A Prototype System for Educational Data Warehousing and Mining[1]

Nikolaos Dimokas, Nikolaos Mittas, Alexandros Nanopoulos, Lefteris Angelis
Department of Informatics,
Aristotle University of Thessaloniki
54124, Thessaloniki, GREECE

Abstract

Universities are encountering growing demands by legislators and communities who are clamoring for valuable information about student achievement and university system accountability. Not only are universities required to measure annual progress for every single student, but government (through ministries for education) aid is directly linked to these results. The department of Informatics of Aristotle university of Thessaloniki has developed a data warehouse solution that assists the analysis of educational data. In this paper we present the design and development of the proposed data warehouse solution, which facilitates better and more thorough analysis of department's data. The proposed system constitutes an integrated platform for a thorough analysis of department's past data. Analysis of data could be achieved with OLAP operations. Moreover, we propose a thorough statistical analysis with an array of data mining techniques, that are appropriate for the examined tasks..

1. Introduction

Universities are encountering growing demands by legislators and communities who are clamoring for valuable information about student achievement and university system accountability. Not only are universities required to measure annual progress for every single student, but government (through ministries for education) aid is directly linked to these results.

The department of Informatics of Aristotle university of Thessaloniki (AUTH) maintains (like all the other departments of university do) an operational database in which the data concerning students, professors, courses, curriculum etc are being stored. The database includes the appropriate data in order the department to operate efficiently and effectively. The data include students' features like names, register numbers and grades, professors' names, courses, directions, etc. However, the operational database suffers from the lack of recording past data. Therefore, every time the database describes the current status of the information concerning the department. That has as consequence, that during the execution of an update or deletion process, the previous information is vanished and there is no appropriate action in order to retrieve it. The functionality of database could be compared with the functionality of database belonging to a bank that is registered the current status of customers. Additionally, the same functionality is appeared in a electronic booking system where the number of available tickets for a specific flight is being registered etc.

According to the above features and taking into account that it would be desirable the storage of great number of information (the cost of storing data has been minimized), data warehouse considered the appropriate solution. In this paper we present the design and development of a data warehouse that facilitates better and more thorough analysis of department's data. The proposed system constitutes an integrated platform for a thorough analysis of department's past data. Analysis of data could be achieved with OLAP operations. Moreover, we propose a thorough statistical analysis with an array of techniques that are appropriate for the examined tasks.

Several universities, especially in US, use data warehouse solutions; see [1] for a complete list. There exist industrial solutions for organizing educational data, for example, [2, 3, 4]. Nevertheless, we believe that for the purposes of national higher education, especially in the advent of significant changes in the forthcoming years, customized solutions should be followed, which assure factors like privacy, maintenance, change of demands, etc.

The first step, during data preparation, is the data extraction from the operational database. The second step is the appropriate transformation of data. In third step, the data are inserted in a temporal database,

[1] Work supported by National Project ΕΠΕΑΕΚ ΙΙ (Ενίσχυση Σπουδών Πληροφορικής στο ΑΠΘ) MIS77020.

which forms the preparation area of data. Finally, data are loaded to the data warehouse.

The rest of this paper is organized as follows: in Section 2 we present the dimensions modeling of the procedures. In Section 3, we present the produced reports and implementations issues and in Section 4 we present the statistical analysis of dataset; finally Section 5 concludes the paper.

2. Dimensional modeling

Initially, we present dimensions modeling of the first procedure that employs the following steps.

1. The most important procedure is the grade procedure, that take place during the operation of a department. Every semester period, students took grades in various semester courses.

2. The level of details is the highest possible one. Thus, every individual grade of a student is being stored. The storage of each individual grade offers us a better flexibility in order to process and analyze the data.

3. According to the selected level of detail, every row of the fact table stores the student's grade in a specific course. From the query: "with which information do we describe each row of the fact table?" it follows that we can use information about examined students, classes, and grades.

4. From the query: "what do we measure in each row of the fact table?" it follows that numerical facts are the mean grade, and the number of times that a student is examined per class.

Firstly, we defined the total set of attributes for each dimension table and discover the relationships among them. We exploited the relationships and constructed the hierarchies among attributes. The hierarchies that have been found for the table "Student" are the followings: "Quinquenniad (from registration date)→ Academic year of registration → Age → ID (AM in greek)" and "Curriculum name → Direction name → ID (AM)". Additionally, the hierarchies that have been found for the table "Course" are the followings: "Curriculum name → Direction name → Semester → Type → Course name → Course's code". Finally, for the table "Grade time" discovered the following hierarchies: "Quinquenniad (from registration date) → Academic year→ Grade period". The star scheme, that is being developed, is presented in Figure 1.

Secondly, we present the second procedure that employs the following steps.

