Informatics in Education, 2008, Vol. 7, No. 1, 31–54

Data Mining Application in Higher Learning Institutions

Naeimeh DELAVARI, Somnuk PHON-AMNUAISUK
Faculty of Information Technology, Multimedia University
Jalan Multimedia, 63100 Cyberjaya, Selangor, Malaysia
e-mail: naeimeh.delavari04@mmu.edu.my, somnuk.amnuaisuk@mmu.edu.my

Mohammad Reza BEIKZADEH
Faculty of Engineering, Multimedia University
Jalan Multimedia, 63100 Cyberjaya, Selangor, Malaysia
e-mail: drbeik@mmu.edu.my

Received: June 2007

Abstract. One of the biggest challenges that higher learning institutions face today is to improve the quality of managerial decisions. The managerial decision making process becomes more complex as the complexity of educational entities increase. Educational institute seeks more efficient technology to better manage and support decision making procedures or assist them to set new strategies and plan for a better management of the current processes. One way to effectively address the challenges for improving the quality is to provide new knowledge related to the educational processes and entities to the managerial system. This knowledge can be extracted from historical and operational data that reside in the educational organization's databases using the techniques of data mining technology. Data mining techniques are analytical tools that can be used to extract meaningful knowledge from large data sets. This paper presents the capabilities of data mining in the context of higher educational system by i) proposing an analytical guideline for higher education institutions to enhance their current decision processes, and ii) applying data mining techniques to discover new explicit knowledge which could be useful for the decision making processes.

Keywords: data mining, explicit knowledge, classification, prediction, association rule analysis, clustering, decision tree, neural network classification, radial basis function, neural network prediction.

1. Introduction

One of the significant facts in higher learning institution is the explosive growth of educational data. These data are increasing rapidly without any benefit to the management. We believe that to manage this difficult task, new techniques and tools for processing the large amount of generated data in business processes and extracting some useful knowledge and information are required. Data mining techniques are analytical tools that can be used to extract meaningful knowledge from large data sets. This paper addresses the capabilities of data mining in higher learning institution by proposing a new guideline of data mining application in education. It focuses on how data mining may help to improve decision making processes in higher learning institution.

In order to propose a new guideline, one must understand the data mining and decision making processes in higher learning institutions. In this regard the literature survey highlights the importance of this technology, what educational system lacks today and how data mining is applied to the current educational system.

This paper is structured as follows. The next section presents a background study of the educational domain and the problems that exist in the current conventional system. Section 3 presents data mining as a key to the current problem in educational system. Section 4 presents a guideline to data mining application in higher learning institution proposed by the authors. Section 5 and 6 present the data analysis and data modeling of data mining application in a university respectively. Section 7 presents the analysis and the results.

2. Background of Educational Systems Based on Indicators

Indicators are agreed measurement scales which identify the quantitative relationships between two variables. They are normally used as numerical values. Indicators are very important in determining the goals and the operational analysis of the educational system (Johnstone, 1976; Johnstone, 1981; Wako, 1988).

The higher learning institution of a country deals with human factors and educating specialists needed by the community, educational promotion, research development, and providing a suitable environment for the country's growth. Thus, the system essentially requires a principle which can express the qualitative characteristics of the higher learning institution to some quantitative values, and facilitates evaluating the functionalities. This principle is summarized into indicators.

To evaluate the different aspects of the higher learning institution, "performance indicator" is used as one of the main educational system indicators (UNESCO, 2006b). Educational performance indicators have been known as the base for educational system methodology improvement. There have been many studies (Oakes, 1986; Scheerens, 1990; Cave *et al.*, 1990) which present the importance of performance indicator as a quality improvement tool in an educational domain. They indicate that performance indicator is vital for educational system improvement. The earlier studies (Yang *et al.*, 1999; Fitz-Gibbon and Tymms, 2002; Van Petegem *et al.*, 2004) state that other than performance indicator, an additional step for supporting educational system improvement, which is built on information from performance indicator, is more important. This step is called educational feedback. The feedback should be up-to-date, valid and reliable.

From the above literature survey in higher learning institution it is possible to derive the following conclusions:

1. Based on the fact that the performance feedback perceived in an educational institution should be accurate, up-to-date, reliable, valid, and toward the goals of educational improvements, therefore more effective strategies should be taken into

account to improve the feedback from an educational domain. Not only the performance indicator is essential for indicating the actual state of an education system, it is also vital to develop a methodology for educational system performance feedback.

2. Improving the feedback of an educational domain implies further analysis and investigation in the forming components of performance indicator. Data mining is able to improve the educational system in each component of the performance indicator. This improves the feedback from the system.

In this study, the components of performance indicator are based on Vlasceanu *et al.* (2004) definition. They argue that performance indicators work efficiently only when they are used as part of a coherent set of input indicator (Human resource, financial resources, sector resources), process indicator (Educating methods, qualitative and quantitative educational improvements, such as registration and dropout rates) and output indicator (alumni and graduates).

In the next section, we present the importance of data mining in education and how it can help to improve the performance indicator and higher learning institutions in general.

3. Data Mining: a Way to Improve Todays Higher Learning Institutions

Data mining is a powerful new technology with great potential in information system. It can be best defined as the automated process of extracting useful knowledge and information including, patterns, associations, changes, trends, anomalies and significant structures from large or complex data sets that are unknown (Han and Kamber, 2001; Two Crows Corporation, 1999; Chen *et al.*, 1996). Many applications areas such as banking (Han and Kamber, 2001), retail industry and marketing (Han and Kamber, 2001; Edelstein, 2000), fraud detection (Chang and Lee, 2000), computer auditing (Teh *et al.*, 2002), biomedical and DNA analysis (Han and Kamber, 2001; Han, 2002; Feldman, 2003), telecommunications (Han and Kamber, 2001; Chang and Lee, 2000), financial industry (Han and Kamber, 2001) have already been advanced through the sturdy techniques of data mining. Another application domain that can take advantage of data mining techniques is higher learning institution.

Nowadays, higher learning institutions encounter many problems which keep them away from achieving their quality objectives. Some of these problems stem from knowledge gap. Knowledge gap is the lack of significant knowledge at the educational main processes such as counseling, planning, registration, evaluation and marketing. For example, many learning institutions do not have access to the necessary information to counsel students. Therefore they are not able to give suitable recommendation to the students. We also observe that there is no accurate grouping of courses to identify which type of course is most appropriate to be offered to which type of students.

Our main idea is that the hidden patterns, associations, and anomalies that are discovered by data mining techniques can help bridge this knowledge gap in higher learning institutions. The knowledge discovered by data mining techniques would enable the higher learning institutions in making better decisions, having more advanced planning in directing students, predicting individual behaviors with higher accuracy, and enabling the institution to allocate resources and staff more effectively. It results in improving the effectiveness and efficiency of the processes.

