Data mining: torturando a los datos hasta que confiesen[*]

Luis Carlos Molina Félix

Coordinador del programa de *Data mining* (UOC)
lmolinaf@uoc.edu

Resumen: El título de este artículo es una explicación informal de la actividad que realiza una tecnología denominada *data mining* (minería de datos). Lo que se pretende con esta tecnología es descubrir conocimiento *oculto* a partir de grandes volúmenes de datos. Desde la década pasada, debido a los grandes avances computacionales, se ha ido incorporando a las organizaciones para constituirse en un apoyo esencial al momento de tomar decisiones. Organizaciones tales como empresas, clubes profesionales deportivos, universidades y gobiernos, entre otros, hacen uso de esta tecnología como *ayuda* en la toma de sus decisiones. Algunos de estos ejemplos serán citados en el presente trabajo.

1. Introducción

Cada día generamos una gran cantidad de información, algunas veces conscientes de que lo hacemos y otras veces inconscientes de ello porque lo desconocemos. Nos damos cuenta de que generamos información cuando registramos nuestra entrada en el trabajo, cuando entramos en un servidor para ver nuestro correo, cuando pagamos con una tarjeta de crédito o cuando reservamos un billete de avión. Otras veces no nos damos cuenta de que generamos información, como cuando conducimos por una vía donde están contabilizando el número de automóviles que pasan por minuto, cuando se sigue nuestra navegación por Internet o cuando nos sacan una fotografía del rostro al haber pasado cerca de una oficina gubernamental.

¿Con qué finalidad queremos generar información? Son muchos los motivos que nos llevan a generar información, ya que nos pueden ayudar a controlar, optimizar, administrar, examinar, investigar, planificar, predecir, someter, negociar o tomar decisiones de cualquier ámbito según el dominio en que nos desarrollemos. La información por sí misma está considerada un bien patrimonial. De esta forma, si una empresa tiene una pérdida total o parcial de información provoca bastantes perjuicios. Es evidente que la información debe ser protegida, pero también explotada.

¿Qué nos ha permitido poder generar tanta información? En los últimos años, debido al desarrollo tecnológico a niveles exponenciales tanto en el área de cómputo como en la de transmisión de datos, ha sido posible que se gestionen de una mejor manera el manejo y almacenamiento de la información. Sin duda existen cuatro factores importantes que nos han llevado a este suceso:

1. El abaratamiento de los sistemas de almacenamiento tanto temporal como permanente.

*** Las transparencias de este artículo se pueden obtener en: http://www.lsi.upc.es/~lcmolina/about.htm[url1].**

2. El incremento de las velocidades de cómputo en los procesadores.

3. Las mejoras en la confiabilidad y aumento de la velocidad en la transmisión de datos.

4. El desarrollo de sistemas administradores de bases de datos más poderosos.

Actualmente todas estas ventajas nos han llevado a abusar del almacenamiento de la información en las bases de datos. Podemos decir que algunas empresas almacenan un cierto tipo de datos al que hemos denominado *dato-escritura*, ya que sólo se guarda (o *escribe*) en el disco duro, pero nunca se hace uso de él. Generalmente, todas las empresas usan un dato llamado *dato-escritura-lectura*, que utilizan para hacer consultas dirigidas. Un nuevo tipo de dato al cual hemos denominado *dato-escritura-lectura-análisis* es el que proporciona en conjunto un verdadero conocimiento y nos apoya en las tomas de decisiones. Es necesario contar con tecnologías que nos ayuden a explotar el potencial de este tipo de datos.

La cantidad de información que nos llega cada día es tan inmensa que nos resulta difícil asimilarla. Basta con ir al buscador Altavista[url2] y solicitar la palabra *information* para ver que existen 171.769.416 sitios donde nos pueden decir algo al respecto. Suponiendo que nos tomemos un minuto para ver el contenido de cada página, tardaríamos entonces 326 años en visitarlas todas. Esto es imposible, y, por lo tanto, existe una clara necesidad de disponer de tecnologías que nos ayuden en nuestros procesos de búsqueda y, aún más, de tecnologías que nos ayuden a comprender su contenido.

