Elementos para una ingeniería de explotación de información

María Florencia Pollo-Cattaneo [1], Ramón García-Martínez [2], Paola Britos [3], Patricia Pesado [4], Rodolfo Bertone [4] *

[1] Universidad Tecnológica Nacional, Facultad Regional Buenos Aires,
Grupo de Investigación en Metodologias de Software,
Medrano 951 (C1179AAQ), Ciudad Autónoma de Buenos Aires, Argentina
[2] Universidad Nacional de Lanús, Departamento Desarrollo Productivo y Tecnológico,
Grupo de Investigación en Sistemas de Información, 29 de Septiembre 3901 (1826)
Remedios de Escalada, Argentina
[3] Universidad Nacional de Río Negro, Grupo de Investigación en Explotación de Información,
San Martín esq. Pellegrini (8430) El Bolsón, Río Negro, Argentina
[4] Universidad Nacional de La Plata, Facultad de Informática, Instituto de Investigaciones en
Informática, Calle 50 esq. Calle 120 (1900) La Plata, Argentina

fpollo@posgrado. frba.utn.edu.ar

Recibido el 20 de Mayo de 2011, aprobado el 20 de Agosto de 2011

Resumen

Los Proyectos de Explotación de Información difieren sustancialmente de los pertenecientes al Software tradicional. Las fases clásicas de desarrollo le son ajenas, al igual que las herramientas involucradas en los procesos de Ingeniería en Software. Un nuevo cuerpo de conocimientos atento a las necesidades de su aplicación industrial, deviene, pues, imprescindible para el avance del nuevo campo disciplinar. En este artículo proponemos: un modelo de negocio, un proceso de educción de requisitos, un método de estimación, una metodología de selección de herramientas, un proceso de transformación de datos y una serie de procesos basados en técnicas de minería de datos.

PALABRAS CLAVE: INGENIERÍA EN SOFTWARE - EXPLOTACIÓN DE INFORMACIÓN

Abstract

The Information Mining Projects have different characteristics compared to traditional software projects. The classic development phases do not apply to the natural phases of Information Mining Projects. Not all Software Engineering tools not apply to these projects. A new body of knowledge is necessary for Information Mining Engineering with a special focus on its use in industry. In this paper we propose: process model, requirement elicitation process, estimation method, a method for selecting the data mining tool, a methodology for transforming the data and, a set of processes for information mining based on the application of different data mining techniques.

KEYWORDS: SOFTWARE ENGINEERING - INFORMATION MINING

***Son coautores del presente trabajo Darío Rodríguez[2], Hernán Merlino[2] , Pablo Pytel [1,2], Juan Vanrell[1]**

Introducción

La Inteligencia de Negocio propone un abordaje interdisciplinario en el que confluyen, entre otras, la Informática, la Matemática y la Economía, su finalidad se centra en generar un conocimiento que contribuya con la toma de decisiones de gestión y generación de planes estratégicos en las organizaciones (Thomsen, 2003). La Explotación de Información, por su parte, constituye la subdisciplina de la Informática que aporta a la Inteligencia de Negocio (Negash & Gray, 2008) las herramientas de análisis y síntesis necesarias para extraer el conocimiento no trivial, alojado (implícitamente) en los datos disponibles en diferentes fuentes de información (Schiefer et al., 2004). Habitualmente, para un experto -o para el responsable de un sistema de información-, no son los datos en sí lo más relevante, sino el conocimiento que se encierra en sus relaciones, fluctuaciones y dependencias. Por ende, identificado el problema de inteligencia de negocio, decidirá la secuencia de procesos de explotación que deben ser ejecutados para obtener una solución adecuada.

Un Proceso de Explotación de Información se define, como un grupo de tareas relacionadas lógicamente (Curtis et al., 1992) que, a partir de un conjunto de información con un cierto grado de valor para la organización, se ejecuta para lograr otro, con un grado de valor mayor que el inicial (Ferreira et al., 2005; Kanungo, 2005). Adicionalmente, existe una variedad de técnicas de minería de datos, en su mayoría provenientes del campo del Aprendizaje Automático (García-Martínez et al., 2003), susceptibles de ser practicadas en cada uno de estos procesos.

Cuando nos encontrábamos en la etapa inicial de nuestro trabajo de investigación observamos que en la bibliografía consultada se repetía con insistencia el uso indiscriminado de los términos "minería de datos" (o *data mining*) y "explotación de información" (o i*nformation mining*) para referirse al mismo cuerpo de conocimientos. Consideramos tal confusión, un error análogo al de utilizar como sinónimos "ciencias de la computación" y "sistemas de información". Por ello, vale aclarar que, por un lado, la minería de datos está relacionada con la tecnología (algoritmos) y, la explotación de información, con los procesos y las metodologías propias de la ingeniería; y que, por otro, la primera se aproxima a las operatorias propias de la Programación, mientras que la segunda se acerca más a los procesos de la Ingeniería de Software.

Es en virtud de este contexto, que postulamos la necesidad de organizar un nuevo cuerpo de conocimientos para la Ingeniería de Explotación de Información, cuyo eje se centre en la problemática derivada de su implementación y uso en la industria. Una de las razones de peso que dan impulso al despegue de esta nueva disciplina, ha sido el descubrimiento de una falta de técnicas asociadas a la ejecución de cada una de las fases de las metodologías de explotación de información vigentes (García-Martínez et al., 2011).

Asimismo, la comprobación de la inadecuación de los métodos y herramientas de la Ingeniería en Software, en tanto no se abocan a los aspectos prácticos requeridos para proyectos de explotación de información; pone de relieve la necesidad del desarrollo y validación de una metodología específica, que pueda asistir a los practicantes del área de software y proveer la ineludible objetividad, racionalidad, generalización y confiabilidad que tales proyectos demandan. Al respecto, durante la última década se han obtenido avances de significativa trascendencia en los siguientes dominios: clasificación de familias de asteroides (Perichinsky et al., 2003), reglas para la identificación de caras humanas (Britos et al., 2005), detección de cambios de consumo de usuarios (Grosser et al., 2005; Britos et al., 2008d), localización de patrones en eventos meteorológicos (Cogliati et al., 2006), predicción de la salud de una comunidad (Felgaer et al., 2006), detección de daños al corazón (Ferrero et al., 2006), registro de uso de sitios web (Britos et al., 2008), selección de protocolos pedagógicos (Britos et al., 2008b), comprobación de malentendidos en programación (Britos et al., 2008), detección de patrones criminales (Valenga et al., 2008e), reconocimiento de patrones de daños en la industria automotriz (Flores et al., 2009), descubrimiento de patrones de deserción en estudiantes universitarios (Kuna et al., 2010a; 2010b), entre otros. Por lo tanto, en base a nuestra experiencia, al campo teórico-conceptual de la Ingeniería de Software y a la luz de los adelantos arriba mencionados, proponemos una batería metodológica para la explota-

ción de información fundada en: un modelo de negocio, un proceso de educción de requisitos, un método de estimación, una metodología de selección de herramientas, un método de transformación de datos y un grupo de procesos para la explotación de información basados en técnicas de minería de datos.

Marco conceptual propuesto

Modelo de proceso para proyectos de explotación de información

La Ingeniería en Software utiliza diversos modelos y metodologías para obtener proyectos de informática con gran nivel de previsibilidad y excelencia. Estos, permiten controlar la calidad final de un producto a desarrollar, estableciendo controles sobre cada una de las etapas que intervienen en el proceso productivo, entendido no sólo como la producción en sí misma, sino también, como las tareas relacionadas con la gestión de un proyecto y de la empresa que lo lleva a cabo.

En el caso de proyectos clásicos, modelos bien probados como CMM (SEI, 2006) o COMPETISOFT -para Pymes- (Oktaba et al., 2007), han sido utilizados con una recurrencia y resultados tales que habilitan su consideración como estables y altamente testeados (en el caso de COMPETISOFT, el más probado ha sido aquel que dio origen al llamado MoProSoft, Oktaba et al., 2005). Sin embargo, son manifiestamente improcedentes y resultan inadecuados para empresas que se dedican a llevar a cabo proyectos de explotación de información, debido a las diferencias que se presentan, mayormente, en la parte operativa de un proyecto. Entre estas, la más evidente se da en los procesos de desarrollo y mantenimiento de software, en los cuales COMPETISOFT define como proceso natural el ciclo de fases de un proyecto de software tradicional (fases de Inicio, Requisitos, Análisis y Diseño, Construcción, Integración, Pruebas y Cierre).

