García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo-Cattaneo, F., Rodríguez, D., Pytel, P., Vanrell. J. (2011)
Towards an Information Mining Engineering. En Software Engineering, Methods, Modeling and Teaching
Sello Editorial Universidad de Medellín. ISBN 978-958-8692-32-6. Páginas 83-99

TOWARDS AN INFORMATION MINING ENGINEERING

García-Martínez, Ramón[1], Britos, Paola[2], Pesado, Patricia[3], Bertone, Rodolfo[3], Pollo-Cattaneo, Florencia[4], Rodríguez, Darío[1], Pytel, Pablo[1,4], Vanrell. Juan[4]

[1] Information Systems Research Group. National University Lanus. rgarcia@unla.edu.ar

[2] Information Mining Research Group. National University Rio Negro. paobritos@gmail.com

[3] School of Computer Science. National University La Plata. {ppesado, pbertone}@lidi.info.unlp.edu.ar

[4] Information Systems Methodologies Research Group. Technological National University at Buenos Aires. fpollo@posgrado.frba.utn.edu.ar

1. INTRODUCTION

Business Intelligence offers an interdisciplinary approach (within which are included the Information Systems) focuses on generating knowledge that supports the management decision-making and generation of strategic plans at organizations (Thomsen, 2003). Information Mining is the sub-discipline of Information Systems which provides to Business Intelligence (Negash & Gray, 2008) the analysis and synthesis tools to extract non-trivial knowledge which is located (implicitly) in the available data from different information sources (Schiefer *et al*., 2004). For an expert, or the person in charge of an information system, normally the data itself is not the most relevant, but it is the knowledge included in their relations, fluctuations and dependencies. Information Mining Process, can be defined as a set of logically related tasks (Curtis *et al*., 1992) that are executed to achieve, from a set of information with a degree of value to the organization, another set of information with a greater degree of value than the initial one (Ferreira *et al*., 2005; Kanungo, 2005). Once the business intelligence problem is identified, the Information Mining Engineer selects the sequence of information mining processes to be run to solve the business intelligence problem. Each Information Mining Process has several data mining techniques that may be chosen to carry on the job. Several of these techniques come from the field of Machine Learning (García-Martínez *et al*., 2003).

In early stages of our research work, we have observed the indiscriminate use of terms "data mining" and "information mining" to refer to the same body of knowledge. We think that is a kind confusion similar to terms “computer-systems” and “information-systems”. Data Mining is related to the technology (algorithms) and Information Mining is related to the processes and methodologies. Data Mining is close to programming and Information Mining is close to software engineering. In this context is an open issue the need of organizing the body of knowledge related to Information Mining Engineering, establishing that data mining is related to algorithms; and information mining is related to processes and methodologies. A new body of knowledge is necessary for the Information Mining Engineering with a special focus on the implementation in the industry. One of the reasons for developing an Information Mining Engineering has been the discovery of a lack of techniques associated to the execution of each of the phases in the information mining methodologies (García-Martínez, *et al*., 2011). Although Software Engineering provides many methods, techniques and tools, they are not useful because they do not care in the practical aspects of the requirement specification of information mining projects. Therefore, it is necessary the development and validation of methods, techniques and tools that will aid the practitioners in the software area and provide the necessary objectivity, rationality, generalization and reliability to the Information Mining Process.

During the last decade we have developed field experience in information mining in the following domains: classification of asteroids family (Perichinsky *et al*., 2003), identification of human faces rules (Britos *et al*., 2005), detection of changes in users consumption (Grosser *et al*., 2005; Britos *et al*., 2008d), pattern discovery in meteorological events (Cogliati *et al*., 2006), prediction in community health (Felgaer *et al*., 2006), detection of breast injuries (Ferrero *et al*., 2006), discovery of web sites usage (Britos *et al*., 2007), selection of pedagogical protocols (Britos *et al*., 2008b), discovery of programming misunderstandings (Britos et al., 2008e), criminal pattern detection (Valenga *et al*., 2008e), discovery of damages patterns in car industry (Flores *et al*., 2009), pattern discovery in university students desertion (Kuna *et al*., 2010a; 2010b), among others. Based on our field experience and the existing body of Software Engineering knowledge, in this paper we propose for Information Mining: a process model (section 2), a requirement elicitation process (section 3), an estimation method (section 4) and, finally, a set of processes for Information Mining based and the different associated data mining techniques (section 5).

