INGENIERÍA DE PROYECTOS DE EXPLOTACION DE INFORMACION

Pollo-Cattaneo, F., Amatriain, H., Rodriguez, D., Pytel, P., Ciccolella, E., Vegega, C., Dearriba , M., Rodiguez Aubert, M., Bose, F., Giordano, L., Britos, P., García-Martínez, R.

Grupo de Estudio en Metodologías de Ingeniería de Software. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional.
Grupo Investigación en Sistemas de Información. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús.
Grupo de Investigación en Explotación de Información. Sede Andina (El Bolsón). Universidad Nacional de Río Negro.

fpollo@posgrado.frba.utn.edu.ar, hamatriain@frba.utn.edu.ar, rgarcia@unla.edu.ar, paobritos@gmail.com

CONTEXTO

Este proyecto de investigación se desarrolla en el marco de la cooperación existente entre el Grupo de Estudio en Metodologías de Ingeniería de Software (GEMIS) de la Universidad Tecnológica Nacional (FRBA), el Grupo de Investigación en Sistemas de Información (GISI) del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús y el Grupo de Investigación en Explotación de Información de la Sede Andina (El Bolsón) de la Universidad Nacional de Río Negro. Articula líneas de trabajo de los proyectos de investigación "Metodología para la Especificación de Requisitos en Proyectos de Explotación de Información" (UTN-FRBA) y "Proyecto 33A081: Sistemas de Información e Inteligencia de Negocio" (UNLa).

RESUMEN

En este proyecto se busca desarrollar y sistematizar el cuerpo de conocimiento asociado a la Ingeniería de Proyectos de Explotación de Información con focalización en su transferencia a la Industria. Las líneas de investigación propuestas buscan proveer a los desarrolladores las siguientes herramientas para proyectos de explotación de información: técnicas de educción y encapsulamiento de requistos, modelo de procesos, modelo de ciclo de vida y mapa de actividades.

INTRODUCCIÓN

La inteligencia de negocio [Morik y Rüping, 2002; Moss, 2003; Moss y Atre, 2003; Stefanovic et al.,2006] propone un abordaje interdisciplinario (dentro del que se encuentra la Informática), que tomando todos los recursos de información disponibles y el uso de herramientas analíticas y de síntesis con capacidad de transformar la información en conocimiento, se centra en generar a partir de estos, conocimiento que contribuya con la toma de decisiones de gestión y generación de planes estratégicos en las organizaciones [Thomsen, 2003. Negash y Gray, 2008].

La Explotación de Información es la sub-disciplina Informática que aporta a la Inteligencia de Negocio [Langseth y Vivatrat, 2003] las herramientas para la transformación de información en conocimiento [Mobasher *et al*., 1999; Srivastava *et al*., 2000; Abraham, 2003; Coley, 2003].

La explotación de información se ha definido como la búsqueda de patrones interesantes y de regularidades importantes en grandes masas de información [Fayad *et al*., 1996]. Al hablar de explotación de información basada en sistemas inteligentes [Evangelos, 1996, Michalski *et*

al., 1998] se refiere específicamente a la aplicación de métodos de sistemas inteligentes, para descubrir y enumerar patrones presentes en la información.

Los Sistemas Inteligentes constituyen el campo de la Informática en el que se estudian y desarrollan algoritmos que implementan algún comportamiento inteligente y su aplicación a la resolución de problemas prácticos [Michalski, 1983; Dejong & Money 1986; Bergadano *et al*., 1992]. Entre los problemas abordados en este campo, está el de descubrir conocimientos a partir de una masa de información [Michalski, 1983; García Martínez, 2004]. Esto resulta una alternativa de solución a problemas que no pueden ser resueltos mediante algoritmos tradicionales, entre los cuales podemos mencionar especificación de condiciones asociadas a diagnósticos técnicos o clínicos, identificación de características que permitan reconocimiento visual de objetos, descubrimiento de patrones o regularidades en estructuras de información (en particular en bases de datos de gran tamaño), entre otros.

Los métodos tradicionales de análisis de datos incluyen el trabajo con variables estadísticas, varianza, desviación estándar, covarianza y correlación entre los atributos; análisis de componentes (determinación de combinaciones lineales ortogonales que maximizan una varianza determinada), análisis de factores (determinación de grupos correlacionados de atributos), análisis de clusters (determinación de grupos de conceptos que están cercanos según una función de distancia dada), análisis de regresión (búsqueda de los coeficientes de una ecuación de los puntos dados como datos), análisis multivariable de la varianza, y análisis de los discriminantes [Michalski *et al*., 1998]. Todos estos métodos están orientados numéricamente. Son esencialmente cuantitativos.