Lo mismo ocurre al evaluar las principales metodologías existentes, ya que se observa la falta de herramientas que permitan soportar de forma completa las fases que, en COMPETISOFT, se encuentran bien definidas y agrupadas en el proceso de administración de proyectos específicos. Por otro lado, entre las metodologías que acompañan el desarrollo de proyectos de explotación de información, se destacan CRISP (Chapman et al., 2000), P3TQ (Pyle D., 2003) y SEMMA (SAS, 2008) que, si bien fueron probadas y tienen un buen nivel de madurez en cuanto al desarrollo del proyecto, dejan de lado aspectos a nivel gestión y empresa. Presentan, también, carencias a la hora de definir las fases relacionadas a la administración, incluso, los pocos elementos administrativos con los que cuenta, se hallan mezclados con los de producción. Además, no consideran las tareas relacionadas con seguimiento, verificación y medición que deberían acompañar al proceso de desarrollo. Claramente, las actividades de administración de proyectos deben llevarse a cabo paralelamente y en procesos separados.

Como solución, proponemos un Modelo de procesos para proyectos de explotación de información (Vanrell et al., 2010) basado en la unión de COMPETISOFT con CRISP-DM, en el que se eliminen todas las fases no necesarias, adaptando aquellas que sean vitales y agregando nuevas, acorde a los aspectos específicos estudiados. Hemos seleccionado CRISP-DM como metodología de referencia a CRISP dado que la comunidad científica ha considerado que, comparativamente, esta última ofrece más instrumentos a nivel operativo que las otras dos antes mencionadas. Las fases propuestas para el Modelo de procesos para administración de proyectos de explotación de información y sus tareas asociadas se pueden observar en la tabla 1.

Las fases propuestas para el Modelo de procesos para desarrollo de proyectos de explotación de información y sus tareas asociadas se pueden observar en la tabla 2.

Proceso de educción de requisitos en proyectos de explotación de información

La primera tarea para la administración de proyectos y el desarrollo de procesos indicados en el Modelo de procesos, es la de buscar y definir los objetivos, criterios de éxito y expectativas del proyecto de explotación de información. En otras palabras, es menester educir los requerimientos que deben ser satisfechos.

La necesidad de adaptar el proceso tradicional de la ingeniería de requerimientos para sistemas-proyecto de explotación de información,

SUBPROCESO	TAREA	SALIDA
Planificación / Entendimiento del negocio	Entendimiento del negocio	▪ Conocimiento del negocio Objetivos del negocio ▪ Criterios de éxito
	Definir el proceso específico basado en la descripción del proyecto y el proceso de desarrollo y mantenimiento	▪ Proceso Especifico (forma parte del Plan de Desarrollo)
	Definir el protocolo de entrega con el cliente	▪ Plan de Entrega
	Definir ciclos y actividades con base en la descripción del proyecto y en el proceso específico	▪ Proceso Especifico (forma parte del Plan de Desarrollo)
	Determinar tiempo estimado para cada actividad	▪ Calendario de actividades (forma parte del Plan de Desarrollo) incorpora el tiempo estimado en el Plan de Proyecto
	Elaborar plan de adquisiciones y capacitación	▪ Plan de Adquisiciones y Capacitación
	Establecer el equipo de trabajo	▪ Equipo de trabajo (forma parte del Plan de Desarrollo)
	Establecer el calendario de actividades	▪ Calendario de actividades (forma parte del Plan de Desarrollo)
	Calcular el costo estimado del proyecto	▪ Costo estimado (forma parte del Plan de Proyecto)
	Evaluación de la situación	▪ Inventario de recursos Requerimientos, suposiciones y restricciones ▪ Riesgos y contingencias (forma parte del Plan de Proyecto nombrado como Plan de Manejo de Riesgos) ▪ Terminología ▪ Costos y beneficios
	Producir un Plan de Proyecto	▪ Plan de Proyecto, incluye ciclos y actividades, tiempo estimado, plan de adquisiciones y capacitación, equipo de trabajo, costo estimado, calendario, plan de manejo de riesgos y protocolo de entrega
	Producir un Plan de Desarrollo	▪ Plan de Desarrollo (incluye descripción del producto y entregables, proceso específico, equipo de trabajo y calendario) Lista inicial de técnicas y herramientas
	Formalizar el inicio de un nuevo ciclo del proyecto	
Realización	Acordar las tareas con el equipo de trabajo	
	Acordar la distribución de información	
	Revisar con el responsable la descripción del producto, el equipo de trabajo y el calendario	
	Revisar cumplimiento del plan de adquisiciones y capacitación	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Administrar subcontratos	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Recolectar reportes de actividades y mediciones y sugerencias de mejora y productos de trabajo	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento Reporte de Mediciones y Sugerencias de Mejora
	Registrar costo real del proyecto	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Revisar el registro de rastreo basado en los productos de trabajo recolectados	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Revisar los productos terminados durante el proyecto	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Recibir y analizar las solicitudes de cambio del cliente	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Realizar reuniones con el equipo de trabajo y cliente para reportar avances y tomar acuerdos	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
Evaluación y Control	Evaluar el cumplimiento del plan de proyecto y plan de desarrollo	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Analizar y controlar los riesgos	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
	Generar el reporte de seguimiento del proyecto	▪ Reporte de Seguimiento / Plan de monitoreo y mantenimiento
Cierre / Entrega	Formalizar la terminación del proyecto o ciclo	▪ Documento de aceptación
	Llevar a cabo el cierre del contrato con subcontratistas	
	Generar el reporte de mediciones y sugerencias de mejora	▪ Reporte de mediciones y sugerencia de mejoras - Lecciones Aprendidas
	Planear la entrega	▪ Plan de entrega (forma parte del Plan de Proyecto nombrado como protocolo de entrega)

Tabla 1. Proceso de administración de proyectos

SUBPROCESO	TAREA	SALIDA
Entendimiento del negocio	Determinar las metas del Data Mining	▪ Metas del Data Mining
		▪ Criterios de éxito del Data Mining
Entendimiento de los datos	Reunir los datos iniciales	▪ Reporte de datos iniciales
	Describir los datos	▪ Reporte de descripción de datos
	Explorar los datos	▪ Reporte de exploración de datos
	Verificar la calidad de los datos	▪ Reporte de calidad de los datos
Preparación de los datos	Tareas preparatorias	▪ Datasets
		▪ Descripción de los Datasets
	Seleccionar los datos	▪ Justificación de inclusión / exclusión
	Limpiar los datos	▪ Reporte de limpieza de datos
	Construir los datos	▪ Atributos derivados
		▪ Registros generados
	Integrar los datos	▪ Datos combinados (combinación de tablas y agregaciones)
	Formatear los datos	▪ Datos formateados
Modelado	Seleccionar la técnica de modelado	▪ Técnica de modelado
		▪ Suposiciones de modelado
	Generar el diseño de test	▪ Diseño de test
	Construir el modelo	▪ Establecimiento de parámetros
		▪ Modelos
		▪ Descripción del modelo
	Evaluar el modelo	▪ Evaluación del modelo
		▪ Revisión de los parámetros establecidos
Evaluación	Evaluar resultados	▪ Evaluación de los resultados de Data Mining respecto a los criterios de éxito
		▪ Modelos aprobados
	Revisar el proceso	▪ Revisión del proceso
	Determinar próximos pasos	▪ Lista de posibles decisiones
		▪ Decisiones

Tabla 2. Proceso de desarrollo de proyectos

está fundamentada en que el análisis de requisitos de estos, difiere significativamente del de los sistemas convencionales. Las metodologías existentes para proyectos de explotación de información fallan a la hora de educir todos los conceptos necesarios durante el conocimiento del negocio: CRISP-DM educe un conjunto de conceptos, P3TQ otro y SEMMA un tercero. En general, se abocan a los relacionados con los objetivos del negocio (evaluando situaciones) y dejan fuera de consideración los relacionados con la determinación de los criterios de éxito y el plan del proyecto.

