A Business Intelligence Model to support the engineering student life cycle analysis Recruitment, Retention-performance, and Graduate process

A. Santoyo-Sanchez[1], José Luis Reyna-Gascón[2],
L. I. Aguirre-Salas[3], Griselda Pérez-Torres[1], and C. A. De Jesús-Velásquez[4]
[1]Department of Computing, Universidad de Guadalajara – CUCEI, Guadalajara, Jalisco, México
[2]IT BI & Analytics department, Tlaquepaque, Freescale Semiconductor, Tlaquepaque, Jalisco, México
[3]Department of Engineering, Universidad de Guadalajara – CUCSUR, Autlán de Navarro, Jalisco, México
[4]Compatibility Validation, Intel Tecnología de México S.A., Tlaquepaque, Jalisco, México
alejandra.santoyo@cucei.udg.mx; jose.reyna@freescale.com; laguirre@cucsur.udg.mx

Abstract. Taking smart decisions during the proper time as well as respond with flexibility to any customer demand, market opportunity or threat, through Business Intelligence (BI), causes that the increase of designing multidimensional solutions become a great solution and alternative to support smart decisions in an organization. In this paper, we present an alternative approach of analyzing the life cycle of the engineering students during their stay at CUCEI, which is a thematic center (campus) of the University of Guadalajara in Mexico dedicated to higher education (Bachelor, Master and PhD Degree). This campus has one of the largest populations within the University of Guadalajara, serving more than 13,000 students [3]. Meaning a way to understand since the student is recruited, and then lives its studies looking to get the expected grades and be part of the graduate's pool; the proposed multidimensional model tries to cover this life cycle.

Keywords: Business Intelligence, Multidimensional Modeling, Students life cycle

1 Introduction

This paper presents a brief overview of how to model the student's life cycle when joining college. This paper intends to be a practical case for the public University of Guadalajara, being focused on one of their campuses, CUCEI which handles all the engineering careers.

The end users typically collects many reports stored in different locations in order to place them in a single repository, sometimes a spreadsheet, and apply their own calculations and also their own charts and metrics. The idea of modeling the business using multidimensional methodologies empower the end user to not waste time on gathering information from many types of sources. The work of the Business Intelligence Analyst is to understand the business and its needs in order to create a semantic layer, from the transactional and operative data that is used to track the day

to day student activities, which translates all these business requirements into a main repository that contains the Business measures and metrics to modeled into a multidimensional environment.

2 Background

The University of Guadalajara is a set of University Network in Jalisco, Mexico, with an historical tradition of more than 2 centuries. It has more than 240,000 students and has presence in all regions of the state, being the most important institution of higher education in western Mexico and the second of the country. Its high school education system covers 101 cities around the state and its higher education (college) is divided in all regions of the state thru 6 main campuses in the metropolitan area and 9 regional campuses around the state [2].

During 2012-2013 47 out of 100 college students studied at University of Guadalajara. This situation has been the trigger for social mobility in all regions of the state, stimulating the local economy, sparking scientific and cultural activities, and moreover, linking every day of hundreds of researches and teachers solving problems of their communities [2].

The main idea of this paper is to obtain a data analysis model to support smart decisions and provide a big picture of the situation, to the key users, that predominates during the students life cycle, which is when the student is recruited, passes an exam to be part of the University, lives its studies looking to get the expected grades and be part of the graduates pool, but turns out that during that period, some students decides to leave their studies due to some reasons. The objective of the Multidimensional model developed in this paper is to support predicting and providing student's situation snapshots during their studies in order to act proactively on students that might give up their career path.

3 Basic Concepts

This section presents the basic concepts related to BI used in this paper. For more details, consult [4-7].

Definition 1. *Business Intelligence* (BI) is the art and science of preparing companies for the future by way of systematic knowledge management process. It is creating knowledge from openly available information by use of a systematic process involving planning, collection, analysis, communication and management, which result in decision-maker action [8].

According to [9] the above definition emphasizes the following: (1) Creation of knowledge from business and project information - availability of information is important in knowledge creation, which will be intelligently used in project management decision making process. (2) The useful application of created knowledge results in better project management decisions. More specifically, BI solutions combine data gathering, data storage, and knowledge management with analytical tools to present complex and competitive information to planners and decision makers

[10]. According to [9] a BI solution has four major components, namely, a Data warehouse, Business Analytics (BI technologies), Business Performance Management (BPM) and a user interface. BI solution proposed is based in the follows components.

First, to providing an integrated historical data, data mart was the reference tool for storage of data to support and to feed the BI system.

