Information Systems & Grid Technologies

Fifth International Conference ISGT'2011

Sofia, Bulgaria, May 27 – 28., 2011.

Proceedings

ST. KLIMENT OHRIDSKI UNIVERSITY PRESS

ISGT’2011 Conference Committees

Co-chairs

- Prof Ivan SOSKOV
- Acad Kiril BOYANOV

Program Committee

- Míchéal Mac an AIRCHINNIGH, Trinity College, University of Dublin
- Pavel AZALOV, Pennsylvania State University
- Marat BIKTIMIROV, Joint Supercomputer Center, Russian Academy of Sciences
- Marko BONAČ, Academic and Research Network of Slovenia
- Marco DE MARCO, Catholic University of Milan
- Milena DOBREVA, University of Strathclyde, Glasgow
- Viacheslav ILIN, Moscow State University
- Vladimir GETOV, University of Westminster
- Jan GRUNTORÁD, Czech Research and Academic Network
- Pavol HORVATH, Slovak University of Technology
- Arto KARILA, Helsinki University of Technology
- Dieter KRANZMUELLER, University of Vienna
- Vasilis MAGLARIS, National Technical University of Athens
- Ivan PLANDER, Slovak Academy of Science
- Dov TE'ENI, Tel-Aviv University
- Stanislaw WRYCZA, University of Gdansk
- Fani ZLATAROVA, Elizabethtown College

Organizing Committee

- Vladimir DIMITROV
- Maria NISHEVA
- Kalinka KALOYANOVA
- Vasil GEORGIEV

Vladimir Dimitrov   Vasil Georgiev (Eds.)

Information Systems & Grid Technologies

Fifth International Conference ISGT'2011

Sofia, Bulgaria, May 27 – 28., 2011.

Proceedings

St. Kliment Ohridski University Press

Preface

This conference was being held for the fifth time in the end of May, 2011 in Sofia, Bulgaria. It is supported by the National Scientific Research Fund, by the University of Sofia "St. Kliment Ohridski" and by the Bulgarian Chapter of the Association for Information Systems (BulAIS). The Organizing Committee consists of scientist from the Faculty of Mathematics and Informatics of the University of Sofia. Traditionally this conference is organized in cooperation with the Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences.

Total number of papers submitted for participation in ISGT'2011 was 34. They undergo the due selection by at least two of the members of the Program Committee. This book comprises 22 papers of 20 Bulgarian authors and 17 foreign authors included in one of the three conference tracks. Responsibility for the accuracy of all statements in each peer-reviewed paper rests solely with the author(s). Permission is granted to photocopy or refer to any part of this book for personal or academic use providing credit is given to the conference and to the authors.

The editors

ISSN 1314-4855

St. Kliment Ohridski University Press

Table of Contents

Information Systems

Intelligent Systems

Distributed Systems

Author Index

Application of Data Mining Techniques to the Students Advising and the Course Plan Construction

Zlatinka Kovacheva,

Higher College of Telecommunications and Post, Sofia, Bulgaria
Higher College of Technology, Muscat, Oman,
zkovacheva@hotmail.com

ABSTRACT. Some of the biggest challenge problems that higher education faces today are related to the advising of students and planning the alumni. It is necessary to predict how many students will enroll in particular course programs, and how to assist the students in order to graduate successfully. One way to effectively address these challenges is using data mining techniques. The aim of this paper is to present some methods of extracting of the relevant information from the existing Academic Database.

Keywords: data mining, advisory, recommendation, filtering, clustering, prediction, association

1 INTRODUCTION

Advising students is an important part of the teaching process in order to provide to all students high quality learning and training opportunities in line with the requirements of the market needs and the specific interests and capabilities of each student. As a result, students have flexibility in controlling the academic subjects they take every semester. It also reflects on the College course plan and timetable.