Data mining is considered as the most suitable technology in giving additional insight into educational entities such as; student, lecturer, staff, alumni and managerial behavior. It acts as an active automated assistant in helping them to make better decisions on their educational activities. The final result is improved decision making processes in higher learning institutions. This improvement would carry the following advantages including; increasing student's promotion rate, retention rate, transition rate, increasing educational improvement ratio, increasing student's success, increasing student's learning outcome, maximizing educational system efficiency, decreasing student's drop-out rate, and reducing the cost of system processes. In the next section, the literature surveys of data mining applications in learning institutions are presented and analyzed based on the components of performance indicator and educational processes.

3.1. *Analysis of Previous Data Mining Applications in Education Systems*

In this section the previous data mining applications in higher learning institutions are analyzed to identify the effect of data mining on which components of performance indicator is. Table 1 provides a summary analysis of the main components.

Table 1

Summary analysis of previous study

ID	Performance Indicator Type	Objective	Data Mining Method	Explicit Knowledge	Educational Main Process	Educational Sub-Process
1	Input indicator	Predicting alumni pledge	Prediction	The pattern of previous graduates contributing to the university activities	Planning	Alumni activity planning
2	Process indicator	Use of Data Mining in CRCT scores	Prediction	The patterns of previous student test score associated with their gender, race, attendance and so on	Evaluation	Course assessment
3		Creating meaningful learning outcome typologies	Cluster analysis	The patterns of previous student's learning outcome		Student assessment
4		Use data mining techniques to develop institutional typologies	Cluster analysis	The pattern of various groups of students		
5		Academic planning and intervention transfer prediction	Prediction	The success patterns of previous students who had previously transferred subjects		
6		Predicting and clustering persisters and non-persisters	Prediction Clustering	The patterns of previous similar student who were persistent or non persistent		
7		Predicting student performance	Classification	Classified pattern of previous students based on their final grade		
8	Output indicator	Improving quality of graduate student by data mining	Association Classification	Characteristic patterns of previous students who took a particular major and The patterns of previous students which were likely to be good in a given major	Counseling	Student major counseling

Project ID 1: Predicting Alumni Pledge

This case study (Luan, 2002a; 2002b) predicts the alumni pledges and helps universities to develop a cost-effective method to identify those alumni most likely to make pledges. Many universities spend lot of money in sending regular mails to their alumni every year, though some of them (alumni) may never contribute at all. The obtained pattern from the data mining technique is able to predict those alumni who make pledges to the university after graduation. The main advantage of this activity to the universities is to spend money on mailing those who are not very interested in contributing. The result is saving money in performing other functionalities.

One of the input indicators in a higher education is estimating the financial resources, cost and expenses. This study attempts to reduce the expenses on alumni activities in educational domain through prediction technique. By reducing the cost of alumni activities, the total cost of university is reduced. This has a great impact on improving this input indicator of a higher education through improving the alumni activity planning process in educational domain.

Project ID 2: Uses of Data Mining in CRCT Scores

This study (Gabrilson, 2003) attempts to analyze the most effective factor in determining student test score in various subjects. It presents that the useful discovered patterns targeting the various relationship of different types of variables are the main factors affecting the students test score. Using data mining prediction technique, these factors are identified. While theses factors are identified, then the attempt would be based on increasing the effect of these factors for new students. It results in better student's test score performance in coming year.

One of the qualitative and quantitative educational improvements of process indicators is transition rate (Wako, 1988). It is defined as the "proportion of pupils progressing from the final grade of one level to the first grade of the next level". Using the standardized test result, the student transition rate from one level to higher is identifiable. High transition rate in a higher educational domain is an indication of a high level of access or transition from one level of education to the next. Also low transition rate can signal problems in the bridging between two cycles or levels of education, due to either deficiencies in the examination system, or inadequate admission capacity in the higher cycle or level of education, or both. For example, if student transition rate is reduced and the standardized test result is also present low trend, therefore we need to determine the reason for result worsening, and try to enrich the factors affecting student's drops. This increment will definitely increase the student's transition rate.

From this case study the effectiveness of data mining in predicting the most effective factors in student test score can be concluded. It results in improving the evaluation-student assessment process. Improving this process has a direct impact in improving transition rate of a higher learning institution.

Project ID 3: Creating Meaningful Learning Outcome Typologies

Another study done by Luan (2002a, 2002b) aims at creating meaningful learning outcome typologies using data mining techniques. The main objective of obtaining typologies of students is to be able to improve students through predefined clusters of behavior.

These grouping can be reached beyond traditional student profiling. Discovering various student typologies in educational domain helps to determine those who quickly are able to pile up their courses and those who take courses for longer period of time. These clusters help universities to better identify the requirements of each group and make better decision on how to behave with them in terms of educating, offering courses and curriculum, required time for teaching and so on. It results in having more student satisfaction of their studies, course offering, and class's periods.

Evaluating student educational achievement is one of the process indicators of educational domain. For example, the rate of female educational achievement (Mashayekh, 1989) can be obtained by dividing female educational achievement over male educational achievement. To keep the educational achievement ratio above the standard level, educational domain need to better understand their student's behavior and characteristics. It necessitates a method to better understand students in their domain, so that they may better identify their requirement and state of their students.

From this case study we can conclude the effectiveness of data mining in developing typologies of students in educational domain. The result has an impact in improving educational achievement of a higher education through improving the evaluation-student assessment process.

Project ID 4: Use Data Mining Techniques to Develop Institutional Typologies

This study (Luan *et al.*, 2004) is very similar to the above study but is more advanced in a sense that it first discovers the factors of student dimension (attributes) using factor analysis, one of the feature reduction methods. Then the result is used to discover the various clusters of students. The third finding is discovering institutional typologies based on the various types of students.

The analysis and the effect of this study on student educational improvement process indicator are similar to the project ID 3 "Creating meaningful learning outcome typologies". Therefore it is not repeated again here.

Project ID 5: Academic Planning and Interventions Transfer Prediction

The study done by Luan (2001, 2002b) presents data mining advantages in predicting students' likelihood of transferability for on time proactive intervention. It notifies the institution the types of students who are most at risk of not transferring to a higher level before they know it. The outcome enables universities in predicting the likelihood of student transferability. Data mining can link student's academic behaviors with their final transfer outcomes. Therefore these kinds of identifications help the universities to pay more attention to those who require more academic assistance by setting extra classes, setting consultation hours with the university's counselors and psychologies.

