A Proposal of a Process Model for Requirements Elicitation in Information Mining Projects

Mansilla, D.[1], Pollo-Cattaneo, M.[1, 2], Britos, P.[3], García-Martínez, R.[4]

[1] Information System Methodologies Research Group. Technological National University. Buenos Aires, Argentina.
[2] PhD Program in Computer Science. Computer Science School. National University of La Plata. Buenos Aires, Argentina.
[3] Information Mining Research Group. National University of Rio Negro at El Bolson, Río Negro, Argentina.
[4] Information Systems Research Group. National University of Lanus, Buenos Aires, Argentina.

dmansilla@educ.ar; fpollo@posgrado.frba.utn.edu.ar; pbritos@unrn.edu.ar; rgarcia@unla.edu.ar

Abstract. A problem addressed by an information mining project is transforming existing business information of an organization into useful knowledge for decision making. Thus, the traditional software development process for requirements elicitation cannot be used to acquire required information for information mining process. In this context, a process of requirements gathering for information mining projects is presented, emphasizing the following phases: conceptualization, business definition and information mining process identification.

Keywords. Process, elicitation, information mining projects, requirements.

1 Introduction

Traditional Software Engineering offers tools and process for software requirements elicitation which are used for creating automatized information systems. Requirements are referred as a formal specification of what needs to be developed. They are descriptions of the system behaviour [1].

Software development projects usually begin by obtaining an understanding of the business domain and rules that govern it. Understanding business domains help to identify requirements at the business level and at product level [2], which define the product to be built considering the context where it will be used. Models such as Context Diagram, Data Flow Diagrams and others are used to graphically represent the business process in the study and are used as validation tools for these business processes. A functional analyst is oriented to gather data about inputs and outputs of the software product to be developed and how that information is transformed by the software system.

Unlike software development projects, the problem addressed by information mining projects is to transform existing information of an organization into useful knowledge for decision making, using analytical tools [3]. Models for requirements elicitation and project management, by focusing on the software product to be developed, cannot be used to acquire required information for information mining processes. In this context, it is necessary to transform existing experience in the use of requirements elicitation tools in the software development domain into knowledge that can be used to build models used in business intelligence projects and in information mining process [4] [5] [6].

This work will describe the problem (section 2), it will present a proposal for a process model for requirements elicitation in information mining projects (section 3), emphasizing in three phases: Conceptualization (section 3.1), Business Definition (section 3.2) and Information Mining Process Identification (section 3.4). Then, a study case is presented (section 4), and a conclusion and future lines of work are proposed (section 5).

2 State of current practice

Currently, several disciplines have been standardized in order to incorporate best practices learned from experience and from new discoveries.

The discipline of project management, for example, generated a body of knowledge where the different process areas of project management are defined. Software engineering specify different software development methodologies, like the software requirements development process [1]. On the other side, related to information mining projects, there are some methodologies for developing information mining systems such as DM [7], P3TQ [8], y SEMMA [9].

In the field of information mining there is not a unique process for managing projects [10]. However, there are several approaches that attempt to integrate the knowledge acquired in traditional software development projects, like the Kimball Lifecycle [11], and project management framework in medium and small organizations [12]. In [13] an operative approach regarding information mining project execution is proposed, but it does not detail which elicitation techniques can be used in a project.

The found problem is that previously mentioned approaches emphasize work methodologies associated with information mining projects and do not adapt traditional software engineering requirements elicitation techniques. In this situation, it is necessary to understand the activities that should be taken and which traditional elicitation techniques can be adapted for using in information mining projects.

3 Proposed Elicitation Requirement Process Model

The proposed process defines a set of high level activities that must be performed as a part of the business understanding stage, presented in the CRISP-DM methodology, and can be used in the business requirements definition stage of the Kimball Lifecycle. This process breaks down the problem of requirement elicitation in information mining projects into several phases, which will transform the knowledge acquired in the earlier stage. Figure 1 shows strategic phases of an information mining project, focusing on the proposed requirement elicitation activities.

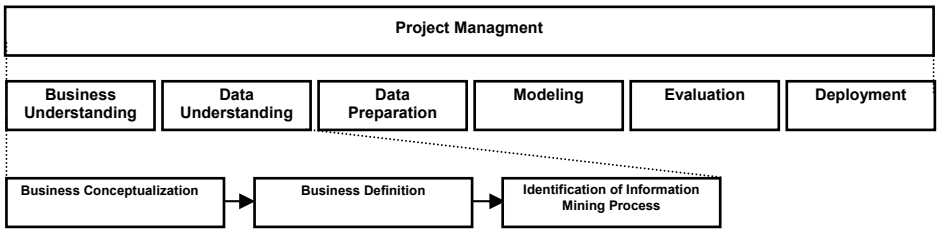

Fig. 1. Information Mining Process phases.

