Research Phases of University Data Mining Project Development

Dorina Kabakchieva[1], Kamelia Stefanova[2], Valentin Kissimov[3],
and Roumen Nikolov[4]

[1] Sofia University „St. Kl. Ohridski“, 125 Tzarigradsko shosse Blvd., Sofia, Bulgaria
dorina@fmi.uni-sofia.bg
[2,3] University of National and World Economy, Sofia, Bulgaria
kamelia@fmi.uni-sofia.bg, vkisimov@gmail.com
[4] State University of Library Studies and Information Technologies, Sofia, Bulgaria
nroumen@abv.bg

Abstract. Educational Data Mining becomes one of the challenging new research fields where data mining methods and tools could help universities in taking timely and data analysis based management decisions, thus contributing to gaining competitive advantages in their successful policy introduction. This paper presents the research activities performed for the implementation of a data mining project initiated in one of the most prestigious Bulgarian universities. The project main goal is to reveal the high potential of data mining applications for university management, referring to the optimal usage of data mining methods and techniques to deeply analyze the collected historical data. That will lead to better understanding the student behavior and building well structured educational process that meets the university policy and supports the management decision making process.

Keywords: educational data mining, university data mining, data mart

1 Introduction

Data Mining is one of the most important steps in the process of Knowledge Discovery in Databases, a relatively new and very quickly developing area of the information industry. It refers to discovering new and potentially useful knowledge, extracted from the large quantities of data, generated and collected during the organizations' every day operations, to support decision making. The effective use of data and information is not only an important factor for the survival and normal functioning of organizations in the modern knowledge-based society, but it is also crucial for remaining competitive and ensuring prosperity.

Educational Data Mining becomes one of the challenging new areas where data mining methods and tools application could help universities in taking timely and content rich management decisions, strongly supported by multi dimensional data analysis, thus contributing to gaining competitive advantages in their successful policy introduction.

The main purpose of this paper is to present the research activities performed for the implementation of a data mining project initiated in one of the most prestigious Bulgarian universities.

2 Educational Data Mining

The implementation of data mining techniques and tools for solving problems within the area of higher education is a new stream in the data mining research field and the software industry that is known as “educational data mining”. The educational data mining research community [1] is constantly growing. It has started by organizing workshops since 2004, then conducting an annual International Conference on Educational Data Mining (the first conference being held in Montreal, Canada, on June 20-21, 2008), and now already having a Journal on Educational Data Mining (the first issue being published in October 2009). There are already a large number of research papers discussing various problems within the higher education sector and providing examples for their successful solutions reached by using data mining.

Two papers are providing an extensive literature review of the educational data mining research field. The first one is published in 2007 by Romero and Ventura [2], providing an overview of the research efforts in the area between 1995 and 2005. Baker and Yacef, the authors of the second paper [3] published in October 2009, review the history and current trends in educational data mining.

Nowadays, higher education institutions are operating in a very complex and highly competitive environment. They are constantly working for gaining competitive advantages over their business competitors. Moreover, modern universities are organizations that are collecting and keeping large volumes of data, referring to their students, the organization and management of the educational process, and other managerial issues. The analysis of these unique types of data by applying data mining methods and tools, and the extraction of new knowledge, could substantially contribute to improving universities’ performance and to better decision making.

Universities, as representatives of innovative institutions, ready to analyze the dynamic processes and successfully manage changes, have been learning how to process the large volumes of data that they are in possession. As mentioned in [2], this data is coming from two types of educational systems – traditional classroom and distance education. In the institutions following the traditional form of education, data is collected at the admission of new students, during the organization and implementation of the educational process, for management issues, etc., and it is usually stored in databases or in a data warehouse. The main sources of data in distance education are web log files, providing information about the learners’ navigation in the web-based education systems. The implementation of data mining techniques for these two types of educational systems differs because of the different data sources and information systems used, and the specific goals and objectives followed. We focus mainly on the traditional form of education in the proposed project because it is most relevant for the Bulgarian educational environment.