2. PROCESS MODEL FOR INFORMATION MINING

Software Engineering uses different models and methodologies in order to carry information technology projects with a high level of predictability and quality. They allow controlling the final quality of each developed product by establishing control points for each of the phases which are part of the production process. Understanding as production process, not only the production itself, but also the tasks related to the project management and the company that developed it. In the case of classical software development projects, there are several well tested models as CMM (SEI, 2006), or the

SMEs model COMPETISOFT (Oktaba *et al*., 2007). These models have been used on several projects and they can be considered as stable and high tested models, in case of COMPETISOFT, the high tested model is MoProSoft (Oktaba *et al*., 2005), which is the base model from which COMPETISOFT was created. However, these models are considered as not adequate for companies dedicated to carry on Information Mining projects because they have different properties, especially on the operation process. The most visible difference is on the software development process and the software maintenance where COMPETISOFT defines as a natural process the typical phases of a traditional software development (i.e. analysis, design, development, integration and testing). On the same line, the most important methodologies for Information Mining projects lack of tools to support completely the project management phases which are well defined on COMPETISOFT and grouped on a specific process. Although the scientific community considers the methodologies CRISP-DM (Chapman *et al*., 2000), SEMMA (SAS, 2008) and P3TQ (Pyle, 2003) as proven for Information Mining projects, they have problems when trying to define the phases related to project management. The elements of project management are mixed with project development process. In other hand, tasks which should follow all the development process such as project monitoring, verification and measurement are not considered in the referenced methodologies. Clearly all the activities related to project management are activities which should be executed at the same time of the project development on a separated process. To solve the detected problems, we propose a Process Model for Information Mining (Vanrell *et al*., 2010a; 2010b) based on a mixture of COMPETISOFT and CRISP-DM. The proposed Process Model has been obtained by removing all the unnecessary phases, by the adaptation of the necessary phases for an Information Mining project and by proposing new phases for specific aspects of information mining projects. CRISP-DM has been selected as a reference methodology because information mining community considers that t contains more quantity of the operation level elements than P3TQ and SEMMA. The proposed phases of the Information Mining Process Model for Project Management and its associated tasks are shown in table 1.

SUB-PROCESS	TASK	OUTPUT
Planification / Business Understanding	Business understanding	▪ Background ▪ Business objectives ▪ Business success criteria
	Definition of the specific process based on the description of the Project and the development and maintenance process	▪ Specific Process (part of the Development Plan)
	Definition of a delivery protocol	▪ Delivery Plan
	Definition of stages and tasks based on the description of the Project and the specific project	▪ Specific Process (part of the Development Plan)
	Determinate estimated time for each activity	▪ Activities calendar (part of the Development Plan) incorporate the estimated time on the Project Plan
	Develop the acquisitions and training plan	▪ Acquisitions and training plan
	Define the work team	▪ Work team (part of the Development Plan)
	Define the activities calendar	▪ Activities calendar (part of the Development Plan)
	Calculate the estimated cost of the project	▪ Estimated cost (part of the Project Plan)
	Assess situation	▪ Inventory of resources ▪ Requirements, assumptions and constraints ▪ Risks and contingencies (part of the Project Plan) ▪ Terminology ▪ Costs and benefits
	Develop a Project Plan	▪ Project Plan, include stages and activities, estimated time, acquisition and training plan, work team, estimated cost, calendar, risks and contingencies plan and deployment plan
	Develop a Development Plan	▪ Development Plan (include a product description and deliveries, specific process, work team and calendar) ▪ Initial list of tools and techniques
	Formalization of the start of a new project cycle	
Realization	Define the tasks with the team	
	Agree on the distribution tasks	
	Review the description of the product, the team and the calendar with the team leader	
	Review the accomplishment of the acquisition and training plan	▪ Monitoring report / Monitoring and maintenance plan
	Manage subcontracts	▪ Monitoring report / Monitoring and maintenance plan
	Collect reports of activities and measurements and improvement suggestions and work products.	▪ Monitoring report / Monitoring and maintenance plan ▪ Measurement report and improvement suggestions
	Register the real cost of the project	▪ Monitoring report / Monitoring and maintenance plan
	Review the track record based on the collected work products.	▪ Monitoring report / Monitoring and maintenance plan
	Review the finished products during the project	▪ Monitoring report / Monitoring and maintenance plan
	Receive and analyze changes request of the client	▪ Monitoring report / Monitoring and maintenance plan
	Realize meetings with the work team and client to report advances and make agreements	▪ Monitoring report / Monitoring and maintenance plan
Evaluation and control	Evaluate the accomplishment of the project plan and development plan	▪ Monitoring report / Monitoring and maintenance plan
	Analyze and control of risks	▪ Monitoring report / Monitoring and maintenance plan
	Generate the monitoring project report	▪ Monitoring report / Monitoring and maintenance plan
Close / Deployment	Formalize the end of the project or cycle	▪ Acceptance document
	Closet he contracts with subcontractors	
	Generate the measurements and improvement suggestions report	▪ Measurements and improvement suggestions report – Lesson learned
	Plan deployment	▪ Deployment plan (part Project Plan)

Table 1. Sub-processes, Tasks and Outputs of the Project Management Process (Vanrell et al., 2010a; 2010b)

The proposed phases of the Information Mining Process Model for Project Development and its associated tasks are shown in table 2.

