Exploiting Student Intervention System Using Data Mining

Samia Oussena, Hyensook Kim
University of West London
London, UK
samia.oussena@uwl.ac.uk, Hyensook.kim@uwl.ac.uk

Tony Clark
Middlesex University
London, UK
Tony.clark@mdx.ac.uk

Abstract—**With the proliferation of systems that are put for the student use, data related to activities undertaken by the student are on the increasing. However, these vast amounts of data on student and courses are not integrated and could therefore not easily queried or mined. Therefore, relatively little data is turned into knowledge that can be used by the institution learning. In the work presented here, different data sources such as student record system, virtual learning system are integrated and analysed with the intention of linking behaviour pattern to academic histories and other recorded information. These patterns built into data mining models can then be used to predict individual performance with high accuracy. The question addressed in the paper is: how can indicators of problems related to student retention produced by data mining be presented in a way that will be effective. A prototype system that integrates data mining with an intervention system based on game metaphor has been build and piloted in the computing school. Early evaluations of the system have shown that it has been well received at all levels of the institution and by the students.**

Keywords-data mining; intervention system; student drop-out; game metaphore

I. Introduction

Whilst student engagement is complex and multi-dimensional, one key aspect for high education institutions is to engage students at a personalised level. There is a proliferation of data related to student activities. Activities might relate to some actions the student has performed such as submitting an assignment or viewing lecture notes. However, these vast amounts of data on student and courses are not integrated and could therefore not easily be exploited. Consequently, little data is turned into knowledge that can be used by the institution learning. Data mining is the field of discovering of implicit and interesting patterns for large data collections [8]. Data mining has been applied to a number of fields including bioinformatics and fraud detection. In recent years, there has been an increased interest in the use of data mining to the educational setting. Data mining has been shown to help predict student educational outcomes [13]. When using Data Mining, the goal is to develop a model, which can infer an aspect of the student academic outcomes, such as passing a module, from a combination of other data that represents student's characteristics. For example, Garbrilson [6] uses data mining prediction techniques to identify the most effective factors in determining student test scores. In [10], authors use a data mining classification technique to predict student's final grades based on their web use. In most of these applications, the results are usually presented to strategic decision makers in graphical form (for example a dashboard), which is interpreted and some intervention programme is then put in place. However the integration of the predictions provided by Data Mining, the presentation of the results produced by applying predictive patterns to live student data and the intervention processes are typically weak. In most cases intervention is provided by humans resulting in an overall process can be very resource intensive. There is therefore a lack of integration between the identification of the problem and the implementation of the intervention actions.

In this paper, we discuss the design of an intervention system based on data mining. We have developed an application that allows the data mining models to be refined as a consequence of intervention actions, as well as involving students in the process. There is evidence that personalising and signposting educational 'moments' contributes to a better learning environment [7]. Although the literature on retention points to the complexity of factors influencing retention, there is evidence that linking social and academic experience, and tailoring the learning environment to individual needs increases an institution's chances of retaining its students [1].

The challenge in designing the application for the student retention has been how to maximise student engagement. Arguably the weakest link in the process arises in ensuring that students at risk are identified as early as possible. Of course, this can be achieved with unlimited resources in the form of tutors who continually monitor raw data sources and who contact students as soon as they detect a problem indicator. However, this is not realistic. Our proposal is to automate the process and to present information to students in a way that will maximise their engagement and therefore reduce the resource burden. Hence our research question is: how can indicators of problems related to student retention produced by data mining be presented to students in a way that will be effective without an unrealistic resource overhead?

The strong widespread appeal of computer and console gaming has motivated a number of researchers to harness the educational potential of gaming [3]. Here, we have looked at using the motivation power of games to encourage students to be involved with the intervention system. Our hypothesis is that: features used by gaming systems can be incorporated into student intervention systems in order to maximise their effectiveness.