The project management layer deals with coordination of different activities needed to achieve the objectives. Defining activities in this layer are beyond this work. This work identifies activities related to the process exposed in [11], and can be used as a guide for the activities to be performed in an information mining project.

3.1 Business Conceptualization Phase.

The Business Conceptualization phase is the phase of the elicitation process that will be used by the analyst to understand the language used by the organization and the specific words used by the business. Table 1 summarizes inputs and outputs of the Business Conceptualization phase.

Table 1. Business Conceptualization phase inputs and outputs.

Phase	Task	Input product		Transformation technique	Output Product	
		Input	Representation		Output	Representation
Business Conceptualization	Business Understanding	Project Definition	Project KickOff	Project Sponsors Analysis	List of users to be interviewed	List of users to be interviewed template.
	Business Process data gathering	List of users to be interviewed	List of users to be interviewed	Interviews Workshops	Gathered Information	Information gathering template
	Business Model Building	Gathered Information	Information gathering Template	Analysis of gathered information	Use Case Model	Use Case Model template

Interviewing business users will define information related problems that the organization has. The first activity is to identify a list of people that will be interviewed. This is done as part of the business process gathering activity.

In these interviews, information related to business process is collected and modeled in use cases. A business process is defined as the process of using the business on behalf of a customer and how different events in the system occur, allowing the customer to start, execute and complete the business process [14]. The Business Analyst should collect specific words used in business processes in order to obtain both a description of the different tasks performed in each function, as well as the terminology used in each use case.

The Use Case modeling task uses information acquired during business data gathering and, as a last activity of this phase, will generate these models.

3.2 Business Definition Phase

This phase defines the business in terms of concepts, vocabulary and information repositories. Table 2 shows inputs and outputs of this phase

Table 2. Business definition phase inputs and outputs

Phase	Task	Input product		Transformation technique	Output Product	
		Input	Representation		Output	Representation
Business Definition	Build Dictionary of Business Vocabulary	Use Case Model	Use Case Model Template	Use Case Analysis	Concepts Dictionary	Concept Dictionary template
	Establish Concept Relationships	Concepts Dictionary	Concept Dictionary template	Entity Relationship model	Concepts relationships	Concept Relationship Model
	Build a Data Repository Map	Concepts relationships	Concept Relationship Model	Documentation Analysis	Data Repository Map	Data Repository Map template

The objective is to document concepts related to business process gathered in the Business Conceptualization Phase and discover its relationships with other terms or concepts. A dictionary is the proposed tool to define these terms. The structure of a concept can be defined as shown in table 3.

Table 3. Concept Structure

Structure element	Description
Concept	Term to be defined
Definition	Description of the concept meaning.
Data structure	Description of data structures contained in the concept
Relationships	A List of Relationships with other concepts
Processes	A list of processes that use this concept

Once the dictionary is completed the map, the analyst begins to analyze the various repositories of information in the organization. It is also important to determinate volume information, as this data can be used to select the information mining processes applicable to the project. The acquired information is used to build a map, or a model, that shows the relationship between the business use cases, business concepts and information repositories. This triple relationship can be used as the start point of any technique of information mining processes.

3.3 Identification of Information Mining Process Phase

The objective of the phase is to define which information mining process can be used to solve the identified problems in the business process. There are several processes that can be used [16], for instance:

- Discovery of behavior rules (DBR)
- Discovery of groups (DOG)
- Attribute Weighting of interdependence (AWI)
- Discovery of membership group rules (DMG)
- Weighting rules of behavior or Membership Groups (WMG)

This phase does not require any previous input, so activities can be performed in parallel with activities related to Business Conceptualization. Table 4 shows inputs and outputs of this phase.

Table 4. Inputs and Outputs of Identification of Information Mining Process Phase

Phase	Task	Input product		Transformation technique	Output Product	
		Input	Representation		Output	Representation
Identification of Information Mining Process	Identify Business Problems	Use Case Model	Use Case Model Template	Documentation Analysis	Problem List	Problem List Template
	Select an information mining process	Problem List Concept Dictionary	Problem List Template Dictionary Template	LEL Analysis	An information Mining process to be applied	

The list of business problems must be prioritized and be written in natural language, using the user's vocabulary. Only important or critical problems are identified.

Analysis of the problems in the list has to be done. This analysis can be done using the model known as "Language Extended Lexicon (LEL)" [17][18] and can be used as a foundation of the work performed in this phase: breaking down the problem into several symbols presented in the LEL model. This model shows 4 general types of symbol, subject, object, verb and state.

