Requirements Engineering Techniques: Considerations for their Adoption in Data Mining Projects

José Gallardo[1], Claudio Meneses[1],
and Óscar Marbán[2],
[1] Departamento de Ingeniería de Sistemas y Computacion, Universidad Católica del Norte.
Av. Angamos 0610, Antofagasta, Chile
{jgallardo, cmeneses}@ucn.cl
[2]Facultad de Informática, Universidad Politécnica de Madrid.
Campus de Montegancedo s/n, Boadilla del Monte, Madrid, España.
omarban@fi.upm.es

Abstract. The correct and complete requisites specification is a key factor to the success for any project. Data Mining (DM) projects constitute decision-making support systems, and therefore the traditional Requirements Engineering techniques cannot be directly applied to them. This on-going research work presents an overview of the main models of the Requirements Engineering (RE) processes, and the most broadly used techniques in the development of the different phases involved in a RE model. Then the key issues that should be considered in the application of these techniques in Requirements Engineering processes for Data Mining projects are discussed. Also a proposition is done on how to structure the requirements in Data Mining projects from three different perspectives, in each one of them the type of information that should be captured is detailed, in order to particularly specify the requirements of DM projects, and generally in the decision-making support systems.

Keywords: Data Mining, Requirements Engineering Techniques, Requirements Elicitation.

1 Introduction

During the last years, a large number of Data Mining projects have been developed and it is expected that in the next decade this quantity will increase to 300%, as estimated in a report from GartnerGroup [6]. However, the execution of this type of projects faces serious problems, for example, they are never finished, or they are out of date or they are out of budget [23]. These problems are similar to those presented in the development of software applications [22], in what was named "software crisis", which was solved with the development of the Software Engineering discipline. In the Data Mining area, as a way of facing the generated problems, mainly due to a lack of standard or methodological guidelines for their development, a group of European companies which are pioneers in this type of projects (Teradata, SPSS, Daimler-Chrysler and OHRA), proposed in 1999 a reference guideline named CRISP-

Data Mining and Information Systems.
Research in Computer Science 22, 2006, pp. 79-90

DM (Cross-Industry Standard Process for Data Mining) [5], that establishes a procedure for the systematic development of this type of projects. CRISP-DM is not the only guideline that has been proposed. Also, there are others, proprietary or open, like the one developed by SAS company, named SEMMA (Sample, Explore, Modify, Model, Assess) [19], DMAMC [10] or the 5 A's [16]. A survey developed by kdnuggets.com [12] shows that CRISP-DM is the most used one.

However, all the proposed methodologies for Data Mining projects development, lack of methods or techniques that allow us to appropriately educe the project requirements. More concretely, a mature process does not exist yet, that can be seen as a solid methodology. Although CRISP-DM establishes a group of activities that should be executed in the project, it does not establish with which techniques or models should be implemented.

In this paper, an overview of the Requirements Engineering (RE) processes and the main techniques used in each phase of the process is done, and a discussion of the main issues to be considered before adopting them in the construction process of the requirements document in a Data Mining project.

2 Models for the Requirements Engineering (RE) Process

Currently, Requirements Engineering is a technique used by many specialists for the construction of the Requirements Document, which should be the starting point for the correct design and implementation of a system, no matter its nature.

In the RE process, certain fundamental activities can be identified that should be developed to build a document for specifying the requirements. These activities are: requirement elicitation, analysis, specification and validation, and they serve as the foundation in the proposal of different models.

In [8] a model of the RE process is proposed, based on the activities of elicitation, specification and validation, as represented in Figure 1. The main elements of the outlined diagram shown in Figure 1 are briefly described in [2]:

- *Elicitation:* It is the process of acquiring the relevant knowledge, necessary to produce the requirements model in a problem domain, by means of the communication with clients, users of the system, and those involved with the project. After having obtained an initial group of requirements, they should be analyzed and represented in a technical language, in order to avoid inconsistencies and ambiguities.
- *Specification:* The elicitation process provides the entrance for the requirements specification process. The product is a specification model, or the models corresponding to different points of view. These models formalize the group knowledge or of the people involved in the project. The requirements specification also has a double purpose: on one hand, it serves as an agreement among the group involved in the project, for the problem to be solved, and on the other hand, it serves as a model for continuing with the following step.
- *Validation:* The validation is the activity that checks if the requirements specification is done according to the clients expectations. In this stage, it takes

place the integration and final validation of what was done in the previous stages, giving as final result the Requirements Document.