As mentioned in the above literature, the transition rate is the level of improvement from one cycle or level to another educational cycle or level. The aim of the educational domain is to keep this indicator high. One way to accomplish this is to be able to predict students' transferability.

From this case study, we conclude that predictions of student's likelihood of transferability assist decision makers with an additional tool to identify those who are less likely to transfer. As a result, this prediction provides a high potential impact on improving the transition rate of an educational domain through improving the student assessment process of the educational domain.

Project ID 6: Predicting and Clustering Persisters and Non-Persisters

A study is done by Luan (2002) to predict student's persistency by considering the students who either re-enroll or doesn't re-enroll the following trimester. Using the prediction techniques, the likelihood of a student's persistency can be determined. The identification of these students is very useful to universities because if they know those who are less likely to persist, then the university can attempt to increase the persistency rate of these students. By knowing students who are not persistence, the faculty can identify the factors affecting their non-persistency. Therefore the university's managerial systems have to make an attempt on improving these factors, which would result in improving the student's persistency rate.

One of the qualitative and quantitative educational improvements in the process indicators of an educational system is the survival rate by grade. It measures the internal efficiency and the holding power of an educational system (UNESCO, 2006a). The definition from (Wako, 1988; Wako, 2003) seems to be comprehensive where they define it as "the percentage of a cohort of pupils enrolled in the first grade of a given level or cycle of education in a given year who are expected to reach each successive grades". It defines the retention of students from grade to grade in educational organizations.

From this study, it can be concluded that the prediction technique is useful in predicting the likelihood of a student persistency. By identifying the students who are less likely to persist, the university intervenes by providing them the academic assist and as a result, student retention rate will be increased. This has a positive impact on improving the survival rate through enhancing the student's assessment process in a higher learning institution.

Project ID 7: Predicting a Student's Performance

Another study (Behrouz *et al.*, 2003) uses the data mining classification technique to predict students final grades based on their web-use feature. By discovering the successful patterns of students in various categories, the university can predict the final grade of each single student. Therefore it helps to identify students at risk early and allow the instructor to provide appropriate advice in a timely manner.

From this case study it can be concluded that data mining is effective in predicting student's performances in the educational domain. The result has an impact in improving the transition rate, and the process indicator of a higher learning institute by improving the student assessment process.

Project ID 8: Improving Quality of Graduate Students by Data Mining

A study is done by Kitsana (2003) to improve the quality of graduate students by data mining. The objective of the project is to discover knowledge from large sets of engineering student's databases records. The discovered knowledge is useful in assisting the development of new curricula, improving of existing curricula and most important, helping students to select the appropriate major. The final result represents the most appropriate major for each single student. The discovered patterns are useful for university counselors or supervisors who are supposed to supervise new student.

One of the output indicators is computed through gross completion rate. This is defined as the "the total number of students completing (or graduating from) the final year of primary or secondary education, regardless of age" (UNESCO, 2006b).

The result obtained from this case study is useful in increasing students' success in their major. It directly causes an increment of students' gross completion rates in every single major. We can conclude that the effectiveness of data mining in predicting the most appropriate major for each single student improve the student course counseling process. Improving this process has a direct impact in improving the gross completion rate of a higher learning institution.

According to Table 1, it is clear that the main focus of many researchers has been to improve the process indicator of an educational domain. It is also observed that there has been much attention on enhancing the main process evaluation. The most emphasis is also on the student assessment sub-process. But the results from previous studies show that to establish advanced higher learning institution there are still many areas hidden from the view of data mining researchers in the educational domain. To better identify these areas in the educational domain, we will present our proposed data mining guideline in the next section. It introduces a plausible area of data mining application in various processes of a higher learning institution.

4. Proposed Analysis Guideline (DM-HEDU) for Application of Data Mining in Higher Learning Institution

In this section, we propose a new analysis guideline to present a roadmap or the plausible area of data mining application in higher learning institution. Its adopted name is DM-HEDU (Data Mining in Higher Education System). As today's higher learning institutions deal with powerful business competitors in a highly competitive environment, they have to look for a new and faster solution to overcome the problems and achieve a high academic standard. Therefore, this guideline may assist the institutions to identify the ways to improve their processes. In the previous literature studies we have not discovered a complete guideline which gathers most of the possible processes to improve a learning institution through data mining.

The idea of our proposed guideline is presented Fig. 1. The importance of tracking DM-HEDU guideline in a higher learning institution can be viewed from four different angels as follows:

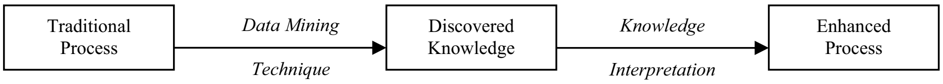

Fig. 1. Basic idea of proposed DM-HEDU guideline.

1) incorrect usage of huge amount of raw data;
2) outsider environmental motivation;
3) internal educational reason;
4) identification of current gaps and further work.

The *incorrect usage of huge amount of raw data* stored in a higher learning institution can be seen where the current university management system and administrators manage the existing educational processes based on the data stored in the university's database systems. But since not all of these data are useful, data mining techniques can help to discover valuable knowledge, patterns, and structures. The result is a set of enhanced and new processes offered to the educational institutions. Our guideline is arranged to demonstrate the identification of the type of knowledge that can be discovered in each process using data mining techniques to the higher learning institutions.

The *outsider environmental motivation* can be observed where the higher educational organizations aim to be ahead of their business competitors. Therefore they first have to be powered by a proper roadmap and to be demonstrated with an exact guideline to reach a higher educational level. Our proposed analysis guideline is designed to serve as a prerequisite to today's higher educational needs. They can use this guideline as an assistance decision support tool.

The *internal educational reason* is considered as the procedure toward improving the educational management system. Our guideline is arranged to demonstrate the managerial system to identify which part of their traditional processes can be improved by data mining technology and how they can achieve their data mining goal through a structured way.

Our DM-UEDU guideline can be used for unifying a common context to *identify the current gaps and further works* for any data mining application in a higher education based on the processes of higher learning institution. It also provides an opportunity for researchers to be known with the existing area of research.

Table 2 shows the main components of this guideline. It signifies the main processes of a generic educational system, sub processes, the related knowledge that can be discovered by data mining techniques, and the appropriate data mining technique that can be used to discover the related knowledge.