Definition 2. *Data warehouse* is a database containing the consolidated data from all available resources, optimized for reporting, analysis and archiving. The Data warehouse integrates and stores data from both internal and external sources [12].

Definition 3. *Data Mart* is basically another form of a Data Warehouse, but it is a subset of information customized based on the business needs, with data summarized, aggregations, metadata, and specific features needed for the analysis [13].

Second, the business analysis was made using Online Analytical Processing (OLAP) systems. These tools are considered as a type of software aiming at fast exploring and analysis of data through a multidimensional approach with several levels of aggregation. The online analysis is a subject oriented, non-volatile and time variant tool with an optimized approach for decision support. The idea of using OLAP solutions is to have it as a long term information asset. This tool does not substitute the transactional and operational systems (OLTP); it is intended to be an additional solution that contributes the decisions support. The OLTP are the source data where OLAP will consult.

Definition 4. *OLAP system* is delivered by Microsoft as part of the Microsoft Dynamics AX solution. The OLAP system collects the performance information and generates reports on measurements. This kind of system is used generally in a business environment as an executive information system detailing the actual states of the subject to be analyzed [14].

Third, BI tools are used to determine customer's current business value (or profile). Then, "BI designer" will apply their knowledge and expertise to offer the best possible option to this customer, while taking into account organizational rules and policies, which is the case of the BI solution proposed. Current BPM comprises modeling, analyzing and monitoring of business processes.

Definition 5. *BPM system* is a system for sensing environmental stimulus, interpreting perceived data, adjudicating the data to be business situations, and making decisions about how to respond the situations [16].

According with [16] in general, there are five representative categories of services in a BPM system: Sense, Detect, Analyze, Decide and Execute [16].

1. "***Sense***" is the stage when a BPM system interacts with business solutions and provides data extraction, transformation, and loading capabilities for the sake of preparing qualified data (metrics) that is to be further monitored and analyzed.

2. "***Detect***" is the stage of detecting business situations and/or exception occurring in the business solutions. An example of situation detection could be lower-than-expected cycle time performance in certain manufacturing process.

3. "***Analyze***" is the stage when a BPM system performs business analytics such as risk-based analysis of resolving business exceptions. The output of this stage often comes with recommendation of potential resolutions to decision makers.
4. "***Decide***" is the stage when a decision maker will make decision about what to respond to business situations. A decision maker can be either human or software agent.
5. "***Execute***" is the stage when a BPM system carries out actions for the purpose of enforcing the decisions made by decision makers. Actions can be of many forms such as sending email or as complicated as a sub-process invocation.

The BI solution proposed in this paper covers the set of categories described in Definition 5, which will be explained in the next chapters, but the idea of contributing with a BI solution has to consider stages since the data extraction, transformation and loading till the analytic solution which will become the tool where decisions makes will consult and trust.

Finally, the interface of the final solution is presented using ad-hoc tools that will allow the user to create their own dashboards and pivots based on a single and unique repository, which will provide accurate information. The user will love to validate and play with the data generated, more over that this information will not require dealing with multiple data sources, the OLAP Data base will provide that unique repository that the user will browse and query. The interface used in this project was Microsoft Power Pivot, a BI tool for Microsoft.

4 Methodology

This section presents our methodology to develop the Business Intelligence model. This consists in two main aspects: Problem definition, source data set preparation and exploring; all this in order to guarantee quality for the BI solution developed, which must be included the follows.

1) Communication. The objective is to involve the client into the project development, generating trust and letting it know any concern or issue during the BI project development. All communication is intended to be in writing means using electronic and physical documents when requested, both by project managers and key users. Therefore, any verbal agreement was reinforced by a written document prepared by agreement of all parties. The meetings always ended to send a draft of the agreement and/or revised conclusions and commitments.

2) Formalization. Models formalize the representation and syntax of the things being modelled. They provide commonality where each component of the model has one and only one meaning.
3) Abstraction. Models allow us to work at multiple levels of abstraction, ranging from very general levels to highly specific. The modelling semantics help to ensure across the levels.

4) Prototypes control. This is in order to re-use any prototype in case a new BI Model needs to be implemented, so everything will be accessible and easy to re-use, these could be Metadata, Cubes or Data marts, and even queries used with parameters.

5) Information Flow. The purpose of this stage is to have all the standard documents of the company in order to check with the BI implemented model and compare if the BI model satisfies the standards company.

6) Control Version. This stage consist on tracking any change done on the code, scripts, BI model, requirements changes, deliveries dates, deliveries priorities.