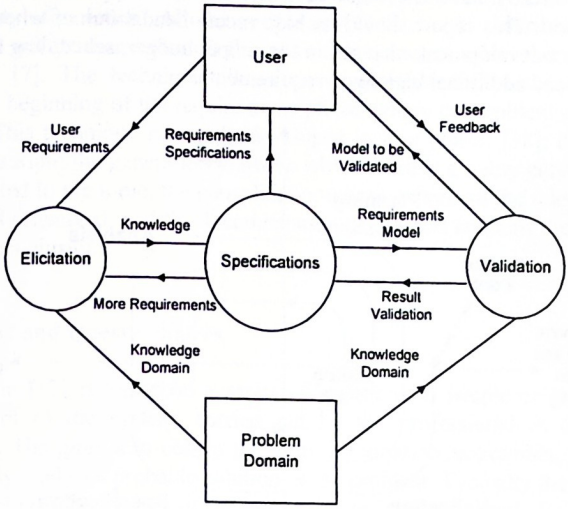

Fig. 1. General diagram of the Requirements Engineering Process

Sommerville [22] proposes another model for the Requirements Engineering process which is represented in Figure 2.

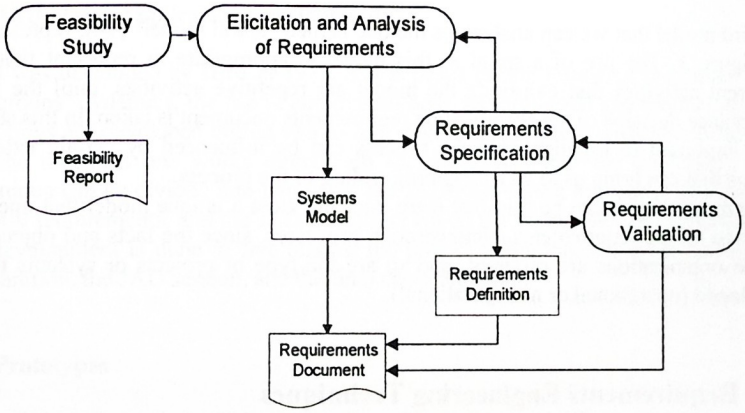

Fig. 2. The Process of Requirements Engineering

This process model incorporates a feasibility study as the first stage in the process, which represents a first approximation that receives as input, a brief description of the system to be developed and how this will be used by the organization. The objective of the feasibility study is to target aspects related to technical and economical feasibility of the project development, under the consideration that the project

contributes to the organizational goals, and the way that the new system may be integrated to the existing systems.

When we already have the required information, we proceed to elaborate the feasibility report. This report should include recommendations of when to continue with the project development, changes in the scope, budget, scheduling of the system development, and additional high level requirements.