As shown in the DM-HEDU guideline, the first column of the table corresponds to the main processes that normally occur in a higher learning institution. We have identified seven main processes in higher educational system, which are "evaluation", "planning", "registration", "consulting", "marketing", "performance" and "examination". Each single process is categorized into some detail sub-processes, which are presented in column two of the DM-HEDU guideline. The third column of the guideline introduces sets of knowledge and information, which are in fact a series of extracted patterns and structures from the masses of data by the use of data mining techniques and algorithms. It means that applying some data mining techniques on the set of data, a series of meaningful knowledge can be discovered. This knowledge helps in enhancing the traditional education processes. The last column of the guideline portrays the most appropriate data mining method for extracting the beneficial knowledge.

Table 2

Portion of DM-HEDU: data mining in higher education

Main Process	Sub-Process	Explicit Knowledge	Data Mining Method
Evaluation	Student assessment	● The success patterns of previous students who previously had transferred subjects	● Prediction
		● The patterns of previous students who were likely to be good in a given major ● The patterns and relationship of various factors affecting the student test score ● Prediction of the likelihood of success	● Prediction
		● The success patterns of previous similar students ● Prediction of likelihood of persistence	● Prediction, ● Clustering
		● The patterns of previous successful and unsuccessful graduates	● Prediction, ● Clustering
		● The patterns of previous students who planned to dropp subject	● Prediction
		● The patterns of previous students who planned for resource allocation	● Prediction
		● The patterns of previous male and female students in test score	● Association
		● The patterns of previous student's learning outcome	● Prediction, ● Clustering
		● The patterns of previous students attendance in accordance with test score	● Association
		● Association of student health information and test score	● Association
	Lecturer assessment	● The characteristic patterns of previous lecturers which were more effective than others	● Prediction, Classification
		● Association between lecturer training and student test score	● Association
	Course assessment	● Cluster of most cost-effective courses to be offered together	● Clustering
		● The patterns of courses who offered previously to different type of students	● Classification ● Association
		● Prediction of factors most affected in test score in various courses	● Prediction
		● The patterns of programs (courses) which produce greatest return and investment in terms of student learning in coming year	● Prediction
	Industrial training assessment	● The patterns of previous training course for different type of student	● Classification ● Association
	Student registration evaluation	● The success patterns of those students who successfully enrolled to the university	● Prediction

To be continued

Continuation of Table 2

Main Process	**Sub-Process**	**Explicit Knowledge**	**Data Mining Method**
Planning	Course Planning	● Classification of courses to the most appropriate time ● Success patterns of courses which were taken together	● Classification ● Clustering
	Academic planning	● The patterns of previous discipline problems in academic planning	● Prediction
	Lecturer time table planning	● The patterns of previous lecturer's class time table ● Prediction of lecturer time table for coming year	● Prediction
	Alumni activities planning	● The pattern of previous graduates contributing in university activities ● Prediction of the likelihood of alumni who continued studies ● Prediction of the likelihood of alumni who find suitable job	● Prediction
Registration	Student course registration	● The patterns of previous students who take various subjects ● Association of student to the most appropriate subject	● Prediction ● Association
Counseling	Student behavioral consulting	● The patterns of previous students behavior in an academic environment	● Clustering
	Major selection consulting	● The characteristic patterns of previous students who took particular major	● Classification ● Association
	Course selection consulting	● Classification of student to various elective subject ● Classification of student to various courses	● Classification
	Program selection counseling	● The patterns of previous student who were good in a given program	● Association ● Classification
Examination	Student examination	● Association between exam level and student mark ● Association between exam level and lecturer class performance	● Association
Performance	Student performance	● Association between student performance and lecturer satisfaction ● Association between student course mark and time and venue of classes	● Association
	Lecturer performance	● Association between lecturer who cancel the class frequently and student test score ● Association between lecturer background and time and his/her performance	● Association
Marketing	University advertising	● The characteristic patterns of previous international lecturer and student which attract to the universities	● Prediction
		● The characteristic patterns of previous local student and lecturer who resign or terminate from local universities	● Prediction

For instance, it is observed in the DM-HEDU guideline that “evaluation” is an educational process. Its main sub-processes are “student assessment”, “lecturer assessment”, “industrial training assessment”, “course assessment”, and “student registration evaluation”. As an example, the traditional “student assessment” sub-process can be enhanced to a new process that can identify the likelihood of student’s transferability. The appropriate technique can be prediction technique stated in column 4 of the table. The discovered knowledge defines the success patterns of previous similar students who successfully transfer the course.

5. Data Analysis and Investigation

In this study, the experiments are conducted on course computer programming II and the pre-requisite of it, Computer Programming I. The new knowledge discovered from these applications help in improving decision-making procedure of a university.

The application is mainly based on the CRISP_DM (Chapman *et al.*, 2000) methodology. The main steps of CRISP are including; *domain understanding, data understanding, data preparation, modeling, evaluation* and *deployment.* In this section, some of the appropriate activities for the first three steps are presented. These three steps prepare and preprocess the data for further analysis and modeling. Next sections will discuss on the rest of the CRISP steps.

Domain Understanding. In this phase the higher education is analyzed and the main data mining objectives are set (Section 5, DM-HEDU guideline)

Data Understanding. In this phase, the required raw data and attributes are collected based on the objectives. According to our data mining goal, the raw data is related to:

1) student demographic and academic knowledge;
2) lecturer demographic and academic knowledge;
3) course information;
4) semester status information.

The data are then described and explored by (i) identifying initial format of data, (ii) the meaning and description of individual attributes and (iii) determining the relation of attributes. The final part verifies the data quality by determining the data completeness and correctness.

Data Preparation. This phase is the final step of directly dealing with data. The dataset produced in this section is used for modeling and the major analysis task of the project. The essence of data preparation is to maximize visibility of the relationship that exists between input and output data sets, which is captured with a modeling tool. Prepared data enables mining technique to generate a better model.

Three main activities done in this phase are as follows:

1) designing the taxonomy of a university educational entities,

2) evaluating data,
3) evaluating variables (feature selection).

By designing the taxonomy, a hierarchy of relations between various students and lecturer knowledge are defined. As an example, prior to the limited space a very small portion of taxonomy related to student demographics information is presented in Fig. 2.

Data evaluation includes resolving the problems with missing values, outliers and redundant variables. In this regard, few activities are done to make sure that the data are in a good quality condition, including: handling missing values, converting continuous variables to discrete variables, constructing new attributes, integrating data, and sampling data.

Our main idea of evaluating the variables is demonstrated in Fig. 3. It presents the suitable technique and approach appropriate to our study to get the most consistence variables.

