Academic Performance: An Approach From Data Mining

David L. LA RED MARTINEZ
Julio C. ACOSTA
Valeria E. URIBE
Alice R. RAMBO
Dpto. Informática, FaCENA, Universidad Nacional del Nordeste (UNNE)
(3400) Corrientes, Argentina

ABSTRACT

The relatively low% of students promoted and regularized in Operating Systems Course of the LSI (Bachelor's Degree in Information Systems) of FaCENA (Faculty of Sciences and Natural Surveying - Facultad de Ciencias Exactas, Naturales y Agrimensura) of UNNE (academic success), prompted this work, whose objective is to determine the variables that affect the academic performance, whereas the final status of the student according to the Res. 185/03 CD (scheme for evaluation and promotion): promoted, regular or free[1].

The variables considered are: status of the student, educational level of parents, secondary education, socio-economic level, and others. Data warehouse (Data Warehouses: DW) and data mining (Data Mining: DM) techniques were used to search pro.les of students and determine success or failure academic potential situations.

Classifications through techniques of clustering according to different criteria have become. Some criteria were the following: mining of classification according to academic program, according to final status of the student, according to importance given to the study, mining of demographic clustering and Kohonen clustering according to final status of the student.

Were conducted statistics of partition, detail of partitions, details of clusters, detail of fields and frequency of fields, overall quality of each process and quality detailed (precision, classification, reliability), arrays of confusion, diagrams of gain / elevation, trees, distribution of nodes, of importance of fields, correspondence tables of fields and statistics of cluster.

Once certain profiles of students with low academic performance, it may address actions aimed at avoiding potential academic failures. This work aims to provide a brief description of aspects related to the data warehouse built and some processes of data mining developed on the same.

[1] *Promoted* refers to pupils that to be exempt from the final exam. *Regularized* refers to students that approve the partial examinations. They test theoretical concepts in the final exam. A student is *free* when he reproves the partial examinations and he should make again the course or he must to make the test out of the course.

Keywords: Database, Data Warehouse, Data Mining, Clustering, Cluster Demographic, Academic Performance, Profiles of students, Operating Systems.

1. INTRODUCTION

Having like reference the official information of UNNE, the academic program of Systems of FaCENA has registered a considerable registration of students of UNNE (2005: 4,42%; 2006: 3,93%; 2007: 3,82%; 2008: 3,53%; 2009: 3,34% (candidates); 2010: 2,79% (candidates)).

It has been the academic program of more number of students in FaCENA (2005: 37,64%; 2006: 34,77%; 2007: 33,23%; 2008: 30,32%; 2009: 26,47% (candidate); 2010: 24,80% (candidate)), that of bigger quantity of new registered in the FaCENA (2005: 33,92%; 2006: 29,89%; 2007: 29,71%; 2008: 38,74%; 2009: 23,70%; 2010: 23,30%) and has produced more graduates in FaCENA (2004: 56,05%; 2005: 41,99%; 2006: 44,02%; 2007: 54,30%; 2008: 46,63%).

These data show the importance of Systems academic program (Bachelor of Information Systems: LSI) of the UNNE FaCENA.

A more detailed analysis allows observing the relatively low graduate percentages regarding new candidate in LSI; these percentages vary if it is considered only the terminal title of grade (Degree in Information Systems) or if it is also considered the intermediate title (University Application Programmer):

- Excluding the intermediate title: 2005: 4.81%, 2006: 5.27%, 2007: 9.49%, 2008: 5.42%.
- Considering the intermediate title: 2005: 21.81% 2006: 20.22% 2007: 18.98% 2008: 15.51%.

These relatively low percentages in the relationship degrees regarding new candidates are also observed considering FaCENA and UNNE in their entirety:

- FaCENA: 2005: 17.62%, 2006: 13.73%, 2007: 10.39%, 2008: 12.89%.
- UNNE: 2005: 22.57%, 2006: 22.11%, 2007: 20.49%, 2008: 22.36%.

That pointed out in the precedent paragraphs allows to affirm that the relationship between graduated and new candidates is

in general relatively low, and specially low if it is considered LSI without the graduated with intermediate title.