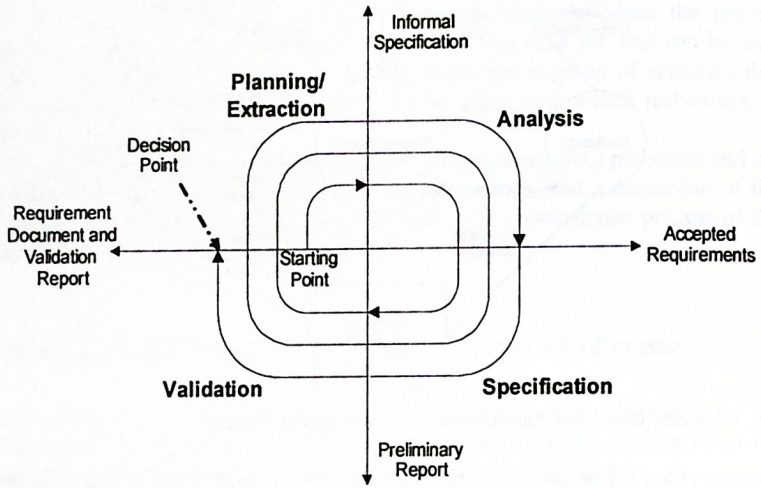

Fig. 3. Spiral Model of the RE process

A third model that we can analyze, is the one named "spiral model" [14], represented in Figure 3. The use of a spiral in this model is appropriate to represent that the different activities that constitute the model are repetitive activities, until the final acceptance decision of the specification requirements document is taken. In this sense, it is important to highlight that the process can be influenced by certain external factors that can bring us to an anticipating ending of the process.

Summarizing, it can be said that there does not exist a unique model that one can apply to all the requirements administration processes, since the facts and objectives of the organizations are different, and so are the type of projects or systems to be developed (operational or not operational).

3 Requirements Engineering Techniques

Independently of which RE process is used, the development of each phase is supported by a set of techniques which have been proposed and applied along time ([2], [4]). It follows a representative, although not exhaustive, review of the RE techniques most commonly used.

3.1 Brainstorming [1]

This technique is broadly used in different areas and it is basically based in the project team creativity stimulation. All the people involved should contribute with ideas, which should not be assessed until the end of the process, when there are no more contributions [7]. The technique allows us to generate different problem views, mainly at the beginning of the requirements phase, where the problem points of view are diffuse. This technique is usually developed in four phases [18]: the meeting or session preparation; the generation phase in which there is a freely contribution to all the ideas related to the topic; the consolidation phase, where all the relevant ideas are identified and organized; and the documentation phase, that contains the main aspects said and the conclusions.

3.2 Interviews and Questionnaires

This technique [13] is based on a series of questions to people or groups that are potential users of the system, carried out by the professional in charge of the requirements. The goal is to collect the more information as possible, which should not necessarily lead to a probable solution of the problem. Typically the questions are at a high level and the success of the use of this technique depends fundamentally in the interviewer's ability to get good answers and to interpret them correctly. In this technique one can identify three phases [17]: interviews preparation, the carry out of them, and the analysis of the data.

3.3 JAD (Joint Application Development)

JAD was developed by IBM in 1977, and is based on the following principles [18]: the groups' dynamics, the use of audiovisual help, the organized and rational process (meetings are developed during two to four days) and the documentation philosophy (in the meetings one works directly on the documents that are generated). This technique can be divided into two parts: the JAD/Plan whose purpose is to elicit and specify requirements, and the JAD/Design, in which the system design is approached. Its development is done in five phases: the project definition, the investigation, the preparation, the JAD session, and the final documentation.

3.4 Prototypes

Technique commonly used in the systems development. It allows the developer to build a model of the system that must be developed in the future [14]. The model is a simulation of the probable system, and subsequently is utilized by the end user. This technique allows getting the required information feedback so as to assess whether the system designed based in the requirements, allows the user to carry out its work in an efficient and effective way.

3.5 Hierarchical Analysis Process

The Hierarchical Analysis Process has as a fundamental objective, to solve quantitative problems, in order to facilitate analytic thought and metrics. This technique is divided into a series of tasks, such as:

- Finding the requirements that will be prioritized.
- Combining the requirements in rows and columns of a matrix.
- Doing comparisons of the requirements in the matrix.
- Adding the columns.
- Normalizing the sum of the rows.
- Calculating the averages.

These steps can be applied easily to a small quantity of requirements, nevertheless, for a large volume this technique is not the most adequate one.