SUB-PROCESS	TASK	OUTPUT
Business understanding	Determine data mining goals	▪ Data Mining goals ▪ Data Mining success criteria
Data understanding	Collect initial data	▪ Initial data collection report
	Describe data	▪ Data description report
	Explore data	▪ Data exploration report
	Verify data quality	▪ Data quality report
Data preparation	Initial tasks	▪ Datasets ▪ Datasets description
	Select data	▪ Rationale for inclusion/exclusion
	Clean data	▪ Data cleaning report
	Construct data	▪ Derived attributes ▪ Generated records
	Integrate data	▪ Merged data
	Format data	▪ Reformatted data
Modeling	Select modeling technique	▪ Modeling technique ▪ Modeling assumptions
	Generate test design	▪ Test design
	Build model	▪ Parameter settings ▪ Models ▪ Model description
	Assess Model	▪ Model assessment ▪ Revised parameter settings
Evaluation	Evaluate results	▪ Assessment of data mining results with respect to business success criteria ▪ Approved models
	Review process	▪ Review of process
	Determine next steps	▪ List of possible actions ▪ Decision
Deployment	Produce final report	▪ Final report ▪ Final presentation

Table 2. Sub-processes, Tasks and Outputs of the Project Development Process (Vanrell et al., 2010a; 2010b)

3. REQUIREMENTS ELICITATION PROCESS FOR INFORMATION MINING PROJECTS

The first task of the Project Management and Project Development processes, included in the proposed Process Model described in section 2, have the objective of finding and defining the objectives, goals and success criteria of the Information Mining Project. It is necessary to elicit the project requirements that should be satisfied.

The need to adapt traditional requirements engineering process for Information Mining systems is based on the premise that the requirements analysis for these types of systems differ substantially from requirements analysis for conventional information systems. Current Information Mining methodologies fail to elicit all the concepts needed during the business understanding phase of Information Mining. CRISP-DM elicits on set of concepts, P3TQ another and SEMMA yet a third. In general, these methodologies attend to concepts related to determining business objectives and assess situations (at least for one methodology) and concepts related to determine data mining goals and project plan production are not attended.

We have proposed a methodology (Britos *et al*., 2008c; Pollo-Cattaneo *et al*., 2009; 2010a) that is more robust than current ones, because it elicits all the necessary concepts to model the Information Mining project's requirements. Once the needed concepts have been identified, it is necessary to establish the steps to elicit those concepts. The proposed structure is similar to those proposed by Software Engineering that allows progressing over the needed concepts to maintain their natural order. In the business understanding phase of any Information Mining methodology we propose an Information Mining project requirements elicitation process of five steps that is shown in Figure 1.

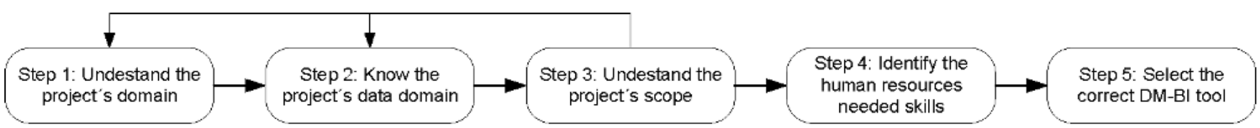

Fig. 1. Process of requirements elicitation (Britos et al., 2008c; Pollo-Cattaneo et al., 2009; 2010a)

The purpose of each step is: "understand the project’s domain" consists in establishing communication channels in ordinary language among persons involved into the Information Mining project; "know the project’s data domain" consists in establishing the project’s requirements; the data needed for those requirements and its location, risks involved in the data and the requirements’ development, the data and requirements’ restrictions, and finally its suppositions; "understand the project’s scope" consists in achieving the Information Mining projects objective, its limitations, expectations and risks; "identify the human resources needed skills" consists in knowing the list of human resources involved, its restrictions, risks and responsibilities; and finally, "select the correct Data Mining tool" consists in selecting an adequate tool according to the information obtained in the earlier steps. To know the project's data domain in terms of requirements goal, the requirements information of data source information, requirements results suppositions, requirements restrictions, attributes involved in requirements, risks and contingency plans; it is necessary to understand the project's domain in terms of definitions, acronyms and abbreviations. To understand the project's scope in terms of project objectives, successful criteria of the project, project expectations, project suppositions, restrictions, risks, and contingency plans; it is necessary to know the project's data domain in terms of requirements goal, the requirements information of data source information, requirements results suppositions, requirements restrictions, attributes involved in requirements, requirements risks and requirements contingency plans. To identify the human resources needed in terms of defining human resources involved; it is necessary to understand the project's scope in terms of project objectives, project successful criteria, project expectations, project suppositions, project restrictions, project risks, and contingency plans. To identify the human resources needed skills in terms of defining human resources involved; it is necessary to select the correct Data Mining tool in terms of tools evaluation. The conceptual dependency among the needed concept is shown in Figure 2.