1. An interesting procedure is the graduation procedure that takes place during the operation of a department. Analyzing these data we can deduce the difficulty of diploma acquisition, the students behavior etc.

2. The level of details is the highest possible one. Thus, the graduation degree for every student is being stored. The storage of each diploma degree offers us a better flexibility in order to process and analyze the data.

3. According to the selected level of detail, every row of the fact table stores the student's graduate degree in a specific course. From the query: "with which information do we describe each row in the fact table?" it follows that this information is the graduated student and the year of graduation.

4. From the query: "what do we measure in each row of the fact table?" it follows that numerical facts are mean degree grade and the number of graduates at each graduation period.

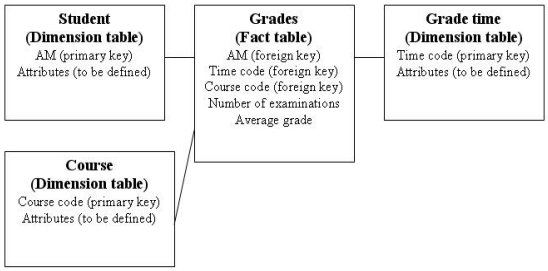

Figure 1. Star scheme for grades procedure

The star scheme, which is being developed for the second procedure, is presented in Figure 2. The hierarchies that have been discovered for the table "Graduate" are the followings: "Quinquenniad (from registration date) → Academic year of registration → Age → ID (AM)" and "Curriculum name → Direction name → ID (AM)". For the table "Graduation time" appears a time hierarchy. This hierarchy includes: "Quinquenniad → Academic year of graduation → Graduation period".

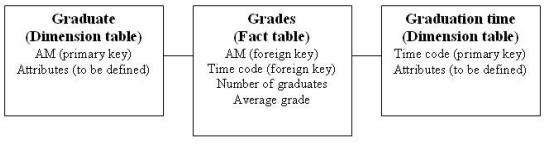

Figure 2. Star scheme for graduate procedure

3. System implementation and reports

After the design of data warehouse that has been presented in the previous section, we developed a great number of reports. The reports revealed some very interesting features that related to the department. In

the interest of space we present a small subset of the queries and reports (Figure 3 and Figure 4) obtained.

1. Average grade per student and grade time. The attributes of students follows the hierarchy; "Curriculum → Direction → AM". The attributes of grade time follows the hierarchy; "Quinquenniad → Academic year → Grade period".

2. Average grade per student and grade time. The attributes of students follows the hierarchy; "Quinquenniad (from registration date) → Academic year of registration → Age → ID (AM)". The attributes of grade time follows the hierarchy; "Quinquenniad → Academic year → Grade period".

3. Average graduation degree per student and graduation time. The attributes of graduated students follow the hierarchy; Quinquenniad → Academic year → Age → ID (AM)". The attributes of graduation time follow the hierarchy: "Quinquenniad → Academic year → Graduation period".

4. Number of graduated students per graduation period: The attributes of graduated students follow the hierarchy; "Quinquenniad → Academic year → Age → ID (AM)". The attributes of graduation time follow the hierarchy: "Quinquenniad → Academic year → Graduation period".

The implementation is divided in to major phases. Initially, it has been implemented a web-based system in order to achieve the preparation of data. Data are extracted from the operational database in a number of Microsoft Excel files and are uploaded to the system. The inserted data are transformed with an automated way in order to produce the final Microsoft Excel file. The constructed file constitutes the input for the data server. The file construction includes the integration of data that appear in the different primitive files. Additionally, the data attributes that are no interest are removed and the files are merged. After the merging, we execute the data transformation. The transformation includes the uniform presentation of data since in many situations the same information is being presented with different ways. Finally, the missing values are filled in with a default value.

Μέσος Όρος Βαθμού	Column Labels			
Row Labels	1992-1997	1997-2002	2002-2007	Grand Total
ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ 1992-1999	7,41	7,22	6,35	7,22
ΓΕΝΙΚΗ ΚΑΤΕΥΘΥΝΣΗ	7,41	7,22	6,35	7,22
ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ 2000-2003		6,50	6,23	6,28
ΓΕΝΙΚΗ ΚΑΤΕΥΘΥΝΣΗ		3,65	4,94	4,91
ΔΙΚΤΥΑ-ΕΠΙΚΟΙΝΩΝΙΕΣ-ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΥΠΟΛΟΓΙΣΤΩΝ		6,29	6,20	6,21
ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ		6,85	6,71	6,75
ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ & ΕΠΙΚΟΙΝΩΝΙΩΝ ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ		6,04	6,13	6,12
ΨΗΦΙΑΚΑ ΜΕΣΑ		6,87	6,78	6,79
Grand Total	7,41	7,01	6,24	6,67

