Capítulo 1

¿QUÉ ES LA MINERÍA DE DATOS?

El primer pensamiento de muchos al oír por primera vez el término "minería de datos" fue la reflexión "nada nuevo bajo el sol". En efecto, la "minería de datos" no aparece por el desarrollo de tecnologías esencialmente diferentes a las anteriores, sino que se crea, en realidad, por la aparición de nuevas necesidades y, especialmente, por el reconocimiento de un nuevo potencial: el valor, hasta ahora generalmente infrautilizado, de la gran cantidad de datos almacenados informáticamente en los sistemas de información de instituciones, empresas, gobiernos y particulares. Los datos pasan de ser un "producto" (el resultado histórico de los sistemas de información) a ser una "materia prima" que hay que explotar para obtener el verdadero "producto elaborado", el conocimiento; un conocimiento que ha de ser especialmente valioso para la ayuda en la toma de decisiones sobre el ámbito en el que se han recopilado o extraído los datos. Es bien cierto que la estadística es la primera ciencia que considera los datos como su materia prima, pero las nuevas necesidades y, en particular, las nuevas características de los datos (en volumen y tipología) hacen que las disciplinas que integran lo que se conoce como "minería de datos" sean numerosas y heterogéneas.

En este capítulo analizaremos estas nuevas necesidades y posibilidades, precisaremos la definición de "minería de datos" dentro del contexto de la "extracción de conocimiento", las disciplinas que la forman y destacaremos los tipos de problemas que se desea tratar y los modelos o patrones que se espera obtener.

1.1 Nuevas necesidades

El aumento del volumen y variedad de información que se encuentra informatizada en bases de datos digitales y otras fuentes ha crecido espectacularmente en las últimas décadas. Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido. Aparte de su función de "memoria de la organización", la información histórica es útil para explicar el pasado, entender el presente y predecir la información futura. La mayoría de las decisiones de empresas, organizaciones e instituciones se basan también en información sobre experiencias pasadas extraídas de fuentes muy diversas. Además, ya que los datos pueden proceder de fuentes diversas y pertenecer a diferentes dominios, parece clara la inminente necesidad de analizar los mismos para la obtención de información útil para la organización.

En muchas situaciones, el método tradicional de convertir los datos en conocimiento consiste en un análisis e interpretación realizada de forma manual. El especialista en la materia, digamos por ejemplo un médico, analiza los datos y elabora un informe o hipótesis que refleja las tendencias o pautas de los mismos. Por ejemplo, un grupo de médicos puede analizar la evolución de enfermedades infecto-contagiosas entre la población para determinar el rango de edad más frecuente de las personas afectadas. Este conocimiento, validado convenientemente, puede ser usado en este caso por la autoridad sanitaria competente para establecer políticas de vacunaciones.

Esta forma de actuar es lenta, cara y altamente subjetiva. De hecho, el análisis manual es impracticable en dominios donde el volumen de los datos crece exponencialmente: la enorme abundancia de datos desborda la capacidad humana de comprenderlos sin la ayuda de herramientas potentes. Consecuentemente, muchas decisiones importantes se realizan, no sobre la base de la gran cantidad de datos disponibles, sino siguiendo la propia intuición del usuario al no disponer de las herramientas necesarias. Éste es el principal cometido de la minería de datos: resolver problemas analizando los datos presentes en las bases de datos.

Por ejemplo, supongamos que una cadena de supermercados quiere ampliar su zona de actuación abriendo nuevos locales. Para ello, la empresa analiza la información disponible en sus bases de datos de clientes para determinar el perfil de los mismos y hace uso de diferentes indicadores demográficos que le permiten determinar los lugares más idóneos para los nuevos emplazamientos. La clave para resolver este problema es analizar los datos para identificar el patrón que define las características de los clientes más fieles y que se usa posteriormente para identificar el número de futuros buenos clientes de cada zona.

