula can also guide us through the MCMS implementation process. Early identification and early intensive intervention can make the difference in whether a student leaves the institution prematurely or not.

DATA SOURCE AND METHODOLOGY

Thames Valley University (TVU) systems (Oizilbash, 2008) contain a large amount of data that can be analyzed and extracted for the data mining system.

Faculties and departments also have important detailed data about courses and modules that are in document form. This section will cover each data source and how they are used in MCMS.

- The student records system contains information about student records, such as student history, test results, and course enrollment. It is the most important data source for our project.
- The e-learning system allows students to access course material for a particular course module and tutors to extend their teaching in the classroom using more interactive techniques. This system can help us monitor the degree of academic integration of students.
- The library system provides informational data that could be used for the academic integration of students.
- The playlist system is hosted on the library system server, but has a separate database. This can help us keep track of how often students borrow books from the recommendation list.
- The online resource system can be used to identify whether a student is a regular user of the system.
- The program specification is a document that provides information about the course. Text mining can be applied to extract the course title, the learning and teaching method, etc.
- The Module Study Guide is a source of textual data that provides information about the module. Contains the details of a module,

including the student assessment strategy, learning outcomes, and reading lists.

- The course marketing system is developed for marketing purposes, which is used to choose a course to advertise and search for a new course. We can get more information about the courses in this system.
- Online testing system allows everyone to take an entry skills check online. It can help us understand the academic background of students.

The MCMS project aims to build a data mining system based on the integration of these TVU systems, which covers the academic perspective, the performance aspect of Tinto's model. For example, e-learning systems can capture data on the learning behavior of students and their interaction with the system. The system also includes wiki and group discussion tools that reveal the interaction of students with their peers and tutors, although these data are rarely used (Oussena, 2008). However, we have difficulty collecting other data for Tinto's model, such as the student's personal development, enjoyment of the subject, and other social perspectives. We believe that the availability of this data will greatly aid our research in the future. Figure 2 shows the architecture of the MCMS system. Data sources cover student enrollments, student outcomes, course / module data, student learning skills and activities, etc. The data sources are then integrated and transformed into a data warehouse. The data warehouse then generates the appropriate data for the mining engine. As Oracle is the most widely used database in our data sources, Oracle 11g is chosen as the project platform, which also integrates with Oracle Data Warehouse Builder and Oracle Data Miner. At this stage, our experiments are primarily based on Oracle 11g. it is chosen as the project platform, which is also integrated with Oracle Data Warehouse Builder and Oracle Data Miner. At this stage, our experiments are primarily based on Oracle 11g. Thames Valley University (TVU) systems (Oizilbash, 2008) contain a large amount of data that can be analyzed and extracted for the data mining system.

Faculties and departments also have important detailed course and module data in document form. This section will cover each data source and how it is used in MCMS.

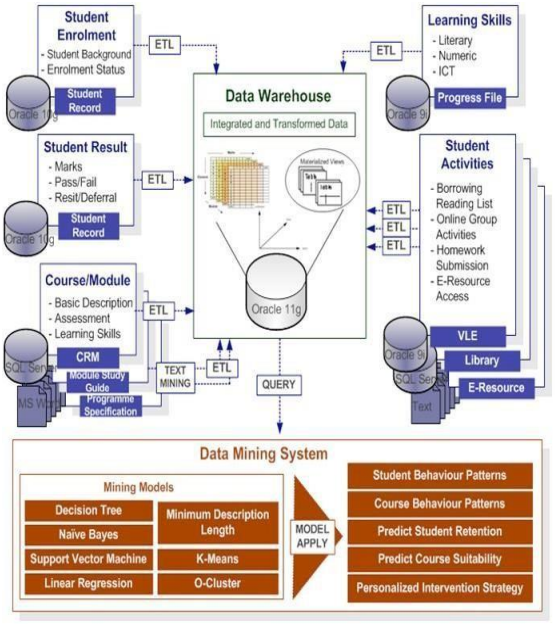

Figure 2: System Architecture of MCMS.

However, in the future, we may combine them with other data mining solutions such as weka (Witten et al., 2005) or develop our own data mining models. In MCMS, a model-driven data integration (MDDI) approach is applied in data integration for our data warehouse. MDDI is a data integration approach that proactively incorporates and uses metadata throughout the data integration process. By coupling data and metadata, MDDI dramatically reduces complexity and provides context-sensitive data integration. Different modeling approaches have been proposed to overcome each difficulty in designing the development of different parts of a data storage system (Mazon et al., 2005 and Fabro et al., 2008).