A set of templates have been defined as products, the complete set of themplates an examples may be seen in (Britos *et al*., 2008c). Each template is associated to each concept. These templates have a detailed description of the concepts to be elicited. The templates allow the concept evolution through the requirements elicitation process. The relation between the elicited concepts as products and the steps of the proposed process to generate them is shown in Table. 3.

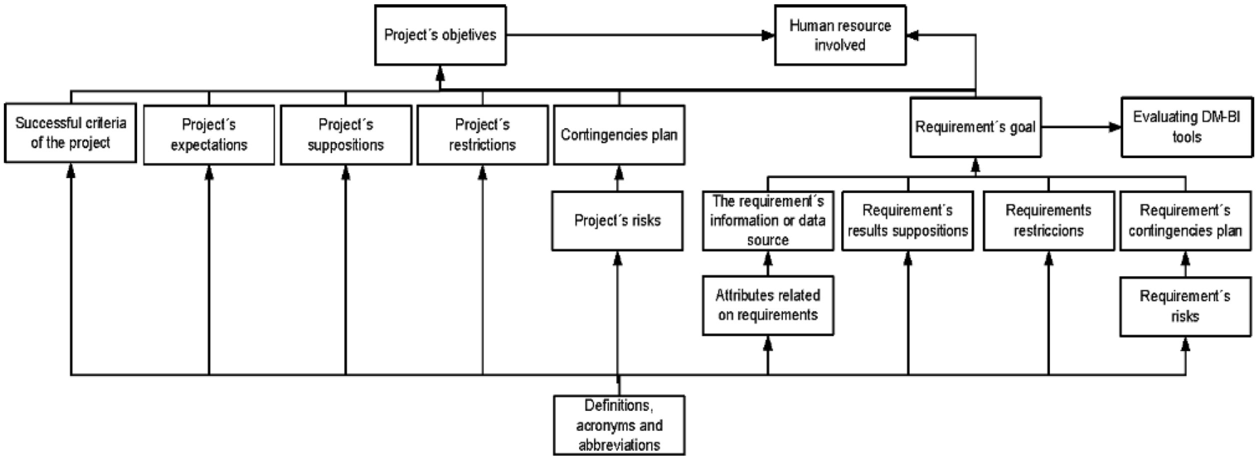

Fig. 2. Cross references of elicited concepts represented by the templates (Britos et al., 2008c)

STEPS \ PRODUCT (concepts to be educed)	Definitions, acronyms and abbreviations	Project's objectives	Successful criteria of the project	Project's expectations	Project's suppositions	Project's restrictions	Project's risks	Contingencies plan	Human resource involved	Requirement's goal	The requirement's information or data source	Requirement's results suppositions	Requirements restrictions	Attributes related on requirements	Requirement's risks	Requirement's contingencies plan	Evaluating DM-BI tools
Understand the project's domain	■								■								
Know the project's data domain	■								■	■	■	■	■	■	■	■	
Understand the project's scope	■	■	■	■	■	■	■	■	■								
Identify the human resources needed skills	■								■								
Select the correct DM-BI tool	■								■	■			■	■			■

Table. 3. Relation among products (elicited concepts) and process steps (Britos et al., 2008c)

3. EMPIRICAL ESTIMATION METHOD FOR INFORMATION MINING PROJECTS

Software Project Management process includes a set of activities that are referred as project planning. Before the project start, estimation shall be performed of: the tasks to be executed, the necessary resources and time that will be elapsed from the beginning until the end of the project (Pressman, 2004). Within the Process Model described in section 2, the task "Calculate the estimated cost of the project" also requires a planning process that allows estimating their times; however, because of the existing differences between a conventional software project and an Information Mining project, the usual methods of estimation are not applicable. The construction of estimation methods of software projects, that achieve predictive results about the resources to be used and that are consistent on the best possible approach to the reality, is an open problem in the field of information systems. Information Mining projects do not escape from this necessity and the history of engineering in general and information-systems in particular, recorded that the first approaches are always of empirical nature.