The relatively low graduation rates for new enrollees referred to in the preceding paragraph, we might consider the "global academic performance" of an Academic Program, College or University, are also observed in many subjects of the LSI, considering "special achievement" or simply "academic performance", the results of student evaluations during the course of a subject, and the final condition achieved by them in the framework of Resolution N° 185/03 CD (evaluation and promotion system): promotions, regular or free. The values of the past years for the Operating Systems (OS) course are as follows:

- Students promoted and regularized of total enrollment: 2003: 12.57%, 2004: 15.23%, 2005: 15.99%, 2006: 9.16%, 2007: 21.26%, 2008: 21.86%, 2009: 9.57%.
- Promoted and regularized students regarding those that surrendered some partial exam[2]: 2006: 21.05% 2007: 32.89% 2008: 34.86% 2009: 13.51%.

It has also been observed that a considerable percentage of students registers to study the course, but after carrying out some activity they leave the course (45.45% during 2009), or sign up to attend and do not pursue, i.e. not engage in any activity (29.19% in the 2009).

In addition, there has been a consistently low turnout of students to tutorials set to support them in carrying out different activities planned (theories, practices, laboratories).

Given the above situation is considered of great importance to carry out an investigation to determine the variables that affect the relatively low academic performance of students in the LSI FaCENA Operating System course, in the UNNE, identify the profiles of successful students (those who promote or regularize the subject), as well as profiles of students not able to (the remaining free status).

Once certain the profiles of students with low academic performance, will be able to be faced spread actions to avoid academic potential failures.

For the determination of the profiles of students it was considered appropriate to use technical of DW and DM techniques.

This article is structured as follows: first, raise the program's main objective, then make a very brief overview of the main concepts involved in terms of DW and DM, then it will briefly display the software used to continue the methodology and presentation of some results, ending with conclusions and future lines of work, acknowledgments and references.

2. MAIN OBJECTIVE

The main objective of this work is to find profiles students through the application of techniques of DM to a DW with data academics, socio-economic and demographic for students of OS of Bachelor's Degree in Information Systems (LSI) of FaCENA of UNNE.

3. CONCEPTUAL REVIEW

A DW is a collection of data oriented issues, integrated, non-volatile, of time variant, which is used for the support of the process of decision-making managerial. It is also a set of integrated data oriented to a field, which vary over time, and that there are not temporary, which bear the process of decision-making of an administration [1], [2], [3], [4].

The DM is the stage of discovery in the process of KDD (Knowledge Discovery from Databases), is the step in the use of specific algorithms that generate a list of patterns from the data development [5], [6], [7], [8].

It is also a mechanism of exploitation, consisting of the search for valuable information in large volumes of data. Is closely linked to the DW since they provide historical information with which the algorithms of mining obtained the necessary information for decision-making [9], [10].

The DM is a set of technical analysis of data that allow draw patterns, trends and regularities to describe and better understand the data and draw patterns and trends to predict future behavior [3], [11], [12], [13].

Figure 1 show the "architecture" of DW.

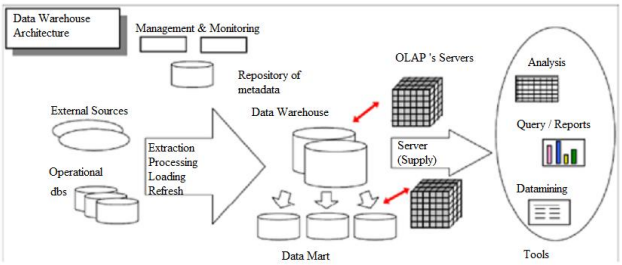

Figure 1: Architecture of a DW.

4. SOFTWARE

It has been used the IBM Data Warehouse Edition (DWE) V. 9.5, which includes the DB2 Enterprise Server Edition (DB2 ESE), the Design Studio (DS) and the Intelligent Miner (IM).

5. METHODOLOGY

This study was conducted on data obtained through surveys carried out to the students in OS, whereas in addition the results of the different instances of evaluation planned during the course of this subject.

Used an integrated environment management that allows the extraction of knowledge in databases and DW through techniques of DM as be clustering, that consists of the partition of a set of individuals in subsets as homogeneous as possible, the goal is to maximize the similarity of individuals of the cluster and maximize the difference between clusters.