Figure 3. Example (in greek) of grades per student and academic year

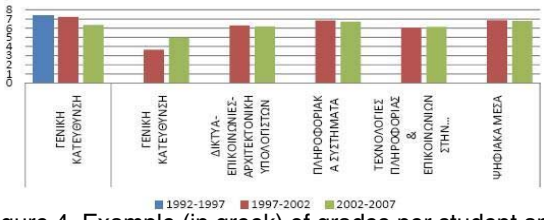

Figure 4. Example (in greek) of grades per student and direction of studies

In the second phase, it has been implemented a middle level database. We also used specific tools, in order to produce data cubes, execute OLAP procedures and present the results

We used Microsoft SQL Server 2005 and the middle level database that has been developed as data server. Additionally, the software package SQL Server 2005 is used during the data storage and update operations. We also exploited OLAP Server (Analysis Services), which is offered with the software platform called Microsoft Visual Studio 2005. In OLAP Server executed the operations that are related to the business logic of the data warehouse application. It is also responsible for the appropriate load of data from the middle level database to data warehouse. OLAP server used as well in order to process the dimensions, to construct the data cubes and to execute the OLAP operations. Finally the software package Microsoft SharePoint Server 2007 is used as presentation server. We have developed a web site for the purposes of presentation of reports. Although the reports could be created and presented with the use of Microsoft Excel, we preferred a web-based solution since it could be achieved better usability and interaction between the user and the system. User could choose the data cubes presentation through the web site.

4. Statistical analysis of the dataset

In this Section, we briefly present some of the results from the statistical analysis of our dataset [5]. The dataset contains both demographic and performance information for 1184 students. Initially, we made all the preliminary descriptive statistical analysis for both the qualitative and quantitative variables of the dataset. Summarizing the results, we can highlight that there are 769 (64.9%) males and 415 (35.1%) females, whereas 1039 (87.8%) students were born in Greece and 145 (12.2%) in another country. Moreover, 488 (41.2%) are graduated students with a mean value 4.56 for the variable Duration that shows the time needed to accomplish their studies.

Secondly, multivariate procedures were applied in

order to identify relationships between the variables. More precisely, we used the contingency (cross-tabulation) tables that form two-way and multi-way tables and we found the joint distribution of the qualitative variables. In addition, statistics and measures were evaluated to identify the association for the aforementioned factors. Furthermore, we performed one-way ANOVA in order to check the impact of every factor on the scale variables (duration of studies and final grade). We also used post hoc tests (Tukey, Tukey's-b, Bonferroni, LSD, Scheffe and Duncan) in order to identify the various homogeneous categories that have to be concatenated in every factor. These tests compare each category of a factor with all the other categories in the same factor and designate the significant differences. An interesting subject is that the final grades are similar for males and females. On the contrary, there seems to exist a statistically significant difference between the males and females for the scale variable Duration (females have smaller mean value).

We also attempted to investigate the relationship between the scale variables Final grade and Duration through the evaluation of Pearson correlation coefficient and scatter plot (Figure 5), whereas the Ordinary Least Squares (OLS) regression was performed to model the aforementioned relationship. The Pearson's correlation coefficient has high negative value (rho=-0.413, sig<0.001) so there is a strong negative correlation between the variables. We built the model setting the Final grade as the response and the Duration as the independent variable. The adjusted R^2 was 0.169, whereas the only independent variable was statistically significant. The Equation of the model is

$$Grade = 8.980 - 0.326 \times Duration$$

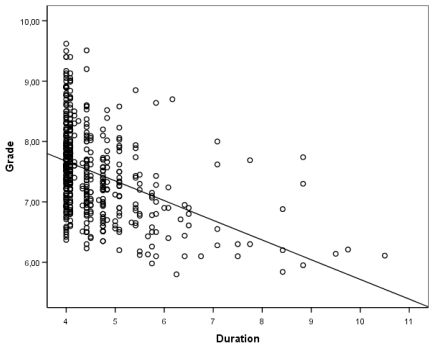

Figure 5. Example (in greek) of grades per student and academic year

An interesting topic of our analysis was the investigation of the distribution of the duration for the students to accomplish their studies and the identification of factors that affect it. For this purpose, we utilized a statistical methodology known from biostatistics as survival analysis (SA) [6].

The main goal is to study the distribution of the time it takes for a critical or terminal event to occur (survival time) for a population of individuals. In our case, the analogue is the distribution of the duration for the undergraduate students to accomplish their studies, estimated from a data sample. SA is especially useful when in the data sample there are cases where the terminal event has not been occurred by the time of the analysis. This is a common situation in a university where there are students that have accomplished their studies (graduate students) and there are students, which have passed in the university but they have not accomplished their studies, yet. The general method can take into account even cases that have been lost to follow-up. In our case, these are the students that have started their studies but afterwards have lost contact and so there is uncertainty about the terminal event (graduation).