El *data mining* surge como una tecnología que intenta ayudar a comprender el contenido de una base de datos. De forma general, los *datos* son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en *información*. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación del confronto entre la información y ese modelo represente un valor agregado, entonces nos referimos al *conocimiento*. En la figura 1 se ilustra la jerarquía que existe en una base de datos entre dato, información y conocimiento (Molina, 1998). Se observa igualmente el volumen que presenta en cada nivel y el valor que los responsables de las decisiones le dan en esa jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así entre la información y el conocimiento. El *data mining* trabaja en el nivel superior buscando patrones, comportamientos, agrupaciones, secuencias, tendencias o asociaciones que puedan generar algún modelo que nos permita comprender mejor el dominio para *ayudar* en una posible toma de decisión.

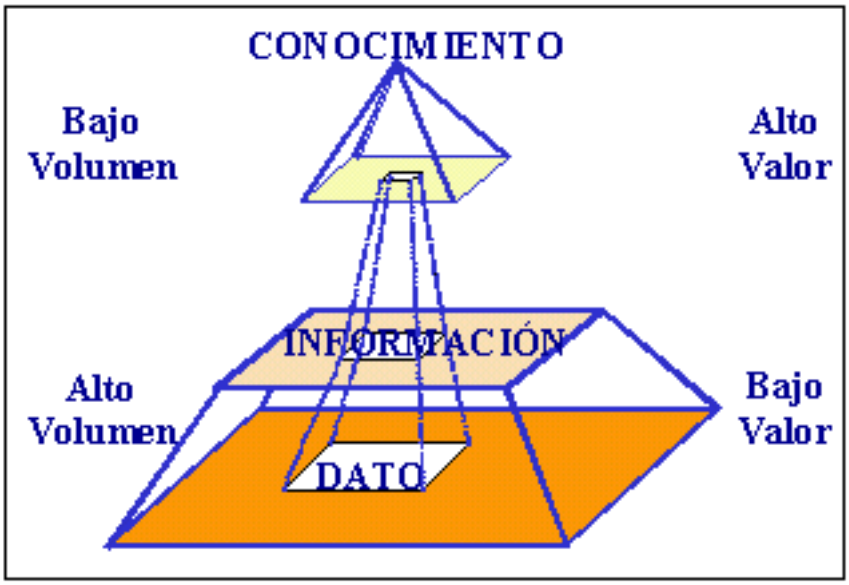

Figura 1. Relación entre dato, información y conocimiento (Molina, 1998).

2. *Data mining*: conceptos e historia

Aunque desde un punto de vista académico el término *data mining* es una etapa dentro de un proceso mayor llamado *extracción de conocimiento en bases de datos* (*Knowledge Discovery in Databases* o KDD) en el entorno comercial, así como en este trabajo, ambos términos se usan de manera indistinta. Lo que en verdad hace el *data mining* es reunir las ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo, principalmente usando como materia prima las bases de datos. Una definición tradicional es la siguiente: "Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos" (Fayyad y otros, 1996). Desde nuestro punto de vista, lo definimos como "la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión" (Molina y otros, 2001).

La idea de *data mining* no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como *data fishing*, *data mining* o *data archaeology* con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de *data mining* y KDD.[3] A finales de los años ochenta sólo existían un par de empresas dedicadas a esta tecnología; en 2002 existen más de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones. Las listas de discusión sobre este tema las forman investigadores de más de ochenta países. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

El *data mining* es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente. Actualmente existen aplicaciones o herramientas comerciales de *data mining* muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta.

3. Más detalles en http://www.kdnuggets.com.[url3].

3. Aplicaciones de uso

Cada año, en los diferentes congresos, simposios y talleres que se realizan en el mundo se reúnen investigadores con aplicaciones muy diversas. Sobre todo en los Estados Unidos, el *data mining* se ha ido incorporando a la vida de empresas, gobiernos, universidades, hospitales y diversas organizaciones que están interesadas en explorar sus bases de datos.