[2] Partial exams: exams where it is evaluated the students during the learning process and in the one studied of the class. It contains the curricular contents taught until the moment.

The demographic cluster is an algorithm developed by IBM and implemented in the IM, component of DWE, environment mentioned above, which automatically resolves the problems of definition of distance / similarity metric, providing criteria to define a optimal segmentation [14], [15], [16], [17], [18], [19].

It was built a Data Warehouse (DW) and Data Mining (DM) techniques were used to search for profiles of the students and to identify potential situations of academic success or failure, using the IBM DWE v.9.5.

Were obtained classifications through (preferably) techniques of clustering, according to various criteria, by e.g.:

- Mining of classification by academic program.
- Mining of classification according final status of the student.
- Mining of classification according to importance given to the study.
- Mining of demographic clustering according final status of the student.
- Mining of Kohonen clustering according final status of the student.

Were analyzed statistics of partition, detail of partitions, details of clusters, detail of fields and frequency of fields, overall quality of each process and quality detailed (precision, classification, reliability), arrays of confusion, diagrams of gain / elevation, trees, distribution of nodes, of importance of fields, correspondence tables of fields and statistics of cluster.

The steps made during the present work have been the following:

- Data collection.
- Treatment and purification of the data.
- Preparation of the database and the corresponding DW on the selected work platform.
- Selection of data mining technique for the study (predominantly clustering).
- Generation of different graphics for the study of the results.
- Analysis of results.
- Obtaining the conclusions.

At this stage it worked with a portion (Data Mart: DM) of DW, whose structure is shown in Figure 2.

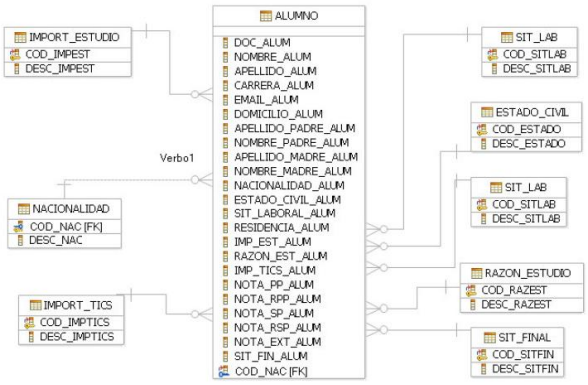

Figure 2: Structure of the used DM, part of DW.

6. RESULTS

Were different classifications using (preferably) clustering techniques according to different criteria of grouping data.

We used the following table of names and meanings of variables.

Variable Name	Meaning
SIT_LABORAL_ALUM	Labour situation of the student
IMP_EST_ALUM	Importance given to the study by the student
RAZON_EST_ALUM	Reason to study according to the student
IMP_TICS_ALUM	Importance given to the ICTs by the student
NOTA_PP_ALUM	Rating first partial exam
NOTA_RPP_ALUM	Rating second round first partial exam
NOTA_SP_ALUM	Rating second partial exam
NOTA_RSP_ALUM	Rating second round second partial exam
NOTA_EXT_ALUM	Rating second round extraordinary exam
SIT_FIN_ALUM	Final status of the student after the coursework

Following shows some results.

Mining classification according to academic program

Figure 3 shows that more than 95% relate to the LSI (Bachelor's Degree in Information Systems) and others to the LS (Bachelor's Degree in Systems), also warn different percentages in regard to the importance attached to the study and ICTs in both groups, as well as the percentages for different reasons for study.

On LSI group, 95% is marital status single, 38% gives more importance to the study that the fun (entertainment) but only 8% more than the work. In this achievement group 20% promoting the matter and 25% regularized matter; 29% said study to learn how comprehensively. The majority percentage corresponds to those who believe that ICTs facilitate learning processes.

On LS group, 100% is marital status single, 38% gives more importance to the study that the fun, 25% studied for approving and 25% does it to learn to learn, this group 65% were free, 25% regularized and 12% promoted.