En contraposición [Britos *et al*., 2005], los métodos basados sistemas inteligentes [Konenko y Kukar, 2007], permiten obtener resultados de análisis de la masa de información que los métodos convencionales no logran tales como: los algoritmos TDIDT, los mapas auto organizados (SOM) y las redes bayesianas. Los algoritmos TDIDT permiten el desarrollo de descripciones simbólicas de los datos para diferenciar entre distintas clases [Quinlan, 1986; 1990]. Los mapas auto organizados pueden ser aplicados a la construcción de particiones de grandes masas de información. Tienen la ventaja de ser tolerantes al ruido y la capacidad de extender la generalización al momento de necesitar manipular datos nuevos [Kohonen, 1982; 1995]. Las redes bayesianas pueden ser aplicadas para identificar atributos discriminantes en grandes masas de información, detectar patrones de comportamiento en análisis de series temporales. [Heckerman *et al*., 1995].

En [Britos, 2008] se señala que se han ido desarrollando metodologías que permiten gestionar la complejidad de los proyectos de explotación de información. La comunidad científica considera metodologías probadas a CRISP-DM, SEMMA y P^3TQ.

La metodología CRISP-DM [Chapman *et al*., 2000] consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general (comprensión del negocio) hasta los más específicos (plan de implementación). Las fases de la metodología CRISP-DM se presentan en la Figura 1.

A la metodología SEMMA [SAS, 2008] se la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos. Las fases de la metodología SEMMA se presentan en la Figura 2.

La metodología P3TQ está compuesta por dos modelos [Pyle, 2003], el Modelo de Negocio y el Modelo de Explotación de Información. El Modelo de Negocio (MII) el cual proporciona una guía de pasos para

FASE	TAREAS COMPONENTES	ACTIVIDADES ASOCIADAS
Comprensión del negocio	Determinar los objetivos del negocio	▪ Background ▪ Objetivos del negocio ▪ Criterios de éxito del negocio
	Evaluar de la situación	▪ Inventarios de recursos ▪ Requisitos, supuestos y requerimientos ▪ Riesgos y contingencias ▪ Terminología ▪ Costos y beneficios
	Determinar objetivos del proyecto de Explotación de Información	▪ Las metas del Proyecto de Explotación de Información ▪ Criterios de éxito del Proyecto de Explotación de Información
	Realizar el Plan del Proyecto	▪ Plan de proyecto ▪ Valoración inicial de herramientas
Comprensión de los datos	Recolectar los datos Iniciales	▪ Reporte de recolección de datos iniciales
	Descubrir datos	▪ Reporte de descripción de los datos
	Explorar de los datos	▪ Reporte de exploración de datos
	Verificar la calidad de datos	▪ Reporte de calidad de datos
Preparación de los datos	Caracterizar el conjunto de datos	▪ Conjunto de Datos ▪ Descripción del Conjunto de Datos
	Seleccionar los datos	▪ Inclusión / exclusión de datos
	Limpiar los datos	▪ Reporte de calidad de datos limpios
	Estructurar los datos	▪ Derivación de atributos ▪ Generación de registros
	Integrar los datos	▪ Unificación de datos
	Caracterizar el formato de los datos	▪ Reporte de calidad de los datos
Modelado	Seleccionar una técnica de modelado	▪ La técnica modelada ▪ Supuestos del modelo
	Generar el plan de pruebas	▪ Plan de pruebas
	Construir el modelo	▪ Configuración de parámetros ▪ Modelo ▪ Descripción del modelo
	Evaluar el modelo	▪ Evaluar el modelo ▪ Revisación de la configuración de parámetros
Evaluación	Evaluar Resultado	▪ Valoración de resultados mineros con respecto al éxito del negocio ▪ Modelos aprobados
	Revisar	▪ Revisión del proceso
	Determinar próximos pasos	▪ Listar posibles acciones
Implementación	Realizar el plan de implementación	▪ Plan de Implementación
	Realizar el plan de monitoreo y mantenimiento	▪ Plan de monitoreo y mantenimiento
	Realizar el informe final	▪ Informe final ▪ Presentación Final
	Realizar la revisión del proyecto	▪ Documentación de la experiencia

