aprenderaprogramar.com

Minería de datos (data mining). Qué es y para qué sirve. (1ª parte) (DV00105A)

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.

MINERÍA DE DATOS: QUÉ ES Y PARA QUÉ SIRVE

Hay diferentes definiciones para minería de datos. Una muy simple sería decir que es el estudio y tratamiento de datos masivos para extraer conclusiones e información relevante de ellos.

Vamos a tratar de explicar para qué sirve la minería de datos dando ejemplos de en qué situaciones se aplica.

a) Supongamos un banco que otorga créditos y va a estudiar la concesión de un crédito. El banco tiene una serie histórica de datos de clientes y concesión de créditos con mxn casillas. Por ejemplo los datos disponibles pueden ser: Cliente / Edad / Estado civil / Trabaja / Nómina / Casa / Hipoteca / ¿Pagó?. Cada una de estas columnas se llama atributo. El campo ¿Pagó? es binario (solo puede tomar como valores sí o no) y es el atributo clave que tiene el banco para estudiar la concesión del crédito. No siempre existe un atributo clave. ¿Para qué le sirve la minería de datos al banco? Pues para decidir si concede el crédito o no: por ejemplo, estudiando y tratando los datos puede llegar a la conclusión de que los varones menores de 20 años que están casados estadísticamente tienen un alto porcentaje de impagos. Si el cliente corresponde a ese perfil la decisión puede ser denegar el crédito.

b) Supongamos un hospital donde hay unos datos de pacientes y un diagnóstico. Se puede tener una tabla de datos que incluya por ejemplo datos como Paciente / Edad / Glóbulos rojos / Glóbulos blancos / Tensión / Azúcar / Diagnóstico. ¿Para qué le serviría la minería de datos al hospital? Pues por ejemplo para hacer un prediagnóstico de la dolencia que con mayor probabilidad pueda tener un paciente en base a sus datos asociados. Un tipo de dolencia se dice que es un dato discreto porque solo puede tomar unos valores concretos (por ejemplo que haya 30 tipos de dolencias). Estudiando y tratando los datos se pueden llegar a conclusiones, por ejemplo que si un paciente tiene más de 60 años, los glóbulos blancos muy altos y el azúcar alto es muy probable que esté desarrollando diabetes. Si el paciente corresponde a ese perfil, la decisión puede ser hacer unas pruebas específicas o poner cierto tratamiento preventivo.

c) Supongamos un hipermercado. La minería de datos para grandes superficies se llama “análisis de cesta de la compra” o Market Basket Analysis. Por cada compra realizada, especialmente las realizadas con tarjeta, podría almacenar datos que le permite conocer los gustos de los clientes, qué es lo que más compran, qué cantidades compran, cómo se correlacionan los productos, etc. En una tabla de datos se podrían tener campos como Cliente / Gasto en leche / Gasto en pan / Gasto en cerveza / Gasto en pañales / Gasto en pescado. ¿Para qué le serviría la minería de datos a la gran superficie? Le permitiría tomar decisiones como que si por ejemplo la leche y el pan están muy correlacionados (cuando se compra leche se suele comprar pan), ambos productos se pueden colocar distanciados dentro del establecimiento para que el cliente recorra más distancia y al ver más productos compre más. Estas técnicas se incluyen dentro de lo que se llama “Marketing Basado en Minería de Datos” (en inglés, CRM, Client Relation Management) y son discutibles, pero de lo que no cabe duda es de que los grandes comercios estudian la psicología y hábitos de los clientes para tratar de sacarles partido. Otro ejemplo sería que si un producto tiene alta demanda se puede ofrecer con descuentos promocionales para servir de gancho y que el cliente acuda a ese hipermercado

d) Supongamos una empresa de desarrollo de software. Un equipo de ingenieros puede desarrollar aplicaciones informáticas y por cada una de ellas se recopilan distintos datos relacionados con la métrica del software (por ejemplo Horas de trabajo / Líneas de código / Número de errores por cada 1000 líneas de código, etc.). ¿Para qué le serviría la minería de datos a la empresa de desarrollo de software? Pues por ejemplo para saber el número de errores que previsiblemente se va a encontrar en un proyecto y el tiempo que puede necesitar para corregirlos, antes de que el proyecto en sí se haya desarrollado completamente.

