ESTIMACIÓN DE PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN ESTUDIO COMPARADO DE MODELOS ANALÍTICOS Y EMPÍRICOS

Pytel, P., Tomasello, M., Rodríguez, D.; Arboleya, H., Pollo-Cattaneo. M., Britos, P., García-Martínez, R.

Grupo Investigación
en Sistemas de Información
Depto. 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

Grupo de Estudio en Metodologías
de Ingeniería de Software
Facultad Regional Buenos Aires.
Universidad Tecnológica Nacional.
Medrano 951 (C1179AAQ) Ciudad Autónoma de Argentina. Buenos Aires Tel +54 11 4867-7511
fpollo@posgrado.frba.utn.edu.ar

Grupo de Investigación
en Explotación de Información
Sede Andina (El Bolsón)
Universidad Nacional de Río Negro
San Martin esq. Pellegrini (8430) El Bolsón. Rio Negro. Argentina. TE + 54 11 02944 49-8939
paobritos@gmail.com

CONTEXTO

Este proyecto de investigación se desarrolla en el marco de la cooperación existente entre 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 Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional, y el Grupo de Investigación en Explotación de Información de la Sede Andina (El Bolsón) de la Universidad Nacional de Río Negro (SAEB-UNRN).

INTRODUCCIÓN

La Explotación de Información (en inglés Information Mining, IM) consiste en la extracción de conocimiento no-trivial que reside de manera implícita en los datos disponibles en distintas fuentes de información [1]. Dicho conocimiento es previamente desconocido y puede resultar útil para algún proceso [2]. Para un experto, o para el responsable de un sistema de información, normalmente no son los datos en sí lo más relevante, sino el conocimiento que se encierra en sus relaciones, fluctuaciones y dependencias. Esta disciplina engloba un conjunto de técnicas de Minería de Datos (Data Mining, DM) encaminadas a la extracción de conocimiento procesable, implícito en el almacén de datos (Data Warehouse, DW) de la organización. Las bases de estas técnicas se encuentran en el análisis estadístico y en los sistemas inteligentes. Con Explotación de Información se aborda la solución a problemas de predicción, clasificación y segmentación [3]. Estos resultados contribuyen con la toma de decisiones de gestión y generación de planes estratégicos en las organizaciones [4].

Por esta razón ha sido necesario disponer de métodos eficientes para la búsqueda de conocimiento en datos mediante el desarrollo de algoritmos y herramientas para la explotación de información. Para el desarrollo de estos algoritmos y herramientas se necesita de una metodología que lo asista. A través de la experiencia acumulada en proyectos de explotación de información se han ido desarrollando metodologías que permiten gestionar esta complejidad de una manera uniforme. Entre estas metodologías, la comunidad científica considera probada a la metodología CRISP-DM.

La metodología CRISP-DM [5] ha evolucionado para resolver los problemas que las organizaciones tienen a la hora de desarrollar proyectos de explotación de información. Así CRISP-DM define los procesos y tareas que se deben realizar para desarrollar en forma exitosa un proyecto de explotación de información. Estos procesos y tareas se encuentran divididos en cuatro niveles de abstracción, organizándolas de forma jerárquica que van desde el nivel más general hasta los casos más específicos (ver Figura 1).

En nivel más general, el proceso está organizado en seis fases, estando cada fase a su vez estructurada en varias tareas generales de segundo nivel o subfases. Las tareas generales se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas. Así, si en el segundo nivel se tiene la tarea general

"limpieza de datos", en el tercer nivel se dicen las tareas que tienen que desarrollarse para un caso específico, como por ejemplo, "limpieza de datos numéricos", o "limpieza de datos categóricos".

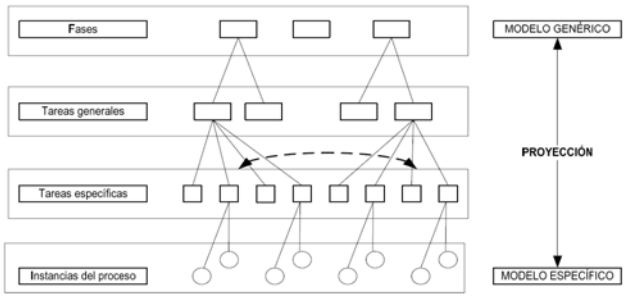

Fig. 1. Esquema de los cuatro niveles de abstracción de la metodología CRISP-DM

El cuarto nivel recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de Explotación de Información específico. En la tabla 1, se detallan las fases que componen la metodología CRISP-DM y se detalla cómo se componen cada una de ellas.

