Estudio comparativo de metodologías para minería de datos

Ing. Juan Miguel Moine
Dra. Ana Silvia Haedo
Dra. Silvia Gordillo

Grupo de investigación en Minería de Datos, UTN Rosario
Facultad de Ciencias Exactas, Universidad Nacional de Buenos Aires
Facultad de Informática, Universidad Nacional de La Plata

juanmiguelmoine@gmail.com, ahaedo@dc.uba.ar, gordillo@lifia.info.unlp.edu.ar

Resumen

La sistematización del proceso de minería de datos es un punto importante para la planificación y ejecución de este tipo de proyecto. Algunas organizaciones implementan el proceso KDD, mientras que otras aplican un estándar más específico como CRISP-DM. Si la organización ha adquirido productos de la empresa SAS, tiene a su disposición una metodología especialmente desarrollada para los mismos, la metodología SEMMA. Por otro lado, la metodología Catalyst (conocida como P3TQ) está ganando cada vez mayor popularidad debido a su completitud y flexibilidad para adaptarse en distintos escenarios.

En este trabajo de investigación se realizará un estudio comparativo entre las distintas metodologías vigentes para proyectos de minería de datos, evaluando las ventajas y desventajas de las mismas en un escenario donde el proyecto tiene como objetivo colaborar a la solución de un problema organizacional.

Palabras clave: *Minería de datos, gestión de proyectos, Knowledge Discovery in Databases, explotación de información, CRISP-DM, SEMMA, Catalyst, P3TQ, metodologías en minería de datos.*

Contexto

Este trabajo se desarrolla en el marco del Proyecto de Investigación y Desarrollo "Análisis comparativo de metodologías para la gestión de proyectos en minería de datos" de la Universidad Tecnológica Nacional, Facultad Regional Rosario.

Introducción

La minería de datos es una disciplina que ha crecido enormemente en los últimos años. Las organizaciones han comprendido que los grandes volúmenes de datos que residen en sus sistemas pueden ser analizados y explotados para obtener nuevo conocimiento a partir de los mismos.

Minería de Datos o Explotación de Información, es el proceso de extraer conocimiento útil, comprensible y novedoso de grandes volúmenes de datos, siendo su principal objetivo encontrar información oculta o implícita, que no es posible obtener mediante métodos estadísticos convencionales. La entrada al proceso de minería está formada generalmente por registros provenientes de bases de datos operacionales o bien bodegas de datos (Datawarehouse).

Los proyectos de explotación de información pueden ser llevados a cabo en distintos escenarios. Según el punto de partida del proceso, es posible clasificarlos en:

- *Escenarios donde se aborda desde la minería de datos una situación organizacional* (un problema o una oportunidad), buscando patrones y relaciones que puedan colaborar con la misma. Este escenario es el más frecuente en el ámbito de las empresas y organizaciones.

- *Escenarios donde el proyecto comienza con un conjunto de datos* y el objetivo es explorarlos para encontrar relaciones interesantes que puedan ser útiles en el dominio de aplicación. En estos casos, algunos autores como Pyle[1], no recomiendan trabajar directamente con los datos sin establecer de antemano la problemática que se aborda, el personal involucrado y las expectativas y necesidades de los usuarios. Este punto resulta de gran importancia para justificar la realización del proyecto, ya que ninguna organización adquirirá una herramienta si no sabe la función que cumplirá.

Los esfuerzos en el área de la minería de datos se han centrado en su gran mayoría en la investigación de técnicas para la explotación de información y extracción de patrones (tales como árboles de decisión, análisis de conglomerados y reglas de asociación). Sin embargo, se ha profundizado en menor medida el hecho de cómo ejecutar este proceso hasta obtener el "nuevo conocimiento", es decir, en las metodologías. Las metodologías permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Ayudan a las organizaciones a entender el proceso de descubrimiento de conocimiento y proveen una guía para la planificación y ejecución de los proyectos.