Academic advisors play an important role in the educational progress of their advisees by monitoring their progress, as well as helping them to clarify their educational goals. Every student is assigned an academic advisor after enrollment in the College. According to the registration plan set up in each semester, the student is personally responsible for registration, under the supervision of the assigned academic advisor.

The academic advisor is responsible to guide the students to choose appropriate courses depending on the level and status of his /her advisee, the audit degree, the subjects offered in the current semester. He has to take care of the prerequisites of a course before advising students to register that particular course, giving preference to the failed courses in the previous semester. If a student has a good GPA, the advisor can recommend him/her to register for more courses next semester, paying attention to the student's interests and skills. He is the one who meets the students regularly and discusses with them their studies and problems.

During this process the advisor faces a lot of questions from his/her advisee whose answers are not unique. For example: "How many courses to choose each semester?"; "Which courses are easier and need less efforts to complete successfully?"; "Which elective course is preferable?"; "What is the best order of the courses offered at one academic level?", etc. The answers of such questions are personally oriented and need analysis and evaluation of the results of the historical practice.

Course plan construction and time table preparation for each semester are very complicated tasks. Their successful implementation depends on the prediction of the number of students attending each course and the resources availability. The problem becomes more difficult when the number of elective courses is increasing. As the goal of the College management is to provide flexible and relevant education, satisfying the society requirements and students preferences, it is recommended to use information technologies in the process of solving such problems.

The College Academic Database contains detailed historical information about the students' preferred courses and the level of their successful completion. In our Higher College of Technology (HCT) we have been collecting such information since 2006. Unfortunately, this huge amount of collected data is not usually used for the advisory purposes because of the lack of the relevant software tools for data analyses. The solution of this problem is in the application of data mining techniques.

2 DATA MINING

One of the most comprehensive definitions of data mining is Gartner Inc.'s as "the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories, and by using pattern recognition technologies, as well as statistical and mathematical techniques." Data mining should be performed on very large or raw datasets using either supervised or unsupervised data mining algorithms.

Data mining enables organizations to use their current reporting capabilities to uncover and understand hidden patterns in vast databases. These patterns are then built into data mining models and used to predict individual behavior with high accuracy. As a result of this insight, institutions are able to allocate resources and staff more effectively. Data mining may, for example, give an institution the information necessary to take action before a student drops out, or to efficiently allocate resources with an accurate estimate of how many students will take a particular course. Data mining encompasses different algorithms that are diverse in their methods and aims. It also comprises data exploration and visualization to present results in a convenient way to users.

Data mining is a powerful tool for academic intervention. Through data mining, a university could, for example, predict which students will or will not graduate. The university could use this information to concentrate academic assistance on those students most at risk. In order to understand how and why data mining works, it is important to understand a few fundamental concepts.

Data mining relies on four essential methods: *classification, categorization, estimation, and visualization.* Classification identifies associations and clusters, and separates subjects under study. Categorization uses rule induction algorithms to handle categorical outcomes, such as "persist" or "dropout," and "transfer" or "stay." Estimation includes predictive functions or likelihood and deals with continuous outcome variables, such as students GPA. Visualization uses interactive graphs to demonstrate mathematically induced rules and scores, and is far more sophisticated than pie or bar charts. Visualization is used primarily to depict three-dimensional geographic locations of mathematical coordinates. Higher education institutions can use classification, for example, for a comprehensive analysis of student characteristics, or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success.

The basic Data Mining techniques include:

2.1. PREDICTION

In prediction, the goal is to develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). Prediction requires having labels for the output variable for a limited data set, where a label represents some trusted "ground truth" information about the output variable's value in specific cases. In some cases, however, it is important to consider the degree to which these labels may in fact be approximate, or incompletely reliable [4].