We have developed an experiment (Rodriguez *et al*., 2010; Pytel *et al*., 2011) which goal is getting the empirical percentage distribution of the quantity of time/work that (in a Information Mining project) takes the execution of each of the tasks associated with the sub-phases of an industry proven Methodology (CRISP-DM). We have focused on projects for small and medium-sized enterprises. By knowing the time spent in any of the sub-phases, we can have an approximation to the times of the other sub-phases and the global estimation of the project. The obtained results (shown in Table 4) include those phases and sub-phases that perform a significant amount of time (more than 50%). The phases of "Business understanding" and "Modeling" use more than 50% of the time of the project, In case of "Business understanding" phase, the sub-phases "Determine business objectives" and "Assess situation" used more than 70% of the time. On the other hand, in the "Modeling" phase, the sub-phase "Build model" requires 62.97% of the time of the phase.

PHASE	% of TIME
Phase 1 Business understanding	20.70
Phase 2 Data understanding	10.90
Phase 3 Data preparation	15.61
Phase 4 Modeling	34.41
Phase 5 Evaluation	7.45
Phase 6 Deployment	10.93

Table 4. Effort (in % of time) of each phase of CRISP-DM methodology (Rodriguez et al., 2010)

5. SET OF INFORMATION MINING PROCESSES AND ASSOCIATED DATA MINING TECHNOLOGIES

Within the Development Management process of the Process Model described in section 2, the tasks of the "Data preparation" and "Modeling" sub-processes uses certain data mining algorithms and techniques in order to process the available information. All sources of information (databases, files, others) that are related to the business intelligence problem are identified and integrated together as a single source of information which will be called integrated data base. We have proposed five Information Mining processes (Britos *et al*., 2008a; Britos y Garcia-Martinez, 2009; Pollo-Cattaneo *et al*., 2010b) described in the following sub-sections: discovery of behavior rules (sub-section 5.1), discovery of groups (sub-section 5.2), discovery of significant attributes (sub-section 5.3), discovery of group-membership rules (sub-section 5.4) and weighting of significant attribute related to behavior or membership rules (sub-section 5.5). Each process has been associated for the usage of the following techniques: TDIDT (Top Down Induction Decision Trees) algorithm (Quinlan, 1986), Kohonen's Self-Organizing Maps (SOM) (Kohonen, 1995) and Bayesian Networks (Heckerman *et al*., 1995).

The proposed Information Mining processes have been validated in the following domains: political alliances, medical diagnosis and user behavior. A full detailed report of these validations can be seen in (Britos, 2008).

5.1. Process of Discovery of Behavior Rules

The process for discovery of behavioral rules applies when it is necessary to identify which are the conditions to get a specific outcome in the problem domain. The following problems are examples among others that require this process: identification of the characteristics for the most visited commercial office by customers, identification of the factors that increase the sales of a specific product, definition of the characteristics or traits of customers with high degree of brand loyalty, definition of demographic and psychographic attributes that distinguish the visitors to a website. For the discovery of behavioral rules from classes attributes in a problem domain that represents the available information base, it is proposed the usage of TDIDT induction algorithms (Britos *et al*., 2008) to discover the rules of behavior for each class attribute. Based on the integrated data base, the class attribute is selected. As a result of applying TDIDT to the class attribute, a set of rules which define the behavior of that class is achieved.

5.2. Process of Discovery of Groups

The process of discovery of groups applies when it is necessary to identify a partition on the available information base of the problem domain. The following problems are examples among others that require this process: identification of the customers segments for banks and financial institutions, identification of type of calls of customer in telecommunications companies, identification of social groups with the same characteristics, identification of students groups with homogeneous characteristics. For the discovery of groups (Kaufman & Rousseeuw, 1990; Grabmeier & Rudolph, 2002) in information bases of the problem domain for which there is no available "a priori" criteria for grouping, it is proposed the usage of Kohonen's Self-Organizing Maps or SOM (Ferrero *et al*., 2006; Britos *et al*., 2008; Britos *et al*.,, 2008). The use of this technology intends to find if there is any group that allows the generation of a representative partition for the problem domain which can be defined from available information bases. Based on the integrated data base, the self-organizing map (SOM) is applied. As a result of the application of using SOM, a partition of the set of records in different groups, that will be called identified groups, is achieved. For each identified group, the corresponding data file will be generated.