En este contexto, formulamos una propuesta metodológica (Britos et al., 2008c; Pollo-Cattaneo et al., 2009; 2010a) más robusta que las existentes, porque busca identificar los elementos que permitan entender el dominio de los requerimientos del proyecto de explotación de información, y establece los pasos para la educción. La estructura planteada es similar a la de la Ingeniería de Software, ya que permite avanzar de manera progresiva a través de los conceptos manteniendo su orden natural. Asimismo, durante la fase de comprensión del negocio, sugerimos un proceso de educción de requisitos para proyectos de explotación de información de cinco pasos, tal y como puede observarse en la Fig. 1.

Descripción de cada paso:

• Comprender el dominio del proyecto: consiste en establecer un lenguaje común entre las personas involucradas.

• Conocer los datos del dominio del proyecto: radica en el establecimiento de los requisitos; los datos necesarios para los requisitos y su localización, los riesgos involucrados en los mismos y sus restricciones.

• Comprender los objetivos del proyecto: alude a la identificación de los objetivos, sus limitaciones, expectativas y riesgos.

• Identificar los recursos humanos involucrados: se refiere al conocimiento de la lista de las personas implicadas, sus restricciones, riesgos y responsabilidades.

• Seleccionar la herramienta adecuada: implica identificar una herramienta adecuada al proyecto, de acuerdo con la información obtenida en los pasos anteriores.

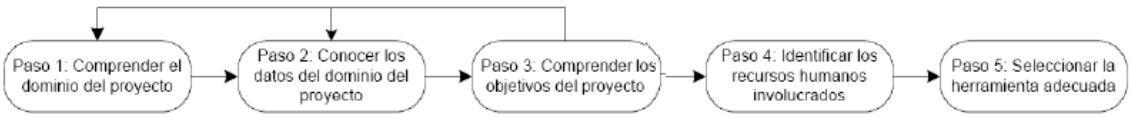

Fig. 1. Proceso de educción de requerimientos

Para conocer los datos del dominio del proyecto en términos de objetivos del requerimiento (requerimiento base o raíz), fuentes de información del requerimiento, suposiciones, restricciones y atributos que involucran al requerimiento, así como los riesgos y su plan de contingencia, es necesario entender el dominio del vocabulario involucrado, en tanto definiciones, acrónimos y abreviaturas.

Para comprender los objetivos del proyecto en términos de metas, criterios de éxito, expectativas, suposiciones, restricciones, riesgos y plan de contingencia, es preciso poseer conocimientos sobre los datos del dominio del proyecto en tanto que objetivos de los requerimientos, fuentes de información, suposiciones, restricciones, riesgos y plan de contingencia de los requerimientos.

Para identificar los recursos humanos involucrados, hace falta estar al tanto de los roles de dichos recursos en el proyecto (para lo cual es necesario comprender los objetivos en términos de metas, criterios de éxito, expectativas, suposiciones, restricciones, riesgos y su plan de contingencia), además de seleccionar la herramienta correcta (con la evaluación en función de las metas del proyecto). En la Fig. 2 se muestra la dependencia de los conceptos entre sí.

Hemos definido un conjunto de plantillas para cada producto involucrado (el conjunto completo de plantillas y sus ejemplos pueden encontrarse en Britos et al., 2008c). Cada plantilla se asocia a un concepto y presenta una descripción detallada de aquellos que son educidos, al tiempo que permite su evolución a través de los requerimientos del proceso de educción. La relación entre los conceptos educidos como productos y los pasos del proceso propuesto a generar por ellos, se muestran en la tabla 3.

Estimación empírica de carga de trabajo en proyectos de explotación de información

La gestión de un proyecto de software comienza con un conjunto de actividades que se denominan Planificación del proyecto, previo a lo cual, deben realizarse una serie de estimaciones: del trabajo a ejecutar, de los recursos necesarios y del tiempo que transcurrirá desde el comienzo hasta el final de su realización (Pressman, 2004). Dentro del Modelo de procesos descripto, la tarea Calcular el costo estimado del proyecto, también requiere una planificación para estimar los tiempos. Sin embargo, debido a la

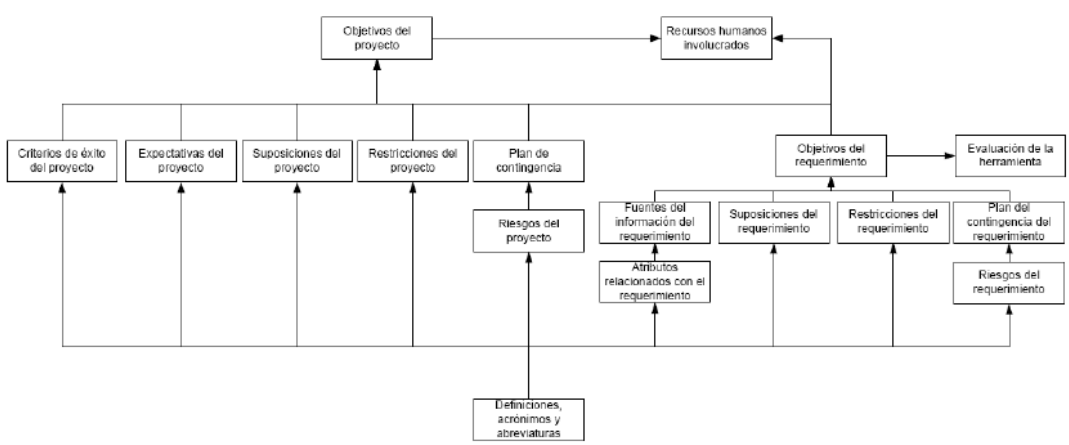

Fig. 2. Referencias cruzadas de los conceptos educidos representados por las plantillas

PASOS \ PRODUCTOS (Conceptos a ser educidos)	Definiciones, acrónimos y abreviaturas	Objetivos del proyecto	Criterios de éxito del proyecto	Expectativas del proyecto	Suposiciones del proyecto	Restricciones del proyecto	Riesgos del proyecto	Plan de contingencia	Recursos humanos involucrados	Objetivos del requerimiento	Fuentes del información del requerimiento	Suposiciones del requerimiento	Restricciones del requerimiento	Atributos del requerimiento	Riesgos del requerimiento	Plan del contingencia del requerimiento	Evaluación de herramientas
Comprender el dominio del proyecto	●								●								
Conocer los datos del dominio del proyecto	●								●	●	●	●	●	●	●	●	
Comprender los objetivos del proyecto	●	●	●	●	●	●	●	●	●								
Identificar los recursos humanos involucrados	●								●								
Seleccionar la herramienta adecuada	●								●	●			●	●			●

Tabla 3. Relación entre los productos y las etapas del proceso

diferencia existente entre software convencional y los proyectos de explotación de información, los métodos típicos convencionales, no son aplicables.

En consecuencia, en el campo de los sistemas de información, se abre la problemática sobre la construcción de métodos de estimación de proyectos de software que, estrictamente ceñidos a la realidad obtenible, logren resultados predictivos sobre los recursos a emplear. Los proyectos de explotación de información no escapan a esta necesidad y la historia de la Ingeniería en general y la Informática en particular, registran que los primeros abordajes son siempre de naturaleza empírica.

En el marco del escenario descripto, se han efectuado experimentos (Rodríguez et al., 2010; Pytel et al., 2011) encauzados a la obtención de una estimación empírica del porcentaje del tiempo del proyecto de explotación de información que insume la ejecución de cada una de las tareas, de las subfases de la Metodología CRISP-DM orientadas a proyectos para pequeños y medianos emprendimientos. A partir de los resultados alcanzados es posible, además, tener una aproximación sobre los tiempos globales del proyecto.