3.6 Use Cases

This technique is based on the definition of certain functionalities that are expected from a system and that allow it to interact with something or someone. These functionalities are called *use cases*. A use case can be defined as: "a textual narrative description of the processes of a business or system" [15], in which the system is considered a black box and from which the actors obtain answers [4]. As actors, will be understood the people or other systems that interact with the system whose requisites are being described [20]. Initially, this technique was proposed in [11]. From this publication, the most recognized specialists in Object-oriented methods have agreed in considering the *use cases* as an excellent way of specifying the external behavior of a system. Due to this, the notation of the *use cases* was incorporated to the standard language of modeling UML (Unified Modeling Language) [3].

4 Considerations to Apply RE Techniques in DM Projects

The establishment of the requirements in a Data Mining project constitutes a fundamental task in order to specify and validate the services that the system should provide, as well as the restrictions with which the work itself should be developed. This process is essential, because the most common and costly errors that have to be corrected are a result of an inadequate Requirements Engineering process. Previously to the requirements specification is necessary to consider the following aspects:

i) To identify and know the objectives of the business (business model). Any Data Mining project should have as a final objective the generation of some type of benefit for the organization, either improving the efficiency of the business processes or discovering new sources of improvement.
ii) To identify the problem domain. In this context, the identification of the problem domain will allow to specify the area in which the Data Mining project will take

place. As an example, a general classification of problems types that a Data Mining project allows us to face are the following:

1. *Marketing.* It considers getting the greater quantity of related information to the business clients in order to establish potential clients, to determine who will buy, when and where, to improve the relationship with the clients, etc. That is, the result of the Data Mining project should allow planning in the best way the future marketing campaigns of the company.
2. *Market basket.* It considers the determination of product purchase patterns in the retail area. These patterns may trigger the initiation of guided promotion campaigns, new physical disposition of the items, multi-item pack offerings, offerings and promotions considering time patterns (e.g., day of the week), etc.
3. *Risks reduction.* In this case, Data Mining will allow an automatic evaluation of risks, base on previous experiences.
4. *Frauds detection.* Users will be able to obtain models that would allow discovering possible frauds in base to anomalous behaviors detection models.
5. *Quality control.* It considers the definition of models that will allow the precise and anticipated detection of faulty products.

iii) To map the business problem into a Data Mining problem. After identifying the domain of the problem, it should be mapped into a Data Mining problem, that is, it should be considered that each type of application of Data Mining is related to one or more type of tasks. The main types of Data Mining tasks are:

1. *Association.* It is basically used to discover relations among attributes. That is, the idea is to discover rules that identify behavior patterns, and it is largely used in the market basket domain.
2. *Time Sequences.* It is similar to the association task, but in this task the time variable is incorporated.
3. *Classification.* This task uses a collection of data to develop a model that will be used to classify new unseen data. The predicted variable is a nominal one.
4. *Regression.* This task is similar to classification, but in this case the predictive variable can take possible unlimited numeric values.
5. *Clustering.* This task is utilized in typical segmentation applications, and consists in a division of the data into collections of related data or groups, in which the data in each identified group are similar, i.e., they share a number of similar characteristics.

iv) To define the generic components that a Data Mining requirement should consider, such as:

1. *Data component.* This component should respond to the question on what data and with what structure (model of data) they are needed, in function of the algorithmic technique that will support the application of the Data Mining task.
2. *Interface component.* It should answer the question about the format in which will be visualized or presented the project results.
3. *Usability and correctness component.* It considers the way in which the project results should contribute to the business and user objectives, and the degree of accuracy that will provide the model.

4. *Understandability Component.* It considers the way in which the Data Mining model can be understood and will allow justifying the achieved results.
5. *Resources Component.* It should consider the available resources for the project, such as personnel (e.g., business expert, data expert, Data Mining team, technical aid), and hardware and software platforms.
6. *Not functional components.* They are reutilization requirements, development environments, results availability and quality (delivery times), security and legislation.