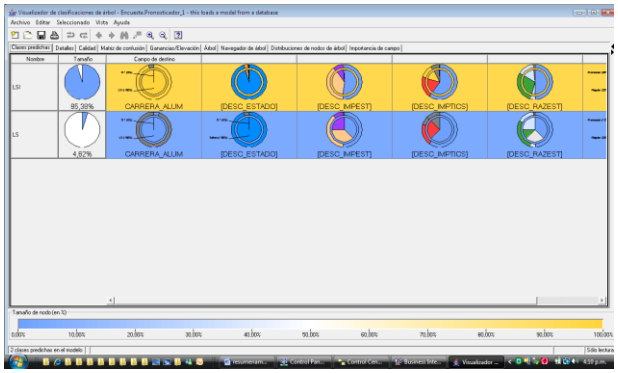

Figure 3: Academic programs.

Mining classification according to final status

Figures 4, 5 and 6 show the classification by final status of students, including pupils from LS and LSI. It notes that the 55.49% was free, 24.86% were able regularize the matter, and 19.65% promoted the matter.

Among the promoted: they give more importance to study and use of ICTs. The free are mostly singles, they provide less importance to study and use of ICTs. The students in regular and promoted status give different reasons to the study that the free. The majority in both groups does not work.

The quality of the model is 0.908 out of 1, this accuracy is 0,794 out of 1, it measures the probability that a prediction is correct, and the classification is 0.91 out of 1, indicating the capacity of the model to correctly sort records based on predicted properties.

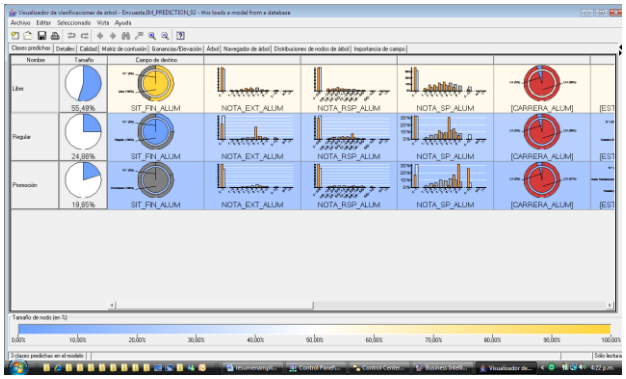

Figure 4: Final status details - 1.

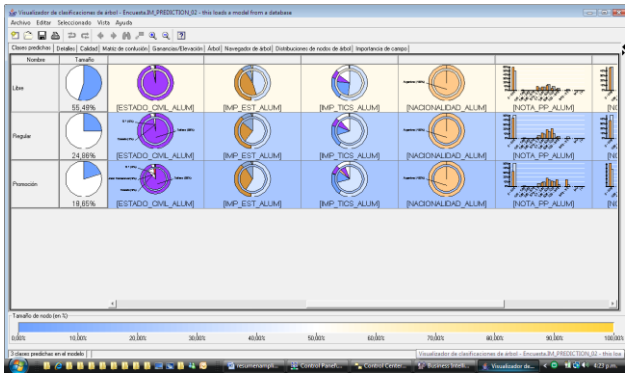

Figure 5: Final status details - 2.

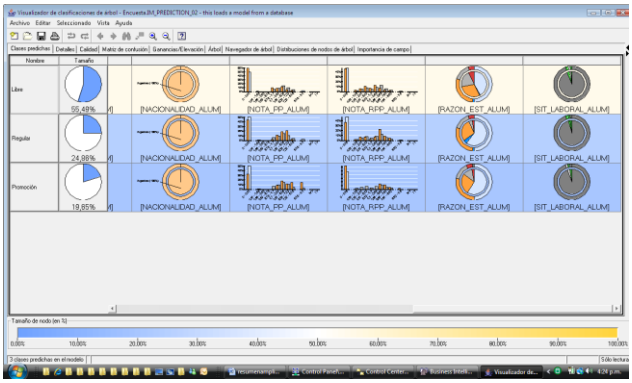

Figure 6: Final status details - 3.

Mining classification according to importance given to the study

In Figure 7 can be seen the classification according to the importance given to the study, in this figure can be seen that 38.15% gives more importance to the fun, 9.25% more than work and 1.16% more than the family.