ISBN: 978-1-61208-162-5

To test out hypothesis we have designed a prototype system that has the following features: we have designed a uniform data model describing a student profile within a teaching and learning environment; data mining techniques are then used to process the information and to produce rules that represent indicators of failure within the educational process; the rules are then processed against live student data in order to raise potential indicators of failure in real-time; gaming systems have been analysed in order to produce a model whereby information can be presented to students as though their learning experience is a game; this model has been implemented in the form of a web application and the events produced by the rules are fed into the gaming model.

The rest of the paper is as follows: Section 2 discusses some of the student retention work; Section 3 discusses the students and the learning models that we have used; Section 4 discusses our gaming environments; Section 5 discusses the design of the application.

II. Background

Several modelling methods have been applied in educational research to predict student's retention. The more widely used models are the Students attrition model (Bean 1980) and the Tinto student integration model [17]. Tinto's model examines factors contributing to a student's decision about whether to continue their higher education. It claims that the decision to persist or drop out is quite strongly predicted by their degree of academic integration, and social integration. Tinto argues that from an academic perspective, performance, personal development, academic self-esteem, enjoyment of subjects, identification with academic norms, and one's role as a student all contribute to a student's overall sense of integration into the university.

Students who are highly integrated academically are more likely to persist and complete their degrees. The same is true from a social perspective. Students, who have more friends at their university, have more personal contact with academics, enjoy being at the university, and are more likely to make the decision to persist. Bean's model appears to use many of the constructs in Tinto's model but the most significant addition is the inclusion of external factors. These include attitude constructs which might have a direct effect such as finance or indirect such as influence of parents and friends' encouragements [2].

There are also other models of student retention. Thomas developed her model "institutional habitus" [18]), based on Tinto's theory, which can be divided into the academic and the social experiences. The academic experience covers attitudes of staff, teaching, learning and assessment. Different learning styles are supported and diverse backgrounds are appreciated. Tutors are friendly, helpful and accessible. Assessment gives students the opportunity to succeed and staffs are available to help. The social experience combines friendship, mutual support and social networks. Thomas noted that one factor in her students' persistence was the fact they felt more at home with their friends.

Recently, data mining models have also been developed for addressing student retention. For example, in [10], the authors use data mining classification techniques to predict students final grades based on their web-use feature. It can identify students at risk early and allow the tutor to provide appropriate advice in a timely manner. Cerrito applied data mining in mathematics courses, her study demonstrates that retention needs to be of concern at all levels of a student's career at the university, not just for the first year students [4]. Students with high entrance scores and low risk factors may also leave the university before graduation. However, most of these projects are lacking the following up intervention. Seidman's has shown that early identification of students at risk as well as maintaining intensive and continuous intervention is the key to increasing student retention [14]. He also explains how universities can prepare their programs and courses so students will have the greatest probability of success both personally and academically.

In this paper, we argue that the intervention process needs to be integrated with the identification process in order to be effective. The data mining process will help with the early identification, followed by early and continuous intervention, as proposed by Seidman. By combining the two processes, we are able to provide an audit trail of the intervention actions that can be then evaluated for their effectiveness.

III. An intervention system based on gaming environment

Gaming, and particularly on-line gaming has become very popular with young people in recent years. In such systems, players can develop a profile based on playing a number of (possibly collaborative) games. The profile can be tailored to the individual in terms of the look and feel and represents achievements in terms of goals attained, points achieved, extra features unlocked etc. When a game is played, there are a number of features that can be attained such as completing a level or defeating a foe. In general, each game attaches points to the different challenges it presents and, although the games are different, the points are in a universal currency (or at least universal to a specific gaming platform). Points awarded to a specific gamer represents their level of achievement in terms of skills attained and challenges overcome. Gamers can compare their aggregate performance against other gamers to produce a league table; relative positions in a league table can be a powerful motivating factor and gamers can spend a great deal of time trying to move their position up the table. In addition, games can compare their performance at a more fine grain level in terms of specific skills and achievements. A gamer may have a specific interest in achieving a given skill because it is transferrable to another game.