Choosing the right variables in data mining is one of the main tasks before applying most of the modeling technique, because the variables can be correlated or redundant. The approach to get these variables is relevance analysis or feature selection. The tech-

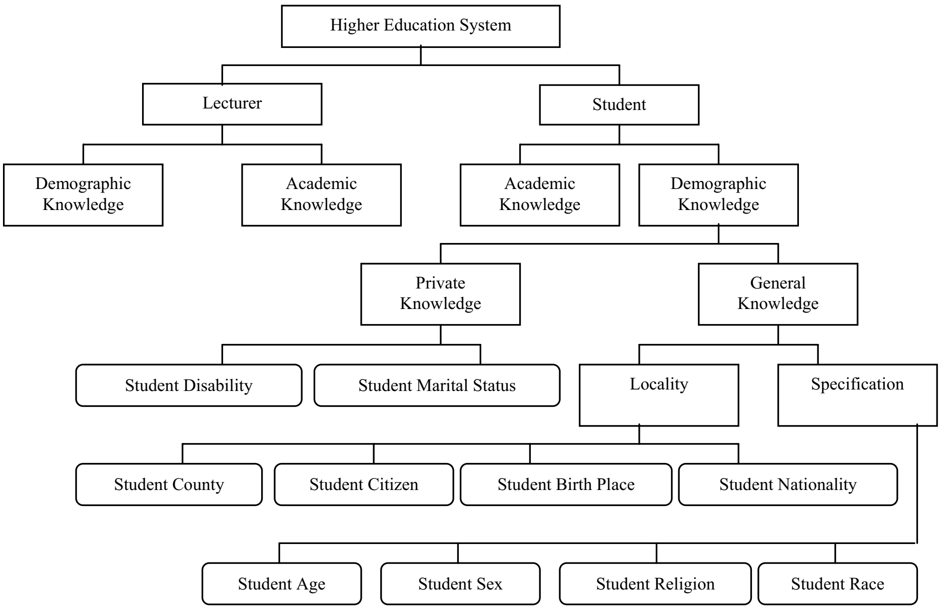

Fig. 2. Small portion of taxonomy related to student demographics information.

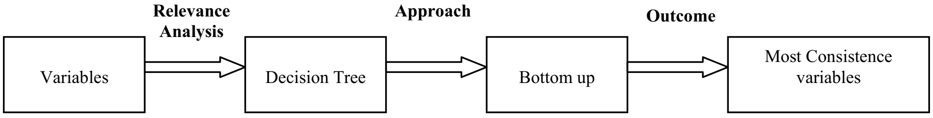

Fig. 3. Evaluating variable technique and approache.

nique used is decision tree. This method is appropriate for handling both categorical and numerical variables values. It is used to get the most consistence variables by presenting them in the various tree levels.

Among two approaches of decision tree including bottom up and top down, the bottom up strategy is selected in this study, because it is more reliable, and we can get the classification trees that describe the class label with only this variables set. The output variables discovered in this phase are going to be used as the input of next phase of CRISP, modeling phase.

6. Data Mining Modeling

In this section the prepared data and attributes from the previous section are used as the input for the development of lecturer, student and course models. This section explains the outcome (explicit knowledge) of the models and the usage of the outcome by managerial decision makers. The knowledge obtained from data mining techniques gives the managerial decision makers the useful information for decision making. The models are classified in two main categories; predictive and descriptive models.

- Descriptive model describes the data set in a concise and summarized manner and presents the interesting general properties of the data. It explains the patterns in existing data, which may be used to guide decisions.
- Predictive model predicts behavior based on historic data and uses data with known results to build a model that can be later used to explicitly predict values for different data (Two Crows Corporation, 1999).

6.1. *Predictive Data Mining Models*

Model A: Predicting Student Success Rate for Individual Student

This model is developed to predict the student success rate for individual students. The explicit knowledge discovered from this model can be used by student management system to consult individual student in successful course taking. Within this procedure, if the students are predicted to be unsuccessful, then they are provided with extra consultation and help to get improved in the course. If they are successful, new personal policies for successful student course taking are set.

Classification and prediction are the two main methods, which are helpful to discover the pattern of successful student. They assist in predicting the performance of students in computer programming II and help to identify who will be poorly carried out in the course before they are even known with it.

For classification method, the experiments are conducted through decision tree using SLIQ algorithm (Mehta *et al.*, 1996) and neural classification. For prediction method the experiments are conducted using neural prediction (Han and Kamber, 2001) and Radial Basis Function (RBF) (Lee *et al.*, 1998). In the following some of the hypotheses[1] discovered from the models are presented.

[1] Hypothesis is the obtained result from the model before validation.

Decision Tree and Neural Network Classification Technique

Using decision tree, the final model has the accuracy of 86.04% and using neural network the final model has the accuracy of 86.8%. In the following some of the meaningful output results are presented.

1. There is a strong correlation between students, whom are younger than 21 years old and taking the course according to the study plan (or earlier), and his success. The student in this group is likely to be successful (97.3% of cases).
2. There is a strong correlation between students who delay on taking the course for long time (more than 3 years), and their success. They are likely to be unsuccessful (100% of cases).
3. There is a strong correlation between students majoring in software engineering who take the course earlier and their success. They are likely to be successful.

Neural Network and RBF Prediction Techniques

This model is applied through hold out and 5-fold cross validation. The result is presented in terms of Global Root Means Square (GRMS) error (Huffman, 1997). The neural network has the error rate of 0.31 and RBF has 0.30 GRMS error. The outcome of prediction techniques cannot be easily understandable and should be interpreted first. Below, conclusion results of some of the hypothesis are presented.

1. There is a strong correlation between the students with high English ability and their success. These students are 82.52% to 84.13% likely to be successful.
2. There is a strong correlation between those students who perform well in pre-requisite of the course, and their success. These students are 82.69% to 86.20% likely to be successful.
3. There is a strong correlation between the students who take the subject while they are not in the first year of study and their success. These students are 71.03% to 73.07% likely to be successful.

Model B: Predicting Student Success Rate for Individual Lecturer

This model is developed to predict the student success rate for individual lecturer. The explicit knowledge discovered from this model can be used by management for general decision-making. In other words it can be used to support policies and procedures, which are set at top-level management. The model is applied through prediction (neural network) and classification (SLIQ and neural network) techniques. In the following some of the obtained results are presented.

Decision Tree and Neural Network Classification Technique

In this model, the lecturer's demographics and academic knowledge and student success/failure as class category are the input variables to the decision tree. The result obtained from decision tree model through 5-fold cross validation has the accuracy of 87.80% and the result obtained from neural classification model has the accuracy of 88.65%.

In the following, some of the hypotheses are presented.

1. There is a strong correlation between the lecturer who is single or has been married for less than 10 years and the success of his students. In this case the students are likely to be successful.
2. There is a strong correlation between the lecturer who has been married for more than 10 years with high academic standard level, and the success of students. In this case the students are likely to be successful.
3. There is a strong correlation between the lecturer who has been married for more than 10 years with low academic standard level, and the success of students. In this case the students are likely to be unsuccessful.