Los resultados obtenidos (ver Tabla 4) destacan aquellas fases y subfases que insumen una cantidad significativa de tiempo. En este sentido, en la Comprensión del Negocio y el Modelado se invierten más del 50% de la duración del proyecto. Y, a su vez, al interior de la primera de ellas, las subfases Determinar los objetivos de negocio y Evaluar la situación, utilizan más del 70% del tiempo pautado. Por otro lado, en Modelado, la subfase Construcción del modelo, requiere el 62,97% del plazo concedido a la totalidad de la ejecución de la fase.

Propuesta para una metodología de selección de herramientas de explotación de información

Una vez que los objetivos del proyecto se encuentran claramente definidos y los recursos humanos identificados, es preciso seleccionar

FASE	% del TIEMPO
Fase 1 COMPRENSIÓN DEL NEGOCIO	20,70
Fase 2 ENTENDIMIENTO DE LOS DATOS	10,90
Fase 3 PREPARACION DE DATOS	15,61
Fase 4 MODELADO	34,41
Fase 5 EVALUACIÓN	7,45
Fase 6 IMPLANTACION	10,93

Tabla 4. Carga de trabajo de cada fase de la Metodología CRISP-DM

la herramienta de explotación de datos. La importancia de su impacto en la organización y la inversión que amerita en términos económicos, hacen que el proceso asociado a su elección sea un tema crítico. Aún más cuando, a lo infrecuente de su realización, se le adiciona la expectativa de un cierto retorno de la inversión. En ocasiones, la adquisición de una herramienta inadecuada, se deriva de la existencia de proveedores diversos y herramientas de uso libre que permiten realizar la tarea sin contar con un método objetivo de selección. Ello trae aparejadas consecuencias tales como: (a) pérdida de tiempo y dinero (b) incremento del riesgo de no cumplir con los objetivos de negocio establecidos. En este contexto, planteamos una metodología (Britos et al., 2005) que apunta a organizar el proceso de selección de una herramienta, para que la organización pueda escoger la que mejor se adapte a sus requisitos, basándose no sólo en cuestiones económicas, sino también en las necesidades propias del negocio.

Esta metodología está compuesta por las siguientes fases:

• Fase 1. Documentar la necesidad:
En esta fase se define y establece el marco general de referencia para la selección de una herramienta de explotación de datos. Para ello, se ponen en consideración las áreas y funciones de la organización que se involucrarán con la herramienta y los objetivos que se pretenden lograr a través de la misma.

• Fase 2. Análisis de la necesidad:
El objetivo de esta fase es documentar las características del negocio que la herramienta debe atender. Así, se intentan describir los requerimientos de la misma que mejor se adapten a los objetivos del negocio de la organización. Ello implica, por ejemplo, no pagar un precio muy elevado por una herramienta que se usará en un 10% de su potencial, que no aporte los métodos necesarios o que resulte obsoleta en el primer intento de ampliación.

• Fase 3. Búsqueda en el mercado:
Esta fase se aboca a la búsqueda de herramientas en el mercado. Tal proceso debe quedar resumido en un informe que detalle los proveedores encontrados.

• Fase 4. Contacto con proveedores:
El objetivo de esta fase es contactar a cada proveedor identificado en la fase anterior y solicitarle información de la herramienta en cuestión.

• Fase 5. Entrevistar posibles candidatos y recopilar información:
En esta fase se concertan entrevistas con cada proveedor con el propósito de completar la información faltante sobre el accionar de éste y los detalles de los productos. Para concluir, se organizan los datos recopilados verificando la homogeneidad de los mismos a efectos de facilitar su comparación y se prepara un informe por cada herramienta.

• Fase 6. Armado del informe de criterios a tener en cuenta:
Esta fase tiene por objetivo desarrollar un informe con criterios de ponderación (ver ejemplo en Fig. 3), que se adecue a las necesidades de la organización y que se constituya en la base de trabajo para las tareas posteriores y para la selección final.

Los criterios del informe son agrupados en cuatro categorías o grupos, que también deberán ser ponderados y que se describen a continuación:

1. Características técnico-funcionales de la herramienta: bajo esta categoría se agrupan los criterios que están ligados a las características técnicas y funcionales de la herramienta.
2. Características del proveedor: incluye aquellos criterios que hacen a la organización proveedora, por ejemplo, evolución y crecimiento, facturación anual, ubicación geográfica, otros clientes y experiencia. Así, se evalúa su solidez, ya que si dejara de existir la organización quedaría un sistema de información sin mantenimiento ni posibilidad de evolución.
3. Características del servicio: se evalúan puntos específicos de la prestación que brinda el proveedor sobre la herramienta.
4. Características económicas: son aquellas relacionadas con costos de licencias y de servicio de mantenimiento de la herramienta.

• Fase 7. Evaluar los candidatos:
En esta fase el equipo debe concertar nuevas entrevistas con los proveedores a efectos de formular una solicitud de propuesta

Encabezado propuesto

Nombre de la herramienta:	
Proveedor: Evaluación:	1 = Malo, 2 = Regular, 3 = Bueno, 4 = Muy bueno

Cuestionario propuesto

Criterios de selección		Descripción	Pond X	Valor Y	Pond X* Y
1. Características técnicos – funcionales					
Metodología / Ciclo de vida soportado		Metodologías / ciclos de vida que soporta la herramienta para la explotación de datos (CRISP-DM, SEMMA, etc.)	3		
Adaptabilidad y flexibilidad para la toma de datos	Desde Bases de Datos	Cantidad de formatos soportado para la toma de datos desde bases de datos diversas.	8		
	Desde fuentes externas (word, excel, etc.)	Cantidad de formatos soportado para la toma de datos.	8		
Facilidad para integrar diferentes técnicas		Posibilidad de integrar diversas técnicas de explotación de datos	5		
Multi-lenguaje		Permite trabajar en distintos idiomas (tomando el idioma ingles como idioma principal).	2		
Técnicas usadas		Cantidad de técnicas que permiten la explotación de datos para el logro de los objetivos del negocio (redes neuronales, redes bayesianas, algoritmos de inducción, etc.)	18		
Herramientas de visualización y informe		Permite visualizar la salida de las distintas técnicas utilizadas en la explotación, así como la generación de informes.	12		

Fig. 3. Ejemplo de Modelo para informe con criterios de ponderación

técnica y económica que complemente el punto anterior.

Para cumplir con el informe, cada criterio se calificará con un valor de 1 a 4, en la columna valor "Y" del cuestionario propuesto (siendo 1=Malo, 2=Regular, 3=Bueno, 4=Muy Bueno). Cada valor en la columna "Y" será multiplicado por el factor en la columna "Pond X" y se colocará en la columna "Pond X*Y". La sumatoria de la columna "Pond X*Y" será multiplicada por la ponderación del grupo y dividida por 100 para obtener la ponderación del grupo en general. Esta operación deberá repetirse para los 4 grupos.

Una vez completo el informe con los datos recolectados, se compararán las ponderaciones resultantes entre las distintas herramienta, para ser presentadas en una reunión de trabajo con el equipo de proyecto convocada a los fines de discutir la evaluación, confrontar los valores obtenidos y seleccionar los candidatos. Al finalizar esta actividad, se deberán elegir los productos de los cuales se pedirá una demostración a los proveedores.

• Fase 8. Demostración del producto:
En este punto los proveedores muestran el producto a los usuarios designados, quienes completan -en cada visita- los cuestionarios confeccionados a tales efectos (ver ejemplo en Fig. 4). Los usuarios calificarán cada criterio indicando en la columna de ponderación ("P") un valor del 0 a 5.

Al finalizar las visitas se recopilan los cuestionarios, se suman los puntajes de cada proveedor otorgado por cada encuestado y se arma un promedio de puntos obtenidos por cada producto. Estos valores se agregan al informe armado en la fase anterior.

• Fase 9. Evaluación de los productos:
Este es el momento de la comparación de los resultados obtenidos (criterios ponderados y demostración de productos) y de la selección del producto que haya sumado el mayor puntaje durante todo el proceso.

