

VIABILIDAD Y ESTIMACIÓN DE PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

Pytel, P., Britos, P., García-Martínez, R.

Grupo Investigación en Sistemas de Información

Departamento de 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 Investigación en Explotación de Información

Sede Andina (El Bolsón)
Universidad Nacional de Río Negro
San Martín esq. Pellegrini (8430) El Bolsón. Río Negro. Argentina.
TE + 54 11 02944 49-8939
paobritos@gmail.com

RESUMEN

La Explotación de Información es la sub-disciplina de la Informática que aporta las herramientas de análisis y síntesis para extraer conocimiento no trivial que se encuentra implícitamente en los datos disponibles. Aunque la Ingeniería en Software provee muchos métodos, técnicas y herramientas para la construcción de software tradicional, estos en general no son aplicables a proyectos de explotación de información por poseer características diferentes. Surge la necesidad del desarrollo y validación de métodos, técnicas y herramientas específicos para las etapas tempranas de estos proyectos, vinculadas con los problemas a resolver y riesgos asociados, así como la obtención de la información necesaria para planificar los recursos y actividades a realizar.

En este contexto, utilizando las metodologías de investigación documental exploratoria, prototipado evolutivo y casos de estudio se plantea el desarrollo en el marco de los proyectos de explotación de información de: un proceso que permita evaluar la viabilidad en sus aspectos multidimensionales, y un proceso que permita estimar los recursos necesarios con especial énfasis en su utilización en el sector de las PyMEs.

Palabras clave: Proyectos de Explotación de Información, Estudio de Viabilidad, Estimación de Esfuerzo, Information Mining.

CONTEXTO

Este proyecto de investigación se enmarca como una 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 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 es la sub-disciplina de la Informática que aporta a la Inteligencia de Negocio [Negash y Gray, 2008] herramientas de análisis y síntesis para extraer conocimiento no trivial que se encuentra implícitamente en los datos disponibles [Schiefer et al., 2004]. Aunque la Ingeniería en Software provee muchos métodos, técnicas y herramientas, estos no son aplicables en general ya que no se ocupan de mismos aspectos prácticos. Por lo tanto, es necesario el desarrollo y la validación de estos elementos que puedan asistir a los practicantes y proveer la necesaria objetividad, generalización y confiabilidad a los Proyectos de Explotación de Información [García-Martínez et al., 2011a]. Si bien existen metodologías para el desarrollo de Proyectos de Explotación de Información (PEI) que se consideran probadas y tienen un buen nivel de madurez, éstas dejan de lado aspectos a nivel gestión de los proyectos y de empresa [Vanrell et al., 2010a; 2010b].

Antes de emprender todo proyecto software, la organización debe decidir si es conveniente realizarlo o no. Para poder tomar esa decisión se necesita conocer el impacto que ese software va

a causar en la organización y sus riesgos asociados [Pressman, 2004]. Para ello es necesario estudiar las características del proyecto a través de una evaluación de la viabilidad técnica del proyecto. Primero se debe recolectar información tanto sobre la organización como del sistema a desarrollar [Sapag, 2000], que luego es procesada para indicar el grado de viabilidad del proyecto.

En los proyectos de construcción de sistemas inteligentes sucede algo similar con la diferencia que la Ingeniería del Conocimiento (INCO) evalúa la viabilidad del proyecto considerando varias dimensiones [López et al., 1991; García-Martínez y Britos, 2004]. Dado las especificaciones iniciales de estos sistemas suelen ser incompletas, imprecisas, inconsistentes y contradictorias, para los proyectos de la INCO se debe considerar si la construcción del sistema es posible, adecuada, justificada y va a tener éxito. En este sentido, la metodología IDEAL [Gómez et al., 1997] propone un método de tipo métrico para medir la viabilidad de un proyecto que incorpora la manipulación de valores lingüísticos a través de su representación mediante intervalos difusos que son procesados y combinados para obtener el valor de viabilidad final del proyecto.