Our proposal is that students can be viewed as gamers and learning outcomes can be viewed as being similar to points awarded when playing games. That being the case, we propose that the same powerful motivating factors that lead gamers to strive to increase their performance (whether relative or absolute) can be applied to students. A number of attributes common to computer games are recognized in fostering active engagement; motivation and a high level of persistence in game play [6]. These include the use of environment that simulates realistic experiences for the player, providing opportunities for identity exploration and play through role play [15]. In a number of games, a player may learn to take on attributes of their avatar [19]. Using avatars may lead to a sense of responsibility towards the character that can lead to educationally relevant outcomes. The other main attribute is the creation of a sense of pride and accomplishment by structuring the game to challenge the player and allow progress.

The above principals have been implemented in the intervention system in order to harness the motivation potential of gaming; including associating academic performance with scores, use of avatar, structure the levels based on learning outcomes and modules and providing a league board.

IV. THE INTERVENTION SYSTEM OVERVIEW

We have built an intervention system that put students as the main actors. Students play a critical role in being successful and subsequently remaining at the university. Studies have indicated that motivation is a prerequisite for student learning [16]. The student can foster this motivation by setting clear and explicit learning goals and understanding the expectation of success. The greater the belief that a task can be accomplished the greater the motivation. In our system, at any point of time the student will be presented with what has been accomplished so far in terms of learning outcomes gained within each module, the modules that have been passed and the marks gained so far. The student is also aware of what is expected in order to acquire his qualification, for example, in terms of modules to be taken, learning outcomes to be gained within a module, and number of assignments. However, this will only make an impact if students are engaged i.e. access and make use of the information. Hence, while modeling the student we looked at both the information predicting their performance as well as information related to their engagement with the system.

The design of the predictive model part is based on two main information channels, as suggested by Tinto, one related to the student and one related to the university. The information related to university is based mainly on information related to courses and their related modules. The design of this part of the model has been constrained by the institution's information that we had access to. The constraints were mainly due to the interpretation of the data protection act. One of the techniques that supported our design was feature importance technique. Here, we have put all the data that we had access to and carried a data mining analysis to help us determine which indicators provide the best assessment of potential student academic performance. As illustrated in Figure 1, the model includes the following types of information:

- Information related to their entry profile such as their qualification and the scores of the tests that they took at their entrance.

- Information on the course that they enrolled on. This will contain a general information related to the course itself such as the modules that are part of the course, the faculty that manages the course, the type of award, and a more specific information related to the student enrolment to the course offering; for example, the year the student has enrolled and how many modules he had to repeat in that course.

- Information related to individual module that the student is enrolled on. Here also it includes a general information related to the module such as the number of assignment for the module, number of credits for the module and the team delivering the module, and information related to the student enrolment to the specific module offering such as the date enrolled and the results for that module.

- The other type of information relates to the interactions the student has with the module resources; such as interaction with the virtual learning site, interaction with library, interaction related to the module assessment such as submission of assignment, results acquired for the assignment.

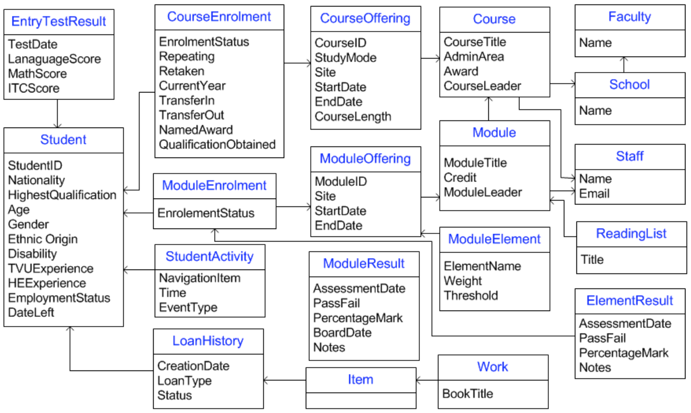


Figure 1. Student model: predictive part

The design of the engagement model part is based on the mapping of student performance to the game attributes. For example, based on their profile, students are associated to one of the groups (Zone) that have been identified by the data mining process. For each of the zone, the data mining process would have identified a threshold for when intervention actions are required. In this model, we have included not only individual performance information but also performance information related to the group, the module and the course.