Podemos decir que "en *data mining* cada caso es un caso". Sin embargo, en términos generales, el proceso se compone de cuatro etapas principales:

1. Determinación de los objetivos. Trata de la delimitación de los objetivos que el *cliente* desea bajo la orientación del especialista en *data mining*.

2. Preprocesamiento de los datos. Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de *data mining*.

3. Determinación del modelo. Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.

4. Análisis de los resultados. Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El *cliente* determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

A continuación se describen varios ejemplos donde se ha visto involucrado el *data mining*. Se han seleccionado de diversos dominios y con diversos objetivos para observar su potencial. Respecto a los modelos inteligentes, se ha comprobado que en ellos se utilizan principalmente árboles y reglas de decisión, reglas de asociación, redes neuronales, redes bayesianas, conjuntos aproximados (*rough sets*), algoritmos de agrupación (*clustering*), máquinas de soporte vectorial, algoritmos genéticos y lógica difusa.

3.1. En el gobierno

El FBI analizará las bases de datos comerciales para detectar terroristas.

A principios del mes de julio de 2002, el director del Federal Bureau of Investigation (FBI), John Aschcroft, anunció que el Departamento de Justicia comenzará a introducirse en la vasta cantidad de datos comerciales referentes a los hábitos y preferencias de compra de los consumidores, con el fin de descubrir potenciales terroristas antes de que ejecuten una acción.[4] Algunos expertos aseguran que, con esta información, el FBI unirá todas las bases de datos probablemente mediante el número de la Seguridad Social y permitirá saber si una persona fuma, qué talla y tipo de ropa usa, su registro de arrestos, su salario, las revistas a las que está suscrito, su altura y peso, sus contribuciones a la Iglesia, grupos políticos u organizaciones no gubernamentales, sus enfermedades crónicas (como diabetes o asma), los libros que lee, los productos de supermercado que compra, si tomó clases de vuelo o si tiene cuentas de banco abiertas, entre otros.[5] La inversión inicial ronda los setenta millones de dólares estadounidenses para consolidar los almacenes de datos, desarrollar redes de seguridad para compartir

4. Ver más en http://www.fcw.com/fcw/articles/2002/0603/news-fbi-06-03-02.asp[url4].
5. Ver más en http://tierra.ucsd.edu/archives/ats-l/2002.06/msg00013.html[url5].

información e implementar nuevo software analítico y de visualización.

3.2. En la empresa

Detección de fraudes en las tarjetas de crédito.

En 2001, las instituciones financieras a escala mundial perdieron más de 2.000 millones de dólares estadounidenses en fraudes con tarjetas de crédito y débito. El Falcon Fraud Manager[6] es un sistema inteligente que examina transacciones, propietarios de tarjetas y datos financieros para detectar y mitigar fraudes. En un principio estaba pensado, en instituciones financieras de Norteamérica, para detectar fraudes en tarjetas de crédito. Sin embargo, actualmente se le han incorporado funcionalidades de análisis en las tarjetas comerciales, de combustibles y de débito.[7] El sistema Falcon ha permitido ahorrar más de seiscientos millones de dólares estadounidenses cada año y protege aproximadamente más de cuatrocientos cincuenta millones de pagos con tarjeta en todo el mundo –aproximadamente el sesenta y cinco por ciento de todas las transacciones con tarjeta de crédito.

Descubriendo el porqué de la deserción de clientes de una compañía operadora de telefonía móvil.

