ProDiGen: minando modelos completos, precisos y simples con un algoritmo genético

Borja Vázquez-Barreiros, Manuel Mucientes, and Manuel Lama

Centro de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela
{borja.vazquez,manuel.mucientes,manuel.lama}@usc.es

Resumen

Un proceso se puede entender como una secuencia de tareas que se llevan a cabo para alcanzar un determinado objetivo. Por ejemplo, en educación, el diseño de aprendizaje es un proceso en el que los alumnos deben realizar una secuencia de actividades —escribir en el foro, hacer un examen, etc.— para poder lograr los objetivos pedagógicos del curso. En general, estos procesos están perfectamente detallados, sin embargo, incluso en estas situaciones, pueden existir diferencias entre lo que está sucediendo en el proceso, y lo que se cree que está a suceder en realidad. Por ejemplo, siguiengo el ejemplo del dominio educativo, los alumnos pueden realizar trabajos adicionales, como puede ser revisar la bibliografía o interactuar entre ellos.

Es así como el descubrimiento de procesos es necesario para obtener información de *qué es lo que está sucediendo en realidad* durante la ejecución del proceso, y *no lo que creemos que está a suceder*. Típicamente, estas técnicas trabajan sobre registros que contienen la información sobre los eventos detectados y almacenados por el sistema de información donde tuvo lugar el proceso. El descubrimiento de procesos tiene como objetivo obtener el flujo de trabajo que mejor representa el comportamiento almacenado en dicho registro. En la última década se han desarrollado decenas de algoritmos que abordan esta problemática de descubrimiento, sin embargo, las técnicas actuales o bien generan modelos difíciles de leer, modelos que no son capaces de representar todo el comportamiento del registro, o bien modelos que no permiten hacer frente a todas las estructuras de control al mismo tiempo.

Este artículo describe ProDiGen (Process Discovery through a Genetic algorithm), un algoritmo de descubrimiento de flujos de trabajo que guía su búsqueda en torno a modelos completos, precisos y simples. ProDiGen se basa en una función de fitness jerárquica que tiene en cuenta la completitud, la precisión y la simpicidad, utilizando dos métricas nuevas para los dos últimos criterios. Además, utiliza heurísticas para optimizar tanto el cruce —teniendo en cuenta los errores del modelo minado— como la mutación —guianda por las dependencias causales del log. ProDiGen se ha validado con 111 registros y se ha comparado con cuatro algoritmos del estado del arte, validándo los resultados con test estadísticos no paramétricos. Los resultados muestran una mejora significativa de ProDiGen respecto al resto de algoritmos utilizados en la comparativa.