Propuesta de un método de transformación de datos orientado al uso de explotación de información

Si avanzamos sobre el proceso de Desarrollo de proyectos del modelo descripto, advertiremos que el subproceso de Preparación de los datos, se ocupa de tomar la información disponible para su manipulación, transformación y presentación con el objetivo de efectuar su procesamiento a través de técnicas de minería de datos. Cuando se trabaja en explotación de información, los datos representan hechos de la vida real, por lo que deben sujetarse a una preparación previa que los disponga para la utilización de la herramienta. Para conocer qué transformaciones se deben realizar y cómo se deben presentar, es menester responder

Encabezado propuesto

Nombre del dataminer:
Fecha:
Proveedor:
Ponderación: 0 = Ítem no evaluado 1 = Ítem evaluado no soportado. 2 = Ítem evaluado soportado de manera incompleta 3 = Ítem evaluado soportado con necesidad de varias modificaciones factibles 4 = Ítem evaluado soportado de manera correcta 5 = Ítem evaluado soportado y provee de valor agregado al trabajo

Cuestionario propuesto

CRITERIOS	P
Compañía con múltiples filiales	
Multiplataforma simultanea	
Multilenguaje - varios idiomas	
Ayudas en pantalla en el idioma de trabajo de la organización	
Manuales en el idioma de trabajo de la organización	
Importación de datos de distintas fuentes	
Cantidad de técnicas de explotación utilizadas para lograr los objetivos del negocio	
Integración entre técnicas	

Fig. 4. Ejemplo de Modelo para cuestionario de usuario

dos preguntas fundamentales: ¿qué solución debemos obtener? y ¿qué técnica de explotación utilizaremos? La primera cuestión se relaciona con las características y la cantidad de información que se quiera manipular, mientras que la segunda cuestión, indica la forma en que se debe presentar la información para la explotación. Por lo tanto, este proceso, lejos de ser automático, comporta la actuación de un ingeniero que debe poner en juego su conocimiento para generar el conjunto de datos necesarios para la aplicación de un modelo de explotación.

En virtud de lo expuesto, (Merlino et al., 2005) ofrecemos un método de transformación de datos orientado a la explotación de información, detallando las características necesarias que debe poseer el entorno de trabajo para la automatización del mismo. Nos referimos al Método Unificado de Transformación (MUT), resultante de la experiencia adquirida en el procesamiento de grandes volúmenes de información sobre distintas plataformas (desde equipos IBM 390 a redes de computadoras con distintas versiones de Microsoft Windows, pasando por AS 400 y versiones de Unix). El principal objetivo de éste no se centra en emprender todas las transformaciones en un solo paso, sino en llevar adelante pequeñas modificaciones sobre los datos, para luego realizar una prueba de regresión completa de lo hecho hasta el momento y, tras una evaluación satisfactoria, reiniciar el ciclo con la siguiente innovación planeada. Para poder concretarlo, recomendamos una serie de fases que se despliegan una vez conocidas las repuestas a las dos preguntas fundamentales, es decir, dónde estoy y a dónde quiero llegar.

Descripción de las fases propuestas para la transformación de datos:

• Fase de análisis de los requerimientos de transformación:
El primer paso consiste en recabar información acerca de qué es lo que se necesita obtener. En otras palabras, conocer el formato que debe tener nuestro conjunto de datos para poder ser ingresado al modelo elegido para la minería de datos.

Con el formato de archivo de ingreso al modelo de datos ya especificado, se abren dos cursos de acción posibles: [a] la indagación por el origen de datos para la creación del archivo solicitado, o [b] la vuelta sobre el modelo de minería de datos seleccionado, con la finalidad de detectar los requisitos del conjunto de datos para su uso, es decir, cantidad de registros

necesarios para su aplicación y cantidad de conjuntos de datos para su validación o entrenamiento. Sea que se opte por la primera o la segunda opción, debe volverse sobre la técnica de la entrevista. Pero, mientras que en el primer caso, se busca detectar el origen de datos para poder acceder a ellos, en el segundo, se espera alcanzar el origen del conjunto de datos de pruebas, para realizar un análisis del modelo requerido y disponible.

En la Fig. 5 se esquematiza el proceso de obtención de requisitos para la transformación de datos.

• Fase de modelo de las transformaciones:
En esta etapa se diseñan las transformaciones necesarias para que los datos tomados del origen lleguen a la estructura requerida por el modelo de minería de datos. Para esto, se utilizan casos de usos que tendrán como actor al "controlador": encargado de generar los eventos para que el flujo de los datos tenga las transformaciones necesarias.

• Fase de codificación:
En este paso se codifican todos los programas esenciales para realizar las transformaciones necesarias para el modelo de minería de datos. El encargado de la codificación recibirá, al menos, las especificaciones de los formatos de entradas y salidas y los casos de uso definidos en la fase anterior. Como el controlador de tareas es independiente del programa que debe ejecutar, puede utilizarse el lenguaje de programación que se desee, siempre y cuando éste pueda ser soportado por la plataforma en la que se ha de trabajar.

• Fase de pruebas:
Una vez finalizada la codificación, se avanza hacia las pruebas de unidad y regresión. Esta etapa no sólo se refiere a la comprobación de los programa encargados de la generación de las transformaciones, sino a la construcción del archivo que proveerá la secuencia de pasos al controlador de tareas.

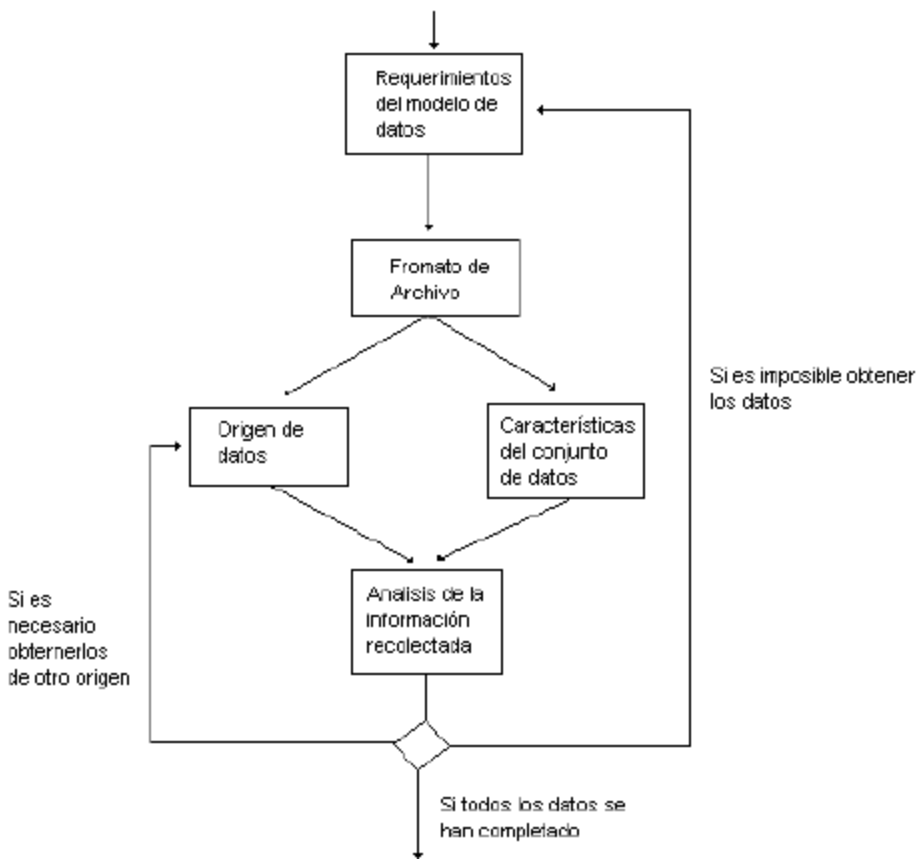

Fig. 5. Proceso de obtención de requisitos para la transformación de datos

• Fase de evaluación:
Con toda la información de las pruebas antes realizadas y, en caso de no encontrarse ninguna anomalía, se analiza el problema y se decide el camino de acción a tomar.

• Fase de nueva iteración:
De lo dicho hasta el momento se deduce la necesidad de generar nuevas iteraciones con cada paso. Así, este proceso se repite hasta finalizar todas las transformaciones involucradas en la satisfacción del Modelo de Minería de Datos.