To define useful information mining process, a decision table is proposed. The table analyses LEL structures, concepts identified in the Business Conceptualization phase, existing information repositories and problems to be solved. All the information is analyzed together and according to this analysis, an information mining process is selected as the best option for the project. Table 5 shows the conditions and rules identified as foundations in this work. An important remark is that subjects discovery refers to concepts or subject that hasn't been identified as part of the business domain

Table 5. Information mining process selection decision table

Condition	R01	R02	R03	R04	R05
Does concepts identified as objects exist in some information repository?	Yes	Yes	Yes	Yes	No
Are there any concept identified as an object that are not stored in any information repository?	No	No	No	Yes	Yes
The action represented by a verb. Denotes the discovery of subject concepts?	Yes	No	No	No	Yes
Are there any concept that can be associated with factors?	Yes	Yes	Yes	No	Yes
Have the required factors been identified?	Yes	Yes	No	No	No
Can we deduce factors based on existing information available on repositories?	Yes	Yes	No	No	No
The action represented by a verb. Associates subjects and objects?	Yes	No	No	Yes	Yes
Is Analysis of factors required to obtain a group of subjects or objects?	No	Yes	Yes	Yes	Yes
Actions					
The technique to be applied is:	**DBR**	**DOG**	**AWL**	**DMG**	**WMG**

The objective of the table is to be able to decide, through the analysis of the information gathered about the business, which information mining process can be applied to the project. An important remark is that this decision table will add new knowledge and new rules, with the end of improving the selection technique criteria. With more

projects used as input and more experience acquired in these projects, the rules proposed on the table can be adjusted and then we can get a better selection choice.

The Process Information Mining Identification phase is the last phase of the process. The following tasks will depend upon project managing process and tasks defined for the project.

4 Proof of concept

A case study is presented next to prove the proposed model.

4.1 Business Description

A real estate agency works mainly with residential properties in the suburban area. It's lead by two owners, partners in equal shares. This real state agency publishes its portfolio in different media, mostly local real estate magazines. Published properties are in the mid range value, up to three hundred thousand dollars. It only has one store, where all the employees work. The following business roles are covered: a real estate agent, salesman, administrative collaborators and several consultants.

4.2 Process Execution

The first step of the process consists in two activities: identify project stakeholders and set the list of people to be interviewed. In this case, with the little business information that we have, we can identify three stakeholders: the owners and the real estate agent.

The second step is to set up the interviews of the stakeholders and gather information related to the business in the study. The following paragraph describes information obtained in the interview.

This agency focuses on leasing and selling real estate. Any person can offer a house for sale. If a house is for sale, the real estate agent will estimate the best offer for the property being sold. When a person has an interest in buying a home, they complete a form with the contact details and characteristics that must meet the property. If there are any properties that meet the requested criteria, they are presented to the customer. The real estate agency considered clients as those who have offered a home to sell or have already begun a process of buying a home offered, and considered interested customers, persons who are consulting on the proposed properties or are looking for properties to buy. If interested customers agree on the purchase of a property will be customers of the estate and begins the process of buying property. The customer contact information and the property details are stored in an Excel file.

In this case, we can identify the following Business Use Cases:

- Sell a property Use Case, action of a person selling a property.
- Buy a property Use Case, action of a person buying a property.
- Show a property managed by the real estate agency Use Case, reflects the action of showing a real estate available for sale to interested parties.

For the Business Definition Phase, the business concept dictionary is created. From gathered information, the concepts shown in table 6 can be identified.

Table 6. Identified Business Concepts

Selling Customer: A person who offers a property for sale	
Structure:	Name and Last Name Contact Information
Relationships:	Property Property appraisal
Busines Process:	Sell a Property

Property Appraisal: Appraisal of property for sale.	
Structure:	Appraisal value (Number) Property ID Transaction Currency
Relationships:	Property Customer
Business Processes	Sell a Property Offer a Property

Identified concepts are analyzed in order to find relationships between themselves. A class model can show the basic relationships between identified concepts in the case.

From the gathered business information, a problem found is that the real estate agent wants to know, when a property is offered for sale, which customers could be interested in buying the property. Following the identification, a LEL analysis is done with each problem on the list. In this case, the analysis finds the symbols presented in table 7.

Table 7. Real Estate agency problems related symbols.

Property [Object]
Idea - It's the object that the real estate agency sells - It has its own attributes
Impact: It is sold to a Customer
To Offer a property [Verb]
Idea - The action of showing a property to a customer.
Impact - The property must satisfy the customer requirements.

Customer [Subject]
Idea - A Person interested in buying a property. - A Person who is selling a property.
Impact: Fills a form with buying criteria.
Interested [Status]
Idea - A Customer state achieved when a property meets his or her requirements
Impact - The property is shown to the interested party.

