

Síntesis Tesis

TESIS DE MAGÍSTER EN INGENIERIA DE SOFTWARE

“Metodología para el descubrimiento de conocimiento en bases de datos: un estudio comparativo”

Juan Miguel Moine

Directora: Dra. Ana Silvia Haedo
CoDirectora: Dra. Silvia Gordillo.

Motivación

Para llevar a cabo en forma sistemática el proceso de descubrimiento de conocimiento en bases de datos, conocido como minería de datos, es necesaria la implementación de una metodología. Actualmente las metodologías para minería de datos se encuentran en etapas tempranas de madurez, aunque algunas como CRISP-DM ya están siendo utilizadas exitosamente por los equipos de trabajo para la gestión de sus proyectos. En este trabajo se establece un análisis comparativo entre las metodologías de minería de datos más difundidas en la actualidad. Para lograr dicha tarea, y como aporte de esta tesis, se ha propuesto un marco comparativo que explicita las características que se deberían tener en cuenta al momento de efectuar esta confrontación.

El objetivo principal de este trabajo de tesis es la construcción de un marco comparativo que sirva como herramienta para poder evaluar y confrontar diferentes metodologías de minería de datos. El objetivo secundario de esta investigación es el análisis mediante un estudio descriptivo-comparativo de las metodologías más difundidas en la actualidad para proyectos de minería de datos, abriendo una discusión sobre qué enfoques deberían ser realmente considerados una metodología.

Aportes de la tesis

En este trabajo de tesis se ha logrado la construcción de un marco comparativo como herramienta para la confrontación de metodologías de minería de datos. El marco propuesto incluye cuatro aspectos donde se analiza el nivel de especificación de las tareas, los escenarios de aplicación, las actividades que componen cada fase del proceso y las actividades destinadas a la dirección del proyecto. Para cada uno de estos aspectos se propone la evaluación de un conjunto de características que deberían estar presentes en una metodología de minería de datos bien definida. El marco comparativo ha sido utilizado para confrontar las metodologías CRISP-DM y Catalyst. Aunque ambos enfoques se encuentran actualmente en etapas tempranas de madurez, han logrado cumplir un gran porcentaje de las características evaluadas. Sin embargo, durante el estudio también se han evidenciado los puntos que se deberían mejorar y seguir desarrollando, como las actividades destinadas a la dirección del proyecto. Si bien en este trabajo se han analizado los cuatro aspectos en función de la proporción de características que se cumplen en cada uno, los resultados podrían ser diferentes si se trabajara con puntajes. Si el usuario del marco comparativo lo considera necesario, puede efectuar dicha valoración según su criterio. Se espera que el resultado de esta tesis sirva como herramienta para que los equipos de trabajo puedan evaluar metodologías de minería de datos, considerando además la posibilidad de complementar las mismas con los conceptos que no estén presentes o bien utilizar el marco como base para construir una propia.

Líneas de I/D futuras

Como línea de trabajo futuro se podría establecer una ponderación de las características evaluadas mediante un sistema de puntajes que permita obtener resultados cuantitativos respecto al desempeño de cada metodología. El desafío de esta investigación radicará en los criterios que se deberían tener en cuenta para establecer estos valores en forma objetiva. Otra línea de investigación podría ser la ampliación del marco comparativo, incorporando nuevos aspectos al mismo.