Propuesta de procesos de explotación de información para problemas de inteligencia de negocio

Continuando con el desarrollo de Proyectos del modelo de proceso descripto, observamos que las tareas del subproceso de Modelado, usan ciertas técnicas y algoritmos de data mining para procesar la información disponible una vez que los datos se encuentran transformados y listos para ser utilizados.

En primer lugar, es necesario identificar todas las fuentes de información (bases de datos, archivos planos, entre otras), para integrarlas formando una sola, que será identificada bajo el término "datos integrados". Hemos definido cinco procesos de explotación de información, (Britos et al., 2008a; Britos y García-Martínez, 2009; Pollo-Cattaneo et al., 2010b) que se describen en las siguientes subsecciones: descubrimiento de reglas de comportamiento, descubrimiento de grupos, descubrimiento de atributos significativos, descubrimiento de reglas de pertenencia a grupos y ponderación de reglas de comportamiento o de pertenencia a grupos. Además, es posible asociar a cada proceso las siguientes técnicas: algoritmos TDIDT -Top Down Induction Decision Trees- (Quinlan, 1986), Mapas Auto Organizados de Kohonen -Self-Organizing Maps o SOM- (Kohonen, 1995) y Redes Bayesianas (Heckerman et al., 1995). Asimismo, los procesos propuestos se validaron en los siguientes dominios: alianzas políticas, diagnóstico médico y comportamiento de usuarios. (Un reporte completo de estas validaciones puede ser encontrado en Britos, 2008.)

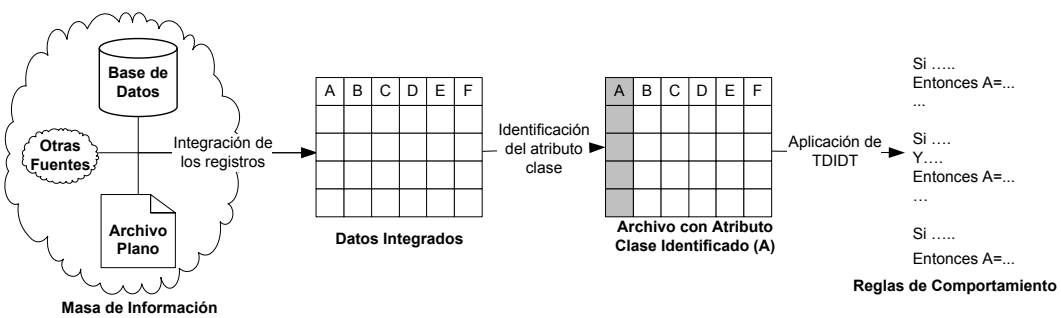

Fig. 6. Esquema y subproductos resultantes de aplicar TDIDT al descubrimiento de reglas de comportamiento

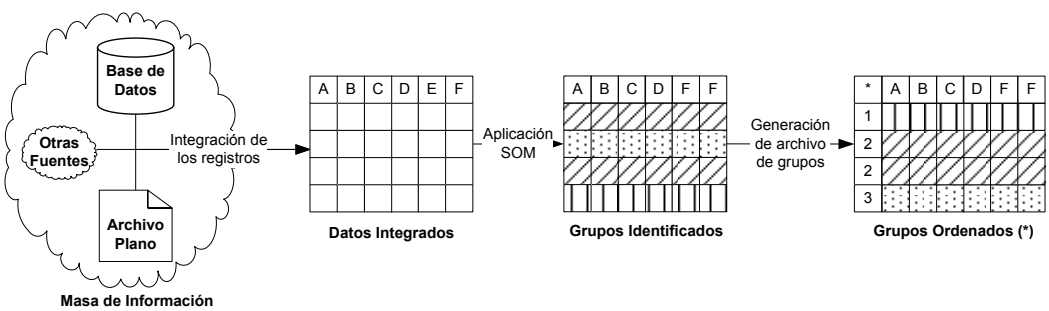

Fig. 7. Esquema y subproductos resultantes de aplicar SOM para el descubrimiento de grupos

Proceso de descubrimiento de reglas de comportamiento

Este proceso toma lugar cuando se requiere identificar cuáles son las condiciones para obtener determinado resultado en el dominio del problema, por ejemplo: caracterización del local más visitado por los clientes, identificación de los factores que inciden en el alza de las ventas de un producto dado, individualización de los rasgos de consumidores con alto grado de fidelidad a la marca, establecimiento de los atributos demográficos y pictográficos que distinguen a los visitantes de un *website*, entre otros.

Para el descubrimiento de reglas de comportamiento definidos a partir de atributos-clase, en el dominio del problema representado por la masa de información disponible, proponemos la utilización de algoritmos de inducción TDIDT (Britos et al., 2008). Este proceso y sus subproductos pueden visualizarse gráficamente en la Fig. 6.

Proceso de descubrimiento de grupos

A partir de este proceso se asiste a la identificación de una partición en la masa de información disponible sobre el dominio del problema. Son ejemplos de problemas que requieren este proceso: identificación de segmentos de consumidores para bancos y financieras, tipificación de llamadas de clientes para empresas de telecomunicación, reconocimiento de grupos sociales con características similares, determinación de grupos de estudiantes con atributos homogéneos, entre otros.

Para el descubrimiento de grupos (Kaufman & Rousseeuw, 1990; Grabmeier & Rudolph, 2002) a partir de masas de información sobre las que no se dispone ningún criterio de agrupamiento a priori (en el dominio del problema) proponemos la utilización de Mapas Auto Organizados de Kohonen -o SOM por su sigla en inglés- (Ferrero et al., 2006; Britos et al., 2008). El uso de esta tecnología explora la existencia de grupos que permitan una partición característica del dominio del problema que la masa de información disponible representa. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Fig. 7.

Proceso de descubrimiento de atributos significativos

Este proceso hace referencia a la identificación de los factores con mayor incidencia (o frecuencia de ocurrencia) sobre un determinado resultado del problema. A modo de ejemplo, algunos problemas que requieren este proceso son: factores con incidencia sobre las ventas, rasgos distintivos de clientes con alto

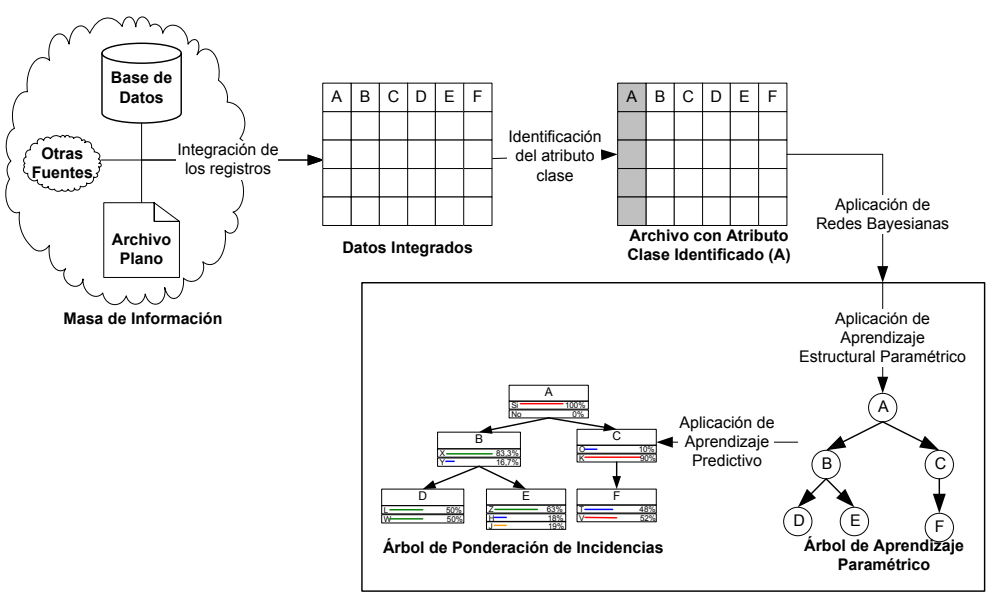

Fig. 8. Esquema y subproductos resultantes de aplicar Redes bayesianas a la ponderación de interdependencia entre atributos