Neural Network Model

The model obtained using neural network has the GRMS error of 0.35. In the following, some of the hypotheses are stated.

1. There is a strong correlation between the high academic standard grade level of a lecturer and the success of the students. The students are likely to be successful in this case.
2. There is a strong correlation between the lecturers majoring in computer science or information technology, and the success of students. The students are likely to be successful in this case.

6.2. *Descriptive Data Mining Modeling*

Model C: Model of Student Course Enrollment

This model is developed to describe the characteristics patterns of successful students who plan to take the course. The discovered explicit knowledge discovered from this model is used for general decision making at top-level management. It either assists in supporting the policies and procedures for student intake procedure or it helps to set new strategies and plan for those who are wishing to take the course.

One of the useful techniques is association rule analysis (Agrawal and Imielinski, 1993). Using this technique the pattern and tendencies of old data are discovered. This pattern describes the relation of successful students with their academic and demographics information. Therefore this knowledge can help to determine new policies for new intakes in this course. In the following the interpretations of some of the hypothesis are presented. Confidence defines the strength of the rule.

1. This hypothesis says that for 70.98% of all students who have successfully passed their perquisite for computer programming II, they are successful with a confidence of 87.65%.
2. This hypothesis says that for 65.49% of all students who are not in the first year, they are successful with a confidence of 79.71%.
3. This hypothesis says that for 58.62% of all students who have successfully passed their prerequisite of the course and they are not in the first year; they are successful with a confidence of 86.67%.
4. This hypothesis says that for 39.22% of all students younger than 21 years, they are successful with a confidence of 98.52%.

5. This hypothesis says that for 37.05% of all students younger than 21 years old and not in the first year, they are successful with a confidence of 98.44%
6. This hypothesis says that for 36.47% of all students younger than 21 years old and with no failure in their prerequisite, they are successful with a confidence of 98.94%.

Model D: Model of Lecturer Course Assignment Policy Making

This model is developed to describe the characteristics pattern of lecturers who plan to take the course. The knowledge discovered from this model can be used for general decision making at top-level management. It can either assists in supporting the current managerial rules and regulations in lecturer course assignment policy making or it helps to set new strategies and plan for managerial decision makers on those lecturers who plan to take the course.

One of the appropriate techniques is association rule analysis. It can be used to discover the patterns and tendencies of old data that relates the academic and demographics information of lecturers with their student's success. In the next page some of the hypotheses are presented.

1. This hypothesis says that for 49.86% of all lecturers majoring in computer science, their students are successful with a confidence of 80.72%.
2. This hypothesis says that the lecturer with high background grade level, and majoring in computer science, the students are successful in 80.58% of cases. This affects 45.99% of the total hypotheses.
3. This hypothesis says that for 38.22% of all lecturers with high academic grade level, the students are successful with a confidence of 93.24%.
4. This hypothesis says that for 39.34% of all lecturers who are active in contributing university activities, then the students are successful with a confidence of 84.02%.

Model E: Model of Lecturer Typologies

This model is also developed to describe the various typologies or segments of lecturers who have taught the course. The appropriate technique is cluster analysis technique. Using this technique, lecturer profiling that yields to high and low achieved students is performed. The various segments of lecturers that deliver students in different level of programming are identified. Two techniques are used including neural clustering and demographics clustering.

Referring to the clusters obtained using demographics and neural clustering; it is observed that most of the clusters obtained using both techniques are similar. In the following the some sample clusters are presented.

Cluster Number 1. This cluster includes the lecturers with high academic grade level and majored in science. They have been very active in university. Their background study was done in research work. They have recently married. The students in this group are 95.74% likely to be successful.

Cluster Number 2. This cluster includes the lecturer with low academic grade level and majored in computer science. They have been very active in university. Their background study was done in research work. They are married for long time. The students in this group are 56% likely to be successful.

Model F: Model of Course Time Planning
This model is developed to describe the patterns of course information for better planning. One of the appropriate solutions for this problem is using association rule analysis to discover the pattern and tendencies of data. These patterns identify the most suitable time of offering course. It can be discovered through identifying the relation of the successful students with the time of taking the course. Therefore higher percentage of successful students in an appropriate time is the representation of successful course planning on time. The obtained hypothesis presents that if the course is offered on the first 7 to 16 first month of student study period, then the student is successful in 91.48% of cases. Therefore this is an important factor for managerial decision makers to propose the course at this appropriate time.

7. Analysis and Discussion

In this section the obtained hypotheses from the above models are analyzed and discussed. The main sources of this analysis are based on the expert knowledge and the main data before any data mining operation. In the following some of the hypotheses are validated for various models and later some improving factors are presented.

Creditability of the Results

The obtained hypotheses from the models should be first validated with the main data set to identify whether the same patterns exists in the data. If the obtained hypothesis and the pattern in the main data set do not match, then the reason of this inconsistency is verified. The hypotheses presented in the previous section are all analyzed with the real data set and they are validated to be meaningful and creditable. In the following some example of the valid and invalid rules are presented.

One of the hypotheses (i.e., a rule generated by a decision tree model) says that, if the students delay on taking the course (according to the study plan) and their English ability level is medium, then they are 100% likely to be successful. It seems to contradict to our commonsense since a 100% success rate seems a bit high. We verify this by comparing the percentage of student success in main data set and preprocessed data set. It is identified that there are missing unsuccessful students that delays on taking the course (according to the study plan) and their English ability level is medium. Therefore this hypothesis is not creditable.

The other hypothesis obtained from the neural and decision tree classification says that there are strong correlation between students, whom are younger than 21 years old and taking the course according to the study plan (or earlier), and his success. The student in this group is likely to be successful. In this hypothesis the percentage accuracy of student success in preprocessed data is similar to the student success in main data set. Therefore this rule is meaningful and creditable.