Regularized and promoted students are among those who give more importance to the study that the fun and work. The free mostly give more importance to study for fun.

Married group predominates among those who give to study more priority than the fun and work. No group gave priority to the study above family.

It notes that among the students to give more importance to the study that the fun and to work are the regularized status and promoted in preponderance within that gives more importance to work. The free status in its most gives more importance to the study that the fun.

The group of married prevails among those who give more priority to the fun and that the work. No group gave more priority to the study that the family.

The quality of the model achieved with this classification is 0.853 out of 1.

Figure 7: Importance given to the study.

Mining classification according to reasons to study

Figure 8 shows the classification considering the reason for studying. Observed that 5.2% studies to approve, 28.32%

studied to learn comprehensively, 15.03% makes it for learning to learn and 1.73% has another motivation for the study.

In the group that studies to learn how to learn are the majority of people regularized and promoted.

Most free students are in the group that studies to approve.

The quality of the model considering the reason for studying is 0.764 out of 1; the precision is 0.939 out of 1 and classification is 0.632 out of 1.

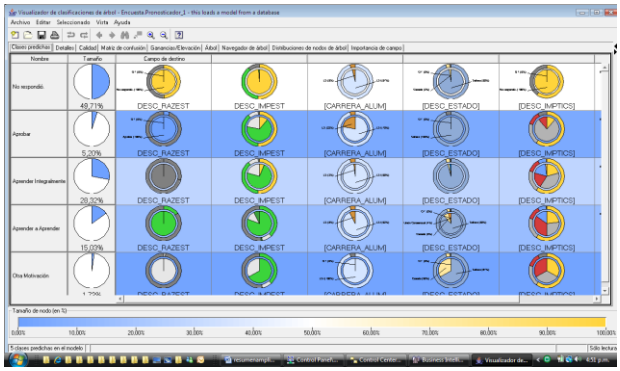

Figure 8: Reasons to study.

Mining classification according to importance given to ICTs

Figure 9 shows the results whereas the importance attributed to ICTs; can be seen to 19.65% consider that they facilitate study, 12.14% which will be essential to its domain, 7.65% are a reality and 0.58% which are fashionable.

Most promoted and regularized considers ICTs to facilitate the process of study and that ICTs are a reality.

The quality of the classification model according to the importance given to ICTs is 0.712 out of 1.

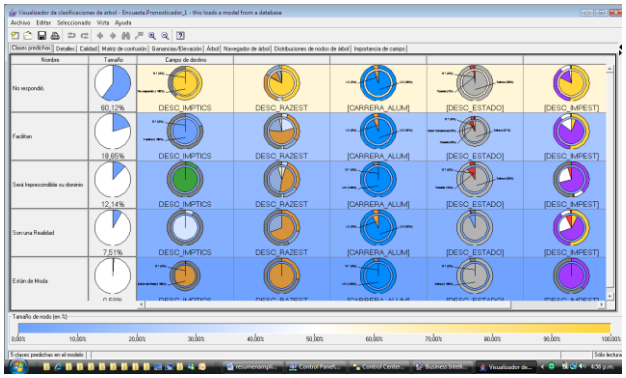

Figure 9: Importance given to ICTs.

Mining classification whereas labour situation

Figure 10 shows the classification according to the labour situation of the student; if you work, what type of work performed. Observed that the 94.22% does not work, 2.31% work in private enterprises, 1.73% work in official entity (State) and 1.73% works as specialized professionals.

Figure 10: Labour situation.

Mining of clustering considering age

Figures 11, 12 and 13 show classification considering the age of the students. Five groups were obtained. Five groups were obtained. The first group has 25.53%, the second 44.68%, the third 4.26%, the fourth 14.89% and the fifth 5.85%. Older students are in the third group; younger students are the first and the fifth group.

The quality of the model according to the age of the students is 0.784 out of 1.