ESTADO ACTUAL DE CONOCIMIENTO SOBRE EL TEMA

Al comenzar la gestión de todo proyecto de software es necesario realizar las actividades que se denominan planificación del proyecto. Esto incluye calcular el costo del proyecto para lo cual es necesario realizar una estimación: del trabajo a ejecutar, de los recursos necesarios y del tiempo que transcurrirá desde el comienzo hasta el final de su realización [6]. En CRISP-DM también se requiere de un proceso de planificación que se encuentra definido dentro de la fase de "Comprensión del Negocio". Sin embargo, CRISP-DM no propone ningún mecanismo para indicar cómo realizar dicha estimación.

Dada las diferencias que existen entre un proyecto convencional de construcción de software y un proyecto de explotación de información, los métodos usuales de estimación no son aplicables ya que los parámetros a ser utilizados son de naturalezas diferentes. Por ejemplo, las herramientas de estimación de software convencional como COCOMO II [7], SLIM [8] o PRICE-S [9] utilizan como parámetros la cantidad de líneas de código, la experiencia del equipo de trabajo, características de la plataforma de desarrollo, entre otras. Sin embargo, en proyectos de explotación de información existen otras características que deben ser consideradas para la estimación que las herramientas de ingeniería en software no consideran, como por ejemplo, cantidad de fuentes de información, nivel de integración de los datos, el tipo de problema a ser resueltos, entre las más representativas de este tipo de proyectos.

FASE	TAREAS COMPONENTES	▪ ACTIVIDADES ASOCIADAS
Comprensión del negocio	Determinar los objetivos de 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 de explotación de información	▪ Las metas del Proyecto de Explotación de Información ▪ Criterios de éxito del Proyecto de Explotación de Información
	Producir el plan del proyecto	▪ Plan de proyecto ▪ Valoración inicial de herramientas
Entendimiento de datos	Recolección inicial de los datos	▪ Reporte de recolección de datos iniciales
	Descripción de los datos	▪ Reporte de descripción de los datos
	Exploración de los datos	▪ Reporte de exploración de datos
	Verificación de calidad de los datos	▪ Reporte de calidad de datos
Preparación de los datos	Tareas preparatorias	▪ Conjunto de Datos ▪ Descripción del Conjunto de Datos
	Seleccionar los datos	▪ Inclusión / exclusión de datos
	Limpieza de datos	▪ Reporte de calidad de datos limpios
	Construcción de datos	▪ Derivación de atributos ▪ Generación de registros
	Integración de los datos	▪ Unificación de datos
	Formato de datos	▪ Reporte de calidad de los datos
Modelado	Selección de la técnica de modelado	▪ La técnica modelada ▪ Supuestos del modelo
	Generación del diseño del ensayo	▪ Plan de pruebas de ensayo
	Construcción del modelo	▪ Configuración de parámetros ▪ Modelo ▪ Descripción del modelo
	Evaluación del modelo	▪ Evaluar el modelo ▪ Revisación de la configuración de parámetros
Evaluación	Evaluar los resultados	▪ Valoración de resultados mineros con respecto al éxito del negocio ▪ Modelos aprobados
	Proceso de revisión	▪ Revisión del proceso
	Determinación de los próximos pasos	▪ Listar posibles acciones
Implantación	Plan de implantación	▪ Plan de implantación
	Plan de vigilancia y mantenimiento	▪ Plan de vigilancia y mantenimiento
	Producción final	▪ Informe final ▪ Presentación Final
	Revisión del proyecto	▪ Documentación de la experiencia

Tabla 1. Tareas de cada fase de la metodología CRISP-DM

Por lo tanto, ha sido necesario desarrollar métodos específicos de estimación de carga de trabajo para proyectos de explotación de información. Entre los métodos existentes, se destacan dos métodos, uno de naturaleza empírica [10] y el otro analítico [11], que serán descriptos a continuación.

