Minería de datos para la Ingeniería del Software

Estamos acostumbrados a escuchar los términos "*data mining*", "*knowledge mining*", "minería de datos" referidos principalmente a las aplicaciones que se realizan en los ámbitos de investigación médica o de entornos financieros. Sin embargo, desde hace ya mucho tiempo se ha trabajado con datos numéricos en el ámbito de la ingeniería del software, especialmente en los aspectos de gestión y planificación. También sabemos que la limitación en la disponibilidad de datos fiables es uno de los principales escollos para la utilización de datos cuantitativos. La realidad actual es que todavía seguimos sin utilizar datos cuantitativos de manera sistemática en muchos aspectos de gestión del software. Mencionamos a continuación algunas referencias importantes a tener en cuenta para quien desee profundizar en esta línea.

Herramientas para la minería de datos

- **Weka** es la herramienta de minería de datos en código abierto más popular. Se compone de tres herramientas: a) Explorer para "probar" rápidamente diferentes algoritmos, preprocesado, selección de atributos y visualización; b) Experimenter para ejecutar múltiples algoritmos y análisis de bases de datos; y c) KnowledgeFlow para el diseño visual de experimentos.
- **Rapidminer** es otra herramienta abierta de minería con un intuitivo y potente interfaz gráfico. Ha sido desarrollada en Java e incluye además Weka.
- **Orange** es otra herramienta de minería de datos, pero en este caso está implementada en Python.
- **R** es principalmente una herramienta estadística pero existen numerosos "*plug-ins*" de minería de datos incluyendo uno para Weka (RWeka). Además existe un entorno gráfico llamado Rattle que facilita la aplicación de ciertos algoritmos de minería.
- Dentro del ámbito comercial nos encontramos con **Clementine**. También dentro del ámbito comercial podemos utilizar las librerías que se pueden encontrar en las herramientas clásicas como **Matlab, Mathematica** para algunos aspectos del análisis de datos.
- **KNIME** es otra herramienta abierta de minería de datos basada en la plataforma Eclipse.
- Además, existen diversas de herramientas para la extracción de datos de repositorios software como paso previo a la aplicación de minería de datos. Una relación de las mismas la podemos encontrar en <http://tools.libresoft.es/>.

Bases de datos disponibles

- **PROMISE** (*Predictor Models in Software Engineering*, <http://promisedata.org/>). Este repositorio contiene múltiples bases de datos clasificadas según su propósito (estimación de defectos, estimación de costes, etc.) junto con los artículos que las han usado. Está relacionado con la conferencia que lleva el mismo nombre.
- **ISBSG**: Es una base de datos clásica, de pago. La experiencia nos ha demostrado que es muy difícil aplicar algoritmos debido a la heterogeneidad de sus datos y a los diversos enfoques de "puntos de función" utilizados, lo que ha resultado en una práctica inutilidad en lo referente a inferencia sobre los datos <http://www.isbsg.org/>.
- **FLOSS** (*Free/libre Open Source Software*) Metrics. <http://www.flossmetrics.org/>. El objetivo de este repositorio es disponer de una gran base de datos con información y métricas de cientos de proyectos software "*Open Source*".
- **Ultimate Debian Database**. Esta base de datos recoge información sobre todos los aspectos de desarrollo de Debian, incluyendo los fuentes de Ubuntu, defectos, migraciones, etc. <http://udd.debian.org/>.

Congresos relacionados

La lista de eventos donde se pueden encontrar ponencias sobre este tema es muy extensa. Mencionamos a continuación algunos específicamente orientados a este tema:

- **PROMISE**. *Predictive Models in Software Engineering*. Conferencia dedicada exclusivamente a los modelos de predicción y análisis de datos <http://promisedata.org/2010/>.
- **MSR** (*Mining Software Repositories*). El mismo título de la conferencia indica que abarca la minería sobre cualquier tipo de datos relacionados con el software. <http://www.msrconf.org>.
- **SCAM** (*Source Code Analysis and Manipulation*). Aunque esta conferencia se centra en la manipulación de código, también incluye entre sus tópicos la minería de datos <http://www2010.ieee-scam.org/>.
- **ESEM** (*Empirical Software Engineering*). Se centra en los estudios empíricos y métricas del software (anteriormente se denominaba *Software Metrics*) pero también se presentan artículos de minería de datos aplicados a proyectos software <http://esem2010.case.unibz.it/>.

Revistas Especializadas

Existe un buen conjunto de revistas internacionales donde se publican resultados de minería de datos en Ingeniería de Software, como son *Empirical Software Engineering Journal*, *Journal of Systems and Software*, *Information and Software Technology*, *IEEE Trans. on Software Engineering*.

Daniel Rodríguez García (Universidad de Alcalá, futuro colaborador estable de ***Novática***), ***Luis Fernández Sanz y Javier Dolado Cosín*** (Coordinadores de la Sección Técnica "*Ingeniería del Software*")

Presunción de inocencia

Hace unos 39 siglos se esculpió el "Código de Hammurabi" (conservado en el Louvre), que es considerado como la primera codificación conocida de derechos y obligaciones humanos. Como referencia, quizá más familiar, Moisés bajó del Sinaí, con las Tablas de la Ley, 5 ó 6 siglos después.

Hace unos 75 años, Isaac Asimov acuñó las "Tres Leyes de la Robótica" (luego evolucionadas a la Roboética) codificando derechos y obligaciones de robots, demonios y otro software de Inteligencia Artificial.

En 2002, Rodney Brooks, director del *MIT Artificial Intelligence Laboratory* pro-nosticaba que (igual que históricamente se han ido reconociendo ciertos derechos a muchos animales y, sobre todo a las mascotas), es plausible que surjan corrientes de reconocimiento de derechos a algunas de esas máquinas, sobre todo a las más antropomorfas y a las que "convivan" en nuestros hogares (robots domésticos).

Aunque los profesionales de los SI (sistemas de información) y las TIC (tecnologías de la información y las comunicaciones) estamos sujetos (por convicción y/o adhesión a un código ético profesional) a ciertos compromisos respecto a las SITIC, eso no obliga en general a otras personas. Y eso permite un pernicioso, generalizado y continuado linchamiento impune de los SITIC, e indirectamente de sus profesionales. Pernicioso, porque nos impone un sambenito a los profesionales; Pero sobre todo porque puede desembocar en un fallo sistémico. Me explicaré.

Pero antes, descendamos a lo concreto. Creo que las siguientes referencias son autoexplicativas y se aceptarán como meros botones de muestra de algo mucho más generalizado (las negritas en la cita han sido introducidas ad-hoc para una mejor exposición).

"*La juez sustituye la fianza de un millón de euros al alcalde de Seseña por otra de 10.000. Emite un auto de rectificación para aclarar un* ***error informático*** *debido a introducir más dígitos de los que procedían*"[1].