Este estudio fue desarrollado en una operadora española que básicamente situó sus objetivos en dos puntos: el análisis del perfil de los clientes que se dan de baja y la predicción del comportamiento de sus nuevos clientes. Se analizaron los diferentes históricos de clientes que habían abandonado la operadora (12,6%) y de clientes que continuaban con su servicio (87,4%). También se analizaron las variables personales de cada cliente (estado civil, edad, sexo, nacionalidad, etc.). De igual forma se estudiaron, para cada cliente, la morosidad, la frecuencia y el horario de uso del servicio, los descuentos y el porcentaje de llamadas locales, interprovinciales, internacionales y gratuitas. Al contrario de lo que se podría pensar, los clientes que abandonaban la operadora generaban ganancias para la empresa; sin embargo, una de las conclusiones más importantes radicó en el hecho de que los clientes que se daban de baja recibían pocas promociones y registraban un mayor número de incidencias respecto a la media. De esta forma se recomendó a la operadora hacer un estudio sobre sus ofertas y analizar profundamente las incidencias recibidas por esos clientes. Al descubrir el perfil que presentaban, la operadora tuvo que diseñar un trato más personalizado para sus clientes actuales con esas características. Para poder predecir el comportamiento de sus nuevos clientes se diseñó un sistema de predicción basado en la cantidad de datos que se podía obtener de los nuevos clientes comparados con el comportamiento de clientes anteriores.

Prediciendo el tamaño de las audiencias televisivas.

La British Broadcasting Corporation (BBC) del Reino Unido emplea un sistema para predecir el tamaño de las audiencias televisivas para un programa propuesto, así como el tiempo óptimo de exhibición (Brachman y otros, 1996). El sistema utiliza redes neuronales y árboles de decisión aplicados a datos históricos de la cadena para determinar los criterios que participan según el programa que hay que presentar.[8] La versión final se desempeña tan bien como un experto humano con la ventaja de que se adapta más fácilmente a los cambios porque es constantemente reentrenada con datos actuales.

3.3. En la universidad

6. Ver más en http://www.fairisaac.com/page.cfm/press_id=325[url6].
7. American Express reporta entre un diez y un quince por ciento de incremento en el uso de sus tarjetas apoyándose en técnicas de *data mining*.
8. Más detalles en http://www.mining.dk/SPSS/Nyheder/nr7case_bbc.htm[url7].

Conociendo si los recién titulados de una universidad llevan a cabo actividades profesionales relacionadas con sus estudios.

Se hizo un estudio sobre los recién titulados de la carrera de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Chihuahua II,[9] en Méjico (Rodas, 2001). Se quería observar si sus recién titulados se insertaban en actividades profesionales relacionadas con sus estudios y, en caso negativo, se buscaba saber el perfil que caracterizó a los exalumnos durante su estancia en la universidad. El objetivo era saber si con los planes de estudio de la universidad y el aprovechamiento del alumno se hacía una buena inserción laboral o si existían otras variables que participaban en el proceso. Dentro de la información considerada estaba el sexo, la edad, la escuela de procedencia, el desempeño académico, la zona económica donde tenía su vivienda y la actividad profesional, entre otras variables. Mediante la aplicación de conjuntos aproximados se descubrió que existían cuatro variables que determinaban la adecuada inserción laboral, que son citadas de acuerdo con su importancia: zona económica donde habitaba el estudiante, colegio de donde provenía, nota al ingresar y promedio final al salir de la carrera. A partir de estos resultados, la universidad tendrá que hacer un estudio socioeconómico sobre grupos de alumnos que pertenecían a las clases económicas bajas para dar posibles soluciones, debido a que tres de las cuatro variables no dependían de la universidad.

3.4. En investigaciones espaciales

Proyecto SKYCAT.

Durante seis años, el Second Palomar Observatory Sky Survey (POSS-II) coleccionó tres terabytes de imágenes que contenían aproximadamente dos millones de objetos en el cielo. Tres mil fotografías fueron digitalizadas a una resolución de 16 bits por píxel con 23.040 x 23.040 píxeles por imagen. El objetivo era formar un catálogo de todos esos objetos. El sistema Sky Image Cataloguing and Analysis Tool (SKYCAT) se basa en técnicas de agrupación (*clustering*) y árboles de decisión para poder clasificar los objetos en estrellas, planetas, sistemas, galaxias, etc. con una alta confiabilidad (Fayyad y otros, 1996). Los resultados han ayudado a los astrónomos a descubrir dieciséis nuevos quásars con corrimiento hacia el rojo que los incluye entre los objetos más lejanos del universo y, por consiguiente, más antiguos. Estos quásars son difíciles de encontrar y permiten saber más acerca de los orígenes del universo.