Fig. 1. Fases Metodología CRISP-DM

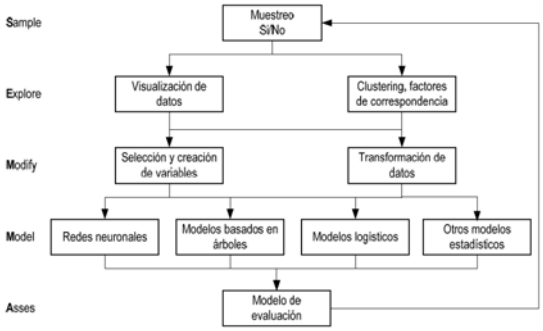

Fig. 2. Fases Metodología SEMMA

el desarrollo y la construcción de un modelo que permita identificar un problema de negocio o la oportunidad del mismo. Las fases de la metodología P3TQ se presentan en la Figura 3.

Las tres metodologías identifican técnicas de explotación de información utilizables. CRISP-DM identifica problemas de inteligencia de negocio y hace una caracterización parcialmente abstracta de los mismos. SEMMA y P^3TQ no identifican problemas de inteligencia de negocio ni formulan una caracterización abstracta de los mismos. CRISP-DM identifica las relaciones entre las técnicas de explotación de información y las variables que modelan los problemas de inteligencia de negocio esbozando parcialmente los procesos a desarrollar.

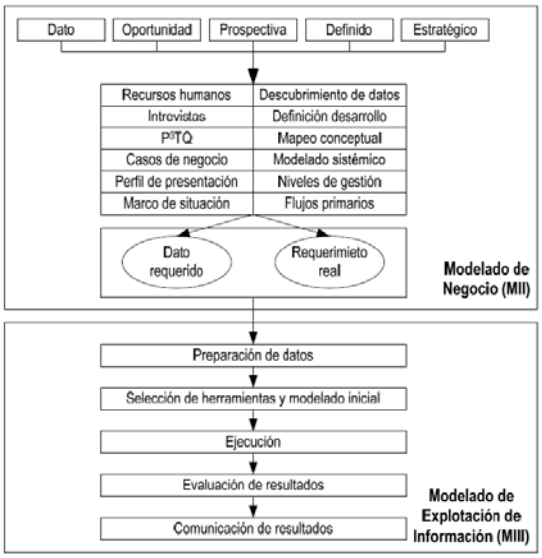

Fig. 3. Fases Metodología P3TQ

SEMMA y P^3TQ no identifican relaciones entre técnicas de explotación de información y problemas de inteligencia de negocio, ni procesos de explotación de información.

En síntesis, las metodologías se centran fuertemente en las técnicas de explotación de información y en la tipificación de los datos sin enfatizar cómo las variables vinculadas a los datos modelan el negocio ni cuáles son los procesos de explotación de información que a partir de aplicar las técnicas al conjunto de valores de las variables, permiten obtener una solución para cada problema de inteligencia de negocio.

LÍNEA DE INVESTIGACIÓN Y DESARROLLO

Este proyecto se articula a través de las siguientes líneas de investigación:

- Aplicación de técnicas de Ingeniería del Conocimiento a la Educción de Requisitos de Proyectos de Explotación de Información. En esta línea ya trabaja un tesista de magister.

- Aplicación de formalismos de Ingeniería del Conocimiento al encapsulamiento de requisitos de Proyectos de Explotación de Información educidos. Esta línea supone extender formalismos existentes como lo propuesto en [Britos *et al*, 2008],
- Modelo de Procesos para Proyectos de Explotación de Información. Esta línea de trabajo, busca construir el modelo a partir de modelos de procesos de software y metodologías de desarrollo de proyectos de explotación de Información de uso habitual en la industria. El proyecto incluye la construcción de métricas de calidad de ingeniería de requerimientos En esta línea ya trabaja un tesista de magister.
- Modelo de Ciclo de Vida Genérico para Proyectos de Explotación de Información. Se toma como base para esta línea los modelos de ciclo de vida conocidos en ingeniería del software y metodologías de desarrollo de proyectos de explotación de información de uso habitual en la industria. En esta línea ya trabaja un tesista de magister.
- Mapa de Actividades para Proyectos de Explotación de Información. Se toman como base los resultados de las líneas de investigación Modelo de Procesos y Modelo de Ciclo de Vida para Proyectos de Explotación de Información.
- Metodología para la Especificación de Requisitos en Proyectos de Explotación de Información. Esta línea busca construir la metodología en interacción con las líneas ingeniería del conocimiento aplicada al encapsula-miento de requisitos educidos, modelo de ciclo de vida y modelo de procesos para proyectos de proyectos de explotación de información.