With the obtained LEL analysis, information repositories and defined business concepts, the information mining process to apply in the project, will be determined. The decision table presented in section 3.3 is used, checking the conditions against the gathered information. The result of this analysis states that the project can apply the process of Discovery Rules of Conduct.

5 Conclusion

This work presents a proposal of a process model for requirements elicitation in information mining projects, and how to adapt in these projects, existing elicitation techniques. The process breaks down in three phases, in the first phase the business is analyzed (Conceptualization phase), later, a business model is built and defined to understand its scope and the information it manages (Business Definition phase), and finally, we use the business problems found and the information repositories that stores business data as an input for a decision tabke to establish which information mining technique can be applied to a specific information mining project (Identification of an Information Mining Process).

As a future line of work, several cases are being identified to support the empirical case proposed, emphasizing the validation of the decision table presented in section 3.3.

6 References

[1] Sommerville, I.,Sawyer, P. 1997. Requirements Engineering: A Good Practice Guide. Chichester, England: John Wiley & Sons.

[2] Soren Lauesen, 2002. Software Requirements. Styles and Techniques. London, Pearson Education.

[3] Pollo-Cattaneo, F., et al. 2010. Proceso de Educción de Requisitos en Proyectos de Explotación de Información. En Ingeniería de Software e Ingeniería del Conocimiento: Tendencias de Investigación e Innovación Tecnológica en Iberoamérica (Editores: R. Aguilar, J. Díaz, G. Gómez, E- León). Pág. 01-11. Alfaomega Grupo Editor. ISBN 978-607-707-096-2.

[4] Pollo-Cattaneo, F., et al. (2010). Ingeniería de Proyectos de Explotación de Información. Proceedings XII Workshop de Investigadores en Ciencias de la Computación. Pág. 172-176.

[5] Pytel, P., et al. (2011). Ingeniería de Requisitos Basada en Técnicas de Ingeniería del Conocimiento. Proceedings XIII Workshop de Investigadores en Ciencias de la Computación. Pág. 426-429. ISBN 978-950-673-892-1.

[6] Chapman, P., et al. (2000). CRISP-DM 1.0 Step by step BIguide. Edited by SPSS. http://tinyurl.com/crispdm

[7] Garcia-Martinez, R., et al (2011). Information Mining Processes Based on Intelligent Systems. Proceedings of II International Congress on Computer Science and Informatics (INFONOR-CHILE 2011). pp. 87-94. ISBN 978-956-7701-03-2.

[8] Pyle, D. (2003). Business Modeling and Business Intelligence. Morgan Kauffmann Publishers.

[9] SAS. 2011. SAS Enterprise Miner: SEMMA. http://tinyurl.com/semmaSAS

[10] Pollo-Cattaneo, F., et al. (2009). Metodología para Especificación de Requisitos en Proyectos de Explotación de Información. Proceedings XI Workshop de Investigadores en Ciencias de la Computación. Pág. 333-335. ISBN 978-950-605-570-7.

[11] Kimball, R., et al. (2011), The Data Warehouse Lifecycle Toolkit. John Wiley & Sons.

[12] Vanrell, J., Bertone, R.,García-Martínez, R., 2010. Modelo de Proceso de Operación para Proyectos de Explotación de Información.

[13] Britos, P., Dieste, O., García-Martínez, R. 2008. Requirements Elicitation in Data Mining for Business Intelligence Projects. Advances in Information Systems Research, Education and Practice. David Avison, George M. Kasper, Barbara Pernici, Isabel Ramos, DewaldRoode Eds. (Boston: Springer), IFIP Series, 274: 139–150.

[14] Jacobson, I., Ericsson, M., Jacobson, A.. 1995. The Object Advantage. Business Process Reengineering with Object Technology. Addison Wesley Publishing Company. p. 98.

[15] García-Martínez, R., et al. (2011). Towards an Information Mining Engineering. En Software Engineering, Methods, Modeling and Teaching. Universidad de Medellín Editorial. ISBN 978-958-8692-32-6. pp. 83-99.

[16] Pollo-Cattaneo, F., et al. 2010. Ingeniería de Procesos de Explotación de Información. En Ingeniería de Software e Ingeniería del Conocimiento: Tendencias de Investigación e Innovación Tecnológica en Iberoamérica (Editores: R. Aguilar, J. Díaz, G. Gómez, E- León). Pág. 252-263. Alfaomega Grupo Editor. ISBN 978-607-707-096-2.

[17] Leite, J.C.S.P.. 1994. Notas de Aula. Material del curso de Ingeniería de Requisitos.

[18] Fresno, M., et al. 1998. Derivación de objetos utilizando LEL y Escenarios en un caso real. http://wer.inf.puc-rio.br/wer98/artigos/89.html. Last Access July 2012