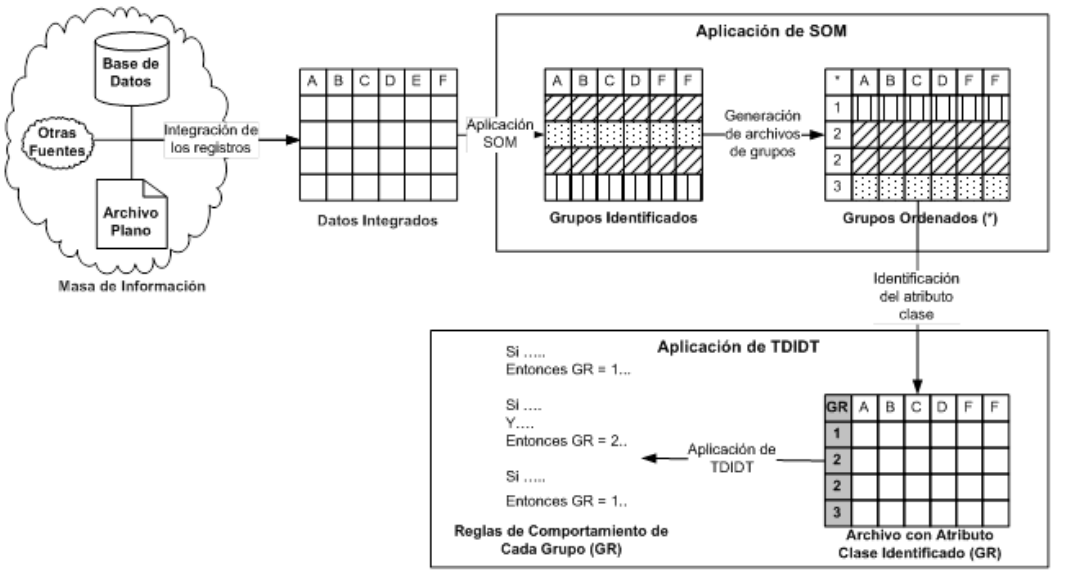

Fig. 9. Esquema y subproductos resultantes de SOM y TDIDT aplicados al descubrimiento de reglas de pertenencia a grupos

grado de fidelidad a la marca, atributos claves que convierten un determinado producto en vendible, características sobresalientes de los visitantes de un *website*, entre otros.

Para ponderar en qué medida la variación de los valores de un atributo, incide sobre la variación del valor de un atributo-clase, indicamos la utilización de Redes bayesianas (Britos et al., 2008). Esta tecnología pretende identificar si existe algún grado de interdependencia entre los atributos que modelan el dominio del problema que la masa de información disponible representa. Este proceso y sus subproductos pueden visualizarse gráficamente en la Fig. 8.

Proceso de descubrimiento de reglas de pertenencia a grupos

En virtud de este proceso se pretende determinar cuáles son las condiciones de pertenencia a cada una de las clases, en una partición desconocida *a priori*, pero presente en la masa de información disponible sobre el dominio del problema. Los siguientes son ejemplos de problemas que requieren este proceso: tipología de perfiles de clientes y su caracterización, distribución y estructura de los datos de un *website*, segmentación etaria de estudiantes y comportamiento de cada segmento, clases de llamadas telefónicas en una región y caracterización de cada una de éstas, entre otros.

Para el descubrimiento de reglas de pertenencia a grupos, postulamos la utilización de mapas auto-organizados (SOM) y, tras la identificación de los grupos, el empleo de algoritmos de inducción (TDIDT) para establecer las reglas de pertenencia a cada uno (Britos et al., 2005; Cogliati et al., 2006a). Este proceso y sus subproductos pueden ser visualizados gráficamente en la Fig. 9.

Proceso de ponderación de reglas de comportamiento o de pertenencia a grupos

Con la aplicación de este proceso se pretende identificar cuáles son las condiciones con mayor incidencia (o frecuencia de ocurrencia) sobre la obtención de un determinado resultado en el dominio del problema, sean éstas las que en mayor medida inciden sobre un comportamiento, o las que mejor definen la pertenencia a un grupo. Son ejemplos de problemas que requieren este proceso: identificación del factor dominante en el alza de las ventas de un producto dado, individuación del rasgo con mayor presencia en los clientes con alto grado de fidelidad a la marca, reconocimiento de la frecuencia de ocurrencia de cada perfil de clientes, determinación del

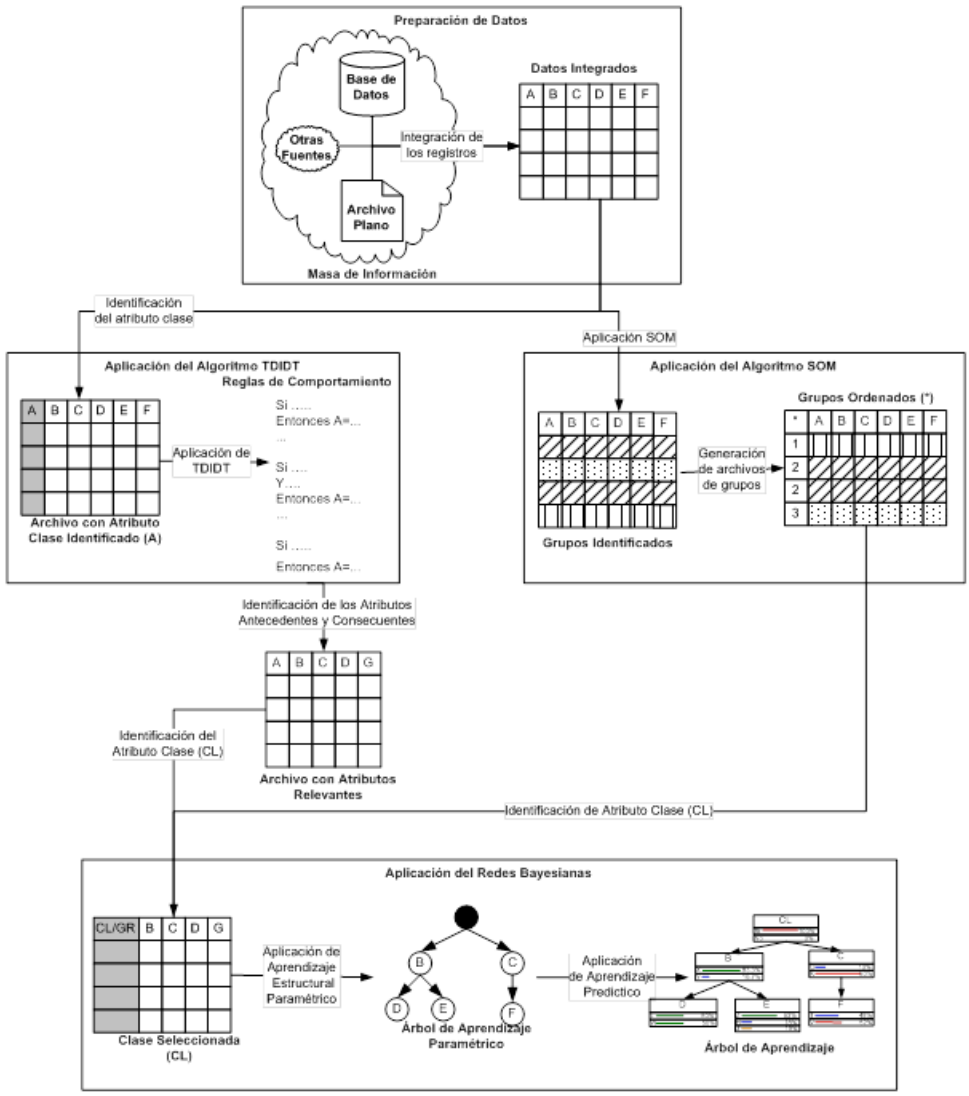

Fig. 10. Esquema y subproductos resultantes de Redes bayesianas aplicadas a la ponderación de reglas de comportamiento o de pertenencia a grupos

tipo de llamada más frecuente en una región, entre otros.