3.5. En los clubes deportivos

El AC de Milán utiliza un sistema inteligente para prevenir lesiones.

Esta temporada el club comenzará a usar redes neuronales para prevenir lesiones y optimizar el acondicionamiento de cada atleta. Esto ayudará a seleccionar el fichaje de un posible jugador o a alertar al médico del equipo de una posible lesión.[10] El sistema, creado por Computer Associates International, es alimentado por datos de cada jugador, relacionados con su rendimiento, alimentación y respuesta a estímulos externos, que se obtienen y analizan cada quince días. El jugador lleva a cabo determinadas actividades que son monitoreadas por veinticuatro sensores conectados al cuerpo y que transmiten señales de radio que posteriormente son almacenadas en una base de datos. Actualmente el sistema dispone de 5.000 casos registrados que permiten predecir alguna posible lesión. Con ello, el club intenta ahorrar dinero evitando comprar jugadores que presenten una alta probabilidad de lesión, lo que haría incluso renegociar su contrato. Por otra parte, el sistema pretende encontrar las diferencias entre las lesiones de atletas de ambos sexos, así como saber si una determinada lesión se relaciona con el estilo de juego de un país concreto donde se practica el fútbol.

9. Equivalente en España a una universidad politécnica.
10. Más detalles en http://www.msnbc.com/news/756968.asp[url8].

Los equipos de la NBA utilizan aplicaciones inteligentes para apoyar a su cuerpo de entrenadores.

El Advanced Scout[11] es un software que emplea técnicas de *data mining* y que han desarrollado investigadores de IBM para detectar patrones estadísticos y eventos *raros*. Tiene una interfaz gráfica muy amigable orientada a un objetivo muy específico: analizar el juego de los equipos de la National Basketball Association (NBA).

El software utiliza todos los registros guardados de cada evento en cada juego: pases, encestes, rebotes y doble marcaje (*double team*) a un jugador por el equipo contrario, entre otros. El objetivo es ayudar a los entrenadores a aislar eventos que no detectan cuando observan el juego en vivo o en película.

Un resultado interesante fue uno hasta entonces no observado por los entrenadores de los Knicks de Nueva York. El doble marcaje a un jugador puede generalmente dar la oportunidad a otro jugador de encestar más fácilmente. Sin embargo, cuando los Bulls de Chicago jugaban contra los Knicks, se encontró que el porcentaje de encestes después de que al centro de los Knicks, Patrick Ewing, le hicieran doble marcaje era extremadamente bajo, indicando que los Knicks no reaccionaban correctamente a los dobles marcajes. Para saber el porqué, el cuerpo de entrenadores estudió cuidadosamente todas las películas de juegos contra Chicago. Observaron que los jugadores de Chicago rompían su doble marcaje muy rápido de tal forma que podían tapar al encestador libre de los Knicks antes de prepararse para efectuar su tiro. Con este conocimiento, los entrenadores crearon estrategias alternativas para tratar con el doble marcaje.

La temporada pasada, IBM ofreció el Advanced Scout a la NBA, que se convirtió así en un patrocinador corporativo. La NBA dio a sus veintinueve equipos la oportunidad de aplicarlo. Dieciocho equipos lo están haciendo hasta el momento obteniendo descubrimientos interesantes.

4. Extensiones del *data mining*

4.1. *Web mining*

Una de las extensiones del *data mining* consiste en aplicar sus técnicas a documentos y servicios del Web, lo que se llama *web mining* (minería de web) (Kosala y otros, 2000). Todos los que visitan un sitio en Internet dejan huellas digitales (direcciones de IP, navegador, galletas, etc.) que los servidores automáticamente almacenan en una bitácora de accesos (*log*). Las herramientas de *web mining* analizan y procesan estos *logs* para producir información significativa, por ejemplo, cómo es la navegación de un cliente antes de hacer una compra en línea. Debido a que los contenidos de Internet consisten en varios tipos de datos, como texto, imagen, vídeo, metadatos o hiperligas, investigaciones recientes usan el término *multimedia data mining* (*minería de datos multimedia*) como una instancia del *web mining* (Zaiane y otros, 1998) para tratar ese tipo de datos. Los accesos totales por dominio, horarios de accesos más frecuentes y visitas por día, entre otros datos, son registrados por herramientas estadísticas que complementan todo el proceso de análisis del *web mining*.