5.3. Process of Discovery of Significant Attributes

The process of discovery of significant attributes applies when it is necessary to identify which are the factors with the highest incidence (or occurrence frequency) for a certain outcome of the problem. The following problems are examples among others that require this process: factors with incidence on the sales, distinctive features of customers with high degree of brand loyalty, key-attributes that characterize a product as marketable, key-features of visitors to a website. Bayesian Networks (Britos *et al*., 2008) allows seeing how variations in the values of attributes, impact on the variation of the value of class attribute. The use of this process seeks to identify whether there is any interdependence among the attributes that model the problem domain which is represented by the available information base. Based on the integrated data base, the class attribute is selected. As a result of the application of the Bayesian Networks structural learning to the file with the identified class attribute, the learning tree is achieved. The Bayesian Networks predictive learning is applied to this tree obtaining the tree of weighting interdependence which has the class attribute as a root and to the other attributes with frequency (incidence) related the class attribute as leaf nodes.

5.4. Process of Discovery of Group-membership Rules

The process of discovery of group membership rules applies when it is necessary to identify which are the conditions of membership to each of the classes of an unknown partition “a priori”, but existing in the available information bases of the problem domain. The following problems are examples among others that require this process: types of customer’s profiles and the characterization of each type, distribution and structure of data of a web site, segmentation by age of students and the behavior of each segment, classes of telephone calls in a region and the characterization of each class. For running the process of discovery of group-membership rules it is proposed to use self-organizing maps (SOM) for finding groups and; once the groups are identified, the usage of induction algorithms (TDIDT) for defining each group behavior rules (Britos *et al*., 2005; Cogliati *et al*., 2006a).

5.5. Process of Weighting of Behavior or Group-membership Rules

Based on the integrated data base, the self-organizing maps (SOM) are applied. As a result of the application of SOM, a partition of the set of records in different groups is achieved which is called identified groups. The associated files for each identified group are generated. This set of files is called "ordered groups". The "group" attribute of each ordered group is identified as the class attribute of that group, establishing it in a file with the identified class attribute (GR). Then is applied TDIDT to the class attribute of each "GR group" and the set of rules that define the behavior of each group is achieved. The procedure to be applied when there are classes/groups no identified includes the identification of all sources of information (databases, files, others), and then they are integrated together as a single source of information which will be called integrated data base. Based on the integrated data base, the self-organizing maps (SOM) are applied. As a result of the application of SOM, a partition of the set of records in different groups is achieved. These groups are called identified groups. For each identified group, the corresponding data file will be generated. This set of files is called "ordered groups". The group attribute of each "ordered group" is identified as the class attribute of that group, establishing it in a file with the identified class attribute (GR). As a result of the application of the structural learning, the learning tree is achieved. The predictive learning is applied to this tree obtaining the tree of weighting interdependence. The root is the group attribute and the other attributes as leaf nodes labeled with the frequency (incidence) on the group attribute.

6. CONCLUSIONS

The history of Information Mining began with the systematization of machine learning algorithms applied to knowledge discovery over a quarter of a century ago. For many years the interest of the research community has been focused more in the algorithms than in the processes. The increasing incorporation of the engineering vision to software projects results in the need for the same type of vision in information mining projects.

In the last decade, we have been developing experience in the field of Information Mining and we have felt the absence of an Information Mining Engineering. During these years we have built, based on Software Engineering principles, a set of tools that have matured as the cornerstones of our own version of an Information Mining Engineering, which was used in Information Mining Projects, developed for small and medium enterprises. This paper seeks to share our experience-based acquired knowledge with the academic community.

7. REFERENCES

Britos, P. (2008a). *Procesos de Explotación de Información Basados en Sistemas Inteligentes*. Tesis Doctoral. Facultad de Informática. Universidad Nacional de La Plata. http://postgrado.info.unlp.edu.ar/Carrera/Doctorado/Tesis/Britos-Tesis%20 Doc toral.pdf. Last access December 2010.

Britos, P. Abasolo, M., García-Martínez, R. y Perales, F. (2005). *Identification of MPEG-4 Patterns in Human Faces Using Data Mining Techniques*. Proceedings 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2005. Pp. 9-10.

Britos, P., Cataldi, Z., Sierra, E., García-Martínez, R. (2008b). *Pedagogical Protocols Selection Automatic Assistance*. Lecture Notes on Artificial Intelligence, 5027: 331-336.

Britos, P., Dieste, O., García-Martínez, R. (2008c). *Requirements Elicitation in Data Mining for Business Intelligence Projects*. In Advances in Information Systems

Research, Education and Practice. David Avison, George M. Kasper, Barbara Pernici, Isabel Ramos, Dewald Roode Eds. (Boston: Springer), IFIP Series, 274: 139–150.

Britos, P., García-Martínez, R. (2009). *Propuesta de Procesos de Explotación de Información*. Proceedings XV Congreso Argentino de Ciencias de la Computación. Workshop de Base de Datos y Minería de Datos. Pp. 1041-1050. ISBN 978-897-24068-4-1.