Hasta no hace mucho, el análisis de los datos de una base de datos se realizaba mediante consultas efectuadas con lenguajes generalistas de consulta, como el SQL, y se producía sobre la base de datos operacional, es decir, junto al procesamiento transaccional en línea (*On-Line Transaction Processing*, OLTP) de las aplicaciones de gestión. No obstante, esta manera de actuar sólo permitía generar información resumida de una manera previamente establecida (generación de informes), poco flexible y, sobre todo, poco escalable a grandes volúmenes de datos. La tecnología de bases de datos ha respondido a este reto con una nueva arquitectura surgida recientemente: el almacén de datos (*data warehouse*). Se trata de un repositorio de fuentes heterogéneas de datos, integrados y organizados bajo un esquema unificado para facilitar su análisis y dar soporte a la toma de decisiones. Esta tecnología incluye operaciones de procesamiento analítico en línea (*On-Line Analytical Processing*, OLAP), es decir, técnicas de análisis como pueden ser el resumen, la consolidación o la agregación, así como la posibilidad de ver la información desde distintas perspectivas.

Sin embargo, a pesar de que las herramientas OLAP soportan cierto análisis descriptivo y de sumarización que permite transformar los datos en otros datos agregados o cruzados de manera sofisticada, no generan reglas, patrones, pautas, es decir, conocimiento que pueda ser aplicado a otros datos. Sin embargo, en muchos contextos, como los negocios, la medicina o la ciencia, los datos por sí solos tienen un valor relativo. Lo que de verdad es interesante es el conocimiento que puede inferirse a partir de los datos y, más aún, la capacidad de poder usar este conocimiento. Por ejemplo, podemos saber estadísticamente que el 10 por ciento de los ancianos padecen Alzheimer. Esto puede ser útil, pero seguramente es mucho más útil tener un conjunto de reglas que a partir de los antecedentes, los hábitos y otras características del individuo nos digan si un paciente tendrá o no Alzheimer.

Existen otras herramientas analíticas que han sido empleadas para analizar los datos y que tienen su origen en la estadística, algo lógico teniendo en cuenta que la materia prima de esta disciplina son precisamente los datos. Aunque algunos paquetes estadísticos son capaces de inferir patrones a partir de los datos (utilizando modelización estadística paramétrica o no paramétrica), el problema es que resultan especialmente crípticos para los no estadísticos, generalmente no funcionan bien para la talla de las bases de datos actuales (cientos de tablas, millones de registros, talla de varios gigabytes y una alta dimensionalidad) y algunos tipos de datos frecuentes en ellos (atributos nominales con muchos valores, datos textuales, multimedia, etc.), y no se integran bien con los sistemas de información. No obstante, sería injusto no reconocer que la estadística es, en cierto modo, la "madre" de la minería de datos. Ésta se vio, en un principio, como una hija rebelde, que tenía un carácter peyorativo para los estadísticos pero que poco a poco fue ganando un prestigio y una concepción como disciplina integradora más que disgregadora. Más adelante trataremos de esclarecer la relación existente entre la minería de datos y la estadística.

Todos estos problemas y limitaciones de las aproximaciones clásicas han hecho surgir la necesidad de una nueva generación de herramientas y técnicas para soportar la extracción de conocimiento útil desde la información disponible, y que se engloban bajo la denominación de minería de datos. La minería de datos se distingue de las aproximaciones anteriores porque no obtiene información extensional (datos) sino intensional (conocimiento) y, además, el conocimiento no es, generalmente, una parametrización de ningún modelo preestablecido o intuido por el usuario, sino que es un modelo novedoso y original, extraído completamente por la herramienta. El resultado de la minería de datos son conjuntos de reglas, ecuaciones, árboles de decisión, redes neuronales, grafos probabilísticos... , los cuales pueden usarse para, por ejemplo, responder a cuestiones como ¿existe un grupo de clientes que se comporta de manera diferenciada?, ¿qué secuenciación de tratamientos puede ser más efectiva para este nuevo síndrome?, ¿existen asociaciones entre los factores de riesgo para realizar un seguro de automóvil?, ¿cómo califico automáticamente los mensajes de correo entre más o menos susceptibles de ser *spam*?

