METODOLOGÍA PARA ESPECIFICACION DE REQUISITOS EN PROYECTOS DE EXPLOTACION DE INFORMACION

F. Pollo-Cattaneo, P. Britos, P. Pesado, R. García-Martínez

Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. UNLP
Centro de Ingeniería de Software Ingeniería del Conocimiento. Escuela de Postgrado. ITBA
Instituto de Investigaciones en Informática LIDI. Facultad de Informática. UNLP - CIC
Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. UBA

{fcattane,pbritos}@itba.edu.ar, ppesado@lidi.info.unlp.edu.ar, rgarciamar@fi.uba.ar

CONTEXTO

El Proyecto articula líneas de trabajo del Proyecto "Aplicaciones de Explotación de Información basada en Sistemas Inteligentes", con financiamiento de la Secretaria de Ciencia y Técnica de la Universidad de Buenos Aires (UBACYT 2008-2010 código I012) y acreditado por Resolución Rector-UBA N° 576/08 con radicación en el Laboratorio de Sistemas Inteligentes de la Facultad de Ingeniería de la Universidad de Buenos Aires.

RESUMEN

Este proyecto tiene como objetivo general definir una metodología de educción de requisitos para proyectos de explotación de información que permita ser integral al ciclo de vida de este tipo de proyectos. Se considera que el proyecto implica investigación básica en la definición del proceso de educción de requisitos de este tipo de proyectos.

Palabras clave: *explotación de información, proyectos, ingeniería de requisitos, especificación.*

1. INTRODUCCION

La Ingeniería de Requerimientos es una fase importante en las metodologías de Ingeniería del Software (IEEE, 1993; Winter & Strauch 2004; Maiden et al. 2004, 2007; Solheim et al. 2005; Jiang & Eberlein 2007) la que permite especificar las necesidades de los clientes Las Metodologías de Explotación de Información buscan organizar los procesos de descubrimientos de patrones en el datawarehouse de la organización. Estas metodologías consideran la especificación de requerimientos como una fase temprana en las actividades de este tipo de proyectos. (Chapman et al. 2000; Pyle 2003).

En (Winter & Strauch 2002; Silva & Freire 2003; Yang and Wu, 2006) se destaca la necesidad que tienen las metodologías de Explotación de Información en focalizarse en la definición de objetivos y sus tareas especialmente en la fase de exploración de los datos; proponen el uso de herramientas conceptuales para la documentación de estos procesos, la construcción de modelos y la búsqueda de patrones. La comunidad dedicada a la Explotación de Información no presta mucha atención a los aspectos a tener en cuenta en la especificación de requerimientos, no identifica técnicas de elicitación ni sugiere plantillas para una documentación sistemática.

En el dominio de la Explotación de Información, durante los procesos de elicitación de requerimiento son identificados los conceptos relativos a la extracción, transformación, agregación y descubrimiento de patrones negocios dentro de la organización.

Una suposición de los Ingenieros en Requerimientos involucrados en proyectos de Explotación de Información es que los recursos humanos involucrados en el proyecto conocen lo suficiente acerca de los requerimientos. Es conocido que en situaciones típicas, clientes y aun usuarios "hablen en otro lenguaje", lo mismo que el equipo de desarrolladores (Maiden et al. 2007). La tarea de traducir a clientes y usuarios es realizados por los Ingenieros en Requerimientos y los Analistas del Negocio utilizando diversas notaciones (Jiang & Eberlein 2007).

En este contexto los stakeholders y los ingenieros de requerimientos trabajan en forma conjunta para identificar "que" y "donde" buscar dentro de los datos de las organizaciones, para proveer las bases del descubrimiento de patrones de negocios. El proceso de elicitación de requerimiento es direccionado comúnmente por el uso de metodologías de explotación de datos (Chapman *et al*, 2000; Pyle, 2003, SAS, 2008), estas mencionan la necesidad de entender en negocio como punto de partida para el desarrollo de este tipo de proyectos.