Algunos modelos conocidos como metodologías son en realidad un modelo de proceso: un conjunto de actividades y tareas organizadas para llevar a cabo un trabajo. La diferencia fundamental entre metodología y modelo de proceso radica en que el modelo de proceso establece qué hacer, y la metodología especifica cómo hacerlo. Una metodología no solo define las fases de un proceso sino también las tareas que deberían realizarse y cómo llevar a cabo las mismas.

En los inicios del año 1996, el modelo KDD (Knowledge Discovery in Databases)[2] constituyó el primer modelo aceptado en la comunidad científica que estableció las etapas principales de un proyecto de explotación de información. Formalmente el modelo establece que la minería de datos es la etapa dentro del proceso en la cual se realiza la extracción de patrones a partir de los datos. Sin embargo actualmente, en la comunidad científica y en la literatura, el término KDD y minería de datos se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento de conocimiento.

A partir del año 2000, con el gran crecimiento que surgió en el área de la minería de datos, surgen tres nuevos modelos que plantean un enfoque sistemático para llevar a cabo el proceso [3]: SEMMA, Catalyst (conocida como P3TQ) y CRISP-DM. Como se puede observar en la Figura 1, CRISP-DM se ha convertido en la metodología más utilizada, según un estudio publicado en el año 2007 por la comunidad KDnuggets (Data Mining Community's Top Resource).

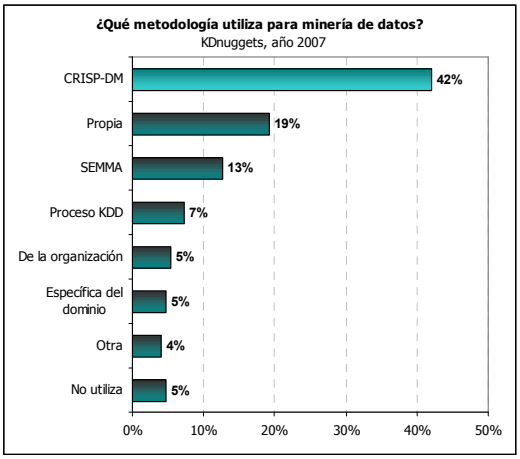

Fig. 1. Encuesta realizada por la KDnuggets en el año 2007

Algunos modelos profundizan en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de minería de datos (como CRISP-DM), mientras que otros proveen sólo una guía general del trabajo a realizar en cada fase (como el proceso KDD o SEMMA).

SEMMA, creada por el SAS Institute, se define como "el proceso de selección, exploración y modelado de grandes

volúmenes de datos para descubrir patrones de negocio desconocidos" [4]. El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Valoración).

La metodología SEMMA se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando. Fue propuesta especialmente para trabajar con el software de minería de datos de la compañía SAS. Este producto organiza sus herramientas (llamadas "nodos") en base a las distintas fases que componen la metodología. Es decir, el software proporciona un conjunto de herramientas especiales para la etapa de muestreo, otras para la etapa de exploración, y así sucesivamente. Sin embargo, el usuario podría hacer uso del mismo siguiendo cualquier otra metodología de minería de datos (como CRISP-DM por ejemplo).

La metodología Catalyst [1], conocida como P3TQ (Product, Place, Price, Time, Quantity), fue propuesta por Dorian Pyle en el año 2003. Esta metodología plantea la formulación de dos modelos: el Modelo de Negocio y el Modelo de Explotación de Información.

El Modelo de Negocio (MII), proporciona una guía de pasos para identificar un problema de negocio (o la oportunidad del mismo) y los requerimientos reales de la organización. Contempla diferentes ámbitos para el proyecto de minería de datos, explicitando acciones específicas según el escenario desde el cual se parte. Para proyectos donde el problema u oportunidad de negocio no está definido, se recomienda comenzar analizando las relaciones P3TQ que existen en la cadena de valor organizacional, es decir, aquellas relaciones precio/lugar/producto/tiempo/cantidad que son importantes para la empresa.

El Modelo de Explotación de Información (MIII), proporciona una guía pasos para la construcción y ejecución de modelos de minería de datos a partir del Modelo de Negocio (MII).

El foco que la metodología Catalyst propone en su Modelo de Negocio sobre la cadena de valor organizacional, hizo que sea difundida en la comunidad científica como metodología "P3TQ", aunque ésta no sea su denominación original.

