aprenderaprogramar.com

Minería de datos (data mining). Qué es y para qué sirve. (2ª parte) (DV00106A)

Sección: Divulgación

Categoría: Tendencias en programación

Fecha revisión: 2029

Autor: César Krall

Resumen: Este artículo explica cuestiones básicas sobre la minería de datos (data mining), desde varios puntos de vista (utilidad empresarial, campo para emprendedores y campo de investigación). Resume y comenta una conferencia impartida por José C. Riquelme (Profesor de la Universidad de Sevilla) en la Escuela de Ingeniería Informática de la Universidad de Sevilla.

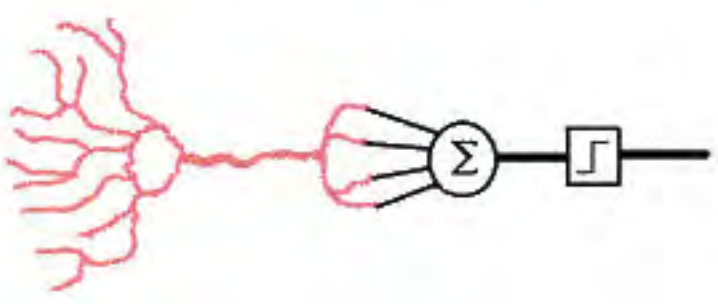
¿QUÉ ES UN MODELO DE MINERÍA DE DATOS?

La minería de datos se aplica a todo tipo de datos imaginable: desde datos numéricos a imágenes de satélite, mamografías, música, archivos de ordenador, imágenes, etc. Podemos decir que "cualquier cosa" constituye un dato. Por tanto la minería de datos tiene infinitas aplicaciones: comerciales, marketing, industria, internet, agricultura, etc.

Con miles de datos, necesitamos limpiarlos (eliminar fragmentos inútiles, repetidos, etc.) y organizarlos, y una vez realizado este proceso decimos que tenemos "Información".

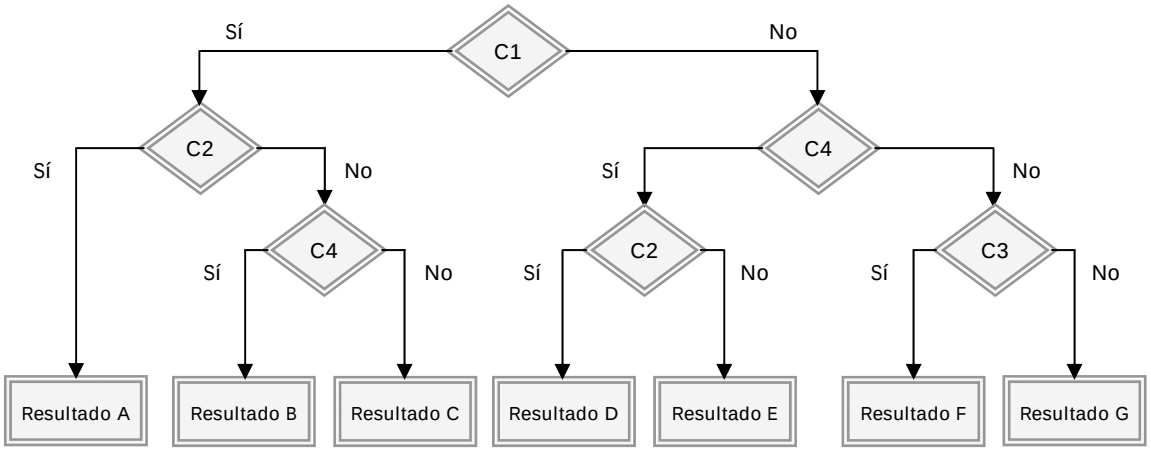
La información hay que tratarla con un modelo para obtener resultados o conclusiones a los que llamamos "Conocimiento". Es decir, el conocimiento es información analizada. Para este análisis hay diferentes modelos de minería de datos. Digamos que un modelo es una forma de aplicar un tratamiento a una cantidad masiva de datos para extraer información de ellos. Podemos citar por ejemplo dos de ellos:

a) Modelo de red neuronal: en una red neuronal tendríamos múltiples nodos que constituyen puntos de entrada de los datos. Esos datos son agrupados y sometidos a un tratamiento mediante un algoritmo que da lugar a que se obtengan unos resultados.



De las redes neuronales suele decirse que son cajas negras, porque el proceso de tratamiento de los datos hasta obtener el resultado no siempre sigue unas pautas lógicas o comprensibles por el ser humano. Sin embargo, su interés radicaría en que son herramientas útiles para realizar predicciones, por lo que son usadas en numerosas aplicaciones.

b) Modelo de árbol de decisión: se trata de la aplicación del conocido procedimiento del "divide y vencerás". Sobre los datos, se van realizando sucesivas bifurcaciones hasta llegar a un resultado. Sigue unas pautas lógicas, por lo que se dice que es una "caja blanca", o proceso comprensible por el ser humano. A modo de anécdota, podemos citar un juego web denominado "Akinator el genio adivino". El juego consiste en que pensamos en un personaje y el sistema nos va haciendo una serie de preguntas: por ejemplo, si es hombre o mujer. Con esta pregunta, se descartan aproximadamente el 50 % de los items en la base de datos. A continuación nos puede preguntar si es un personaje vivo, con lo cual descarta otro porcentaje significativo. En base a bifurcaciones, se llega finalmente al personaje en la base de datos que corresponde con el que habíamos pensado y se produce la "adivinación".