Factors Affecting the Reliability of Model

Reduction in the total number of data. There are few reasons that the result from modeling techniques may not conclude to real data. One of the main possible reasons is due to some negative impacts of data preprocessing in the third phase of CRISP_DM that the total number of data is reduced. In the following some of them are stated:

1. Handling missing value phase
   In this study, two methods for handling missing values are applied. The first approach is encoding the values. If this phase cannot be applied to some variables then the records containing missing values are discarded. The missing records in this phase may have some negative impact on the outcome of the modeling result.
2. Incompleteness and low quality of important attribute value
   One of the important attribute values is the grade of students in the course so that their success and failure can be obtained from it. This attribute is very important since among all techniques it plays an important role in the modeling phase. Some of the records in this phase have to be discarded because this important attribute is not completely entered by the administrators.
3. Data integration and database application
   In data integration phase the data from various sources are merged to produce the information about students and lecturers. In this regards the full outer join operation causes some of the records to be discarded in data preparation phase.
4. Inconsistency and value error among attributes
   The other reason that some of the modeling results do not confirm with the real life situation is inconsistency and value error among the attribute values. There are many reasons for these inconsistencies, including wrong entering the values of attributes by the administrators over a period of time, like “state” and “country” do not match. Or when calculating the age of a lecturer, it is observed that the lecturer is still young, but the pension date of him/her has already passed.
   The modeling result is based on the values and attributes in the input data. Therefore a wrong input produces a wrong modeling result too. So in our modeling, we discard those attributes and values which may lead to wrong result. As a result, the whole information about one educational entity is removed from the dataset.

Reduction in the total number of attributes. There are many reasons which cause reduction in the number of attributes. We have identified this through the expert knowledge. It means the attributes which are important from an expert’s point of view are not observed in the modeling result. In the following some of the reasons are stated:

1. Feature Selection in data preparation phase
   Based on the attribute values, in relevant analysis phase those attributes with major importance and more consistency are identified. Some of those attributes may be considered important from an expert point of view but they may not be considered important in relevance analysis and vice versa. The ignorance of attributes in relevance analysis is based on the low quality of attributes and their value. For example, most of the data in an attribute has the same value for all the values.
2. Medium quality of attributes and their values

There are some attributes which most of the values appropriate to them are not filled properly or there is a lack of knowledge from the administrators about the meaning of attribute values which has been designed earlier. These attributes have to be discarded though they may be important from an expert point of view.

Unavailability of the Attributes

1. Non-extractability of many attributes from the main data source
   There are some attributes which are important from an expert's point of view, but they are not extractable from the database.
2. Unavailability of some attributes in the main database
   There are some important attributes important to our data mining analysis, but they are not stored and unavailable in the university database. In the following section, a number of these attributes relating to lecturer, student and course knowledge are determined.

Suggestion to Improve the Model's Quality

Based on the expert knowledge it is suggested that the main data base system may be improved by adding the extra attributes. In the following, they are suggested and categorized into the groups.

Student's background knowledge (pre-university academic information):

- average of mathematical subjects such as; calculus, finite math, algebra;
- average of theoretical subjects such as family and society, management, economic, accounting;
- major or degree;
- average of computer science mark.

Student's course knowledge:

- average of mid-grade exam mark;
- attendance rate;
- average of project and assignment mark;
- carried mark itself (to know the performance during the course).

Student's demographics knowledge:

- under any treatment which require consideration;
- any psychological problem;
- family monthly income;
- student monthly cost.

Lecturer academic knowledge:

- lecturer's research activity, journals, book publishing in the teaching area;
- conferences, seminars and workshops in the area of teaching;
- lecturer's attendance rate;
- lecturer's year of experience in teaching the course.

Lecturer's demographic knowledge:

- any disability;
- under any treatment which require consideration.

Course knowledge:

- Are the course materials based on the standard?
- Are there any standard number of assignment and project for the course?
- Is there enough pre-requisite as the basis of the course?

8. Conclusion and Further Works

This study was an attempt to enhance the traditional educational process via data mining technology. The advantages and suitability of this system in higher learning institution has been discussed in detail. The main idea of this analysis is organized into DM-HEDU guideline proposed by the authors, which targets the superior advantages of data mining in higher learning institution. The DM-HEDU is used to analyze the current works of data mining in education and identify the existing gaps and further works. It also provides an opportunity for researchers to learn the existing area of study for data mining in education.

The other main contributions of this study discusses on how the various data mining techniques can be applied to the set of educational data and what new explicit knowledge or models are discovered. The models are classified based on the type of techniques used, including predictive and descriptive. The obtained rules from each model are translated into plain English as a factor to be considered by the managerial system to either support their current decision makings or help them to set new strategies and plan to improve their decision making procedures.

The final results have been analyzed and validated with real situations in a university. The factors affecting the anomalies have been discusses in detail. The final result from each model using various techniques, presented that they are all performing similarly.

As a further work, we would like to enhance other data mining processes in higher learning institution by referring to DM-HEDU analysis guideline. These processes are according to first class priorities of the universities. Other work can be generating student and lecturer models for the other type of course offered in the university. Since the application of data mining brings a lot of advantages in higher learning institution, it is recommended to apply these techniques in other academic institution such as schools, language institutions, institutions for special students and private collages.

References

Agrawal, R.T. and Imielinski, A.S. (1993). Mining association rule between sets of item in large database. In *Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D. C., 207–216.

Cave, M., Kogan, M. and Hanney, S. (1990). The scope and effects of performance measurement in British higher education. In F.J.R.C. Dochy, M.S.R. Segers and W.H.F.W. Wijnen (Eds.), *Management Information and Performance Indicators in Higher Education*. Van Gorcum and Comp, B.V., Assen/Maastricht, 48–49.

Chang, W.H.T. and Lee, Y.H. (2000). Telecommunications data mining for target marketing. *Journal of Computers*, **12**(4), 60–74.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). *CRISP-DM 1.0*: *Step-by-Step Data Mining Guide*.

Chen, M.S., Han, J. and Yu, P.S. (1996). Data mining: an overview from a database perspective. *IEEE Transaction on Knowledge and Data Engineering*.

Delavari, N. and Beikzadeh, M.R. (2004). A new analysis model for data mining processes in higher educational systems. In *MMU International Symposium on Information and Communications Technologies 2004 in Conjunction with the 5th National Conference on Telecommunication Technology 2004*. Putrajaya, Malaysia.

Delavari, N., Beikzadeh, M.R. and Shirazi, M.R.A. (2004). A new model for using data mining in higher educational system. In *5th International Conference on Information Technology Based Higher Education and Training*: *ITEHT '04*. Istanbul, Turkey.

Delavari, N., Beikzadeh, M.R. and Phon-Amnuaisuk, S. (2005). Application of enhanced analysis model for data mining processes in higher educational system. In *6th International Conference on Information Technology Based Higher Education and Training*. Santo Domingo, Dominant Republic.

Edelstein, H. (2000). Building profitable customer relationships with data mining. *SPSS White Paper-Executive Briefing*. Two Crows Corporation.

Feldman, R. (2003). *Mining the Biomedical Literature using Semantic Analysis and Neural Language Processing Techniques*, a link analysis approaches. ClearForest Corporation. New York.

Fielden, J., and Abercromby, K. (2000). *UNESCO Higher Education Indicators Study: Accountability and International Co-operation in the Renewal of Higher Education*. Georgia Professional Standards. UNESCO, Paris.