Britos, P., Grosser, H., Rodríguez, D., García-Martínez, R. (2008d). *Detecting Unusual Changes of Users Consumption*. In Artificial Intelligence in Theory and Practice II, ed. M. Bramer, (Boston: Springer), IFIP International Federation for Information Processing Series, 276: 297-306.

Britos, P., Jiménez Rey, E., García-Martínez, E. (2008e). *Work in Progress: Programming Misunderstandings Discovering Process Based On Intelligent Data Mining Tools*. Proceedings 38th ASEE/IEEE Frontiers in Education Conference. Session F4H: Assessing and Understanding Student Learning. ISBN 978-1-4244-1970-8.

Britos, P., Martinelli, D., Merlino, H., García-Martínez, R. (2007). Web Usage Mining Using Self Organized Maps. *International Journal of Computer Science and Network Security,* 7(6):45-50. ISSN 1738-7906.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step Data Mining Guide*. http://www.crisp-dm.org/ CRISPWP-0800.pdf. Last access April 2011.

Cogliati, M., Britos, P., García-Martínez, R. (2006). *Patterns in Temporal Series of Meteorological Variables Using SOM & TDIDT*. In Artificial Intelligence in Theory and Practice, ed. M. Bramer, (Boston: Springer), IFIP International Federation for Information Processing Series, 217: 305-314.

Curtis, B., Kellner, M., Over, J. (1992). *Process Modelling*. Communications of the ACM, 35(9): 75-90.

Felgaer, P., Britos, P. and García-Martínez, R. (2006). *Prediction in Health Domain Using Bayesian Network Optimization Based on Induction Learning Techniques*. International Journal of Modern Physics C 17(3): 447-455.

Ferreira, J., Takai, O. & Pu, C. (2005). *Integration of Business Processes with Autonomous Information Systems: A Case Study in Government Services*. Proceedings Seventh IEEE International Conference on E-Commerce Technology. Pp. 471-474.

Ferrero, G., Britos, P. & García-Martínez, R., (2006). *Detection of Breast Lesions in Medical Digital Imaging Using Neural Networks*. In IFIP International Federation for Information Processing, Volume 218, Professional Practice in Artificial Intelligence, eds. J. Debenham, (Boston: Springer), Pp. 1-10.

Flores, D., Garcia-Martinez, R. Fernandez, E., Merlino, H., Rodriguez, D., Britos, P. (2009). *Detección de Patrones para la Prevención de Daños y/o Averías en la Industria Automotriz*. Proceedings XV Congreso Argentino de Ciencias de la Computación.

Workshop de Base de Datos y Minería de Datos. Págs. 1021-1030. ISBN 978-897-24068-4-1.

García Martínez, R., Servente, M. y Pasquini, D. 2003. *Sistemas Inteligentes*. Editorial Nueva Librería. ISBN 987-1104-05-7

García-Martínez, R., Lelli, R., Merlino, H., Cornachia, L., Rodriguez, D., Pytel, P. & Aboleya, H. (2011). *Ingeniería de Proyectos de Explotación de Información para PYMES*. Proceedings XIII Workshop de Investigadores en Ciencias de la Computación (in pret).

Grabmeier, J., Rudolph, A. (2002). *Techniques of Cluster Algorithms in Data Mining*. Data Mining and Knowledge Discovery, 6(4): 303-360.

Grosser, H., Britos, P. y García-Martínez, R. (2005). *Detecting Fraud in Mobile Telephony Using Neural Networks.* Lecture Notes in Artificial Intelligence 3533:613-615

Heckerman, D., Chickering, M. & Geiger, D. (1995). *Learning bayesian networks, the combination of knowledge and statistical data*. Machine learning 20: 197-243.

Kanungo, S. (2005). *Using Process Theory to Analyze Direct and Indirect Value-Drivers of Information Systems*. Proceedings of the 38th Annual Hawaii International Conference on System Sciences. Pp. 231-240.

Kaufmann, L. & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons Publishers.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer Verlag Publishers.

Kuna, H., García Martínez, R. Villatoro, F. (2010a). Pattern Discovery in University Students Desertion Based on Data Mining. *Advances and Applications in Statistical Sciences Journal*, 2(2): 275-286. ISSN 0974-6811.

Kuna, H., García-Martínez, R., Villatoro, F. (2010b). *Identificación de Causales de Abandono de Estudios Universitarios. Uso de Procesos de Explotación de Información*. Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología 5: 39-44.

Negash, S. & Gray, P. (2008). *Business Intelligence*. In Handbook on Decision Support Systems 2, eds. F. Burstein y C. Holsapple (Heidelberg, Springer), Pp. 175-193.