A. Problem Definition

Today, engineering schools worldwide have a relatively high attrition rate. Typically, about 35% of the first-year students in various engineering programs do not make it to the second year. Of the remaining students, quite often drop out or fail in their second or third year of studies [22]. The purpose of this investigation is to identify the factors that serve as good indicators of whether a student will drop out or fail the program. In order to establish early warning indicators/measures thru the usage of analytics tools being fed by a multidimensional model that can cover the student life cycle.

Next, building a strong semantic layer and a multidimensional model that allow decision makers and users have an unified and consolidated information repository where they can be served, as a self-service instance, with accurate information and obtain consolidated data about many factors and measures such as the status in terms of new students accepted into the campus, performance of the students during their studies and a graduates situation, in each period cycle and career. All these in order to get a big picture of the features that can affect the academic performance of each student by career, looking forward to avoid student's desertion.

B. Data set preparation and exploring.

The idea of this step of the methodology is to know and get familiar with the current transactional reports and information that is currently being in use in CUCEI.
These reports provide information about the status of new student's recruitment until they are graduated, having information and measures about the overall number of all the high-school graduates that applied to be part of the college campus, other measures such as current assignments that the student is having during a cycle and some others such as the status and situation of the recent and non-recent graduates.

This stage of the methodology also looks to define extraction, depuration and data cleaning, eliminating non-needed or adding needed attributes that the user will be using as navigators thru the information. The Figure 1 shows these steps mentioned. The access granted by the University was thru their administrative information tool called SISECA (Tracking System for Quality Educational Program)[17-18], which contains a repository where the reports can be retrieved and used to analyze what time of measures it is used and the dimensions that the information is navigated.

Approximately 50% of the information gathered, to build the BI solution proposed, was obtained using the SISECA repository, being accessed thru SIIAU information system (Integrated Information System and University Administration). The other 50% of the information obtained came from a repository where the student's tutorial activities are tracked (which consist on a set of meetings with his/her tutor and follow up their career path), this repository is called SIT, which is the information system where the tutorials activities are tracked [21-23].

The next stage during the execution of the methodology is to focus on exploring the source data sets, meaning that figuring out what it is needed to retrieve in each one of the reports used at SISECA and SIT information systems, getting familiar with them and eventually consider them as a data validator against the multidimensional solution implemented.

At the end of the design phase objectives were examined and detailed scope of the system, the choice of architecture, and resolution of major risks.

The following logical data flow represents and resumes the steps mentioned on this part of the methodology, the stages of each design phase: Sources, Business Intelligence (ETL, Semantic Layer, and Analytics) and the BI Partner community that will be served with the information obtained.

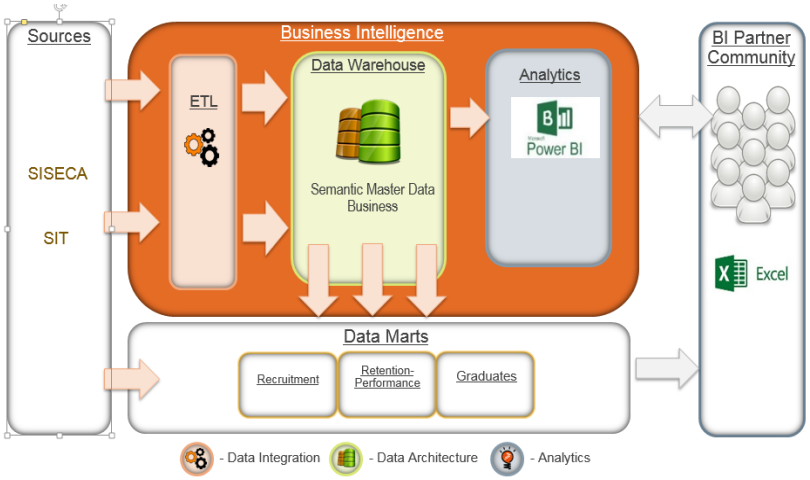

Fig. 1. Business Intelligence Logical Data Flow.

5 Practical Case

Based on the transactional reports provided by SISECA and SIT, the idea of design a Multidimensional solution under a constellation approach was created, including 3 starts schemas that could cover the life cycle of the engineering student, stated under the background section of this paper, for more details about Multidimensional solutions consult [5].

The idea of having a multidimensional solution is to focus all the information to be analyzed in one single repository that can be easy to obtain and retrieve, cross joining

measures that normally are presented or showed separately during the transactional reports and which possible can be difficult to compare using this type of solutions. And also get shared dimensions and measures to obtain better decisions based on the information offered in the model.