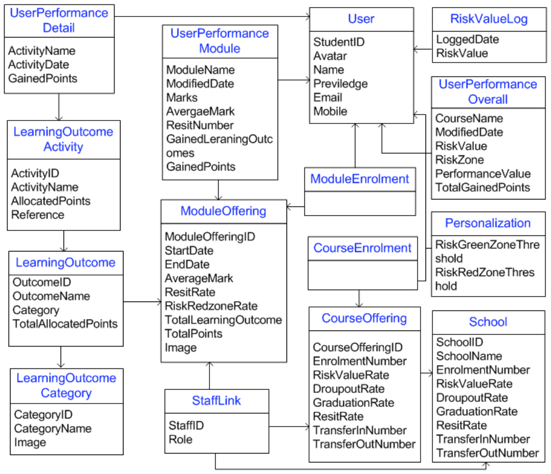


Figure 2. Student model: engagement part

As illustrated in Figure 2, the types of information included in the model are the following:

- Personal information such as their Avatar, their preferred mode of communication.
- Information related to points that they have acquired and need to acquire.
- Information related to module; these include all the learning outcomes that are associated to it, the activity that allows the achievement of the learning outcome and the number of points associated to it. They will get predefined points for doing activities such as accessing blackboard, borrowing a book from library, or submitting an assignment. The blackboard visits in a day for one particular module will be counted as 1 visit. If the student visits blackboard to access resources for another module on the same day, another visit will be added related to that module. Similarly they will obtain points when they obtain marks for each of the assessment. They will also unlock learning outcomes for the modules when they do the activities related to the learning outcome. They will unlock all the learning outcomes for a module when they pass the module.

The architecture of the system is illustrated in Figure 3. The system has been built using Oracle technology [11] and includes three main components; the data warehouse component, the data-mining component and the intervention application. We have used oracle workflow system to execute data warehouse update and run data mining engine. Currently the workflow runs every Friday. The intervention reporting system runs on its own scheduler after data mining process.

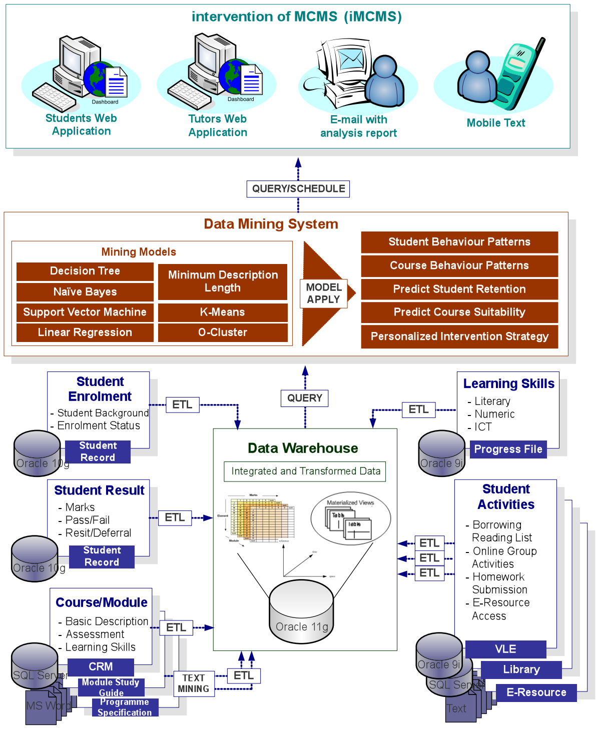


Figure 3. MCMS System Architecture

A. *The data warehouse component*

The data warehouse component takes the data provided by the institution data sources and builds the data warehouse; i.e. the different dimensions of the cubes. The data sources used for the system are discussed in the next section. We have defined four cubes in this study; one for students, one for student activities, one for modules and the other one for courses.