La metodología CRISP-DM (Chapman *et al*, 2000) consisten en 4 niveles de abstracción jerárquicos organizados desde las áreas generales a los casos específicos. El proceso esta divide en 6 fases, cada una de las cuales tienen subfases. Las tareas generes son proyectadas como una, donde las acciones deben ser desarrolladas para situaciones descriptas específicamente. Como consecuencia, nos encontramos con tareas muy generales, por ejemplo "limpieza de datos"; para lo cual se cuenta con un

tercer nivel en el cual se desarrollan tareas específicas para esos casos, como por ejemplo "limpieza de datos numéricos", o "limpieza de datos categóricos". Un cuarto nivel recolecta las acciones del grupo, decisiones y resultados específicos del proyecto de explotación de información. La metodología CRISP-DM presenta dos documentos diferentes como herramientas de ayuda durante el desarrollo del proyecto: el modelo de referencia y la guía de usuario. El modelo re referencia describe en términos generales las fases, tareas generales y salidas previstas en cada una de ellas. La guía de usuarios brinda en detalle la documentación acerca de la aplicación del modelo de referencia en proyectos de explotación de datos; también sugiere lista de validación acerca de cada una de las fases

La metodología P^3TQ (Product, Place, Price, Time, Quantity) consta de dos partes (Pyle, 2003): [a] Modelado (PI): provee una guía paso a paso para desarrollar y construir un modelo de negocio, problema u oportunidad. El modelado depende de las circunstancias de negocios que sean señaladas, en primer lugar identifica 5 escenarios para esta fase. Principalmente provee una lista de acciones a ser completadas dependiendo del escenario planteado; y [b] Data Mining (PII): provee una guía paso a paso de cómo llevar adelante el proceso de explotación de información para el modelo indicado en la fase anterior (PI). Esta fase consiste en una serie de acciones que deben ser completados en orden. Para los diversos modelos se deben realizar un conjunto de tareas al mismo tiempo, el proceso de explotación pasa de una actividad a otra. Cada una de las partes está basada en 4 tipos de "cajas de actividades"; *cajas de acciones*: indican una o más actividades requeridas en los próximos pasos a realizar; *cajas de descubrimiento*: que proveen acciones de exploración necesarias para realizar la acción y decidir que realizar en el próximo paso; siempre contienen una "acción de descubrimiento" la cual tiene resultados asociados, interpretaciones y posibles problemas; *cajas técnicas*: proveen información suplementaria acerca de las recomendación de los pasos descriptos en las cajas de acciones o de descubrimiento; y *cajas de ejemplo*: proveen de una descripción detallada de cómo usar una técnica especifica.

SEMMA es una metodología orientada a seleccionar, explotar y modelar un gran conjunto de datos; destinado al descubrimiento de patrones de negocio (SAS, 2008). El proceso comienza con la extracción de una muestra de los datos para los cuales el análisis es aplicado. Con la muestra seleccionada, la metodología propone explotar los datos en orden a simplificar el modelo. Una tercera fase está orientada seleccionar el algoritmo de explotación de datos más adecuado. La cuarta fase está orientada a ejecutar el algoritmo seleccionado con la muestra. La última fase consiste en la evaluación de los resultados a través del contraste con modelos estadísticos o nuevas muestras.

En este contexto, las metodologías no cubren adecuadamente la fase de elicitación de requerimiento, los conceptos necesarios ni su correspondiente documentación.