En este contexto, se identifica la necesidad de evaluar la viabilidad de un PEI antes de comenzar. No todos los proyectos finalizan con éxito [Edelstein y Edelstein, 1997; Strand, 2000] y un 60% de los proyectos terminan en fracasos [Gondar, 2005] debido, entre otras razones, a problemas no detectados al comienzo del mismo. Sin embargo, dada que las características de los PEI son diferentes a los proyectos de software tradicional y los de la INCO, no es posible reutilizar los modelos propuestos para estos tipos de proyectos.

Por otro lado, una vez que se considera que un sistema es viable para su construcción, es preciso realizar las actividades de planificación del proyecto. Estas actividades necesitan una estimación del trabajo a ejecutar, de los recursos necesarios y del tiempo que transcurrirá desde el comienzo hasta su fin [Pressman, 2004]. El proceso de estimación está formado proceso continuo en la organización para poder llegar a una predicción fiable que

debe ser usada y consultada a lo largo de todo el proyecto [Agarwal y Kumar, 2001]. Se puede decir que toda técnica de estimación pertenece a una de estas tres categorías [Bielak, 2000]: modelos analíticos, técnicas basadas en teorías y técnicas empíricas.

Los PEI no escapan a la necesidad de estimar el esfuerzo necesario para la construcción del sistema. Pero, dadas las diferencias existentes entre un proyecto de software tradicional y un PEI, los métodos usuales de estimación no son aplicables [Pytel et al., 2011]. Luego de una búsqueda documental se ha encontrado dos métodos específicos de estimación para PEI. El primero de tipo empírico [Rodríguez et al., 2010] utiliza registros de proyectos concluidos con el esfuerzo de desarrollo medido en tiempo/hombre para obtener la distribución porcentual de la carga de trabajo. En cuanto al otro método denominado DMCoMo [Marbán, 2003; Marbán et al., 2003], es un modelo analítico que permite estimar los meses/hombre usando una serie de factores de costo vinculados a las características de los PEI. Sin embargo, como indican sus autores, el método se considera confiable sólo para estimar el esfuerzo de proyectos que se encuentren en el rango de esfuerzo de 90 a 185 meses/hombre. Si el esfuerzo se encuentra fuera de este rango, el comportamiento del método es desconocido.

En este segundo contexto, todavía no se dispone de un proceso de estimación confiable para PEI pequeños que son los que usualmente requieren las Pequeñas y Medianas Empresas (PyMEs). Dado que estas empresas poseen características especiales, se cree necesario disponer de un método de estimación que las habilite a predecir el costo de la implementación de PEI para asistir la toma de decisiones [García-Martínez et al., 2011b].

OBJETIVOS E HIPOTESIS DE INVESTIGACION

Este proyecto se inscribe en una línea de investigación que busca desarrollar procesos específicos para ser utilizados en las primeras etapas de los Proyectos de Explotación de Información.

Entre los supuestos (o hipótesis) que guían el proyecto se encuentran:

Hipótesis I: Los proyectos de Explotación de Información poseen características muy distintas a las de los proyectos de desarrollo de software tradicionales y de los sistemas inteligentes. En consecuencia, los modelos para el estudio de viabilidad y estimación de recursos existentes en estos sistemas no pueden ser aplicables a proyectos de explotación de información.

Hipótesis II: Los riesgos y problemas encontrados durante el desarrollo de un proyecto de explotación de información pueden ser reducidos mediante una evaluación temprana de las características del proyecto.

Hipótesis III: Es posible disponer de un proceso de estimación confiable y robusto para proyectos de explotación de información con las características que son normalmente requeridas por las PyMEs.

Objetivo General: En el marco de proyectos de explotación de información, este proyecto se enfocará a establecer: [a] un proceso que permita evaluar la viabilidad en sus aspectos multidimensionales; y [b] un proceso que permita estimar los recursos necesarios con particular énfasis en su utilización en PyMEs.