Normalmente, el *web mining* puede clasificarse en tres dominios de extracción de conocimiento de acuerdo con la naturaleza de los datos:

1. *Web content mining* (minería de contenido web). Es el proceso que consiste en la extracción de conocimiento del contenido de documentos o sus descripciones. La localización de patrones en el texto de los documentos, el descubrimiento del recurso

11. Ver http://domino.research.ibm.com/comm/wwwr_thinkresearch.nsf/pages/datamine296.html[url9].

basado en conceptos de indexación o la tecnología basada en agentes también pueden formar parte de esta categoría.

2. *Web structure mining* (minería de estructura web). Es el proceso de inferir conocimiento de la organización del WWW y la estructura de sus ligas.

3. *Web usage mining* (minería de uso web). Es el proceso de extracción de modelos interesantes usando los *logs* de los accesos al web.

Algunos de los resultados que pueden obtenerse tras la aplicación de los diferentes métodos de *web mining* son:

- El ochenta y cinco por ciento de los clientes que acceden a */productos/home.html* y a */productos/noticias.html* acceden también a */productos/historias_suceso.html*. Esto podría indicar que existe alguna noticia interesante de la empresa que hace que los clientes se dirijan a historias de suceso. Igualmente, este resultado permitiría detectar la noticia sobresaliente y colocarla quizá en la página principal de la empresa.

- Los clientes que hacen una compra en línea cada semana en */compra/producto1.html* tienden a ser de sectores del gobierno. Esto podría resultar en proponer diversas ofertas a este sector para potenciar más sus compras.

- El sesenta por ciento de los clientes que hicieron una compra en línea en */compra/producto1.html* también compraron en */compra/producto4.html* después de un mes. Esto indica que se podría recomendar en la página del producto 1 comprar el producto 4 y ahorrarse el costo de envío de este producto.

Los anteriores ejemplos nos ayudan a formarnos una pequeña idea de lo que podemos obtener. Sin embargo, en la realidad existen herramientas de mercado muy poderosas con métodos variados y visualizaciones gráficas excelentes. Para más información, ver Mena (1999).

4.2. *Text mining*

Estudios recientes indican que el ochenta por ciento de la información de una compañía está almacenada en forma de documentos. Sin duda, este campo de estudio es muy vasto, por lo que técnicas como la categorización de texto, el procesamiento de lenguaje natural, la extracción y recuperación de la información o el aprendizaje automático, entre otras, apoyan al *text mining* (minería de texto). En ocasiones se confunde el *text mining* con la recuperación de la información (*Information Retrieval* o IR) (Hearst, 1999). Ésta última consiste en la recuperación automática de documentos relevantes mediante indexaciones de textos, clasificación, categorización, etc. Generalmente se utilizan palabras clave para encontrar una página relevante. En cambio, el *text mining* se refiere a examinar una colección de documentos y descubrir información no contenida en ningún documento individual de la colección; en otras palabras, trata de obtener información sin haber partido de algo (Nasukawa y otros, 2001).

Una aplicación muy popular del *text mining* es relatada en Hearst (1999). Don Swanson intenta extraer información derivada de colecciones de texto. Teniendo en cuenta que los expertos sólo pueden leer una pequeña parte de lo que se publica en su campo, por lo general no se dan cuenta de los nuevos desarrollos que se suceden en otros campos. Así, Swanson ha demostrado cómo cadenas de implicaciones causales dentro de la literatura médica pueden conducir a hipótesis para enfermedades poco frecuentes, algunas de las cuales han recibido pruebas de soporte experimental. Investigando las

causas de la migraña, dicho investigador extrajo varias piezas de evidencia a partir de títulos de artículos presentes en la literatura biomédica. Algunas de esas claves fueron:

- El estrés está asociado con la migraña.
- El estrés puede conducir a la pérdida de magnesio.
- Los bloqueadores de canales de calcio previenen algunas migrañas.
- El magnesio es un bloqueador natural del canal de calcio.
- La depresión cortical diseminada (DCD) está implicada en algunas migrañas.
- Los niveles altos de magnesio inhiben la DCD.
- Los pacientes con migraña tienen una alta agregación plaquetaria.
- El magnesio puede suprimir la agregación plaquetaria.