2. LINEAS DE INVESTIGACION y DESARROLLO

Para construir el conocimiento asociado al presente proyecto de investigación, se seguirá un enfoque de investigación clásico [Kumar, 1996; Creswell, 2003; Marczyk, *et al*, 2005] en el que se han identificado métodos y materiales necesarios para desarrollar el proyecto. Se dispone de los materiales: [a] metodologías CRISP-DM (Chapman, *et al*, 2000), P^3TQ (Pyle, 2003) y SEMMA (SAS, 2008) para identificar los conceptos necesarios para la educción de conocimiento en proyectos de explotación de información, [b] norma IEEE 830-1993 (IEEE, 1993) sobre educción de requisitos para la ingeniería del software, [c] manual SWEBOOK (IEEE, 2004), sobre educción de requisitos en ingeniería del software, [d] técnicas de educción de conocimientos para sistemas inteligentes (García-Martínez, *et al,* 2004). Y se prevé utilizar el método basado en las siguientes tareas: [a] se buscará identificar los conceptos necesarios a educir en proyectos de explotación de información a través de revisión bibliográfica, estudio de casos y consulta a expertos, [b] identificación de la relación existente entre los diversos conceptos a educir, [c] identificación de técnicas de elicitación de requisitos para poder obtener los conceptos de este tipo de proyecto a través de revisión bibliográfica y consulta con expertos, [d] se propondrán: plantillas para la educción de conceptos, técnicas de elicitación para los conceptos a educir, cuestionarios modelos para facilitar la educción de conceptos.

En este proyecto se realizarán investigaciones sobre:

[a] Las distintas metodologías para proyectos de explotación de información.
[b] Los conceptos necesarios a ser educidos para estas metodologías.
[c] Las técnicas de educción de conocimiento aplicables a la identificación de requisitos de proyectos de software.
[d] La fiabilidad de las técnicas identificadas.
[e] La relación entre los conceptos a ser educidos con la metodología propuesta.

3. RESULTADOS OBTENIDOS/ESPERADOS

Se ha avanzado en la identificación de conceptos que deben ser educidos en el dominio del proyecto de explotación de información e identificado la necesidad de definir procesos que definan como llevar esa educción y herramientas que den soporte a la documentación de la misma [Britos *et al*, 2008].

4. FORMACION DE RECURSOS HUMANOS

En el marco de este proyecto se esta desarrollando una tesis doctoral y tres tesis de ingeniería informática (una de ellas defendida y dos en curso).

5. BIBLIOGRAFIA

Britos, P., Dieste, O., García-Martínez, R. (2008). *Requirements Elicitation in Data Mining for Business Intelligence Projects*. IFIP Series, 274: 139–150.

Chapman P, Clinton J, Keber R, Khabaza T, Reinartz T, Shearer C, Wirth R (2000) *CRISP-DM 1.0 Step by step BIguide Edited by SPSS*. http://www.crisp-dm.org/CRISPWP-0800.pdf Acceso Marzo 2008.

García Martínez, R. y Britos, P. (2004). *Ingeniería de Sistemas Expertos*. Editorial Nueva Librería. ISBN 987-1104-15-4

IEEE (1993) *Standard IEEE 830-1993: Recommended Practice for Software Requirements Specifications*. Institute of Electronic and Electrical Engineers Press.

IEEE (2004) *Guide to the Software Engineering Body of Knowledge*. IEEE Comp. Society Press

Jiang L, Eberlein A (2007) *Selecting Requirements Engineering Techniques based on Project Attributes - A Case Study*. 14th Annual IEEE ECBS: 269-278

Maiden N, Ncube C, Robertson S (2007) *Can Requirements Be Creative? Experiences with an Enhanced Air Space Management System* Proceedings 29th ICSE: 632-641

Maiden N, Robertson S, Gizikis A (2004) *Provoking Creativity: Imagine What Your Requirements Could be Like*. IEEE Software 21(5): 68-75

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

SAS (2008) SAS *Enterprise Miner: SEMMA* http://www.sas.com/ technologies/analytics/data-mining/miner/semma.html. Acceso Marzo 2009

Silva F, Freire J (2003*) DWARF: An Approach for Requirements Definition and Management of Data Warehouse Systems*. RE´03: 75-84

Solheim H, Lillehagen F, Petersen S, Jorgensen H, Anastasiou M (2005) *Model-driven visual requirements engineering* Proceedings RE´05: 421-428

Winter R, Strauch B (2002) *A Method for Demand-driven Information Requirements Analysis in Data Warehousing Projects*. HICSS-36:231-239

Yang Q, Wu X (2006) *10 Challenging Problems in Data Mining Research*. Int. J. Inf. Tech. & Decis. Mak. 5(4):597–604