Para la ponderación de reglas de comportamiento o de pertenencia a grupos proponemos la utilización de redes bayesianas. Lo cual puede realizarse a través de dos procedimientos diferentes, dependiendo de las características del problema a resolver. En el caso de que se contara con clases/grupos identificados, se acude a algoritmos de inducción TDIDT para descubrir las reglas de comportamiento de cada atributo clase, tras lo cual se utilizan Redes bayesianas para saber cuál de los atributos establecidos como antecedente de las reglas, tiene mayor incidencia sobre el atributo establecido como consecuente. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Fig. 10.

Conclusiones

La historia de la Explotación de Información comenzó con la sistematización de algoritmos de aprendizaje automático aplicados al descubrimiento de conocimientos hace más de un cuarto de siglo. Durante muchos años el interés de la comunidad científica estuvo enfocado más en los algoritmos que en los procesos. La evolución disciplinar que ha ido de una concepción artesanal a una concepción ingenieril en Proyectos de Software convencional, ha te-

nido su paralelo en la evolución de Proyectos de Explotación de Información. La experiencia que acumulamos en la última década nos ha permitido identificar la falta de una Ingeniería acorde a sus necesidades. Durante estos años hemos desarrollado, basados en principios de la Ingeniería en Software, una serie de herramientas que han madurado como las bases principales de nuestra versión de una Ingeniería de Explotación de Información, la cual ha sido usada en proyectos para pequeños y medianos emprendimientos. Este trabajo realiza una presentación sistemática de las herramientas desarrolladas para dicho proyecto, entre las que se encuentran: un Modelo de procesos, un Proceso de educción de requisitos, un Modelo de estimación empírica de carga de trabajo, una Metodología en selección de herramientas, un Método de transformación de datos y un conjunto de procesos de explotación de información para inteligencia de negocios.

Referencias

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

BRITOS, P.; ABASOLO, M.; GARCÍA-MARTÍNEZ, R.; 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. 9, 10.

BRITOS, P.; CATALDI, Z.; SIERRA, E.; GARCÍA-MARTÍNEZ, R. (2008). Pedagogical Protocols Selection Automatic Assistance. (2008b). Lecture Notes in on Artificial Intelligence, 5027: 331, 336.

BRITOS, P.; DIESTE, O.; GARCÍA-MARTÍNEZ, R. (2008c). 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 y Dewald Roode Eds. (Boston: Springer), IFIP Series, 274: 139, 150.

BRITOS, P.; FERNÁNDEZ, E.; OCHOA, M.; MERLINO, H.; DIEZ, E.; GARCÍA MARTÍNEZ, R. (2005) Metodología de Selección de Herramientas de Explotación de Datos. II Workshop de Ingeniería del Software y Bases de Datos. XI Congreso Argentino de Ciencias de la Computación. 113, 123.

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. 1041, 1050.

FELGAER, P.; GARCÍA-MARTÍNEZ, R. (2008) Bayesian Networks Optimization Based on Induction Learning Techniques. In Artificial Intelligence in Theory and Practice II, ed. M. Bramer, (Boston: Springer).

BRITOS, P.; GROSSER, H.; RODRÍGUEZ, D.; GARCÍA-MARTÍNEZ, R. (2008d) Detecting Unusual Changes of Users Consumption. In Artificial Intelligence in Theory and Practice II, Op. Cit, 276: 297, 306.

BRITOS, P.; JIMÉNEZ REY, E.; GARCÍA-MARTÍNEZ, E. (2008e) Work in Pro-gress: Programming Misunderstandings Discovering Process Based On Intelli-gent Data Mining Tools. Proceedings 38th ASEE/IEEE Frontiers in Education Conference.

BRITOS, P.; MARTINELLI, D.; MERLINO, H.; GARCÍA-MARTÍNEZ, R. (2007) Web Usage Mining Using Self Organized Maps. International Journal of Com-puter Science and Network Security, 7(6):45, 50.

Carnegie Mellon University, Software Engineering Institute (2006) Session F4H: Assessing and Understanding Student Learning.

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. April 2011.

COGLIATI, M.; BRITOS, P.; GARCÍA-MARTÍNEZ, R. (2006) Patterns in Temporal Series of Meteorological Variables Using SOM & TDIDT. Volume 217, En Artificial Intelligence in Theory and Practice, ed. M. Bramer, IFIP (Boston: Springer), 217: 305, 314.

CURTIS, B.; KELLNER, M.; OVER, J. (1992) Process Modelling. Communications of the ACM, 35(9): 75, 90.

FELGAER, P.; GARCÍA-MARTÍNEZ, R. (2008) Bayesian Networks Optimization Based on Induc-

tion Learning Techniques. En Artificial Intelligence in Theory and Practice II, Op. Cit.
FELGAER, P.; BRITOS, P.; 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. 471, 474.
FERRERO, G.; BRITOS, P.; GARCÍA-MARTÍNEZ, R., (2006) Detection of Breast Lesions in Medical Digital Imaging Using Neural Networks, IFIP, Volume 218, Professional Practice in Artificial Intelligence, eds. J. Debenham, (Boston: Springer). 1, 10.
FLORES, D.; GARCÍA-MARTÍNEZ; 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. 1021, 1030.
GARCÍA-MARTÍNEZ, R.; SERVENTE, M.; PASQUINI, D. (2003) Sistemas Inteligentes. Editorial Nueva Librería.
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 (en Prensain 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.; 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 net-works, 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. 231, 240.
KAUFMANN, L.; ROUSSEEUW, P. (1990) Finding Groups in Data: An Introduc-tion to Cluster Analysis. John Wiley & Sons Publishers.
KOHONEN, T. (1995). Self-Organizing Maps. Springer Verlag Publishers.
MUSEN, M.; FERGERSON, R.; GROSSO, W.; NOY, N.; CRUBEZY, M.; GENNARI, J. (2000) Component-Based Support for Building Knowledge-Acquisition Systems. Reporte SMI- 2000-0838. Stanford Medical Informatics. Universidad de Stanford.
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.
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.
MERLINO, H.; BRITOS, P.; IERACHE, J.; DIEZ E.; GARCÍA-MARTÍNEZ, R. (2005) Un Método de Transformacion De Datos orientado al uso de Explotación de Información. II Workshop de Ingeniería del Software y Bases de Datos. XI Congreso Argentino de Ciencias de la Computación. 22, 32
NEGASH, S.; GRAY, P. (2008) Business Intelligence. En Handbook on Decision Support Systems, 2 ed., Eds. F. Burstein y C. Holsapple (Heidelberg, Springer), 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.; 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. Río de Janeiro (Brasil). 308, 315.
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. 467, 469.
POLLO-CATTANEO, F.; BRITOS, P.; PESADO, P.; GARCÍA-MARTÍNEZ, R. (2010a) Proceso de Educción de Requisitos en Proyectos de Explotación de Información y (2010b) Ingeniería de Procesos de Explotación de Información. En Ingeniería de Software e Ingeniería del Conocimiento: Tendencias de Investigación e Innovación Tecnológica en Iberoamérica, Eds: R. Aguilar, J. Díaz, G. Gómez, E- León. Alfaomega Grupo Editor. 1,11 y 252, 263.
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. (1990) Learning Logic Definitions from Relations. (1986) Induction of decision trees. Machine Learning, 5:239, 2661(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. 664, 673.
SAS, (2008) SAS Enterprise Miner: SEMMA.
SCHIEFER, J.; JENG, J.; KAPOOR, S.; CHOWDHARY, P. (2004) Process Infor-mation Factory: A Data Management Approach for Enhancing Business Process Intelligence. Proceedings 2004 IEEE International Conference on E-Commerce Technology. 162, 169.
SEI (2006). CMMI-DEV for Development, Vers. 1.2. Software Engineering Institute Carnegie Mellon University.
SAS, (2008). SAS Enterprise Miner: SEMMA.
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. 31, 39.
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. 674, 682.
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. 167, 171.