Estas claves sugieren que la deficiencia de magnesio podría representar un papel en algunos tipos de migraña, una hipótesis que no existía en la literatura y que Swanson encontró mediante esas ligas. De acuerdo con Swanson (Swanson y otros, 1994), estudios posteriores han probado experimentalmente esta hipótesis obtenida por *text mining* con buenos resultados.

5. Conclusiones

Nuestra capacidad para almacenar datos ha crecido en los últimos años a velocidades exponenciales. En contrapartida, nuestra capacidad para procesarlos y utilizarlos no ha ido a la par. Por este motivo, el *data mining* se presenta como una tecnología de apoyo para explorar, analizar, comprender y aplicar el conocimiento obtenido usando grandes volúmenes de datos. Descubrir nuevos caminos que nos ayuden en la identificación de interesantes estructuras en los datos es una de las tareas fundamentales en el *data mining*.

En el ámbito comercial, resulta interesante encontrar patrones ocultos de consumo de los clientes para poder explorar nuevos horizontes. Saber que un vehículo deportivo corre un riesgo de accidente casi igual al de un vehículo normal cuando su dueño tiene un segundo vehículo en casa ayuda a crear nuevas estrategias comerciales para ese grupo de clientes. Asimismo, predecir el comportamiento de un futuro cliente, basándose en los datos históricos de clientes que presentaron el mismo perfil, ayuda a poder retenerlo durante el mayor tiempo posible.

Las herramientas comerciales de *data mining* que existen actualmente en el mercado son variadas y excelentes. Las hay orientadas al estudio del web o al análisis de documentos o de clientes de supermercado, mientras que otras son de uso más general. Su correcta elección depende de la necesidad de la empresa y de los objetivos a corto y largo plazo que pretenda alcanzar. La decisión de seleccionar una solución de *data mining* no es una tarea simple. Es necesario consultar a expertos en el área con vista a seleccionar la más adecuada para el problema de la empresa.

Como se ha visto a lo largo del este artículo, son muchas las áreas, técnicas, estrategias, tipos de bases de datos y personas que intervienen en un proceso de *data mining*. Los negocios requieren que las soluciones tengan una integración transparente en un ambiente operativo.

Esto nos lleva a la necesidad de establecer estándares para hacer un ambiente interoperable, eficiente y efectivo. Esfuerzos en este sentido se están desarrollando actualmente. En Grossman y otros (2002) se exponen algunas iniciativas para estos estándares, incluyendo aspectos en:

- Modelos: para representar datos estadísticos y de *data mining*.
- Atributos: para representar la limpieza, transformación y agregación de atributos usados como entrada en los modelos.
- Interfaces y API: para facilitar la integración con otros lenguajes o aplicaciones de software y API.
- Configuración: para representar parámetros internos requeridos para construir y usar los modelos.
- Procesos: para producir, desplegar y usar modelos.
- Datos remotos y distribuidos: para analizar y explorar datos remotos y distribuidos.

En resumen, el *data mining* se presenta como una tecnología emergente, con varias ventajas: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas de negocios; por otro, ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios. Además, no hay duda de que trabajar con esta tecnología implica cuidar un sinnúmero de detalles debido a que el producto final involucra "toma de decisiones".

