INGENIERÍA DE PROYECTOS DE EXPLOTACION DE INFORMACION PARA PYMES

García-Martínez, R., Lelli, R., Merlino, H., Cornachia, L., Rodriguez, D., Pytel, P., Arboleya, H.

Grupo Investigación en Sistemas de Información
Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús.
29 de Septiembre 3901 (1826) Remedios de Escalada, Lanús. Argentina. Tel +54 11 6322-9200 Ext. 194

rgarcia@unla.edu.ar

CONTEXTO

Este proyecto de investigación es continuación del "Proyecto UNLa 33A081: Sistemas de Información e Inteligencia de Negocio" y se enmarca en la Línea de Investigación en Ingeniería de Explotación de Información que desarrollan de manera conjunta el Grupo de Investigación en Sistemas de Información (GISI) del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús; el Grupo de Estudio en Metodologías de Ingeniería de Software (GEMIS) de la Universidad Tecnológica Nacional (FRBA), y el Grupo de Investigación en Explotación de Información de la Sede Andina (El Bolsón) de la Universidad Nacional de Río Negro.

RESUMEN

Los proyectos de explotación de información poseen características muy distintas a las de los proyectos de desarrollo de software tradicionales. Las clásicas etapas de análisis, diseño, desarrollo, integración y testeo, no encajan con las etapas naturales de los procesos de desarrollo de este tipo de proyectos. En consecuencia, herramientas de la Ingeniería de Software clásica tales como la ingeniería de requerimientos, los modelos de procesos, los ciclos de vida e incluso los mapas de actividades no son aplicables a este tipo de proyectos.

En este contexto, este proyecto busca desarrollar y sistematizar el cuerpo de conocimiento de la Ingeniería de Proyectos de Explotación de Información con focalización en su transferencia a la Industria, particularmente al sector PyMEs.

Utilizando las metodologías de investigación documental exploratoria, prototipado evolutivo y casos de estudio se plantea a través de objetivos específicos, el desarrollo de los siguientes artefactos de Ingeniería de Proyectos de Explotación de Información: [a] una batería de técnicas de educción y formalismos de documentación de requerimientos; [b] un modelo de procesos y las métricas asociadas; [c] un modelo de ciclo de vida; y [d] un mapa de actividades.

AVANCES SOBRE EL TEMA

En el Proyecto UNLa-33A081 se trabajaron sobre las áreas de fundamentos de la explotación de información y sus aplicaciones.

En el área de fundamentos: [a] se formuló una propuesta de procesos de explotación de información (Britos y García-Martínez, 2009); [b] se abordó el problema de descubrimiento automático de reglas de negocio (Rancan et al., 2010); [c] se identificaron las bases para una ingeniería de proyectos de explotación de información (Pollo-Cattaneo et al., 2010); y [d] y se planteó la necesidad de disponer de un modelo de procesos de explotación de información (Vanrell et al., 2010).

En el área de aplicaciones de explotación de información se exploraron: [a] la identificación de causales de abandono de estudios universitarios (Kuna et al., 2009; 2010b); [b] la prevención del estrés de los suelos y fatiga de soja (Sanson et al., 2009); [c] la caracterización de problemas de aprendizaje (Jiménez Rey et al., 2009); [d] la detección de patrones para la prevención de daños y/o averías en la industria automotriz (Flores et al., 2009); y [e] la identificación en auditoría de sistemas de datos

faltantes, con ruido e inconsistentes (Kuna et al., 2010a).

OBJETIVOS E HIPOTESIS DE INVESTIGACION

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. El propósito de la investigacion es sentar las bases para el desarrollo de una ingeniería de proyectos de explotación de información centrado en las particularidades de la PyMEs.

Entre los supuestos (o hipotesis) que guian el proyecto se pueden citar:

Hipótesis I: Existen metodologías de explotación de información que destacan la importancia del planeamiento de una elicitacion de requerimientos a lo largo de todo el proyecto de una manera ordenada, documentada, consistente y trazable. Sin embargo, el abordaje clásico de la ingeniería del software no se ajusta del todo a los proyectos de explotación de información porque descuida los aspectos particulares de especificación de requerimientos para este tipo de proyectos. De hecho las técnicas clásicas no son aplicables al proceso de identificación de conceptos para entender el dominio de los proyectos de explotación de información ni a la educción de sus requerimientos, ni al desarrollo la documentación asociada.

Hipótesis II: Los proyectos de explotación de información poseen características muy distintas a las de los proyectos de desarrollo de software tradicionales. Las clásicas etapas de análisis, diseño, desarrollo, integración y testeo no encajan con las etapas naturales de los procesos de desarrollo de este tipo de proyectos. En consecuencia, los ciclos de vida para software (p.ej.: cascada, prototipado, ó espiral); los modelos de procesos software (p.ej.: IEEE 1074 o MOPROSOFT); y la natural derivación de ambos: los mapas de actividades para proyectos software no son aplicables a este tipo de proyectos.