The application of data mining in educational systems can be addressed for contribution to different stakeholders – students, educators, administrators, [2] managers, governmental institutions. Students could be supported by being recommended different learning resources, activities, tasks or even different learning paths. Educators can get more objective feedback and insight about the educational process that could help them to improve the content of the courses, to select adequate methods for the content provision, to differentiate students, based on their needs in guidance and monitoring, taking into consideration their learning abilities and peculiarities. Administrators could benefit from getting the right information at the moment when it is needed, thus supporting the decision-making process with adequate analysis. University managers could be assisted in their strategic tasks by receiving deep multi dimensional analysis that could reveal trends and opportunities for improving the effectiveness and efficiency of the university management.

The problems in the higher education institutions, that are most often attracting the attention of researchers and becoming the reasons for initiating data mining projects, are focused mainly on retention of students, improving institutional effectiveness, enrollment management, targeted marketing, and alumni management. Using data mining for predicting student drop-out, finding the factors predicting student failure, maximizing student retention, are considered in a large number of research papers [4,5,6,7,8,9,10,11]. The improvement of student performance and institutional effectiveness is the research focus in [12,13,14,15,19]. The development of enrollment prediction models by applying different data mining methods is discussed in [16,17]. Targeted marketing based on data mining models characterizing the best performing students at a university is the research topic in [2,3,16,17]. Data mining application for alumni management is addressed in [7,14].

3 University Data Mining Research Project

Universities today are challenged to meet the society and business requirements for adequately educating the future specialists that could fulfill the needs for advanced professional knowledge, technological expertise and open minded human behavior. Universities are becoming strong pillars of the knowledge-based society development and the most comprehensive institutions not only for delivering sophisticated knowledge and skills, but also for introducing innovative management approaches, supported by progressive methods and techniques.

These main directions insist on implementation of flexible universities management and optimal performance results. In order to survive within the rapid changing environment today and make salutary analysis for discovering promising trends, universities could much benefit from implementing the data mining methods to process the collected data.

The research goal that the initiated project explores is devoted to revealing the high potential of data mining applications for university management, referring to the optimal usage of data mining methods and techniques to deeply analyze the collected historical data. That will lead to better understanding the student behavior and building well structured educational process that meets the university policy and supports the management decision making process.

The main project goal will be achieved through the fulfillment of the following specific objectives:

- To provide a clear picture of the students admitted at the University and to find out what are the characteristic features of the successfully admitted students;
- To analyze the students' profile development and to characterize the students according to their performance at the University;
- To analyze the university management decision making process, based

on the analysis of the most successful in student performance years, compared to those years showing the worse students results.

The main research questions that will be answered during the project implementation are:

- Who become students at the University?
- Who are the students that develop successfully throughout the educational process, who fail and who could perform better?
- What are the most successful in student performance years and what are the management factors that could be associated with that success?

The results achieved when answering the first research question will be used for targeting future university campaigns to those potential university candidates that best match the defined profile of successful students. The outcomes of the second research question will be useful for improving institutional effectiveness by better implementation of the educational process, considering different student profiles, learning capabilities and preferences. The answers to the third research question would support the university management to take important administrative decisions concerning the organization of the educational process.

3.1 The Research Approach

In theory and practice different approaches exist for the realization of a data mining project. The most famous of them are usually associated with the vendors of the data mining software, e.g. the 5A's approach (Assess, Access, Analyze, Act, Automate) of SPSS Clementine, the SEMMA approach (Sample, Explore, Modify, Model, Assess) of SPSS Enterprise Miner, etc. In 1999-2000, a consortium of data mining software vendors and end-users developed the CRISP-DM (Cross-Industry Standard Process for Data Mining) model as a non-propriety, freely available, and application-neutral standard for data mining projects.

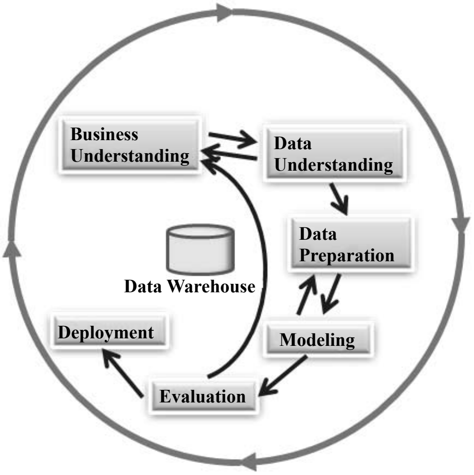

Fig. 1. The CRISP-DM (Cross-Industry Standard Process for Data Mining) Approach.