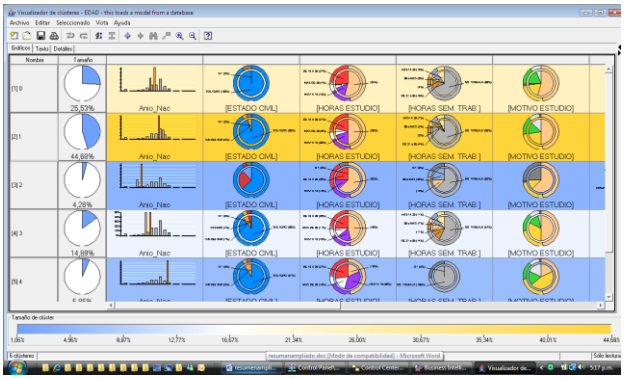

Figure 11: Age - Details 1.

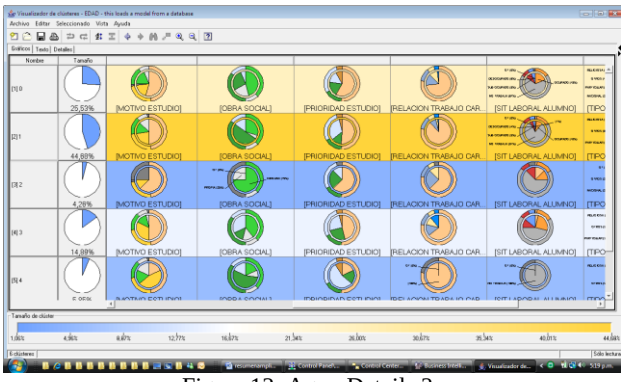

Figure 12: Age - Details 2.

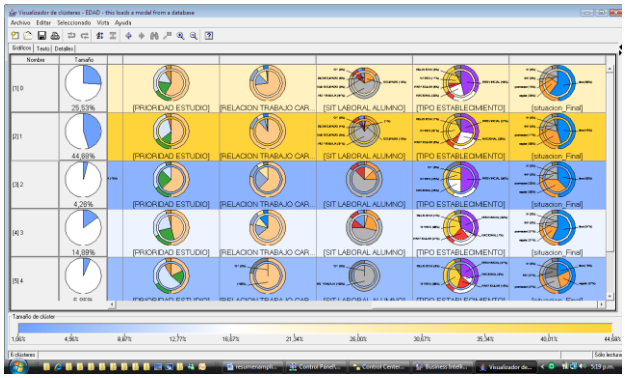

Figure 13: Age - Details 3.

7. CONCLUSIONS AND FUTURE LINES OF WORK

It has been proved the great advantages of the use of technologies and latest generation software that support multiplatform systems. Were several data mining models regarding various aspects of the students of OS; this enabled to discover the profile of these students.

Were obtained preliminary students profiles, highlighting the following aspects:

- Free students are mostly single, lesser extent divorced, most do not work, they say give more importance to study for fun and that the family, most consider ICTs to facilitate the study and your domain is important, the majority said that you studying to learn comprehensively.
- Regularized students are mostly single and consensual union[3], give more importance to the study for fun and work, most consider ICTs facilitate learning, which will be essential to your domain and they are a reality, most studies to learn integrally and learning to learn, most do not work.
- Promoted students are mostly single, most gives more importance to study for fun and work, most considered ICTs facilitate learning and will be essential to your domain and to a lesser extent, are fashionable, mostly say study to comprehensively learn, learning to learn and to a lesser extent only to approve, the majority doesn't work.

The students who:

- Study to pass: they give more importance to the study that the fun and work, are single, considering that ICTs facilitate learning and will be essential to its domain, the majority was free.
- Study to comprehensively learn: they give more importance to the study that the fun and work, most are unmarried and younger % married and divorced, mostly consider ICTs facilitate learning, which will be essential to your domain and they are a reality, the majority was free, but almost half was regular or promoted.
- Study to learn how to learn: give more importance to study that work, fun and family, they are mostly single but increases the % married and consensual union, most considered the domain of the ICTs is essential, the majority was free but it is important the % of regular or promoted.

The majority in all groups does not work, but the percentage of workers between the promoted is greater than between the free and the regular.

Intends to develop the following future lines of work:

- Using DM algorithms based on neural networks, Bayesian networks and decision trees.
- Applying DM techniques used on other DW of students of other subjects and academic programs to compare the results.

[3] Consensual union refers to a couple that live together without married.