Objetivo General: El objetivo de este proyecto es sistematizar el cuerpo de conocimiento existente y sentar las bases para comenzar a desarrollar el faltante asociado a la Ingeniería de Proyectos de Explotación de Información con focalización en su transferencia al sector PYMES. Las líneas de trabajo propuestas buscan proveer a los profesionales del área de sistemas las siguientes herramientas para proyectos de explotación de información: técnicas de educción y encapsulamiento de requisitos, modelo de procesos, modelo de ciclo de vida y mapa de actividades.

Objetivos específicos vinculados a Hipótesis I:

1.- Relevar las distintas metodologías para proyectos de explotación de información existentes y los conceptos necesarios a ser educidos para estas metodologías, identificando las diversas técnicas de educción de conocimiento y la fiabilidad de las mismas técnicas para la educción de los conceptos necesarios en proyectos de explotación de información.

2.- Determinar la aplicabilidad de técnicas y formalismos de la ingeniería del conocimiento a la educción y el encapsulamiento de requisitos de proyectos de explotación de información.

Objetivos específicos vinculados a Hipótesis II:

3.- Desarrollar un Modelo de Procesos para Proyectos de Explotación de Información con particular énfasis en su utilización en PyMEs. El desarrollo incluye la construcción de métricas asociadas a los procesos.

4.- Desarrollar un Modelo de Ciclo de Vida Genérico para Proyectos de Explotación de Información.

5.- Desarrollar un Mapa de Actividades para Proyectos de Explotación de Información.

FUNDAMENTACION Y PERTINIENCIA

a) Marco general en el que se contextualiza la investigación

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, que se ha definido como la búsqueda de patrones interesantes y de regularidades importantes en grandes masas de información (Fayad et al., 1996), 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).

b) Planteamiento investigativo propuesto

En (Britos, 2008) se señala que se han ido desarrollando metodologías de desarrollo de proyectos de explotación de información, entre estas, la comunidad científica considera metodologías probadas a CRISP-DM, SEMMA y P3TQ. 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). 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. 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 el cual proporciona una guía de pasos para el desarrollo y la construcción de un modelo que permita identificar un problema de negocio o la oportunidad del mismo.

Mientras que entre las fortalezas de estas metodologías se encuentran: [a] la identificación de problemas de inteligencia de negocio, [b] la caracterización parcialmente abstracta de los mismos, [c] la identificación de las relaciones entre las técnicas de explotación de información y las variables que modelan los problemas de inteligencia de negocio, y [d] el planteo parcial de los procesos a desarrollar; son señalables entre sus debilidades las siguientes: [a] se centran fuertemente en las técnicas de explotación de información y en la tipificación de los datos, [b] no determinan como las variables vinculadas a los datos modelan el negocio, y [c] no identifican cuáles son los procesos de explotación de información, ni el modelo asociado, 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.

c) Relevancia

c.1) Interés cientifico: 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, el proyecto promueve el desarrrollo y la validación de métodos, técnicas y herramientas, conllevando a una mejora en el campo de la Ingeniería de Software. Los métodos con abordaje ingenieril permiten dotar al proceso de desarrollo de: objetividad, sistematicidad, racionalidad, generalidad y fiabilidad, contribuyendo al avance del conocimiento científico mediante el uso de técnicas consistentes.

c.2) Interés social: El sesgo previsto del proyecto para el sector PyMEs habilita que los resultados puedan ser transferidos a la industria del software con radicación en la zona de influencia de la UNLa, generando las bases para una industria de servicios de inteligencia de negocios orientados a a proveer información cualitativa que contribuya a la toma de dicisiones en los niveles de gestión de la industria y el comercio regional.

c.3) Interés Educativo: En Febrero del 2010, en el ámbito del Departamento de Desarrollo Productivo y Tecnológico se dictó el curso asistemático "Tecnologías de Explotación de Información" con la participación de nueve estudiantes de la Licenciatura en Sistemas. Se ha previsto que este curso evolucione hacia una asignatura optativa del ciclo superior de la Licenciatura en Sistemas de la UNLa.

Los resultados parciales y finales tendrán un impacto sobre la actualización de los contenidos de las asignaturas Ingeniería de Software III y Sistemas y Organizaciones de la Licenciatura de Sistemas de la UNLa.