Para obtener un modelo empírico de estimación de proyectos software convencional, es necesario primero contar con registros de proyectos concluidos, con sus tamaños estimados en forma temprana y el esfuerzo de desarrollo (sin incluir la puesta en marcha ni su mantenimiento) medido en tiempo / hombre.

Considerando esto, en [10] se busca obtener empíricamente la distribución porcentual de la carga de trabajo en proyectos de explotación de información a través de experimentos utilizando siete proyectos para pequeños y medianos emprendimientos. Los experimentos consistieron en solicitar a varios grupos de estudiantes avanzados en la carrera de Ingeniería en Sistemas de Información el desarrollo de un proyecto de pequeño o mediano tamaño de explotación de información. Cada grupo debía registrar el tiempo utilizado para desarrollar cada tarea de cada subfase de la metodología CRISP-DM. Una vez finalizados los experimentos, los tiempos registrados se integraron para obtener una estimación del porcentaje de tiempo del proyecto de explotación de información que insume la ejecución de cada una de las tareas asociadas a las subfases de la Metodología CRISP-DM. En la tabla 2 se muestran los resultados de porcentaje de tiempo obtenidos para cada una de las fases.

FASE	% de TIEMPO
Fase 1 - Comprensión Del Negocio	20,70
Fase 2 - Entendimiento De Los Datos	10,90
Fase 3 - Preparación De Datos	15,61
Fase 4 - Modelado	34,41
Fase 5 - Evaluación	7,45
Fase 6 - Implantación	10,93

Tabla 2. Carga de trabajo (en % de tiempo) de cada fase de la Metodología CRISP-DM

Nótese que las fases de "Comprensión del Negocio" y "Modelado" insumen más del 50% del tiempo del proyecto, utilizando "Modelado" el 34,41% y "Comprensión del Negocio" el 20,70%. Esto es útil, ya que al conocer el tiempo utilizado en alguna de las subfases, se puede tener una aproximación a los tiempos de las otras subfases y a la estimación global del proyecto.

Por otro lado, los métodos analíticos de estimación de proyectos software convencional utilizan ecuaciones matemáticas para obtener el costo o esfuerzo de desarrollo medido normalmente en meses / hombres. Para realizar este cálculo son utilizados un conjunto de parámetros que describen las características del proyecto a los y que se denominan factores de costo (o cost drivers en inglés).

En este contexto, en [11] se propone un método analítico de estimación para proyectos de explotación de información el cual se denomina Matemático Paramétrico de Estimación para Proyectos de Data Mining (en inglés Data Mining Cost Model, o DMCoMo) [12]. El método DMCoMo es un modelo de estimación de esfuerzo paramétrico de la familia de COCOMO [7] que permite estimar los meses-hombre que serán necesarios para desarrollar un proyecto de explotación de información desde su concepción hasta su puesta en marcha. Para realizar la estimación, DMCoMo define una serie de factores de costo los cuales están vinculados a las características más importantes de los proyectos de explotación de información. Estos se clasifican en seis categorías. En la tabla 3 se describen los factores de costo utilizados por DMCoMo con su categoría.

Una vez que los valores de los factores de costo son definidos, se ingresan estos valores en las ecuaciones suministradas por el método. DMCoMo dispone de dos ecuaciones, una que utiliza 23 factores de costo como variables que puede ser utilizada cuando el proyecto está bien definido y otra de 8 factores de costo como variables que puede utilizarse cuando no todos los datos del proyecto se encuentran definidos. Como resultado de ingresar los valores a la ecuación correspondiente, se obtiene la cantidad de meses-hombre correspondiente al proyecto.

Luego de aplicar DMCoMo en 15 proyectos reales de explotación de información y comparar el esfuerzo estimado obtenido por el método y el esfuerzo real del proyecto. Los resultados se pueden ver en la tabla 4.