B. *Data sources*

When analyzing the data sources, we have identified information that will give us a good understanding of the students' profile; i.e. data related to information before their entry to the university (student background), data related to their interaction in the university; including their goals (student interaction), and data related to their results. The key sources used are as follows:

- The **student record system** relates to the profile of the student prior to joining the institution. Student profile will include information such as their entrance level, their ethnicity, literacy and numeracy test entrance score. In addition the student's assessment profile including marks and the number of re-sits taken is maintained by the record system.
- The **library system** and the **reading list system** captures information about a student's book loan activities and links this to the reading lists set for the modules that the student is registered for.

- The **online learning system** and **e-library** records how often the student logs into the system and the use of the various pages in the *virtual learning environment* (VLE). In our case the VLE can capture the number of hits on individual pages and document downloads.
- The **module study guides** provide information such as reading lists and the schedule of assessments for each module. The **course specification** provides information such as regulations about options and assessment hurdles. The study guides and course specifications also provide lists of learning outcomes that can be gained by the students when they pass individual modules or module elements. Individual assessments are broken down into different types. Institution departments that own course components are recorded in addition to the tutor responsible for delivering the module.
- The **marketing system** and **entrance test system** provides information about entry requirements and whether students have had any additional tutoring on entry. For example overseas students may be offered extra tuition in English and the test system will record whether they took up this offer and, if so, the results.

C. *The data mining component*

Educational data mining is a newly emerging discipline and there have been reports of a number of demonstration applications in Spanish [13] and American [9] universities, and particularly in distance-learning institutions [11]. Data Mining has already proved to be successful in e-commerce and bio-informatics, where results are achieved through the use of associators, classifiers, clusterers, pattern analysers, and statistical tools. In the educational context, data mining provides analysis of the students' behaviour, navigation, frequency, and length of interaction with the e-Learning system that can identify patterns of behaviour and associations. It can classify students into groups depending on their learning behaviour rather than just ability. At the same time it also identifies students exhibiting atypical behaviour that needs early intervention and feedback. The overall process that we have used in this system is illustrated below. At each iteration, data is fed to the data mining process, in order to identify potential drop-out students and their performance. The intervention process will then identify specific intervention actions for these students.

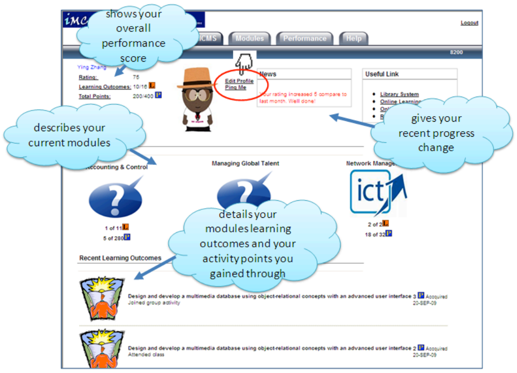
The data mining process that we have adopted includes three main stages: finding the features' relations, data grouping, and making the prediction [20] The first stage involves applying features' importance and associate rules to find the correlation among data features. This stage helps eliminate any data that is unlikely to make any impact on subsequent tasks. For example, we found that age and gender is not related to students' performance (results or drop-out), whereas the VLE interaction is related to students' performance.

The next stage involves classification of data and extraction rules and patterns. Here we identify groups of students that have shown related features and will require similar intervention. Examples of such groups include first year full time students, students transferring from other institutions, non-UK students, students with low library usage or post-graduate students. The third and final stage includes applying regression to the identified groups in order to predict future behaviours i.e. allowing us to predict the potential students that are at risk of dropping out, or their academic performance requires attention. Here we identify students that are underperforming or the ones that have shown an improvement in their performance.

The data mining models have been built based on three years of historical data before being integrated into live feed data. The data were divided in two sets, one for building the models and one for validating the model. Once the models have been validated, the data warehouse is updated every week. The prediction stage of the process is conducted on the updated data leading to new alerts and update of the data.