5 A Model of the RE Process for Data Mining Projects

Based on the considerations mentioned before, now is the turn of applying a RE model to construct the requisite document. Determining the needs for the project development is a complex process, and there doesn't exist unique and standard techniques that provide a framework for development that guarantees good results. Considering that a Data Mining project is essentially a decision support system, in which is important the adequate problem domain understanding, the construction of the business and data models, and the basic elements seen in the RE processes mentioned before, it is proposed a RE process model for data mining projects which is shown in Figure 4.

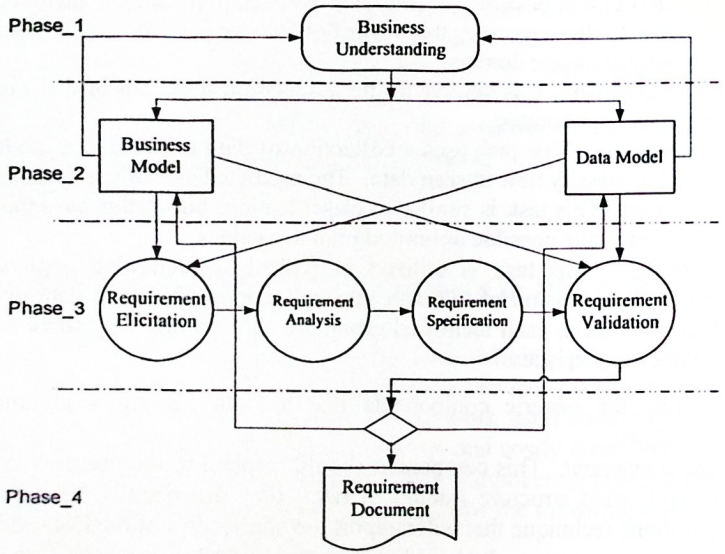

Fig. 4. RE Model Process proponed for Data Mining Projects.

This model is structured in the sequence of phases that should be carried out in order to generate the requisite document. The first phase, business understanding, considers the knowledge of the problem domain, the contexts, the organizational structure, the

decision-making levels, and the own vocabulary of the business domain. In the second phase, the business decisional model and the data model are built, which are the input necessary to tasks to be performed in the third phase (requirements elicitation, analysis, specification, and validation). The fourth and final phase corresponds to the construction of the requisite document.

The selection of RE techniques to use for the execution of each phase of the RE proposed model, and the success of the results reached will depend finally on the clients/users, and on the development team and its experience in similar projects.

6 Structuring Data Mining Requirements

Data Mining projects involve different stakeholders, who express the purpose of the project, indicate the direction of the activities, and define the expectations in terms of information goals for the organization, in a similar way to the development of a project based on data analysis, such as a Data Warehouse system [21]. Therefore, the requisites generation in Data Mining projects should consider different perspectives, corresponding to different stakeholders, each one associated to different abstraction levels in the organization. Thus, we can distinguish three levels of abstraction in the generation of requisites for Data Mining: business point of view, user perspective, and technical staff (development team) view.

From the business point of view, the requisites of Data Mining should consider the identification of at least the following information:

- Business goals
- Business opportunities
- Business necessities to be satisfied
- Stakeholders and the generated value for them
- Criteria to decide to start / non-start the project
- Problem domain background
- An appropriate business case to illustrate the project impact

From the user perspective, we should be able to identify and describe the tasks that the user should be able to perform or improve using the project results. Among other information, the user requirements should consider:

- The user specific goals
- Business questions that may be answered with the project
- The process and business rules related to the user requirements
- The opportunities to improve certain business processes
- The way that the results should be used (user case)
- The way that the results should be tested (test case)
- The user profiles associated to the project results.