Objetivos específicos vinculados a Hipótesis I:

1. Identificar los principales procesos para estudios de viabilidad para proyectos de la Ingeniería en Software e Ingeniería del Conocimiento.
2. Identificar los principales procesos de estimación para proyectos de la Ingeniería en Software e Ingeniería del Conocimiento.
3. Identificar las principales características de los proyectos de explotación de información que deben ser consideradas en ambos procesos.

Objetivos específicos vinculados a Hipótesis II:

4. Desarrollar un proceso para evaluar la viabilidad a través de la ponderación de las características y dimensiones de un proyecto de explotación de información.

Objetivos específicos vinculados a Hipótesis III:

5. Desarrollar un proceso para estimar los recursos necesarios para desarrollar un proyecto de explotación de información en el ámbito de las PyMEs.

FUNDAMENTACION E IMPORTANCIA

a) *Relevancia científica:*

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 los procesos de las fases iniciales planteadas en las metodologías identificadas. El proyecto promueve el desarrollo y la validación de procesos iniciales con sus respectivos métodos, técnicas y herramientas, conllevando a una mejora en el campo de la Ingeniería de Explotación de Información. 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 la definición de técnicas consistentes.

b) *Relevancia 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 proveer información cualitativa que contribuya a la toma de decisiones en los niveles de gestión de la industria y el comercio regional.

c) *Relevancia Educativa:*

Se prevé la propuesta de una asignatura optativa en el 5to año de la Licenciatura en Sistemas de la UNLa denominada "Tecnologías de Explotación de Información". Además, los resultados parciales y finales también tendrán un impacto sobre la actualización de los contenidos de la asignatura "Ingeniería de Software III" de 4to año de la Licenciatura en Sistemas de la UNLa.

METODOLOGÍA DE TRABAJO

Para realizar construir el conocimiento asociado al presente proyecto de investigación, se seguirá un enfoque de investigación clásico identificando métodos y materiales necesarios para desarrollar el proyecto:

Las tareas a realizar para desarrollar los procesos de evaluación de la viabilidad y de estimación para proyectos de explotación de información serán las siguientes:

- [i] Se realizará una investigación documental exploratoria tratando de identificar conceptos de interés para el establecimiento de la viabilidad de un proyecto de explotación de información, tratando de identificar posibles dimensiones y características asociadas.
- [ii] Se buscará identificar relaciones existentes entre las diversas características y dimensiones a ponderar.
- [iii] Se desarrollará mediante la metodología de prototipado evolutivo un Modelo de Evaluación Viabilidad de Proyectos de Explotación de Información que vincule y pondere características y dimensiones identificadas previamente.
- [iv] Se desarrollarán los instrumentos de interpretación correspondientes al Modelo de Evaluación Viabilidad de Proyectos de Explotación de Información obtenido.
- [v] Se contrastará los resultados del Modelo de Evaluación de Viabilidad de Proyectos de Explotación de Información con datos pertenecientes a casos reales de PEI realizados en el Grupo de Investigación en Sistemas de Información.
- [vi] Se realizará una investigación documental exploratoria tratando de identificar elementos de interés para la estimación de un proyecto de explotación de información.
- [vii] Se desarrollará mediante la metodología de prototipado evolutivo un Modelo de Estimación de Proyectos de Explotación de Información.
- [viii] Se desarrollarán los instrumentos de interpretación correspondientes al Modelo de Estimación de Proyectos de Explotación obtenido.
- [ix] Se realizarán pruebas de concepto y simulación de Montecarlo para validar los modelos desarrollados.
- [x] Se contrastará los resultados del Modelo de Estimación de Proyectos de Explotación de Información con datos pertenecientes a casos reales de PEI realizados en el Grupo de Investigación en Sistemas de Información.