CATEGORÍA	GENERADOR DE COSTO
FACTORES DE COSTO RELACIONADOS A LOS DATOS	• CANTIDAD DE TABLAS (NTAB) • CANTIDAD DE TUPLAS DE LAS TABLAS (NTUP) • CANTIDAD DE ATRIBUTOS DE LAS TABLAS (NATR) • GRADO DE DISPERSIÓN DE DATOS (DISP) • PORCENTAJE DE VALORES NULL (PNUL) • GRADO DE DOCUMENACIÓN DE LAS FUENTES DE INFORMACIÓN (DMOD) • GRADO DE INTEGRACIÓN DE DATOS EXTERNOS (DEXT)
FACTORES DE COSTO RELACIONADOS A LOS MODELOS	• CANTIDAD DE MODELOS A SER CREADOS (NMOD) • TIPO DE MODELOS A SER CREADOS (TMOD) • CANTIDAD DE TUPLAS DE LOS MODELOS (MTUP) • CANTIDAD Y TIPO DE ATRIBUTOS POR CADA MODELO (MATR) • CANTIDAD DE TÉCNICAS DISPONIBLES PARA CADA MODELO (MTEC)
FACTORES DE COSTO RELACIONADOS AL DESARROLLO DE LA PLATAFORMA	• CANTIDAD Y TIPO DE FUENTES DE INFORMACIÓN DISPONIBLES (NFUN) • DISTANCIA Y MEDIO DE COMUNICACIÓN ENTRE SERVIDORES DE DATOS (SCOM)
FACTORES DE COSTO RELACIONADOS A LAS TÉCNICAS Y HERRAMIENTAS	• HERRAMIENTAS DISPONIBLES PARA SER USADAS (TOOL) • GRADO DE COMPATIBILIDAD DE LAS HERRAMIENTAS CON OTROS SOFTWARE (COMP) • GRADO DE ENTRENAMIENTO DE LOS USUARIOS DE LAS HERRAMIENTAS (NFOR)
FACTORES DE COSTO RELACIONADOS AL PROYECTO	• CANTIDAD DE DEPARTAMENTOS INVOLUCRADOS EN EL PROYECTO (NDEP) • GRADO DE DOCUMENTACIÓN QUE ES NECESARIO GENERAR (DOCU) • CANTIDAD DE SITIOS DONDE SE REALIZARÁ EL DESARROLLO Y SU GRADO DE COMUNICACIÓN (SITE)
FACTORES DE COSTO RELACIONADOS AL EQUIPO DE TRABAJO	• GRADO DE FAMILIARIDAD CON EL TIPO DE PROBLEMA (MFAM) • GRADO DE CONOCIMIENTO DE LOS DATOS (KDAT) • ACTITUD DE LOS DIRECTIVOS (ADIR)

Tabla 3. Factores de costo utilizados por DMCoMo

Nótese que para la ecuación que utiliza 23 factores de costo como variable se posee una desviación estándar del error de ±16,908 meses/hombre, mientras que para la de 8 factores de costo la desviación estándar del error es de ±23,105 meses/hombre.

	Ecuación con 23 factores de costo	Ecuación con 8 factores de costo
Error Mínimo	-17,559	-23,979
Error Máximo	31,498	50,216
Error Medio	11,097	8,972
Error Medio Absoluto	18,025	20,066
Desviación Estándar del Error	16,908	23,105

Tabla 4. Comparación del esfuerzo estimado y el esfuerzo real

Sin embargo, se debe aclarar que DMCoMo se considera confiable para estimar el esfuerzo de proyectos de explotación de información que se encuentren en el rango de esfuerzo de 90 meses / hombre a 185 meses / hombre. Si el esfuerzo del proyecto se encuentra fuera de este rango, el comportamiento del método es desconocido.

OBJETIVOS DE INVESTIGACIÓN

Dados los dos métodos analizados para la estimación de proyectos de explotación de información, este proyecto de investigación buscará continuar la línea de trabajo comenzado en [10] y [12] para evaluar y comparar los resultados de ambos métodos de estimación en proyectos de emprendimientos de diferentes tamaños y características.

METODOLOGÍA DE TRABAJO

El desarrollo de este proyecto utilizará la metodología propia de la investigación experimental, de generación de experimentos, el estudio de los resultados obtenidos con técnicas de análisis comparativo y de síntesis de comparaciones.