From the development people side (data miners), the requisites must have enough information about the specific tasks to be performed by them, the data sources, as well as the results attributes or characteristics. The technical requisites must be

explicitly linked to business and user goals, specified before. Technical requisites should consider mainly:

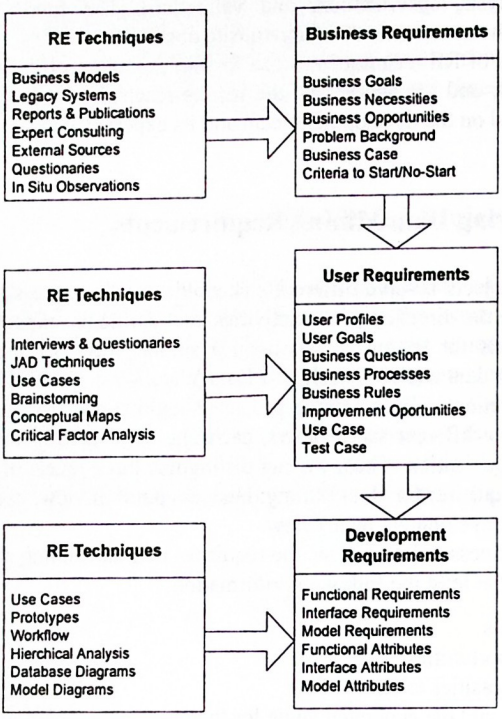

Fig. 5. Requisites structure for Data Mining and the corresponding RE techniques

a) *Functional Requirements,* which considers the tasks specification as extraction, loading, integration, cleansing and data transformation, as well as the identification of the type of task, data exploration, model generation to be carried out, and the process and results documentation.
b) *Interface Requirements,* that considers the characteristics specification of the interfaces to be developed during the project, such as database, software, and hardware interface
c) *Model Requirements,* which considers the identification of the type of model, its characteristics of understandability, ways to evaluate correctness and consistency of the generated model.
d) *Functional attributes,* that considers the definition of operational attributes, of performance, and of security of necessary tasks to be carried out by the development team of the project.
e) *Interface attributes,* that considers the definition of usability conditions and of form of the different interfaces required in the project.

f) *Model attributes,* that considers the establishment of criteria to generate the models and to evaluate its validity and acceptance.

Thus, we can summarize the different perspectives of the stakeholders involved in the definition of a requirement, to define a format for requirements acquisition in Data Mining projects, and identify associated RE techniques for each different vision or stakeholder perspective, as shown in Figure 5.

6 Conclusions

A Data Mining project has an exploratory nature, of decisional type and not of operational type that seeks to contribute value to the business through its data. Normally, the decision making processes are not well structured, with the consequent and inherent difficulty for modeling them. A Data Mining project is essentially approached to the comprehension and exploration of data and therefore, the requirements should be focused in determining the way the data influence in making decisions or in which way they impact in the decisions that are taken. Thus, to establish a unique criterion for the selection of the most appropriate techniques for its application in the different phases of the Requirements Engineering process is somewhat complex.

Our proposal constitutes a first step in order to establish a process model for the requirements definition in this type of project, under the consideration, that the different development processes models for Data Mining projects, such as CRISP-DM [5], SEMMA [19] or DMAMC [10], in spite of the fact that they present the requirements capture like a task to be performed, they do not indicate how to carry out this task, what techniques to utilize, neither the formats for the outputs that are proposed. Further work needs to be done in order to evaluate whether the proposed techniques are appropriate and useful to capture requisites in Data Mining projects.

The application of RE techniques in the proposed model should probable consider adaptations, and other aspects such as learning facility, cost, quality, completeness, time restrictions for its applicability, available personnel, etc.

Finally, we consider that it doesn't exist a valid and unique Requirements Management process and technique that can be applied in all the organizations and in all type of projects. Each organization should select or develop its process, in agreement with the type of product to be generated, to the organizational culture, and the level of experience and ability of the people involved in the Requirements Engineering process. Nevertheless, the identification of different perspectives of the same problem (obtaining of requirements in a Data Mining project) seeks to structure the process of RE, such that particular techniques can be applied to each perspective/stakeholder identified.

The next steps in this research work consider a detailed and formal description of the process here proposed the definition of criteria to evaluate the effectiveness of the proposed model, and its application to real Data Mining projects, in order to validate the model. This will allow us to illustrate, to refine, and to validate the proposed structure of Data Mining requirements.