D. *Intervention Application Component*

This component presents the results of the data mining models and implements the intervention system. We have used different views for different stakeholders. One view is targeted towards improving students' performance and providing them with a holistic and a detailed view of their performance. Students at risk of dropping out receive an intervention message via email or SMS. The other view is to support the academics in implementing their intervention policies. The data here is presented at different levels of abstractions depending on their role (tutor, programme leader, and head of school). Any of the intervention content is generated according to a predefined and personalised intervention rule.



Ying Zhang
Rating: 75
Learning Outcomes: 10/16 L
Total Points: 200/400 P

Your rating increased 5 compare to last month. Well done!

1 of 11 L
5 of 280 P

2 of 2 L
18 of 32 P

Design and develop a multimedia database using object-relational concepts with an advanced user interface 3 P Acquired 20-SEP-09
Joined group activity

Design and develop a multimedia database using object-relational concepts with an advanced user interface 2 P Acquired 20-SEP-09
Attended class

Figure 4. Student Intervention Application

Students are encouraged to use avatars that have characters that they would like to relate to. Being associated with a strong avatar might motivate them to acquire high scores. Student activities are mapped to module learning outcomes activities in order to convert student performance as counted points. For example, if a student borrowing a book from the reading list, he or she would get points for his or her module and if he or she submits an assignment, he or she gets points as well. If the students reach the target points for each learning outcome of a module, they would see their mission achievement records in the web application and would encounter the next challenge for further improvement such as a usual role-playing game. We have also implemented a league table where scores can be compared within the same group (Zone). A screenshot of the implemented student web application is shown in Figure 4:

Students are alerted by email or text if they are in at-risk zone. In the message, they are presented with their achievements so far in terms of the points. The message also advises them with a number of actions that they need to undertake in order to get out of the zone. A student that has taken the right action and has managed to get out of the zone is also alerted. The message that they receive is a congratulatory message that invites them to review their achievements and show them how they can unlock the next level.

The design of the messages and the actions that the students are invited to undertake are associated with each of the groups that were identified in the second stage of the data mining process. For each group, based on the pattern that describes the group, a template for the message is designed, with specific actions that will lead to correcting the low scores in the particular interaction, such as accessing the module VLE content. The templates are then personalised with live data specific to each student. The intervention messages are common to all modules, however each module leader is encouraged to refine the rules that will lead to alerts and customise the intervention messages and actions.

V. DISCUSSION AND CONCLUSION

Through this paper, we have explored the benefitting factors of using a gaming environment in order to engage and motivate students in terms of retention. The prototype system has been used to integrate the data mining process with the intervention system. The data mining process is used as two main functions; primarily to predict those students who are most likely to drop-out early, and secondly to group students into specific categories that can be targeted with personalised interventions if it is predicted that a drop-out is imminent. Certain activities and accomplishments merit a specific number of points to be added on the appropriate student's game profile. The student unlocks each level in the gaming process through achieving a set score for each learning outcome. This idea of levels is a great source of motivation to students, as well as the idea of a 'leader board' in which students' are encouraged to become competitive. This also enforces the gaming environment, making 'winning' appear more appealing than in a classic university environment.

The system has been piloted for a semester in the computing school and the institution is planning to pilot it in more schools in the next academic year. The early evaluation of the system has mainly been done through presentation and focus groups. The questions that we tried to address in these focus groups are for example; how do student perceive the intervention application impact on learning and their academic performance? Are the predictions of the data-mining models accurate? Is the gaming metaphor a good motivation tool? How do tutors and university staff perceive the impact of the intervention application on the student learning and academic performance? What improvements were made to some of the learning and teaching processes? The early results have shown that the system has been well received and the prediction have been very similar to what the tutor expected. We are planning to do a more in depth quantitative data analysis in order to track student engagement and survey their perceptions of the environment.