Para el desarrollo de las tareas propuestas se utilizarán los siguientes materiales:

- Metodologías CRISP-DM [Chapman, et al., 2000], P3TQ [Pyle, 2003] y SEMMA [SAS, 2008] y trabajos previos de miembros del Grupo de Investigación en Sistemas de Información [Rodríguez et al., 2010, García-Martínez et al., 2011a; 2011b] para identificar los conceptos necesarios a los proyectos de explotación de información.
- Métodos de estimación usuales en Ingeniería del Software [Putman, 1978; Boehm, 1981a; 1981b; Albrecht, 1983; Fairley, 1992; Boehm et al., 1995; Boehm et al., 2000; Putnam et al., 2000].
- Métodos de evaluación de viabilidad usuales en Ingeniería de Conocimiento [Gómez et al., 1997; García-Martínez y Britos, 2004].
- Reportes de datos de proyectos de explotación de información reales recolectados por miembros del Grupo de Investigación en Sistemas de Información (GISI) del Departamento de Desarrollo Productivo y Tecnológico (Universidad Nacional de Lanús).
- Repositorios Científicos accesibles por Internet.
- Acceso a bibliotecas digitales de IEEE, ACM y SCOPUS.

RESULTADOS OBTENIDOS/ESPERADOS

Como resultado de este proyecto, se esperan los siguientes aportes:

- Definición de un proceso que permita evaluar la viabilidad de un proyecto de explotación de información.
- Definición de un proceso que permita estimar los recursos necesarios para desarrollar un proyecto de explotación de información.

FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo se encuentra formado por un investigador formado, un investigador en formación y dos alumnos avanzados de la carrera Licenciatura en Sistemas de la UNLa. En el marco de este proyecto se desarrolla una Tesis de Doctorado y dos Trabajos de Fin de Carrera.