Once we have the data to identify the characteristics of students who have failed in previous semesters and years, data mining can find the profile of students who have failed. For example, the rules for selecting and associating characteristics can help us find the main characteristics that may be linked to school dropout. Classification and grouping can identify potential "at risk" students. Text mining and NLP will be used to implement our intervention strategy. This early and intensive intervention can be continually measured to see if it has made a difference in student retention rates. Data mining can also be applied to course data. For example, we could find the modules that are important for a specific course, since they can cause the dropout of more students, this will help the university to evaluate the suitability of the module, prepare programs and courses so that students have the highest probability of success, both personally and academically.

EXPERIMENTS

We intend to collect three-year historical data from university systems, but we currently only have one-year data. For the course marketing system, we have 5,458 records, which include 1,881 courses; 5,352 course offerings; 7 schools and 7 faculties. For the student registration system, there are 4,22.3 students, 5,352 course enrollments. For the library system there are 144,604 borrowers, 3,150,816 loans, 630,190 articles, 435,113 books, and 45,900 classifications. For the playlist system, there are 552 courses, 1540 lists, and 7084 list entries. For e-learning systems, there are 2,460 module offerings and 2,021,334 online activities.

For each data source, UML models have been developed. A mapping model has also been developed that describes how to integrate data from multiple sources. The Data Warehouse ETL (Extract, Transform, Load) process is designed in the Oracle Data Warehouse Builder, and finally the results are entered into the Oracle Data Miner. The data mining process is shown by Figure 3.

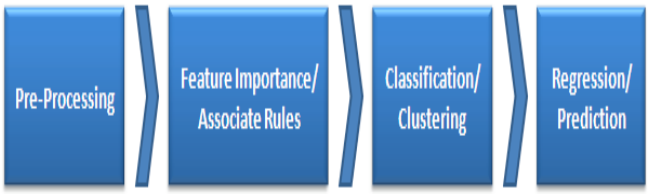

Figure 3: Data Mining Process of MCMS.

To increase student retention, we need to understand why students drop out. Some experiments are used to evaluate Tinto's model. Our student data includes: average grade

(AVGMARK), online learning systems information (BB_USAGE), library information (LIBRARY_USAGE),nationality (UK), university entry certificate(ENTRYCERTIFICATE), course award (COURSE_AWARD), current study level (CURRENT_STUDYLEVEL), study mode(STUDYMODE), postgraduate or undergraduate (PG_UG), resit number (RESIT_NUM),current year (CURRENT_STUDYYEAR), age(AGEGROUP), gender (SEX), race(RACE) and etc.. Oracle Data Mining provides a function called Attribute Importance that uses the Minimum Description Length (MDL) algorithm to rank attributes by importance in determining the target value. As shown in Figures 4 and 5, positive values represent that the characteristic is more important for attrition than characteristics with negative values. So, we can see that if the student drops out it is not related to her background, such as age, gender, race, etc. (as shown in figure 4), but related to academic activities, such as how often you use the learning system or the library system and what grade the student has (as shown in Figure 5).

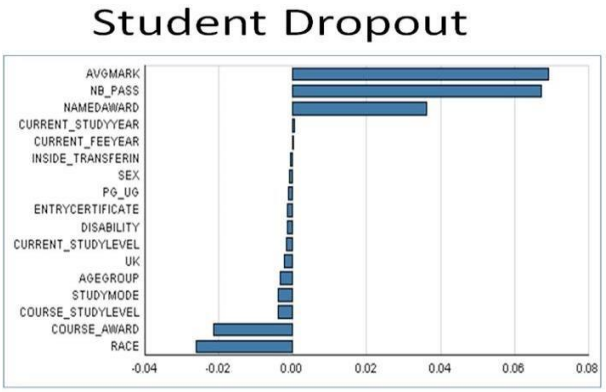

Figure 4: Dropout and student background.