Oktaba, H., Alquicira Esquivel, C., Ramos, A. S., Martínez Martínez, A., Quintanilla Ozorio, G., Ruvalcaba López, M., López Lira Hinojo, F., Rivera López, M. E., Orozco Mendoza, M. J., Fernández Ordoñez, Y. & Flores Lemus, M. A. (2005). Modelo de Procesos para la Industria de Software. Secretaría de Economía de México.

Oktaba, H., Garcia, F., Piattini, M., Ruiz, F., Pino & F.J., Alquicira, C. (2007). *Software Process Improvement: The COMPETISOFT Project*. Computer 40(10): 21-28.

Perichinsky, G., Servetto, A., García-Martínez, R., Orellana, R. y Plastino, A.2003. *Taxomic Evidence Applying Algorithms of Intelligent Data Minning Asteroid Families*. Proceedings International Conference on Computer Science, Software Engineering, Information Technology, e-Bussines & Applications. Pp. 308-315. Río de Janeiro (Brasil). ISBN 0-9742059-3-7.

Pollo-Cattaneo, F., Britos, P. Pesado, P., García-Martínez, R. (2009). *Metodología para Especificación de Requisitos en Proyectos de Explotación de Información*. Proceedings XI Workshop de Investigadores en Ciencias de la Computación. Pp. 467-469. ISBN 978-950-605-570-7.

Pollo-Cattaneo, F., Britos, P., Pesado, P., García-Martínez, R. (2010a). *Proceso de Educción de Requisitos en Proyectos de Explotación de Información.* In Ingeniería de Software e Ingeniería del Conocimiento: Tendencias de Investigación e Innovación Tecnológica en Iberoamérica, Eds: R. Aguilar, J. Díaz, G. Gómez, E- León. Pp. 01-11. Alfaomega Grupo Editor. ISBN 978-607-707-096-2.

Pollo-Cattaneo, F., Britos, P., Pesado, P., García-Martínez, R. (2010b). *Ingeniería de Procesos de Explotación de Información*. In Ingeniería de Software e Ingeniería del Conocimiento: Tendencias de Investigación e Innovación Tecnológica en Iberoamérica, Eds: R. Aguilar, J. Díaz, G. Gómez, E- León. Pp. 252-263. Alfaomega Grupo Editor. ISBN 978-607-707-096-2.

Pressman, R. (2004). *Software Engineering: A Practitioner's Approach*. Editorial Mc Graw Hill.

Pyle, D. (2003). *Business Modeling and Business intelligence*. Morgan Kaufmann Publishers.

Pytel, P., Tomasello, M., Rodríguez, D.; Arboleya, H., Pollo-Cattaneo. M., Britos, P., García-Martínez, R. *Estimación de Proyectos de Explotación de Información Estudio Comparado de Modelos Analíticos y Empíricos*. Proceedings XIII Workshop de Investigadores en Ciencias de la Computación (in press).

Quinlan, J. (1986). *Induction of decision trees*. Machine Learning, 1(1): 81-106.

Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. (2010). *Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información*. Anales del XVI Congreso Argentino de Ciencias de la Computación. Pp. 664-673. ISBN 978-950-9474-49-9.

SAS, (2008). SAS Enterprise Miner: SEMMA. http://www.sas.com/technologies/analytics /datamining/miner/semma.html. Last access June 2008.

Schiefer, J., , Jeng, J., Kapoor, S. & Chowdhary, P. (2004). *Process Information Factory: A Data Management Approach for Enhancing Business Process Intelligence*. Proceedings 2004 IEEE International Conference on E-Commerce Technology. Pp. 162-169.

SEI (2006). *CMMI-DEV for Development, Vers. 1.2*. Software Engineering Institute Carnegie Mellon University. http://www.sei.cmu.edu/reports/06tr008.pdf. Last access April, 2011.

Thomsen, E. (2003). BI's Promised Land. Intelligent Enterprise, 6(4): 21-25.

Valenga, F., Fernández, E., Merlino, H., Rodríguez, D., Procopio, C., Britos, P., García-Martínez, R. 2008. *Minería de Datos Aplicada a la Detección de Patrones Delictivos en Argentina*. Proceedings VII Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento. Pp. 31-39. ISSN 1390-292X.

Vanrell, J., Bertone, R., García-Martínez, R. (2010). *Modelo de Proceso de Operación para Proyectos de Explotación de Información.* Anales del XVI Congreso Argentino de Ciencias de la Computación. Pp. 674-682. ISBN 978-950-9474-49-9.

Vanrell, J., Bertone, R., García-Martínez, R. (2010). *Un Modelo de Procesos de Explotación de Información*. Proceedings XII Workshop de Investigadores en Ciencias de la Computación. Pp. 167-171. ISBN 978-950-34-0652-6.