Prediction has two key uses. In some cases, prediction methods can be used to study what features of a model are important for prediction, giving information about the underlying construct. This is a common approach in programs of research that attempt to predict student educational outcomes without predicting intermediate or mediating factors first [15]. In a second type of usage, prediction methods are used in order to predict what the output value would be in contexts where it is not desirable to directly obtain a label for that construction (for example, in previously collected repository data, where desired labeled data may not be available, or in contexts where obtaining labels could change the behavior being labeled, such as modeling affective states, where self-report, video, and observational methods all present risks of altering the construction being studied).

Broadly, there are three types of prediction: classification, regression, and density estimation. In classification, the predicted variable is a binary or categorical variable. Some popular classification methods include decision trees, logistic regression (for binary predictions), and support vector machines. In regression, the predicted variable is a continuous variable. Some popular regression methods include linear regression, neural networks, and support vector machine regression. In density estimation, the predicted variable is a probability density function. Density estimators can be based on a variety of kernel functions, including Gaussian functions. For each type of prediction, the input variables can be either categorical or continuous; different prediction methods are more effective, depending on the type of input variables used.

2.2. CLUSTERING

In clustering, the goal is to find data points that naturally group together, splitting the full data set into a set of clusters. Clustering is particularly useful in cases where the most common categories within the data set are not known in advance. If a set of clusters is optimal, within a category, each data point will in general be more similar to the other data points in that cluster than data points in other clusters. Clusters can be created at several different possible grain-sizes: for example, students could be clustered together (to investigate similarities and differences between them), or student actions could be clustered together (to investigate patterns of behavior) [1], [3], [6].

Clustering algorithms can either start with no prior hypotheses about clusters in the data (such as the k-means algorithm with randomized restart), or start from a specific hypothesis, possibly generated in prior research with a different data set (using the Expectation Maximization algorithm to iterate towards a cluster hypothesis for the new data set). A clustering algorithm can postulate that each data point must belong to exactly one cluster (such as in the k-means algorithm), or can postulate that some points may belong to more than one cluster or to no clusters (such as in Gaussian Mixture Models).

The goodness of a set of clusters is usually assessed with reference to how well the set of clusters fits the data, relative to how much fit might be expected solely by chance given the number of clusters, using statistical metrics such as the Bayesian Information Criterion.

2.3. RELATIONSHIP MINING

In relationship mining, the goal is to discover relationships between variables, in a data set with a large number of variables. This may take the form of attempting to find out which variables are most strongly associated with a single variable of particular interest, or may take the form of attempting to discover which relationships between any two variables are the strongest.

Broadly, there are four types of relationship mining: association rule mining, correlation mining, sequential pattern mining, and causal data mining. In association rule mining, the goal is to find if-then rules of the form that if some set of variable values is found, another variable will generally have a specific value. For example, a rule might be found of the form {student is frustrated, student has stronger goal of learning than goal of performance} _ {student frequently asks for help}. In correlation mining, the goal is to find (positive or negative) linear correlations between variables. In sequential pattern mining, the goal is to find temporal associations between events – for example, to determine what path of student behaviors leads to an eventual learning event of interest. In causal data mining, the goal is to find whether one event (or observed construction) was the cause of another event (or observed construction), either by analyzing the covariance of the two events or by using information about how one of the events was triggered [18]. For example, if a pedagogical event is randomly chosen using automated experimentation and frequently leads to a positive learning outcome, a causal relationship can be inferred [4], [12].

2.4. DISTILLATION of DATA for HUMAN JUDGMENT

Another area of interest within educational data mining is the distillation of data for human judgment. In some cases, human beings can make inferences about data, when it is presented appropriately, that are beyond the immediate scope of fully automated data mining methods.

The methods in this area of educational data mining are information visualization methods – however, the visualizations most commonly used are often different from those most often used for other information visualization problems, owing to the specific structure, and the meaning embedded within that structure, often present in educational data [9], [10].