The CRISP-DM model is considered a standard approach for data mining projects and it is widely used during the last ten years. For these reasons, it is chosen as a research approach to be followed for the realization of the initiated data mining project at the University. It is a cyclic approach (see Fig.1), including six main phases– Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment. There are a number of internal feedback loops between the phases, resulting from the very complex non-linear nature of the data mining process and ensuring the achievement of consistent and reliable results.

The software tools that will be used for the project implementation are the data mining open source software packages WEKA (v3.6) and RapidMiner (v5.0).

3.2 The Project Implementation

The initiated University data mining project has started recently and the research work is still in the initial stage. The activities performed and the results achieved during the first two phases are briefly presented below.

During the *Business Understanding Phase* an extensive literature review was performed in order to study the existing problems at higher education institutions that have been solved by the application of data mining techniques and methods in previous research projects. Formal interviews with representatives of the University management at university, faculty and departmental levels were also conducted, for finding out the specific problems at the University which have not yet been solved but are considered very important for the improvement of the University performance and for more effective and efficient management. Some insights were gathered from informal talks with lecturers, students and representatives of the administrative staff (IT experts and managers). Based on the outcomes of the performed research, the project goal and objectives, and the main research questions were formulated.

The stated project goal and objectives, and the formulated research questions, were then transformed into specific data mining tasks. These are mainly tasks for classification and description that could be solved by using a great variety of data mining methods (decision trees, regression, neural networks, Bayesian classification, K-nearest neighbour method, etc). The logical architecture, on which the realization of the data mining solutions will be based, is presented on Fig.2.

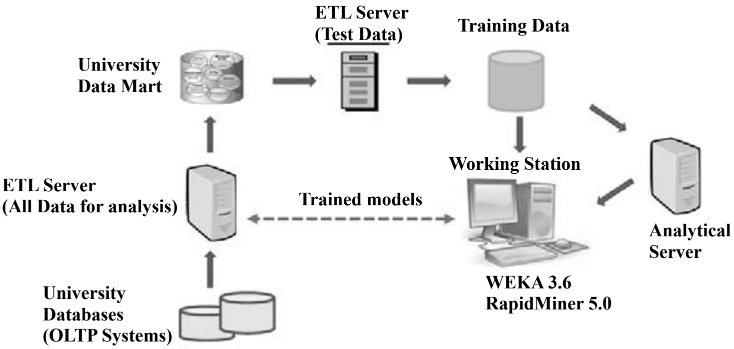

Fig. 2. Logical Architecture of a Data Mining Solution.

The main components of the logical architecture are the OLTP (On-Line Transaction Processing) Systems (University Databases) where the university's operational raw data is collected, the data warehouse (University Data Mart) in which the university's historical data is stored and organized in a specific manner that is providing opportunities for extensive data analysis, and the analytical environment including the various tools for data analysis (e.g. queries and reports, OLAP tools, data mining tools), and Graphical User Interface (GUI) for data visualization and graphical representation of the achieved results.

The project implementation is currently in its second *Data Understanding Phase*. This phase started with a study of the application process for student enrollment at the University, including the formal procedures and application documents, in order to identify the types of data collected from the university applicants and stored in the university databases in electronic format. The rules and procedures for collecting and storing data about the academic performance of the university students, including drop-outs and stop-outs, were reviewed as well. The research continued with discussions with representatives of the administrative staff that is responsible for the university data collection, storage and maintenance. University data is basically stored in two databases. All the data related to the university admission campaigns is stored in the first database, including personal data of university applicants (names, addresses, secondary education scores, selected admission exams, etc.), data about the organization and performance of the admission exams, scores achieved by the applicants at the admission exams, data related to the final classification of applicants and student admission, etc. All the data concerning student performance at the university is stored in the second database, including student personal and administrative data, the grades achieved at the exams of the different subjects, etc.

Standard reports are usually generated from the two databases, supporting the University administrative management. However, the available data is not being subjected to extensive analysis for getting an overall picture about the university performance and a better insight about the educational and management processes. There is no data mart or data warehouse being developed yet, and that is an important prerequisite for the effective analysis of the available historic data at the university.