La metodología Catalyst, en sus dos modelos, está compuesta por una serie de pasos llamados "boxes". El concepto es que luego de llevar a cabo una acción, se deben evaluar los resultados y determinar cuál es el próximo paso (box) a seguir. La secuencia y la interacción entre los distintos pasos permiten una flexibilidad muy grande, y una amplia variedad de caminos posibles.

CRISP–DM, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining. Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación [5]. La sucesión de fases, no es necesariamente rígida. Cada fase es descompuesta en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, pero en ningún momento se propone como realizarlas. Es decir, CRISP-DM establece un conjunto de tareas y actividades para cada fase del proyecto pero no especifica cómo llevarlas a cabo.

Líneas de investigación/desarrollo

En el marco de este proyecto se investigará:

- Las distintas metodologías y modelos de proceso vigentes para proyectos de minería de datos.
- Las similitudes y diferencias entre cada modelo. Se tendrán en cuenta no sólo las etapas que los componen, sino también aspectos clave para la gestión de

proyectos, como gestión del tiempo, gestión del riesgo y costos.

- Ventajas y desventajas de cada metodología en un escenario de aplicación. El caso particular que se estudiará será aquel donde se comienza con un problema de negocio a partir del cual el proyecto de minería de datos tiene el objetivo de encontrar patrones y relaciones que aporten nuevo conocimiento para la solución del mismo.

Resultados y Objetivos

En la actualidad, son escasos y poco difundidos los estudios que comparan los modelos mencionados, enfocados en aspectos principalmente descriptivos (comparación de las fases que los componen) y no en un estudio comprensivo-comparativo, que contemple aspectos tales como:

- Grado en el que se incorporan actividades para la gestión del proyecto (como gestión del riesgo, de costos, de Recursos Humanos).
- Nivel de detalle de las tareas que componen cada fase, abriendo una discusión sobre qué modelos pueden ser realmente considerados una metodología.
- Viabilidad de cada modelo para la aplicación en diferentes escenarios (ya sea partiendo de un conjunto de datos o abordando una situación o problema organizacional).

Como objetivo de este trabajo se pretende la construcción de un marco comparativo que permita confrontar los distintos modelos, y evaluar la adecuación de los mismos en escenarios donde el proyecto de minería de datos tiene por objetivo colaborar en la solución de un problema organizacional.

Formación de los Recursos Humanos

En el marco de este proyecto de investigación se está realizando una tesis de Maestría en Ingeniería de Software en la Universidad Nacional de La Plata, por medio del Programa de Becas de Posgrado de la Universidad Tecnológica Nacional.

Referencias

1. Pyle, Dorian (2003). “Business Modeling and Data Mining”. Morgan Kaufmann Publishers.
2. Fayyad, Usama (1996). “Advances in Knowledge Discovery and Data Mining”. MIT Press.
3. Britos Paola (2008). “Procesos de explotación de información basados en sistemas inteligentes”. Universidad Nacional de La Plata, Argentina.
4. SAS Institute. “Data Mining and the Case for Sampling” (www.sasenterpriseminer.com/documents /SAS-SEMMA.pdf). Último acceso Julio 2010.
5. Chapman, P., Clinton, J., Keber y otros (2000). “CRISP-DM 1.0 Step by step guide”. SPSS (www.crisp-dm.org/CRISPWP-0800.pdf). Último acceso Julio 2010.
6. Fayyad, Usama y otros, (1996). “The KDD process for extracting useful knowledge from volumes of data”. ACM vol. 39 (11).
7. Azevedo Ana (2008). "KDD, SEMMA AND CRISP-DM: a parallel overview". AIDIS 2008.
8. Pollo-Cattaneo F. y otros (2010). “Ingeniería de Proyectos de explotación de información”. WICC 2010. ISBN 978-950-34-0652-6
9. Mariscal Gonzalo y otros (2010). “A survey of data mining and knowledge discovery process models and methodologies”. The Knowledge Engineering Review, Vol. 25:2, 137–166.