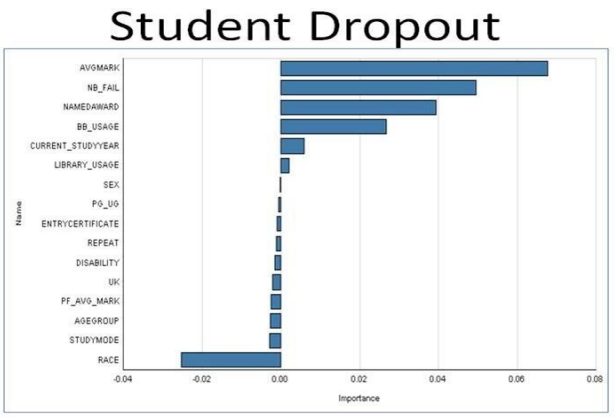

Figure 5: Dropout and student academic activities.

We also identify patterns that describe a group of students. There are some interesting findings, such as that students transferring to other institutions are mostly college students, they enroll with a lower certificate, but they get higher grades. While students transferring from another institution are mostly international students, enroll with a higher certificate, use the library and online learning system less, and score lower. These results can help us understand student behavior and then point to an effective intervention strategy to achieve better educational outcomes.

An experiment is also carried out to predict early school leaving based on the student's profile. The data are divided into training group and evaluation group in a ratio of 2: 1. Three algorithms are chosen: Naive Bayes (Harry, 2004), Support Vector Machine (Cristianini et al., 2000) and Decision Tree (Quinlan, 1986). Different settings are tested for each algorithm to find the optimal result. Since we don't want to give a negative prediction error for a true positive goal, it

is much worse to give a positive error for a true negative goal, so we increase the cost of false negatives in the cost matrix. As shown by Table 1, Naive Bayes achieved the highest prediction accuracy while the Decision tree with lowest one.

Table 1: PREDICTION RESULTS.

Accuracy	Naïve Bayes	Support Vector Machine	Decision Tree
Negative Accuracy	85.9 %	78.7%	71.2%
Positive Accuracy	93.1 %	88.3%	91.4%
Average Accuracy	89.5 %	83.5%	81.3%

CONCLUSION

In this article, we conclude how to use data mining to improve student retention. For MCMS, the information embedded in the data warehouse is historical data from past students and characteristics associated with current and future prospective students. We use this information to build the model of the student likely to drop out. These students can then be divided into different groups based on their risk value. Once these students have been identified, several methods can be used to improve retention:

Universities should develop an intervention program that focuses on specific problems. For example, Tinto cites five conditions that best promote retention (Tinto 2000):

REFERENCES

1. Committee of Public Accounts, 2001-02. Fifty-eighth Report of Session Improving Student Achievement and WideningParticipation in Higher Education in England, HC 588.
2. Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press.
3. Gabrilson, S., Fabro, D. D. M., Valduriez, P., 2008. Towards the efficient development of model transformations using model weaving and matching transformations, Software and Systems Modeling 2003. Data Mining with CRCT Scores. Office of information technology, Geogia Department of Education.
4. Han, J. W., Kamber, M., 2006. Data Mining: Concepts and Techniques, 2nd Edition, The Morgan Kaufmann Series in Data Management Systems, Gray, J. Series Editor,
Morgan Kaufmann Publishers.
5. Harry, Z., 2004. The Optimality of Naive Bayes, FLAIRS2004 conference.
6. Herzog, S., 2006. Estimating student retention and degree- completion time: Decision trees and neural networks vis-à-vis regression, New Directions for Institutional Research, p.17-33.
7. Kim, H., Zhang, Y., Oussena, S., and Clark, T., 2009. A Case Study on Model Driven Data Integration for Data
8. Centric Software Development, In Proceedings of ACM First International Workshop on Data-intensive Software Management and Mining.
9. Luan, J., 2002. Data mining and knowledge management in higher education – potential applications. In Proceedings of AIR Forum, Toronto, Canada.
10. Mazon, J. N., Trujillo, J., Serrano, M., Piattini, M., 2005. Applying MDA to the development of data warehouses. DOLAP 2005
11. Minaei-Bidgoli, B., Kortemeyer, G., Punch,W.F., 2004. Enhancing Online Learning Performance: An Application of Data Mining Methods, In Proceeding of Computers and Advanced Technology in Education.

BIOGRAPHIES

NAMITA AWASTHI
P.G Student,
Department of Computer Science & Engineering,
KIT Kanpur AKTU, U.P India.

VIMAL KUMAR AWASTHI
Assistant Prof., Department of Computer Science & Engineering,
KIT Kanpur AKTU, U.P India.