Since the data warehouse or a data mart is one of the main components of the logical architecture of any data mining solution, the project implementation continues with the development of a data mart where the university data, which will be used for the implementation of the formulated data mining tasks, is organized.

4 Creation of the UNSS Data Mart

There are several reasons for starting the initiated data mining project by building a university data mart. First of all, only the university data that is needed for the purposes of the data mining analysis, and is relevant for the achievement of the formulated project goal and objectives, will be incorporated in the data mart. This data is actually collected and stored in the two available university databases described above. The data mart creation will make possible the integration of that data in a single data source and its structuring in formats that are suitable for data analysis. Another important reason for the data mart development is the need for careful data preprocessing which is an important prerequisite for achieving reliable results from the data mining analyses. One of the most important tasks within the data preprocessing phase includes data cleaning, referring to treating missing values, mistakes, inconsistencies, redundancies, etc. The software tools for data mart development usually support data cleaning functions and could be successfully used for ensuring clean data for the analyses. Last but not least, when the data needed for the analyses is stored in a university data mart, the risks of data damage due to continuous interaction with the operational University databases will be minimized.

Having in mind the formulated project goal, objectives and research questions, the following data will be organized within the developed university data mart:

- *Student personal data*: faculty number, gender, age at enrollment (will be calculated based on the birth date or the unique personal identification number), constant address (city/village, municipality, region, country), secondary school (city, profile), year of secondary education graduation.
- *Student pre-university data*: scores achieved at the admission exams, successful admission exam, and total score at admission.
- *Student performance data*: length of university education (based on the enrollment and graduation dates), form of education, specialty, total graduation score.

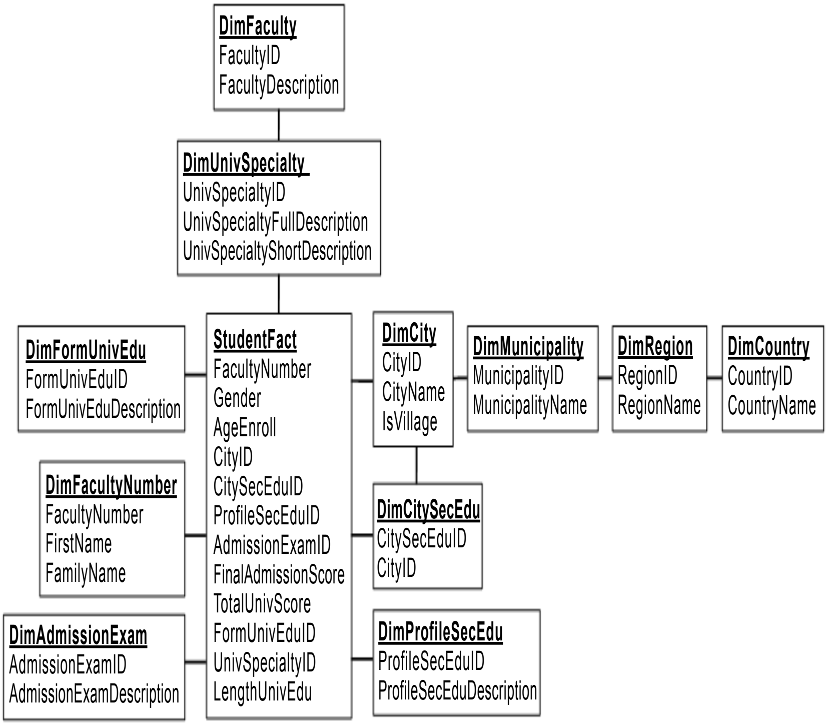

Fig. 3. University Data Mart Structure.

The next step in the university data mart development is the organization of the data into the data mart components: measures, dimensions, hierarchies, and attributes. The snow-flake schema has been chosen for building the structure of the University data mart. It is preferred to the star schema because it has all the advantages of good relational design - each level of a hierarchy is stored in a separate dimension table, there is no duplicate data and it is, therefore, easier to maintain. The developed University data mart structure is presented on Fig.3.