And it is because of the 3 stages that a student lives during their stay at the campus, (recruitment, retention-performance, graduates), it was decided to create 3 main fact tables. The Table 1 summarizes the semantic model implemented. Note that each one of the fact tables is joined with share dimensions, which basically are the attributes where the end user will navigate and obtain the needed information to take the proper decision. While as the Figure 2 show the constellation schema (multidimensional model design) which includes the 3 semantic layers described under table 1, that act as 3 stars schemas, resulting a one constellation schema (the union of multiple stars). The idea is that these 3 Fact tables can have shared dimensions as well as information unified.

Table 1. Fact tables descriptions.

FACT	PURPOSE
Recruitment (Fact_Recruitment)	DataMart focused in knowing the situation in the beginning of a cycle by campus and the number of students, prospects students and grades obtained during the admission test.
Retention Performance. (Fact_RetentionPerformance)	DataMart focused on obtaining the right metrics related with each student situation in regards its performance during each cycle time, campus and the different careers.
Graduates (Fact_Graduates)	DataMart focused in knowing the situation of the graduated students who obtain or did not obtain their grades.

Some interesting statistical numbers to be shared, which were obtained from the BI solutions are the following.

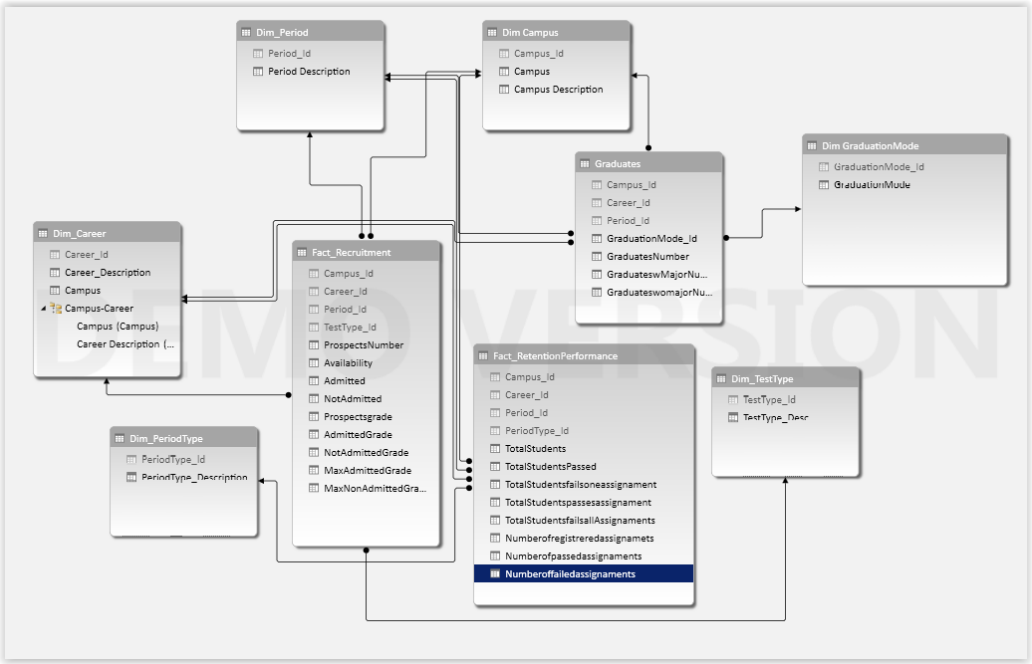

Fig. 2. Constellation Schema implemented in the Project.

- In every period, CUCEI has an admission rate for engineering careers; in overall the admission rate to the campus for all of them is 50% [3] see Figure 3a.
- The highest admission rate is from careers such as: Computer Engineering, Chemistry, Electronic Engineering, among others [3] sees Figure 4.
- The score average of the applicants at CUCEI was 59.7 out of 100, but the score average of the admitted applicants was 68.7 out of 100 where the highest of the non-admitted applicants was 81.7 out of 100 [3] see Figure 3b and 3c.
- Only 52.4% of the students in CUCEI obtain acceptable scores in all of their assignments, 3.7% of the students fail all their assignments and 43.9% of the students fail some assignments [3] see Figure 5.
- The average of enrolled assignments per students is 5.9 and 2.1 assignments are failed per students [3]. 29.6% of the students who did not get an acceptable score and were removed, asked for a new opportunity to enroll again and take the assignment. This occurred between the 1st and the 3rd semester. [3]. 35.9% occurred during the 4th and 6th semester. 22.7% occurred during the last stage of the career.