Resumiendo lo expuesto hasta ahora podemos decir que la funcionalidad de la minería de datos puede ser:

a) Predictiva (p.ej. caso del banco, hospital): sirve para predecir cosas.

    i. En base a una clasificación: por ejemplo si el cliente pagará o no pagará, o el tipo de dolencia que puede tener un paciente.
    ii. En base a una regresión: por ejemplo calcular el tiempo previsible que se empleará en corregir los errores de un desarrollo de software.

b) Descriptiva:

    i. Agrupamiento (clustering): clasificar individuos en grupos en base a sus características. Por ejemplo, clasificar pacientes del hospital en base a los datos de sus analíticas.
    ii. Reglas de asociación: conocer cómo se relacionan los datos o campos. Por ejemplo conocer en el hipermercado que un cliente que compra leche muy probablemente comprará también pan.
    iii. Secuenciación: intentar predecir el valor de una variable en función del tiempo. Por ejemplo la demanda de energía eléctrica.

CAMPOS DE APLICACIÓN DE LA MINERÍA DE DATOS

La minería de datos tiene muchos campos de aplicación pues puede ser útil en prácticamente todas las facetas de la actividad humana. Vamos a indicar algunas cuestiones relevantes sobre la posible aplicación de la minería de datos:

a) La minería de datos tiene utilidad empresarial: las empresas pueden optimizar procesos y mejorar sus productos y ventas utilizando minería de datos.
b) Existen pocos especialistas o empresas especializadas en minería de datos. Teniendo en cuenta su importancia, es un campo de trabajo para emprendedores.
c) La minería de datos es una disciplina que se está desarrollando cada vez con mayores capacidades gracias al avance en tecnología y a la cada vez más alta capacidad de computación de los ordenadores. Constituye un campo amplio de investigación en el que cada vez trabajan más investigadores y equipos de investigación.

METODOLOGÍA DE LA MINERÍA DE DATOS

Un trabajo de minería de datos podríamos decir que típicamente consta de las siguientes partes:

1. **Entendimiento del problema:** se trata de hablar con el cliente, conocer sus necesidades, conocer su negocio o actividad, conocer qué datos relevantes tiene disponibles y cuáles serían necesarios pero no están disponibles, etc.
2. **Entendimiento de los datos:** hay que saber qué significan los datos, si son continuos o discretos, qué tipo de valores toman, qué utilidad futura pueden tener y saber si están bien capturados o no.
3. **Preparación de datos:** se trata de reflexionar sobre cómo guardar los datos. Típicamente hablaremos de tablas con filas y columnas, pero hay que ver cómo se organizan las tablas, cómo se interrelacionan entre ellas, etc. En definitiva organizar los datos para poder sacarles partido.
4. **Modelamiento:** una vez se tienen los datos organizados hay que definir los algoritmos que se van a utilizar para tratar los datos. Una vez tratados, los datos nos devolverán información útil.
5. **Evaluación:** los resultados obtenidos deben de ser sometidos a comprobación, verificar que están libres de errores, ratificar que son útiles para los objetivos perseguidos, etc.
6. **Despliegue funcional-comercial:** una vez se tiene automatizada la captura y tratamiento de datos para obtener unos resultados, se desarrollan herramientas, normalmente en forma de aplicaciones informáticas que permiten generar alertas, informes, estadísticas, etc. que tienen una utilidad directa para la toma de decisiones y sistema de información del cliente.

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.

Continuación del artículo en aprenderaprogramar.com: DV00106A (busca este código en el buscador de la web)