5 Conclusions and Future Steps

A research data mining project is initiated at one of the most prestigious Bulgarian Universities in order to reveal the opportunities that data mining applications could provide for improving university management effectiveness and for achieving optimal performance results. The project implementation is based on the CRISP-DM approach considered standard for data mining project development and fulfillment. The rationale for the project initiation and the research work performed during the first two phases of the project implementation, Business Understanding and Data Understanding, are presented in the paper. The future steps are related to the implementation of the next phases, including population of the developed University data mart, data preprocessing, the application of data mining methods using the open source software packages WEKA 3.6 and RapidMiner 5.0, evaluation of the developed models, and deployment.

References

1. Educational Data Mining Research Community, http://www.educationaldatamining.org/
2. Romero, C., Ventura, S. Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications 33, 2007, pp.135-146
3. Baker, R., Yacef, K. (2009). The State of Educational Data mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, Vol.1, Issue 1, Oct. 2009, pp.3-17
4. Dekker, G., Pechenizkiy, M., Vleeshouwers, J. (2009). Predicting Students Drop Out: A Case Study. Conference Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09), 1-3 July 2009, Cordoba, Spain, pp.41-50
5. Gao, H. (2005). Who Are the Students Who Left? Answers from the Answer Tree: A First-Year Retention Study from a Private 4-year College. AIR 2005 Forum, 29 May – 1 June 2005, San Diego
6. Herzog, S. (2006). Estimating Student Retention and Degree-Completion Time: Decision Trees and neural Networks Vis-à-vis Regression. New Directions for Institutional Research, No.131, Fall 2006, Wiley Periodicals Inc., pp.17-33
7. Luan, J. (2004). Data Mining Applications in Higher Education. SPSS Executive Report, 2004
8. Shyamala, K., Rajagopalan, S. (2007). Mining Student Data to Characterize Drop out Feature Using Clustering and Decision Tree Techniques. International Journal of Soft Computing 2 (1), 2007, pp.150-156
9. Superby, J. Vandamme, J., Meskens, N. (2006). Determination of Factors influencing the achievement of the first-year university students using data mining methods. Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006). Jhongli, Taiwan, pp.37-44

10. Wang, M. (2007). Using Data Mining Techniques to Predict Student Development and Retention. Presentation included in the SAS Conference Proceedings: Midwest SAS User Group 2007, 2007-10-28/2007-10-30, Des Moines, Iowa, USA
11. Yu, C., DiGangi, S., Jannasch-Pennell, A., Lo, W., Kaprolet, C. (2007). A Data-Mining Approach to Differentiate Predictors of Retention. Educational Resources Information Center (ERIC), ERIC No.ED496657, February 2007. Available at: http://www.eric.ed.gov
12. Kotsiantis, S., Pintelas, P. (2004). A Decision Support Prototype Tool for Predicting Student Performance in an ODL Environment. International Journal of Interactive Technology and Smart Education (ITSE), Vol. 1, Issue 4, Nov. 2004, pp 253-263
13. Kumar, N., Uma, G. (2009). Improving Academic Performance of Students by Applying Data Mining techniques. European Journal of Scientific Research, ISSN 1450-216X, Vol.34, No.4 (2009), pp.526-534
14. Luan, L. (2002). Data Mining and Knowledge management in Higher Education – Potential Applications. Presentation at AIR Forum, Toronto, Canada, 2002. Available at: www.cabrillo.edu/services/pro/oir_reports/DM_KM2002AIR.pdf
15. Minaeli-Bidgoli, B., Kashy, D., Kortemeyer, G., Punch, W. (2003). Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System LON-CAPA. 33rd ASEE/IEEE Frontiers in Education Conference, 5-8 Nov 2003, Boulder, CO
16. Nandeshwar, A., Chaudhari, S. (2009). Enrollment Prediction Models Using Data Mining. Available at: http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf
17. Noel-Levitz White Paper (2008). Qualifying Enrollment Success: Maximizing Student Recruitment and Retention Through Predictive Modeling. Noel-Levitz, Inc., 2008
18. Shyamala, K., Rajagopalan, S. (2006). Data Mining Model for a Better Higher Educational System. Information Technology Journal 5 (3), 2006, pp.560-564
19. Vialardi, C., Bravo, J., Shafti, L., Ortigosa, A. (2009). Recommendation in Higher Education Using Data Mining Techniques. Conference Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09), 1-3 July 2009, Cordoba, Spain, pp. 190-199