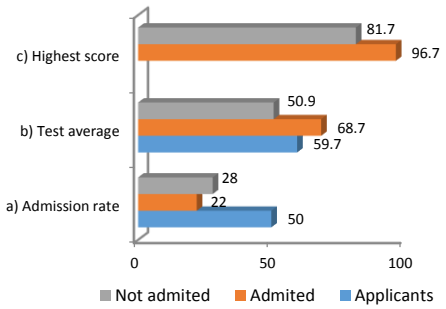

Fig. 3. a) Admission Rate average every period, b) Test average score per type of applicant, and c) The highest score per type of applicant.

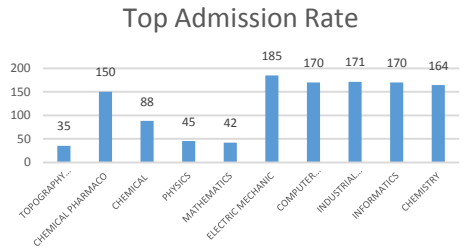

Fig. 4. Top admissions rate.

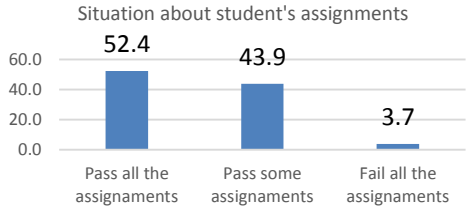

Fig. 5. Average situation about student's assignments.

All the reasons and statistical numbers obtained thru the current transactional reports, surveys and currently the BI system, are tracked at SIT software application system, as an user interface where the tutor can save all their inputs and comments in regard an specific student where economical and psychological aspects are tracked and are not obtained directly from SISECA (main transactional system). Physic and mathematical assignments were the most difficult ones to obtain an acceptable grade and that happened during the first stage of the student career. Some students mention that they do not feel prepared enough in the first semesters of the career due to the assignments commented above.

Additionally, other reasons that affect student's desertion are, in descendent order. 1) Economical and family situation. 2) Family dysfunction. 3) Healthy family problems and students. 4) The long distance the campus is away from the student residence. 5) Not the correct chosen career.

Therefore it is necessary to intensify the work to meet a good performance in each student, for instance the usage of learning objects, design and develop teaching materials, among other things aiming to support vulnerable groups of students to succeed in their career path.

The numbers presented here, from the BI environment results, lead us to propose some strategies that are aligned to CUCEI in order to meet the particular needs improving current conditions in the field of teaching and learning. 1) Systematize studies of relevance and quality. 2) Establish a formal CUCEI graduates monitoring

program and consulting employers. 3) Having evaluated educational programs, with satisfactory results of the assessment bodies. 4) Establish a program of renovation and certification for teachers in teaching and learning environments. 5) Increase the number of teachers with desirable profile. 6) Update and improve infrastructure and sources of information for teaching. 7) Improve infrastructure to support teaching, particularly laboratories and computer services. 8) Establish systems and procedures to give awards, certificates, and assessments to teachers in a timely manner. 9) Establish specific support programs for groups with different conditions. 10) Improve individual student support. 11) Implement the tutorial staff takes place in its various forms. 12) Seek for sport activities.

6 Conclusion and Future Work

Information technology teams need to be aligned into the business looking to be a strategic partner that could contribute on transforming and growing the business and more over making the life of the users easy in terms of information systems.

Analytics skills are needed by the Information Technologies teams in order to provide foundation for delivery of BI applications. Analysis concentrates on understanding business needs for data and information. Designing focuses on translating business information needs into data structures that are adaptable, extensible and sustainable. The core skills to develop include information needs analysis, specification of business metrics and data modeling. A solid understanding of data warehousing concepts, architectures and processes is also essential.

With all these said the end user as a third party validates the BI proposed architecture. Future work will be the use of mathematical models that Petri nets for validation, verification and performance analysis. Also expand the range of use of business process modeling.

Acknowledgment
This work was supported by Instituto Jalisciense de la Juventud (IJJ) in the scope of the project 2011A2986, period April to December 2014, between Universidad de Guadalajara, Jalisco, México and the IJJ, Jalisco, México.

The scholarship obtained by IJJ allowed José Luis Reyna Gascón to obtain his Master Degree in Informatics, sponsored by Jalisco Government. The pay back, due to this support, was to collaborate with CUCEI in creating a multidimensional modeling for their current workflow in terms of Recruitment, Retention-performance, and Graduate process of the students, which is explained thru this paper.