Con esta base:

a) Se producirá un relevamiento de proyectos de explotación de información y sus principales características a ser consideradas por cada proyecto de acuerdo a su tamaño.
b) Se realizará la estimación de proyectos mediante la utilización del método analítico DMCoMo. Al efecto se realizará:
   1. Se desarrollará una herramienta software que tendrá como objetivo generar un banco de pruebas de proyectos de explotación de información con sus características determinadas aleatoria-mente dentro de un marco definido por el usuario (por ejemplo, el tamaño del proyecto) al que le aplicará las fórmulas de estimación del método DMCoMo.
   2. Se utilizará la herramienta software desarrollada para generar el banco de prueba.
   3. Se analizará e integrarán los resultados obtenidos para cada tamaño de proyecto.
c) Se realizará la estimación de proyectos mediante la utilización del método empírico. Al efecto se realizará:
   1. Se entrenará a estudiantes avanzados en la carrera de Ingeniería en Sistemas de Información en la identificación de objetivos de negocio, la identificación de los procesos de explotación de

información asociados y la documentación que se debe desarrollar durante la aplicación de la metodología CRISP-DM. Los alumnos serán divididos en grupos de trabajo con capacidades homogéneas.

2. Se asignará cada uno de los proyectos de explotación de información relevados en el paso (a) a un grupo de los formados en el paso (c.1).
3. Se desarrollará por cada grupo el proyecto de explotación de información con registro del tiempo utilizado para desarrollar cada tarea de cada subfase de la metodología CRISP-DM.
4. Se integrará los resultados del experimento de los tiempos obtenidos por cada grupo.

d) Se analizarán y compararán los resultados obtenidos por ambos métodos para proyectos con características y tamaño similares.
e) Se intentará establecer cuál método es recomendable de acuerdo a las características y tamaño de los proyectos.
f) Se identificará las aportaciones del proyecto para darle difusión mediante comunicaciones a congresos generales o del área de la Explotación de Información.

RESULTADOS OBTENIDOS/ESPERADOS

Como resultado de este proyecto, se espera tener una caracterización de los métodos de estimación estudiados en términos de su confiabilidad dado el tipo de proyecto de explotación de información en el que se está tratando de valorar el esfuerzo demandado.
Se espera determinar para qué tipo de proyectos es utilizable cada método, y en particular, si pueden ser usados complementariamente.

FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo se encuentra formado por dos investigadores formados, tres investigadores en formación y dos alumnos avanzados de las carreras Ingeniería en Sistemas de Información y Licenciatura en Sistemas. En su marco se desarrolla una tesis de Maestría y dos Trabajos de Fin de Carrera.

BIBLIOGRAFÍA

[1] Schiefer, J., , Jeng, J., Kapoor, S., Chowdhary, P. (2004). *Process Information Factory: A Data Management Approach for Enhancing Business Process Intelligence*. Proceedings 2004 IEEE International Conference on E-Commerce Technology. Pág. 162-169.

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

[3] Umapathy, K. (2007). *Towards Co-Design of Business Processes and Information Systems Using Web Services*. Proceedings 40th Annual Hawaii International Conference on System Sciences. Pág. 172-181.

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

[5] Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. 2000. *CRISP-DM 1.0 Step by step BIguide. Edited by SPSS.* http://www.crisp-dm.org/ CRISPWP-0800.pdf. Ultimo acceso 01/06/08.

[6] Pressman, R. 2004. *Software Engineering: A Practitioner's Approach*. Editorial Mc Graw Hill.

[7] B.W. Boehm, C. Abts, A.W. Brown, S. Chulani, B.K. Clark, E. Horowitz, R. Madachy, D. Reifer, B. Steece, *Software Cost Estimation with COCOMO II*, Prentice-Hall, Englewood Cliffs, NJ, 2000.

[8] L.H. Putnam Sr., D.T. Putnam, L.H. Putnam Jr., M.A. Ross, *Software Lifecycle Management (SLIM) Training. SLIM Estimate Exercises with Answers*, Quantitative Software Management, Mc Lean, VA, 2000.

[9] LLC PRICE Systems, *PRICE S Reference Manual Version 3.0*, Lockheed-Martin, 1998.

[10] Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. *Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información.*

[11] Marbán, O. *Modelo Matemático Paramétrico de Estimación para Proyectos de Data Mining (DMCoMo)*, Ph.D. Thesis, Facultad de Informática, Universidad Politécnica de Madrid, June 2003.

[12] Oscar Marbán, Ernestina Menasalvas, Covadonga Fernández-Baizán. *A cost model to estimate the effort of data mining projects (DMCoMo)* Information Systems 33: 133–150