¿CÓMO ESCOGER UN MODELO DE MINERÍA DE DATOS?

No hay un modelo óptimo de tratamiento de datos. Por tanto, el modelo a elegir depende de las circunstancias y necesidades. Factores a tener en cuenta son la efectividad del modelo para dar resultados de calidad, y el si resulta necesario o no que sea comprensible para el ser humano.

En el caso de escoger una red neuronal, las operaciones que se aplican a los datos hay que determinarlas. ¿Cómo se hace esto? Digamos que "entrenando" a la red neuronal (a esto se le llama machine learning o aprendizaje automático) a través de algoritmos de optimización de forma que dados unos datos de entrada, vamos informando al sistema de si el resultado es más o menos bueno. En sucesivas iteraciones, el sistema puede alcanzar un grado de perfeccionamiento adecuado para su explotación comercial.

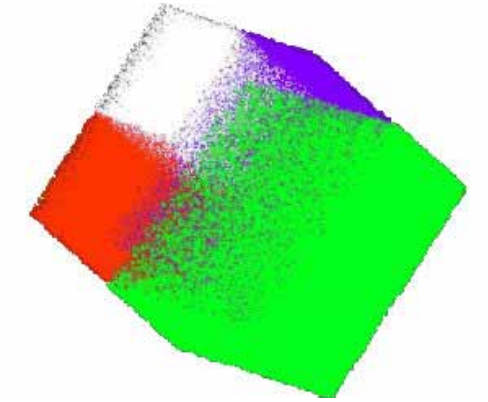
LAS BASES DE DATOS Y LA MINERÍA DE DATOS

Las bases de datos han sido sin duda una herramienta fundamental que ha permitido la evolución de la ciencia de la minería de datos. De hecho, a veces se usa el término "KDD (Knowledge Discovery in Databases o Descubrimiento de Conocimiento en Bases de Datos) como sinónimo de minería de datos.

Las bases de datos puede decirse que son una de las tres patas en que se apoya la minería de datos, y que son: 1. Bases de datos 2. Estadística y 3. Algoritmia

VISUAL DATA MINING O MINERÍA DE DATOS VISUAL

Una aplicación curiosa de la minería de datos es obtener imágenes representativas para realizar el análisis de datos. Esto permite mostrar lo que ocurre con miles de datos de forma gráfica.



GRAN HERMANO (BIG BROTHER) Y LA MINERÍA DE DATOS

La evolución de la informática y de las bases de datos permiten el almacenamiento masivo de datos de una manera que hace unos años hubiera sido inimaginable. Esto ha hecho posible el desarrollo de la telefonía móvil (¿Has pensado cuántos datos maneja una compañía de telefonía móvil que a cada usuario envía una carta con sus datos personales y los datos de las llamadas realizadas incluyendo número llamado, duración de la llamada, tarifa, etc.?) o de los buscadores de internet (al fin y al cabo google nació y basa buena parte de su negocio en la minería de datos).

Las grandes tecnológicas como Oracle o Microsoft (con SQLServer) tienen herramientas para hacer minería de datos. También hay herramientas para la minería de datos en el ámbito del software libre.

Ahora bien, esta extraordinaria capacidad tiene obviamente un peligro, porque es como el Gran Hermano descrito por George Orwell en su novela “1984”. Una forma de tener “intervenidas” las comunicaciones y pensamientos de las personas. Digamos que la minería de datos permite saberlo todo (o casi todo): ¿Qué música se escucha? ¿De qué se habla? ¿Qué prefiere la gente? ¿Qué preocupa a la gente? Si esto te parece exagerado, piensa en la cantidad de datos que almacena FaceBook. Las redes sociales tienen una componente peligrosa porque permiten saber todo lo que hace, le gusta o no le gusta a la gente.

Y aún más, pueden utilizarse para tratar de ir contra supuestos enemigos como “terroristas”. A través de minería de datos, podría tratar de establecerse si un usuario de FaceBook es un potencial terrorista evaluando si responde al patrón de terrorista.

CONCLUSIONES Y ALGO SOBRE EL FUTURO DE LA MINERÍA DE DATOS

La minería de datos es algo más allá de la estadística tradicional (cálculo de medias, análisis de varianza, etc.). Mientras que en Estados Unidos su sanidad usa ya técnicas de minería de datos, en la mayoría de los países los sistemas sanitarios se apoyan aún en la estadística tradicional de principios de siglo XX.

Esto obviamente irá cambiando, y es un ejemplo del enorme potencial que adquirirá la data mining en los años venideros.

También hay frenos al desarrollo del data mining. En muchos casos las empresas son muy celosas de sus datos y resultados en minería de datos. Por eso es frecuente que rechacen colaborar o contratar trabajos de minería de datos con las universidades porque tienen auténtico pánico a que la competencia pueda hacerse con sus datos.

Como conclusión, podría decirse que la minería de datos está en pleno auge y aún mucha gente no es consciente de la importancia que tiene. Su avance se constata, por ejemplo, en las ofertas de empleo, donde cada vez con mayor frecuencia aparecen términos como "Análisis de datos", "CRM", "Data Mining", "Clustering", etc.

REFERENCIAS Y MÁS INFORMACIÓN

Este artículo resume y comenta la conferencia pública impartida por José C. Riquelme, profesor de la Universidad de Sevilla, en el marco de las "Jornadas Imaginática: La informática del futuro", que tuvieron lugar en la Escuela Técnica Superior de Informática de la Universidad de Sevilla (España) y a las que tuvimos la oportunidad de asistir.