References

1. CUCEI crece en investigación, obras y equipos de laboratorios, Universidad de Guadalajara, consulting in 27-11-2014, http://www.udg.mx/es/noticia/cucei-crece-en-investigacion-obras-y-equipo-delaboratorios
2. Plan de Desarrollo Institucional 2014-2013, Universidad de Guadalajara, consulting in 27-11-2014, http://www.copladi.udg.mx/sites/default/files/pdifinal1_0.pdf
3. Reportes de Admisión a pregrado 2014A, Coordinación de Control Escolar, Universidad de Guadalajara. Internal Document stored in SISECA, www.siseca.udg.mx.

4. Curko K., Pejic Bach M. Business Intelligence and Business Process Management in Banking Operations, Proceedings of the ITI 29th Int. Conf. on Information Technology Interfaces, 2007, 57-62.
5. Kimball, R., The Data Warehouse Toolkit, Wiley Computer Publishing, United States of America, 2002.
6. Liautaud, B., Hammond, M.: e-Business Intelligence, Turning Information into Knowledge into Profit, McGraw-Hill, New York, (2001).
7. Osterfelt, S.; Business Intelligence: The Intelligent Customer, DM Review [http://wdmreview.com], 2000.
8. J.L. Calof, and B. Skinner, "Competitive intelligence for government officers: a brave new world", Optimum, vol. 28, issue 2. pp. 38-42, 1998.
9. L. Cheng, and P. Cheng, " Integration: Knowledge management and business intelligence", IEEE Computer Society, Fourth International Conference on Business Intelligence and Financial Engineering, pp. 307-310, 2011.
10. Foulquier, T.; Perreault, L.; Caron, C.; Levesque, J., "Combining BI Technologies with Microgeomatics at a Clothing Retailer", Proc. on 44th Hawaii International Conference on System Sciences, HICSS, 2011, pp. 1 – 10.
11. B.S.Sahay, and J. Ranjan, "Real time business intelligence in supply chain analytics". Information Management & Computer Security, vol. 16, issue 1, pp. 28-48, 2008.
12. Tvrdiková, M.; Koubek, O., "Support of e-business by business intelligence tools and data quality improvement", IEEE Proc. on International Multiconference on Computer Science and Information Technology (IMCSIT), 2010, pp.271 – 278.
13. Sung Ho Ha ; Sang Chan Park, "Data modeling for improving performance of data mart", IEEE Proc. on International Conference on Engineering and Technology Management, Pioneering New Technologies: Management Issues and Challenges in the Third Millennium, IEMC, 1998, pp. 436 – 441.
14. Selmeci, A.; Orosz, I.; Gyorok, G.; Orosz, T., "Key Performance Indicators used in ERP performance measurement applications", Proc. on 10th IEEE Jubilee International Symposium on Intelligent Systems and Informatics, SISY, 2012, pp. 43 – 48.
15. Yamamoto, R.; Yamamoto, K.; Ohashi, K.; Inomata, J., "Development of a Business Process Modeling Methodology and a Tool for Sharing Business Processes", Software Engineering Conference, Asia-Pacific, Notes in Computer Science, 679-687, 2005.
16. Jun-Jang Jeng, "Service-Oriented Business Performance Management for Real-Time Enterprise", Proc. on the 8th IEEE International Conference on E-Commerce Technology, and Enterprise Computing, E-Commerce, and the 3rd IEEE International Conference on E-Services, 2006, pp. 28.
17. Kin Fun Li; Rusk, D.; Song, F., "Predicting Student Academic Performance", Proc. on Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, CISIS, 2013, pp.2733.
18. CIEP-Universidad de Guadalajara, consulting 11-27-2014, http://ciep.cga.udg.mx/contenidos/54/
19. SISECA-Universidad de Guadalajara, consulting 11-27-2014, www.siseca.udg.mx.
20. SIIAU-Universidad de Guadalajara, consulting 11-27-2014, http://www.siiau.udg.mx/siiau
21. CIEP-Universidad de Guadalajara, consulting 11-27-2014, http://ciep.cga.udg.mx/contenidos/15/
22. SIT-Universidad de Guadalajara, consulting 11-27-2014, http://tutoria.udg.mx/
23. Sistema Integral de Tutorías, CUEI- Universidad de Guadalajara, consulting 11-27-2014, http://tutorias.cucei.udg.mx/