RESULTADOS OBTENIDOS/ESPERADOS

La necesidad de desarrollar una ingeniería de proyectos de explotación de información surge del relevamiento efectuado en el campo metodológico, en el que se identifica la carencia de técnicas asociadas a la ejecución de cada una de las fases planteadas en las metodologías identificadas.

En este contexto, este proyecto busca desarrollar un conjunto de herramientas para la etapa temprana del proyecto de explotación de información, focalizando las acciones en la formalización de requisitos, la adopción de un ciclo de vida, la estructuración según un modelo de procesos y el ordenamiento del proyecto a través de un mapa de actividades.

FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo se encuentra formado por dos investigadores formados y por tres investigadores en formación. En el marco de este proyecto se están desarrollando: una tesis doctoral y tres tesis de maestría. Se prevee incorporar adicionalmente, dos tesistas de maestría.

BIBLIOGRAFÍA

Abraham, A. (2003). *Business Intelligence from Web Usage Mining.* Journal of Information & Knowledge Management, 2(4): 375-390.

Bergadano, F., Matwin, S. Michalski, R.S., Zhang, J. (1992). *Learning Two-Tiered Descriptions of Flexible Concepts: The POSEIDON System.* Machine Learning 8: 5-43.

Britos, P. (2008). *Procesos de Explotación de Información Basados en Sistemas Inteligentes*. Tesis Doctoral. Facultad de Informática. Universidad Nacional de La Plata.

Britos, P., Dieste, O., García-Martínez, R. 2008. *Requirements Elicitation in Data Mining for Business Intelligence Projects*. En 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., Hossian, A., García Martinez, R., Sierra, E. 2005. *Minería de Datos Basada en Sistemas Inteligentes*. Nueva Librería.

Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000). *CRISP-DM 1.0 Step by Step BI Guide*. Edited by SPSS.

Cooley, R. (2003). *The Use of Web Structure and Content to Identify Subjectively Interesting*

Web Usage Patterns. ACM Transactions on Internet Technology, 3(2): 93-116.

DeJong, G., Mooney, J. (1986). *Explanation-Based Learning: An Alternative View*, Machine Learning, 1: 145-176

Evangelos, S., Han, J, (1996). *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (editores). AAAI Press.

Fayad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uhturudsamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*, (editors). AAAI Press.

García Martínez, R. y Britos, P. (2004). *Ingeniería de Sistemas Expertos*. Editorial Nueva Librería.

Kononenko, I. y Kukar, M. (2007). *Machine Learning and Data Mining. Introduction to Principles and Algorithms*. Horwood Publishing.

Langseth, J., Vivatrat, N. (2003). *Why Proactive Business Intelligence is a Hallmark of the Real-Time Enterprise: Outward Bound*. Intelligent Enterprise 5(18): 34-41.

Michalski, R. (1983). *A Theory and Methodology of Inductive Learning*. Artificial Intelligence, 20: 111-161.

Michalski, R. Bratko, I. Kubat, M. (1998). *Machine Learning and Data Mining, Methods and Applications* (Editores) John Wiley & Sons.

Mobasher, B, R Cooley and J Srivastava (1999). *Creating adaptive web sites through usage-based clustering of URLs*. Proceedings Workshop on Knowledge and Data Engineering Exchange, Pág. 19-25.

Morik, K., Rüping, S. (2002). *A Multistrategy Approach to the Classification of Phases in Business Cycles***.** Lecture Notes in Computer Science, 2430: 307-318.

Moss, L. (2003). *Nontechnical Infrastructure of BI Applications*. DM Review 13(1): 42-45.

Moss, L., Atre, S. (2003). *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications*. Addison-Wesley Information Technology Series.

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

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

SAS, (2008). *SAS Enterprise Miner: SEMMA*. http://www.sas.com/technologies/analytics/datamining/miner/semma.html. Ultimo acceso Junio 2008.

Srivastava, J., Cooley, R., Deshpande, M., Tan, P. (2000). *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*. SIGKDD Explorations, 1(2): 12-23.

Stefanovic, N., Majstorovic. V.,, Stefanovic, D. (2006). *Supply Chain Business Intelligence Model*. Proceedings 13th International Conference on Life Cycle Engineering. Pág. 613-618.

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