Data is distilled for human judgment in educational data mining for two key purposes: identification and classification. When data is distilled for identification, data is displayed in ways that enable a human being to easily identify well-known patterns that are nonetheless difficult to formally express. For example, one classic educational data mining visualization is the learning curve, which displays the number of opportunities to practice a skill on the X axis, and displays performance (such as percent correct or time taken to respond) on the Y axis. A curve with a smooth downward progression that is steep at first and gentler later indicates a well specified knowledge component model.

Alternately, data may be distilled for human labeling, to support the later development of a prediction model. In this case, sub-sections of a data set are displayed in visual or text format, and labeled by human coders. These labels are then generally used as the basis for the development of a predictor. This approach has been shown to speed the development of prediction models of complex phenomena such as gaming the system by around 40 times, relative to prior approaches for collecting the necessary data [5].

2.5. DISCOVERY with MODELS

In discovery with a model, a model of a phenomenon is developed via prediction, clustering, or in some cases knowledge engineering (within knowledge engineering, the model is developed using human reasoning rather than automated methods). This model is then used as a component in another analysis, such as prediction or relationship mining [4].

In the prediction case, the created model's predictions are used as predictor variables in predicting a new variable. In the relationship mining case, the relationships between the created model's predictions and additional variables are studied. This can enable a researcher to study the relationship between a complex latent construction and a wide variety of observable constructions.

Often, discovery with models leverages the validated generalization of a prediction model across contexts. Generalization in this fashion relies upon appropriate validation that the model accurately generalizes across contexts.

3 APPLICATION of DATA MINING to the STUDENTS ADVISING

The most frequent questions that students ask the advisors, whose answer is not always unique are of the type:

- How many courses should I take in each semester? (Or a related question:
  What is better for me – to take more courses in the first semester, or in the second or in summer?)
- Is it better for me to study Course A before Course B or it doesn't matter?
- When I have a chance to repeat one course to increase my GPA, which one should I repeat?
- I failed three courses and I can choose only two of them to repeat in summer. Which two do I have to choose?
- I want to finish the level earlier, but I'm afraid that if I take more courses in the summer semester, I will not be able to complete them successfully? What can you advise me?
- I need a specific GPA to pass to the next level and I can choose 2 courses to repeat. Which courses should I repeat?
- To increase my GPA next semester, what will you advise me – to repeat one course or to concentrate my efforts to complete other courses more successfully?
- I can't register this semester for one of the courses because all sections are already full. I have to change it with another, but which one should I choose?
- Which specialization do I have to choose on the higher diploma level, according to my presentation on the first two levels?

The answer of such questions is difficult and depends on the students' success and interests. But it is recommended to extract from the College database the results of the students with similar characteristics during a long time period and advise the student according to these historical results.

One solution is to apply *association rule mining, creating a system of production rules*.

The College registration system has been collecting information since 2006. The database of the system stores a huge amount of data for the students and the courses they have been taken in the chronological order (by the number of the semester), including their results (passed or failed) and grades (from A to F) (cf. Fig.1).

For each student of the College, the system provides his/her personal data, the courses he/she had taken each semester, the teacher and the credit hours of each course, the mark and grade of the student for each course. If a student repeated one course one or more times, all his marks and grades for this course are stored. After each semester, the GPA and cumulative GPA (up to this moment) are calculated.

For each course, the credit hours, the passing grade and all prerequisite courses are available.

This information is structured for the semesters of all levels of the educational process – certificate, diploma, higher diploma and bachelor level.