Fitz-Gibbon, C.T. and Tymms, P. (2002). Technical and ethical issues in indicator systems: doing things right and doing wrong things. *Education Policy Analysis Archives*, **10**.

Gabrilson, S. (2003). *Data Mining with CRCT Scores*. Office of information technology, Geogia Department of Education.

Han, J. and Kamber, M. (2001). *Data Mining*: *Concepts and Techniques*. Simon Fraser University, Morgan Kaufmann publishers.

Han, J. (2002). How can data mining help bio-data analysis. In *BIOKDD02*: *Workshop on Data Mining in Bioinformatics*.

Huffman, J.G. (1997). Estimates of root-mean-square random error for finite samples of estimated precipitation. *Journal of Applied Meteorology*, **36**(9), Maryland.

Johnstone, J.N. (1976). *Indicators of the Performance of Educational Systems*. UNESCO: International Institute for Educational Planning, Paris.

Johnstone, J.N. (1981). *Indicators of Education Systems*. Paris.

Lee, S., Cho, S. and Wong, M.P. (1998). Rainfall prediction using artificial neural networks. *Journal of Geographic Information and Decision Analysis*, **2**(2), 233–244.

Luan, J. (2001). Data mining and knowledge management, a system analysis for establishing a tiered knowledge management model (TKMM). In *Proceedings of Air Forum*, Toronto, Canada.

Luan, J. (2002a). Data mining and knowledge management in higher education – potential applications. In *Proceedings of AIR Forum*, Toronto, Canada.

Luan, J. (2002b). *Data Mining Application in Higher Education*. SPSS Executive Report.

Luan, J., Zhao, C.M. and Hayek, J. (2004). Use data mining techniques to develop institutional typologies for NSSE. *National Survey of Student Engagement*.

Mehta, M., Agrawal, R. and Rissanen, J. (1996). *SLIQ: A Fast Scalable Classifier for Data Mining*. IBM Almaden Research Center.

Oakes, J. (1986). *Educational Indicators: A Guide for Policymakers*. Center for Policy Research in Education, Rutgers University, New Brunswick.

Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. *School Effectiveness and School Improvement*, **1**, 61–80.

Teh, Y.W., Mustaffa, K.M., Zaitun, A.B. and Lee (2002). Data mining in computer auditing. In *Proceedings of the 2002 Informing Science*, Cork, Ireland, June 19–21.

Two Crows Corporation (1999). *Introduction to Data Mining and Knowledge Discovery*, third edition. U.S.A.

UNESCO Nairobi Cluster (2006a). *Analysis and Data Requirements of Core Indicators for Monitoring Education for All Goals*. Kenya.

UNESCO (2006b). *National Education Sector Development Plan*. A result-based planning handbook, January.

Van Petegem, P., Vanhoof, J., Daems, F. and Mahieu, P. (2004). Benchmarking the quality of school education: enhancing the impact of indicators. Accepted for publication in *Assessment in Education*.

Vlãsceanu, L., Grünberg, L. and Pârlea, D. (2004). *Quality Assurance and Accreditation: A Glossary of Basic Terms and Definitions*. Bucharest, UNESCO-CEPES. Papers on higher education.

Waiyamai, K. (2003). *Improving Quality of Graduate Students by Data Mining*. Kasetsart University, Bangkok, Thailand.

Wako, T.N. (1988). *Basic Indicators for Education Systems. A Manual of Methodology*. Ministry of Education, Addis Ababa.

Wako, T.N. (2003). *Basic Indicators of Educational System's Performance, National Educational Statistics Information Systems.* (NESIS)/ UNESCO/ ADEA, Harare, Zimbabawe.

Yang, M., Goldstein, H., Rath, T. and Hill, N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education*, **25**, 469–483.

N. Delavari received Canadian matriculation pre-university degree majored in science from Ontario Canada in 2000. She received BSc (Hons) in Faculty of Information Technology majored in information system engineering from Multimedia University (MMU) in Malaysia. She is completing Msc (IT) in Faculty of Information Technology from the same university by the end of 2006. The topic of her interest is application of various data mining techniques in higher education system.

M.R. Beikzadeh received BEng degree in electronic engineering from Beheshti University in Iran, the master in artificial intelligence from Computer Science Department and, and the PhD degree in artificial intelligence and VLSI design from Electronic System Engineering Department of Essex University, UK in 1992. Until 2001 he was managing and directing various research projects as well as directing research projects in Iran Telecommunication Research Center (ITRC) meanwhile he was teaching in Tehran Universities. He joined Multimedia University in Malaysia in 2001 and his research interests include temporal logic, fuzzy logic, data mining and other Artificial Intelligence (AI) techniques and their applications in e-learning, education and meaning-based systems.

S. Phon-Amnuaisuk received his BEng from King Mongkut Institute of Technology (Hons) and PhD in artificial intelligence from the University of Edinburgh. He is currently an associate dean for the Faculty of Information Technology, Multimedia University, Malaysia where he is the chairman for Centre of Artificial Intelligence and Intelligent Computing. He has also served as a committee member in many editorial boards and research grant screening committees.

Duomenų gavybos taikymai aukštojo mokslo įstaigose

Naeimeh DELAVARI, Mohammad Reza BEIKZADEH, Somnuk PHON-AMNUAISUK

Šiuo metu, aukštosios mokyklos, sprendžia vieną didžiausių problemų – kaip patobulinti sprendimų valdymo kokybę. Kuo mokymasis sudėtingesnis, tuo sudėtingesnis ir šis procesas. Mokymo įstaigos ieško veiksmingesnių technologijų geresniam sprendimų valdymui ir palaikymui, imasi kurti naujų strategijų ir veiksmų planus geresniam dabartinių procesų valdymui. Tobulinant kokybę vienas iš veiksmingų būdų sprendžiant problemas yra suteikti valdymo sistemai naujų žinių, kurios būtų siejamos su mokymo procesais. Šias žinias galima išgauti iš ankstesnių ir šiuo metu operuojamų duomenų mokymo organizacijos duomenų bazėje, taikant duomenų gavybos technologijos metodus. Duomenų gavybos metodai – tai analitinės priemonės, kurias taikant galima išgauti daug vertingų žinių iš didesnių duomenų bazių. Šiame straipsnyje aptariami du aukštojo mokslo sistemos duomenų gavybos būdai: 1) aukštojo mokslo įstaigoms siūlant analitinius nurodymus ir stiprinant jų dabartinių sprendimų procesus, 2) naudojant duomenų gavybos technologijas naujų tikslių žinių sprendimų paieškos procesams gerinti.