REFERENCIAS

- Agarwal, R. y Kumar, M. (2001) Estimating software projects. *Software engineering Notes*, 26(4), pp. 60-67.
- Albrecht, A. (1983) Software Function, Source Lines of Code, and Development Effort Prediction: A Software Science Validation. *IEEE Trans. Software Eng.*, 9(6).
- Bielak, J. (2000) Improving Size Estimates Using Historical Data. *IEEE Software*, 17(6): 27-35.
- Boehm, B. (1981a). *Software Engineering Economics*. Editorial Prentice Hall.
- Boehm, B. (1981b). *Software Engineering Economics*. *IEEE Transactions on Software Engineering*, 10(1):4-21.
- Boehm, B., Clark, B., Horowitz, E., Westland, C., Madachy, R., Selby, R. (1995) Cost Models for Future Software Life Cycle Process: COCOMO 2.0. *Ann. of Software Eng. Special Volume on Software Process and Product Measurement*, J.D. Arther and S.M. Henry, eds., Science Publishers, Amsterdam, The Netherlands, Vol. 1, pp. 45-60.
- Boehm, B., Abts, C., Brown, A.W., Chulani, S., Clark, B.K., Horowitz, E., Madachy, R., Reifer, D., Steece, B. (2000) *Software Cost Estimation with COCOMO II*. Prentice-Hall, Englewood Cliffs, NJ.
- 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. <http://www.crisp-dm.org/CRISPWP-0800.pdf> Acceso Marzo 2008.
- Edelstein, H. A., Edelstein, H. C. (1997) *Building, Using, and Managing the Data Warehouse*. Data Warehousing Institute. Prentice Hall PTR, 1st edition.
- Fairley, R. E. (1992) Recent Advances in Software Estimation Techniques. *Proc. 14th Int'l Conf. Software Eng.*, ACM Press, New York.
- García Martínez, R. y Britos, P. (2004). *Ingeniería de Sistemas Expertos Nueva Librería*. ISBN 987-1104-15-4.
- García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo-Cattaneo, F., Rodríguez, D., Pytel, P., Vanrell, J. (2011a). Towards an Information Mining Engineering. En *Software Engineering, Methods, Modeling and Teaching* (Editor: Carlos Zapata Jaramillo). Sello Editorial Universidad de Medellín. ISBN 978-958-8692-32-6. Páginas 83-99.
- García-Martínez, R., Lelli, R., Merlino, H., Cornachia, L., Rodríguez, D., Pytel, P., Arboleya, H. (2011b). *Ingeniería de Proyectos de Explotación de Información para PYMES. Proceedings XIII Workshop de Investigadores en Ciencias de la Computación*. Pág. 253-257. ISBN 978-950-673-892-1.
- Gómez, A., Juristo, N., Montes, C., Pazos, J. (1997) *Ingeniería del Conocimiento*. Editorial CERA, Madrid.
- Gondar, J. E. (2005) *Metodología Del Data Mining*. Number 84-96272-21-4. Data Mining Institute, S.L.
- Kalos, M.H.; Whitlock P.A. (1986) *Monte Carlo Methods*. Vol I . Basics. John Wiley & Sons. New York.
- López, D., García-Martínez, R., Marsiglio, A. (1991) Factibilidad de Construcción de Sistemas Basados en el Conocimiento. *Anales del II Simposio de Inteligencia Artificial y Robótica*. Páginas 63-73. UNLu, Buenos Aires.
- Marbán, O. (2003) *Modelo Matemático Paramétrico de Estimación para Proyectos de Data Mining (DMCoMo)*. Ph.D. Thesis, Facultad de Informática, UPM.
- Marbán, O., Menasalvas, E., Fernández-Baizán, C. (2003) A cost model to estimate the effort of data mining projects (DMCoMo). *Information Systems* 33: 133-150.
- Negash, S. y Gray, P. (2008). *Business Intelligence*. In *Handbook on Decision Support Systems 2*, ed.eds. F. Burstein y C. Holsapple (Springer), Pág. 175-193.
- Pressman, R. (2004). *Software Engineering: A Practitioner's Approach*. Editorial Mc Graw Hill.
- Putnam, L.H. (1978) A General Empirical Solution to the Macro Software Sizing and Estimating Problem. *IEEE Trans. Software Eng.*, Apr. 1978, pp. 345-361.
- Putnam Sr., L.H., D.T. Putnam, D.T., Putnam Jr., L.H., Ross, M.A. (2000). *Software Lifecycle Management (SLIM) Training. SLIM Estimate Exercises with Answers, Quantitative Software Management*, Mc Lean, VA.
- Pyle, D. (2003) *Business Modeling and Business intelligence*. Morgan Kaufmann
- Pytel, P., Tomasello, M., Rodríguez, D., Arboleya, H., Pollo-Cattaneo, M., Britos, P., García-Martínez, R. (2011). *Estimación de Proyectos de Explotación de Información. Estudio Comparado de Modelos Analíticos y Empíricos. Proceedings XIII Workshop de Investigadores en Ciencias de la Computación*. Pág. 295-299. ISBN 978-950-673-892-1.
- 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*. Pág. 664-673. ISBN 978-950-9474-49-9.
- Sapag, N. (2000) *Preparación y evaluación de proyectos*. 4ª Edición. McGraw Hill.
- Strand, M. (2000) *The Business Value of Data Warehouses - Opportunities, Pitfalls and Future Directions*. Phd. Thesis, Department of Computer Science, Univ. of Skövde.
- SAS (2008) *SAS Enterprise Miner: SEMMA* <http://www.sas.com/technologies/analytics/datamining/miner/semma.html>.
- Schiefer, J., Jeng, J., Kapoor, S. y 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.Pp. 162-169.
- Vanrell, J., Bertone, R., García-Martínez, R. (2010a). *Modelo de Proceso de Operación para Proyectos de Explotación de Información*. *Anales del XVI CACIC*. Pp. 674-682. ISBN 978-950-9474-49-9.
- Vanrell, J., Bertone, R., García-Martínez, R. (2010b). *Un Modelo de Procesos de Explotación de Información*. *Proce. XII WICC* Pp. 167-171. ISBN 978-950-34-0652-6.