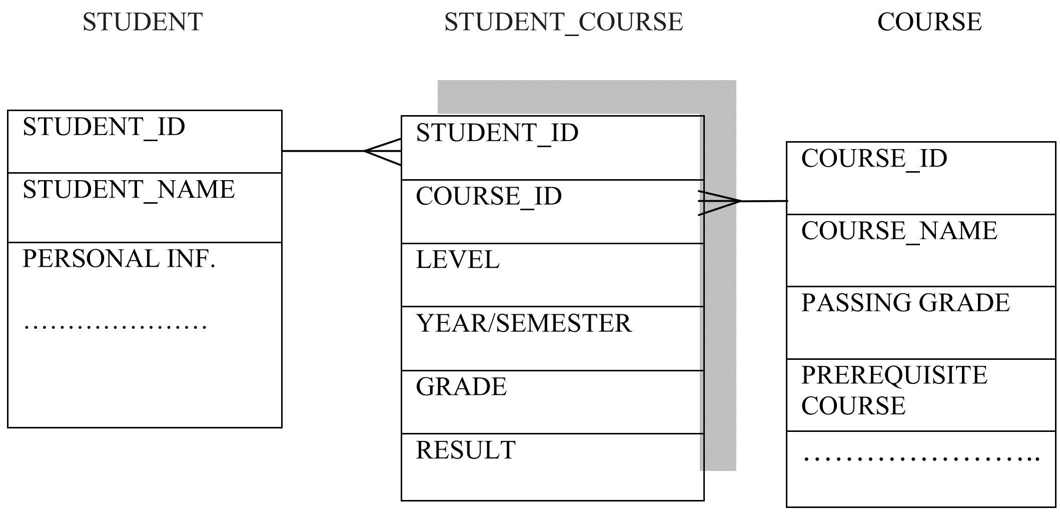

Fig.1. Information provided by the College database

On the base of this information it is recommended to create a Rule Based Filtering System, applying information search and retrieve filtering techniques. Analyzing different students' profiles, we can discover interesting models of students' behavior and performance.

For example:

From the group of students with similar GPA before taking a Course A (giving priority to the prerequisite courses) we can estimate the accuracy of successful completion of the same Course A for a STUDENT X. The production rule is of the following type:

IF Course name = A
and starting the new semester CGPA of Student X $\in [3, 3.5]$
THEN Student X will complete Course A successfully with accuracy 98%

From the group of students who enrolled more than 5 courses with similar GPA before starting a new semester we can estimate the accuracy of successful completion of all enrolled courses for a STUDENT X. The production rule is of the following type:

IF Number of enrolled courses > 5
and starting the new semester CGPA of Student X $\in [2, 2.5]$
THEN Student X will complete all enrolled courses successfully with accuracy 65%

From the group of students who enrolled Course A before Course B and completed Course A with similar grade we can estimate the accuracy of successful completion of Course B after Course A for a STUDENT X. The production rule is of the following type:

IF Grade of Course A is greater than C and starting the new semester CGPA of Student X $\in [2.5,\ 3]$

THEN Student X will complete Course B successfully with accuracy 75%

By analogy, from the group of students who enrolled Course B before Course A and completed Course B with similar grade we can estimate the accuracy of successful completion of Course A after Course B for a STUDENT X.

Comparing both results, we can advise the student for the order of enrollment of both courses.

From the group of students with similar CGPA who repeated Course A, we can estimate the accuracy of increasing the CGPA for a STUDENT X. The production rule is of the following type:

IF The student X repeats Course A

and starting the new semester CGPA of Student X $\in [1.5,\ 2]$

THEN Student X will increase his GPA next semester with accuracy 95%

The structure of this process is illustrated on Fig.2.

A similar idea is used by Cesar Vialardi, et al. [19].