References

1. Arango J. "Tormenta de Ideas", Colombia. Universidad EAFIT, 2002. D [en línea], disponible en: http://www.eafit.edu.co/tda/boletin/TORMENTA%20DE%20IDEAS.htm
2. Bahamonde J.M., Rossel R. "Un Acercamiento a la Ingeniería de Requisitos", Universidad Técnica Federico Santa María, 2003.
3. Booch, G., J. Rumbaugh, y I. Jacobson. "The Unified Modeling Language User Guide", Addison–Wesley, 1999.
4. Choque Guillermo. "Ingeniería de Requisitos", artículo de divulgación, Ingeniería de Software, Universidad Mayor de San Andrés, 2003.
5. Chapman P., (NCR), Clinton J., (SPSS) Kerber R., (NCR), Khabaza T. (SPSS), Reinartz T. (DaimlerChrysler), Shearer C. (SPSS), and Wirth R. (DaimlerChrysler). "CRISP-DM 1.0 step-by-step data mining guide", Technical report, 2000.
6. Dilauro, L. "¿What's nest in monitoring technology?", Data Mining finds a calling in centers", May 2000.
7. Gause, D. C. and G. M. Weinberg. "Exploring Requirements: Quality Before Design". Dorset House, 1989.
8. Jaap Gordijn, "Value-based Requirements Engineering Exploring Innovative e-Commerce Ideas", VRIJE UNIVERSITEIT, 2003.
9. Houghton Mifflin Company. "The American Heritage Dictionary of the English Language", 3rd Edition, Houghton Mifflin Company, Electronic Version. 1992.
10. http://www.isixsigma.com, "consulta sobre metodología 6-Sigma" [en línea], disponible en: http://www.isixsigma.com/sixsigma/six_sigma.asp.
11. Jacobson, I., M. Christerson, P. Jonsson, y G. Övergaard. "Object–Oriented Software Engineering: A Use Case Driven Approach ", Addison–Wesley, 4ta. edición, 1993.
12. Kdnuggets, http://www.kdnuggets.com, "consulta sobre metodologías utilizadas en Data Mining" , http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm
13. Komer, P. "Dirección de la Mercadotecnia", Séptima Edición. España. Prentice Hall, 1993.
14. Kotonya G. and Sommerville I. "Requirements Engineering. Processes and techniques", USA. J. Wiley, 1998.
15. Larman C. "UML y Patrones, introducción al análisis y diseño orientado a objetos", Ed. Prentice Hall, 1999.
16. Martínez de Pisón Ascacibar, F.J. "Optimización mediante técnicas de minería de datos del ciclo de recocido de una línea de galvanizado", Tesis Doctoral, Universidad de La Rioja, Servicio de Publicaciones, 2003.
17. Piattini, M. G., Calvo-Manzano, J. A., Cervera, J., Fernández, L. "Análisis y Diseño Detallado de Aplicaciones Informáticas de Gestión". Rama, 1996.
18. Raghavan, S., G. Zelesnik, y G. Ford. "Lecture Notes on Requirements Elicitation", Educational Materials CMU/SEI–94–EM –10, Software Engineering Institute, Carnegie Mellon University, 1994. http://www.sei.cmu.edu
19. SEMMA, http://www.sas.com/technologies/analytics/datamining/miner/semma.html
20. Scheneider, G. & Winters. Applying Use Cases: a Practical Guide. Addison–Wesley, 1998.
21. Schiefer, J. List, B., Bruckner, R.M., A Holistic Approach for Managing Requirements of Data Warehouse Systems, Proceedings of the Eighth Americas Conference on Information Systems, 2002.
22. Sommerville, I., "Ingeniería de Software", 6ta. Edición, Ed. Addison Wesley, 2002.
23. Zornes A., META Group Research-Delta Summary, "The Top 5 Global 3000 Data Mining Trends for 2003/04", Enterprise Analytics Strategies, Application Delivery Strategies, Delta, 2061, March 26, 2003.