In terms of further improving the system there are a few problems that must be targeted. For example, in each iteration, the data mining models must be manually retrained. This is a hindrance and can be rectified by implementing an algorithm that creates self-adapting models. Another area to expand upon is the range of data sources available. We hope to add new data sources, such as financial data, timetabling and data relating to their social involvement in universities. This has not initially been possible due to data constraints. Currently, another part of the system that has not been discussed in this paper, involves the monitoring of courses and modules. This has given us some insight into the performance of the education processes and one of our planned works is to investigate the relationship between the performance of the education processes and the student performance. The other area for improvement will involve creating an additional gaming feature (i.e. settings), providing a user-friendly interface enabling students to access key information, such as course information, university regulations, and their personal timetable. This idea of integrating learning and games could be further developed through literacy and numeracy features to support their main subject. This is once again could be incorporated into the gaming environment, similar to many educational games and consoles available in the current market.

REFERENCES

[1] K. Anagnostopoulou, and D. Parmar, Practical guide: bringing together e-learning & student retention, London: Cats Ltd. ISBN: 978-1-85924-301-5.

[2] J. Bean, Dropouts and turnover: The synthesis of a causal model of student attrition. Research in Higher Education, 12, pp. 155-187.

[3] S. DeCastell and J. Jenson, Serious play, Journal of curriculm Studies 35(6), pp. 649-665

[4] P. Cerrito, Data Mining Student Performance in Mathematics Courses, CinSUG Third AnnualOne Day Conference. Committee of Public Accounts, (2002) Fifty-eighth Report of Session Improving Student Achievement and Widening Participation in Higher Education in England, HC 588,2001-02.

[5] S. Gabrilson, Data Mining with CRCT Scores. Office of information technology, Geargia Department of education.

[6] R. Garris, R. Ahlers, and J. Driskell, Games, motivation, and learning: A research and practice model. Simulation and Gaming: An Interdisciplinary Journal, 33(4), pp. 441–467.

[7] L. Harvey, S. Drew, and M. Smith, The First-year Experience: A Review of Literature for the Higher Education Academy, http://www.heacademy.ac.uk/research/Harvey_Drew_Smith.pdf

[8] W. Kloesgen and Zytkow, Handbook of Knowledge Discovery and Data Mining, Oxford University Press, Oxford, 2002.

[9] J. Luan, Data mining and knowledge management in higher education –potential applications. In Proceedings of AIR Forum, Toronto, Canada.

[10] B. Minaei-Bidgoli, G. Kortemeyer and W. Punch, Enhancing Online Learning Performance: An Application of Data Mining Methods, From Proceeding of Computers and Advanced Technology in Education .

[11] E. Mor and J. Minguillón, E-learning personalization based on itineraries and long-term navigational behavior, Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters,pp. 264-265 May 19-21, New York, NY, USA .

[12] Oracle http://www.oracle.com/technology/products/bi/odm/index.html [accessed on 25/05/2011]

[13] C. Romero and S. Ventura, Data mining in e-learning, Southampton, UK: Wit Press.

[14] A. Seidman, Retention Revisited: RET = E Id + (E + I + C)Iv. College and University, 71(4), pp-18-20.

[15] K. Squire, H. Jenkins, W. Holland, H. Miller, A. O'Driscoll and K. Tan, Design principles of next-generation digital gaming for education. Educational Technology, 700 33, pp.17–23.

[16] M. Svinicki, A. Hagen and D. Meyer, How to Reaserch on Learning Strengthens instruction, Teaching on Solid Ground: Using Scholarship to Improve Practice, Jossey-Bass publishers, San Francisco, CA.

[17] V. Tinto, Dropout from Higher Education: A Theoretical Synthesis of Recent Research, Review of Educational Research vol.45, pp. 89-125,1975.

[18] L. Thomas, Student retention in higher education: the role of institutional habitus", Journal of Education Policy, Vol. 17 No. 4, August, pp. 423-442, 2002.

[19] N. Yee, and J. Bailenson, The Proteus effect: The effect of transformed self-representation on behavior. Human Communication Research 33, pp. 271-290.

[20] Y Zhang, S. Oussena, T. Clark. and H. Kim, Use Data Mining To Improve Student Retention in Higher Education – A Case Study, to be presented in 12th International Conference on Enterprise Information Systems(ICEIS) pp. 190-197.