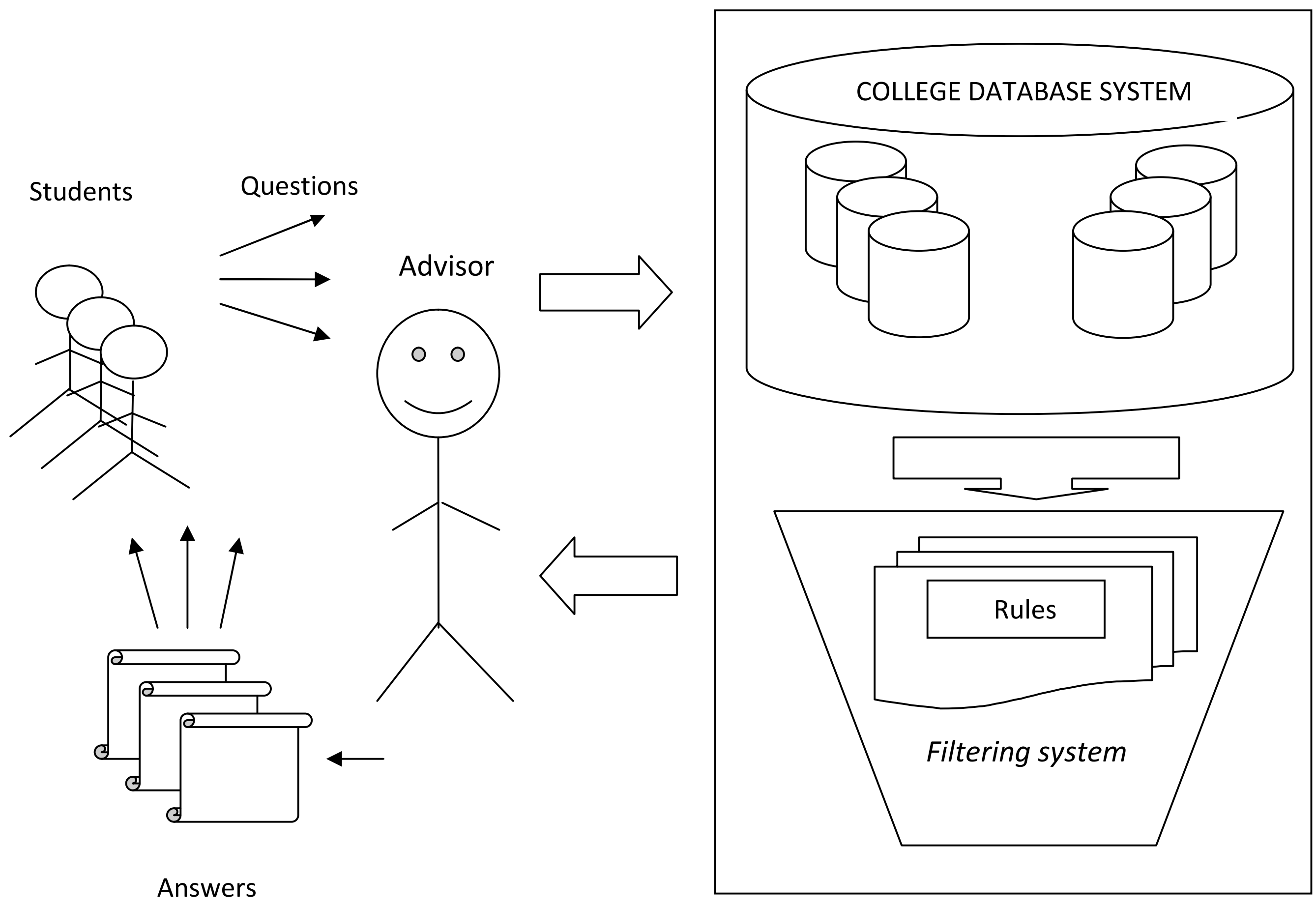

Fig.2. Structure of the advisory process using Rule Based Filtering System

4 CONCLUSION

Students' advising is a very responsible duty. The variety of students' interests and skills makes the decisions taking process very complicated and individually oriented. The successful implementation of this process depends on the experience of the advisors, on their knowledge of many subjects, as well as on their ability of understanding the students' particularities and goals. The proposed Data Mining method can save a lot of time of the advisors and help them in the process of taking the right decisions concerning students' performance in the College, their successful completion of all levels of education and their future professional realization.