METODOLOGIA DE TRABAJO

Para el Objetivo Específico 1 se propone realizar una investigación documental exploratoria vinculada a los conceptos de interés indicados en la fase de entendimiento del negocio de las metodologías específicas de explotación de información y su correspondiente educción con variaciones de técnicas usuales en el contexto ingeniería de requerimientos clásica. Para el Objetivo Específico 2 se propone realizar una investigación exploratoria sobre la utilización de las técnicas de modelado de conocimientos fácticos, tácticos y estratégicos de la Metodología IDEAL de Ingeniería del Conocimiento; a la descripción del problema de negocio con el propósito de documentar el relevamiento de los conceptos de interés identificados en la fase de entendimiento de negocio. Para el Objetivo Específico 3 se propone desarrollar mediante la metodología de prototipado evolutivo un modelo de procesos para proyectos de explotación de información con base en la fusión del modelo MOPROSOFT/COMPETISOFT y el propuesto por CRISP-DM y aplicar estudio de casos para su validación. El proceso de fusión se realizará de manera evolutiva y requerirá: [a] la obtención de las fuentes de información asociadas a cada caso dividiéndolas en la que se utilizaran para la prueba de concepto de los procesos a desarrollar y las que se utilizarán para la validación de dichos procesos, [b] la construcción de procesos de explotación de información de prueba para los casos de estudio en inteligencia de negocio identificados, su refinamiento y generalización; y [c] la validación de los procesos desarrollados y contrastación con los resultados históricos de cada caso. Para el Objetivo Específico 4 se propone desarrollar mediante la metodología de prototipado evolutivo un modelo de ciclo de vida para proyectos de explotación de información que sea una variante de la Espiral de Böehm. Para el refinamiento del los distintos prototipos de ciclo de vida se utilizarán los casos identificados para el Objetivo Específico 3. Para el Objetivo Específico 5 se propone desarrollar mediante la metodología de prototipado evolutivo la articulación del modelo de proceso para proyectos de explotación que surge del Objetivo Específico 3 y el modelo de ciclo de vida que surge del Objetivo Específico 4. Para el refinamiento del los distintos prototipos de mapa de actividades para proyectos de explotación de información se utilizarán los casos identificados para el Objetivo Específico 3.

RESULTADOS OBTENIDOS/ESPERADOS

Se estima que este proyecto formule aportes en el campo de la ingeniería de proyectos de explotación de información mediante el desarrollo de un conjunto de herramientas para la etapa temprana del proyecto de explotación de información.

Se esperan obtener los siguientes artefactos conceptuales para proyectos de explotación de información: (1) Batería de técnicas de educción y formalismos de documentación de requerimientos, (2) Modelo de Procesos, (3) Modelo de Ciclo de Vida; y (4) Mapa de Actividades.

FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo se encuentra formado por cuatro investigadores formados y por dos investigadores en formación. En el marco de este proyecto se están desarrollando: dos tesis doctorales y tres tesis de maestría. Se ha previsto la incorporación de dos becarios alumnos con beca de Universidad.

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., 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. Págs. 1041-1050. ISBN 978-897-24068-4-1.

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.

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. y Britos, P. (2004). Ingeniería de Sistemas Expertos. Editorial Nueva Librería.

Jiménez Rey, E., Rodriguez, D., Britos, P., García-Martínez. R. (2009). Caracterización de Problemas de Aprendizaje Basada en Explotación de Información. Proceedings XI Workshop de Investigadores en Ciencias de la Computación. Pág. 627-629. ISBN 978-950-605-570-7.

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

Kuna, H., Caballero, S., Rambo, A., Meinl, E., Steinhilber, A., Pautch, G., García-Martínez, R., Villatoro, F. (2010a). Avances en Procedimientos de la Explotación de Información Para la Identificación de Datos Faltantes, con Ruido e Inconsistentes. Proceedings XII Workshop de Investigadores en Ciencias de la Computación. Pág. 137-141.

Kuna, H., García Martínez, R. Villatoro, F. (2010b). 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. (2009). 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. ISSN 1681-5653.

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.

Pollo-Cattaneo, F., Amatriain, H., Rodriguez, D., Pytel, P., Ciccolella, E., Vegega, C., Dearriba, M., Rodríguez Aubert, M., Bose, F., Giordano, L., Britos, P., García-Martínez, R. (2010). Ingeniería de Proyectos de Explotación de Información. Proceedings XII Workshop de Investigadores en Ciencias de la Computación. Pág. 172-176.

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

Rancan, C., Pesado, P., García-Martínez, R. (2010). Issues in Rule Based Knowledge Discovering Process. Advances and Applications in Statistical Sciences Journal (ISSN 0974-6811), 2(2): 303-314. ISSN 0974-6811.

Sanson, E., Britos. P., Rodriguez, D., García-Martínez, R. (2009). Clasificación Automática para la Prevención del Estrés de los Suelos y la Fatiga de Soja en el Noroeste Argentino. Proceedings XI Workshop de Investigadores en Ciencias de la Computación. Pág. 333-335. ISBN 978-950-605-570-7.

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

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. Pág. 167-171.