As we have already mentioned, we also utilized traditional statistical analysis for the duration of the undergraduate students to accomplish their studies. In the traditional statistical analysis (ANOVA), the aforementioned cases (no-graduate students) are ignored. The durations are considered as measures of only the graduate students while the durations of students that have not accomplished their studies are overlooked as missing values. However, it is possible to exploit all available information, even from no-graduate students by setting the problem in the framework of a more general methodology formulated for time variables.

The time to the terminal event of our interest (duration of studies) is a positive random variable, denoted by T. The probability density function is $f(t)$ and the cumulative distribution function is

$$F(t) = P(T \le t) = \int_0^t f(u)du$$

The duration function (generally survival function) is defined as:

$$S(t) = P(T > t) = 1 - F(t)$$

This is the probability of a student to accomplish the studies longer than time t or else the probability that the student will accomplish the studies after time t. The graphical representation of the non-increasing function $S(t)$ is called the duration curve (generally survival curve) and shows the course of the duration. A steep curve indicates short duration while a gradual or flat curve is an indication for longer duration. In our

study, we utilized the Kaplan-Meier estimate which is the most common method of evaluating the survival function from a sample of individuals.

As we have already mentioned, there are 1184 students, 488 (41.2%) of which are graduates and 696 (58.8%) have not accomplished their studies (censored cases), yet. In Figure 6, we can see the corresponding survival function for the dataset. The survival function has value 1 for the range [0, 4], since the obligatory duration for every student is 4 years. The median survival time is 4.5 years (dashed line) which means that half of the students are expected to accomplish theirs studies up to the 4.5 years. By observing the shape of the curve, we can see that the slope of the curve changes first around the 3.8 years and then around the 5 years.

Finally, we investigate the distribution of the duration for the factor "Gender". In Figure 7, we can see the curves corresponding to the two levels of factor "Gender". It is clear that the Female students accomplish their studies earlier (median=4.08 years) than the Male students (median=4.83 years). So, the SA which utilizes the entire dataset signifies the findings of the ANOVA procedure.

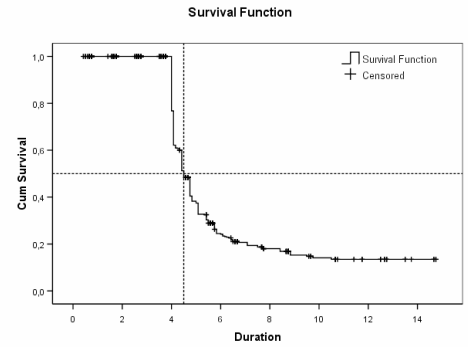

Figure 6. Survival function for Duration

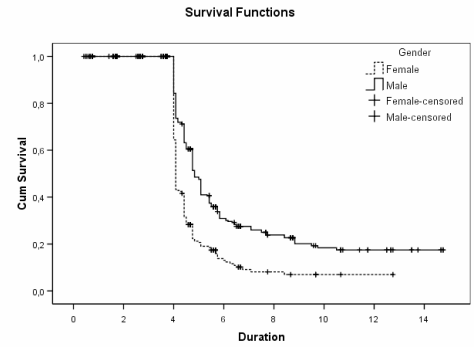

Figure 7. Survival function of Duration for Gender

5. Conclusions

During the recent years, universities are encountering growing demands by legislators and communities who are clamoring for valuable information about student achievement and university system accountability. Moreover, besides universities, other institutions (like ministries of education) need to measure progress. The department of Informatics of Aristotle university of Thessaloniki developed a data warehouse solution that assists the analysis of educational data. In this paper we present the design and development of the proposed data warehouse solution, which facilitates better and more thorough analysis of department's data. The proposed system constitutes an integrated platform for a thorough analysis of department's past data. Analysis of data could be achieved with OLAP operations.

Moreover, we proposed a thorough statistical analysis with an array of data mining techniques and advanced statistical hypothesis tests, models and methods that are appropriate for the examined tasks. The overall statistical framework we developed enables the longitudinal study of the students' performance in the department and particularly facilitates the search for factors that may affect their performance. We believe that our results will be encouraging and will assist the continuation of this effort.

Acknowledgement

We would like to thank Mr. Leonidas Zagkaretos for his help in the implementation of the described system.

6. References

[1] http://dheise.andrews.edu/dw/DWData.htm
[2] ibm.com/industries/education
[3] http://www.educause.edu/
[4] http://www.ecs-eduk12.com/datawarehouse.html
[5] D.J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures,* Third Edition Chapman & HAL/CRC 2004.
[6] M.K.B. Parmar, and D. Machin, *Survival Analysis. A Practical Approach,* Wiley 1995.