Lista de URL:

[url1]:http://www.lsi.upc.es/~lcmolina/about.htm

[url2]:http://www.altavista.com

[url3]:http://www.kdnuggets.com

[url4]:http://www.fcw.com/fcw/articles/2002/0603/news-fbi-06-03-02.asp

[url5]:http://tierra.ucsd.edu/archives/ats-l/2002.06/msg00013.html

[url6]:http://www.fairisaac.com/page.cfm/press_id=325

[url7]:http://www.mining.dk/SPSS/Nyheder/nr7case_bbc.htm

[url8]:http://www.msnbc.com/news/756968.asp

[url9]:http://domino.research.ibm.com/comm/wwwr_thinkresearch.nsf/pages/datamine296.html

Bibliografía:

BRACHMAN, R.J.; KHABAZA, T.; KLOESGEN, W.; PIATETSKY-SHAPIRO, G.; SIMOUDIS, E. (1996). "Mining business databases". *Communications of the ACM*. Vol. 39, pág. 42-48.

BRODLEY, C.E.; LANE, T.; STOUGH, T.M. (1999). "Knowledge discovery and data mining". *American Scientist*. Vol. 86, pág. 55-65.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (ed.) (1996). *Advances in knowledge and data mining*. Cambridge (Massachussets): AAAI/MIT Press.

FAYYAD, U.; HAUSSLER, D.; STOLORZ, P. (1996). "Mining scientific data". *Communications of the ACM*. Vol. 39, pág. 51-57.

FELDMAN, R.; DAGAN, I. (1995). "Knowledge discovery in textual databases (KDT)". En:

Proceedings of the 1st international conference on knowledge discovery. ACM.

GROSSMAN, R. L.; HORNIK, M.F.; MEYER, G. (2002). "Data mining standards initiatives". *Communications of ACM*. Vol. 45 (8), pág. 59-61.

HEARST, M. (1999). "Untangling text data mining". En: *Proceedings of 37th annual meeting of the association for computational linguistics*. Universidad de Maryland.

KOSALA, R.; BLOCKEEL, B. (2000). "Web mining research: a survey". *SIGKDD Explorations: Newsletter of the special interest group on knowledge discovery and data mining*. ACM Press. Vol. 2 (1).

MENA, J. (1999). *Data mining your website*. Digital Press.

MOLINA, L.C. (1998). *Data mining no processo de extração de conhecimento de bases de dados*. Tesis de máster. São Carlos (Brasil): Instituto de Ciências Matemáticas e Computação. Universidad de São Paulo.

MOLINA, L.C.; RIBEIRO, S. (2001). "Descubrimiento conocimiento para el mejoramiento bovino usando técnicas de data mining". En: *Actas del IV Congreso Catalán de Inteligencia Artificial*. Barcelona, pág. 123-130.

NASUKAWA, T.; NAGANO, T. (2001). "Text analysis and knowledge mining system". *IBM Systems Journal, knowledge management*. Vol. 40 (4).

RODAS, J. (2001). "Un ejercicio de análisis utilizando rough sets en un dominio de educación superior mediante el proceso KDD". Documento interno. Barcelona: Departamento de Lenguajes y Sistemas Informáticos, Universidad Politécnica de Cataluña.

SWANSON, D.R.; SMALHAISER, N.R. (1994). "Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease". *Neuroscience research communications*. Vol. 15, pág. 1-9.

WAY, J.I.; SMITH, E.A. (1991). "The evolution of synthetic aperture radar systems and their progression to the EOS SAR". *IEEE transactions on geoscience and remote sensing*. Vol. 29 (6), pág. 962-985.

ZAIANE, O.R.; HAN, J.; LI, Z-N.; CHEE, S.H.; CHIANG, J.Y. (1998). "MultiMedia-Miner: a system prototype for multimedia data mining". En: *Proceedings of international conference on management of data*. ACM SIGMOD. Vol. 27 (2), pág. 581-583.

Enlaces relacionados:

- Formación en la UOC
  http://www.uoc.edu/masters/esp/cursos/especializacion/208_id.html
- KDnuggets
  http://www.kdnuggets.com/
- KDcentral
  http://www.kdcentral.com/
- *Data Mining and Knowledge Discovery. An International Journal*
  http://www.digimine.com/usama/datamine/
- Departamento de Lenguajes y Sistemas Informáticos. Grupo de Soft Computing
  http://www.lsi.upc.es/~webia/soft-comp.html
- Página de Luis Carlos Molina Félix
  http://www.lsi.upc.es/~lcmolina/