REFERENCES

1 .Amershi, S., Conati C.: Automatic Recognition of Learner Groups in Exploratory Learning Environments, Proceedings of ITS 2006, 8th International Conference on Intelligent Tutoring Systems (2006)

2. Agrawal R, Imielinski T., Swami A.: Mining Association Rules between Sets of Items in Large Databases, ACM SIGMOD Conference on Management of Data, pp. 207-216 (1993)

3. Arabie. P., Hubert J., De Soete G.: Clustering and Classification (Eds.), World Scientific Publishers Company, London (1996)

4. Baker, R.S.: Data Mining for Education. McGaw, B., Peterson, P.,Baker, E. (Eds.), International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier, (in press)

5. Baker, R.S., de Carvalho, A.M. :Labeling Student Behavior Faster and More Precisely with Text Replays, Proceedings of the First International Conference on Educational Data Mining, pp.38-47 (2008)

6. Beal, C., Qu, L. , Lee H.: Classifying learner engagement through integration of multiple data sources, Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-2007)

7. Bravo J., Vialardi C., Ortigosa A: A Problem-Oriented Method for Supporting AEH Authors through Data Mining, Proc. of International Workshop on Applying Data Mining in E-learning (ADML'07) held at the Second European Conference on Technology Enhanced Learning (EC-TEL 2007), pp. 53-62 (2007)

8. Erdogan S., Z., Timor, M.: A Data Mining Application in a Student Database, Journal of Aeronautics and Space Technologies, July 2005, vol. 2, Number 2, pp. 53-57 (2005)

9. Hershkovitz, A., Nachmias, R. :Developing a Log-Based Motivation Measuring Tool, Proceedings of the First International Conference on Educational Data Mining, pp.226-233, (2008)

10. Kay, J., N. Maisonneuve, Yacef K. and Reimann P.: The Big Five and Visualisations of Team Work Activity, Proceedings of Intelligent Tutoring Systems (ITS06), M. Ikeda, K. D.Ashley & T-W. Chan (eds). Taiwan. Lecture Notes in Computer Science 4053, Springer-Verlag, pp.197-206 (2006)

11. Luan J, :Data Mining Applications in Higher Education, SPSS Executive report, hctp//www. spss.com http://www.spss.ch/eupload/File/PDF/Data%20Mining%20Applications%20in%20Higher%20Education.pdf

12. Mostow, J.: Experience from a Reading Tutor that listens: Evaluation purposes, excuses, and methods, C. K. Kinzer & L. Verhoeven (Eds.), Interactive Literacy Education: Facilitating Literacy Environments Through Technology, pp. 117-148. New York: Lawrence Erlbaum Associates, Taylor & Francis Group, (2008)

13.Romero C., Ventura S.: Data Mining in E-Learning, WIT Press, Ashurst Lodge, Ashurst, Southampton, SO40 7AA, UK, http://www.witpress.com/contents/c1523.pdf

14. Romero C., Ventura S.: Educational Data Mining, a Survey from 1995 to 2005, Expert Systems with Applications, 2007, 33(1), pp. 135-146 (2007)

15. Romero, C., . Ventura S., Espejo P.G., Hervas C.: Data Mining Algorithms to Classify Students, Proceedings of the First International Conference on Educational Data Mining, 8-17, (2008).

16. Sarwar B., Karypis G., Konstan, J., Riedl, J. : Item – Based Collaborative Filtering Recommendation Algorithms, Proceedings of the 10-th International Conference on World Wide Web, pp. 285-295 (2001)

17. Schafer J.B. : The application of data-mining to recommender system, J. Wang (Eds.), Encyclopedia of data warehousing and mining, Hershey, PA, Idea Group, pp. 44-48 (2005)

18. Scheines, R., Sprites, P., Glymour, C., Meek, C. :Tetrad II: Tools for Discovery. Lawrence Erlbaum Associates: Hillsdale, NJ (1994)

19. Vialardi C., Bravo J., Shafti L., Ortogosa Al.: Recommendation in Higher Education using Data Mining Techniques, Educational Data Mining (2009)