8. ACKNOWLEDGMENTS

This work falls within the Research Project "The uneven development of ICTs in the teaching – learning process of Operating Systems in FaCENA of UNNE", accredited by the Science and Technology Secretariat of UNNE as PI-120-07 (Res. N° 369/08 CS).

The software used, Data Warehouse Edition V. 9.5, which includes DB2 Enterprise Server Edition, Design Studio and Intelligent Miner, have been obtained from the IBM Argentina S.A. company in the framework of the IBM Academic Initiative and agreements made between the IBM and the FACENA of UNNE (Agreement of 18/06/04 D, Res. N° 1417/04 D, Res. N° 858/06 CD).

9. REFERENCES

[1] W. H. Inmon, **Data Warehouse Performance**, John Wiley & Sons, USA, 1992.

[2] W. H. Inmon, **Building the Data Warehouse**, John Wiley & Sons, USA, 1996.

[3] A. Simon, **Data Warehouse, Data Mining and OLAP**, John Wiley & Sons, USA, 1997.

[4] J. C. Trujillo, M. Palomar & J. Gómez, **Applying Object-Oriented Conceptual Modeling Techniques To The Design of Multidimensional Databases and OLAP Applications**. First International Conference On Web-Age Information Management (WAIM.00). Lecture Notes in Computer Science 1846:83-94, 2000.

[5] U. M. Fayyad, G. Grinstein & A. Wierse, **Information Visualization in Data Mining and Knowledge Discovery**, Morgan Kaufmann, Harcourt Intl., 2001.

[6] U. M. Fayyad, G. Piatetskiy-Shapiro, P. Smith, U. Ramasasmy, **Advances in Knowledge Discovery and Data Mining**, AAAI Press / MIT Press, USA, 1996.

[7] J. Han & M. Kamber, **Data Mining: Concepts and Techniques**, Morgan Kaufmann, 2001.

[8] D. J. Hand, H. Mannila & P. Smyth, **Principles of Data Mining**, The MIT Press, USA, 2000.

[9] J. M. Gutiérrez, **Data Mining, Extracción de Conocimiento en Grandes Bases de Datos**, España, 2001.

[10] IBM Software Group, **Enterprise Data Warehousing whit DB2: The 10 Terabyte TPC-H Benchmark**, IBM Press, USA, 2003.

[11] A. Berson & S. J. Smith, **Data Warehouse, Data Mining & OLAP**, Mc Graw Hill, USA, 1997.

[12] W. J. Frawley, G. Piatetsky-Shapiro & Ch. J. Matheus, **Knowledge Discovery in Database, An Overview**, AI Magazine, 1992.

[13] C. J. White, **IBM Enterprise Analytics for the Intelligent e-Business**, IBM Press, USA, 2001.

[14] J. Grabmeier. & A. Rudolph, **Techniques of Cluster Algorithms in Data Mining version 2.0**, IBM Deutschland Informationssysteme GmbH, GBIS (Global Business Intelligence Solutions), Germany, 1998.

[15] C. Baragoin, R. Chan, H. Gottschalk, G. Meyer, P. Pereira & J. Verhees, **IBM International Technical Support Organization Enhance Your Business Applications, Simple Integration of Advanced Data Mining Functions**, IBM Press, 2002.

[16] Ch. Ballard, J. Rollins, J. Ramos, A. Perkins, R. Hale, A. Dorneich, E. Cas Milner & J. Chodagam, **Dynamic Warehousing: Data Mining Made Easy**, IBM International Technical Support Organization, IBM Press, USA, 2007.

[17] Ch. Ballard, A. Beaton, D. Chiou, J. Chodagam, M. Lowry, A. Perkins, R. Phillips & J. Rollins, **Leveraging DB2 Data Warehouse Edition for Business Intelligence**, IBM International Technical Support Organization, IBM Press, USA, 2006.

[18] O. Maimon & L. Rokach, **Data Mining and Knowledge Discovery Handbook**, 2/E, Springer, USA, 2010.

[19] C. Pérez López & D. Santín González, **Minería de Datos: Técnicas y Herramientas**, Thomson Paraninfo S. A., España, 2007.