1.2 El concepto de minería de datos. Ejemplos

En [Witten & Frank 2000] se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ser automático o semi-automático (asistido) y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización.

Por lo tanto, dos son los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos...), y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. En muchos casos la utilidad del conocimiento minado está íntimamente relacionada con la comprensibilidad del modelo inferido. No debemos olvidar que, generalmente, el usuario final no tiene por qué ser un experto en las técnicas de minería de datos, ni tampoco puede perder mucho tiempo interpretando los resultados. Por ello, en muchas aplicaciones es importante hacer que la información descubierta sea más comprensible por los humanos (por ejemplo, usando representaciones gráficas, convirtiendo los patrones a lenguaje natural o utilizando técnicas de visualización de los datos).

De una manera simplista pero ambiciosa, podríamos decir que el objetivo de la minería de datos es convertir datos en conocimiento. Este objetivo no es sólo ambicioso sino muy amplio. Por tanto, de momento, y para ayudar al lector a refinar su concepto de minería de datos, presentamos varios ejemplos muy sencillos.

1.2.1 Ejemplo 1: análisis de créditos bancarios

El primer ejemplo pertenece al ámbito de la banca. Un banco por Internet desea obtener reglas para predecir qué personas de las que solicitan un crédito no lo devuelven. La entidad bancaria cuenta con los datos correspondientes a los créditos concedidos con anterioridad a sus clientes (cuantía del crédito, duración en años...) y otros datos personales como el salario del cliente, si posee casa propia, etc. Algunos registros de clientes de esta base de datos se muestran en la Tabla 1.1.

IDC	D-crédito (años)	C-crédito (euros)	Salario (euros)	Casa propia	Cuentas morosas	...	Devuelve-crédito
101	15	60.000	2.200	sí	2	...	no
102	2	30.000	3.500	sí	0	...	sí
103	9	9.000	1.700	sí	1	...	no
104	15	18.000	1.900	no	0	...	sí
105	10	24.000	2.100	no	0	...	no
...	...	...	...	...	...	...	...

Tabla 1.1. Datos para un análisis de riesgo en créditos bancarios.

A partir de éstos, las técnicas de minería de datos podrían sintetizar algunas reglas, como por ejemplo:

SI Cuentas-Morosas > 0 **ENTONCES** Devuelve-crédito = no
SI Cuentas-Morosas = 0 **Y** [(Salario > 2.500) **O** (D-crédito > 10)] **ENTONCES** Devuelve-crédito = sí

El banco podría entonces utilizar estas reglas para determinar las acciones a realizar en el trámite de los créditos: si se concede o no el crédito solicitado, si es necesario pedir avales especiales, etc.

1.2.2 Ejemplo 2: análisis de la cesta de la compra

Éste es uno de los ejemplos más típicos de minería de datos. Un supermercado quiere obtener información sobre el comportamiento de compra de sus clientes. Piensa que de esta forma puede mejorar el servicio que les ofrece: reubicación de los productos que se suelen comprar juntos, localizar el emplazamiento idóneo para nuevos productos, etc. Para ello dispone de la información de los productos que se adquieren en cada una de las compras o cestas. Un fragmento de esta base de datos se muestra en la Tabla 1.2.

Idcesta	Huevos	Aceite	Pañales	Vino	Leche	Mantequilla	Salmón	Lechugas	...
1	sí	no	no	sí	no	sí	sí	sí	...
2	no	sí	no	no	sí	no	no	sí	...
3	no	no	sí	no	sí	no	no	no	...
4	no	sí	sí	no	sí	no	no	no	...
5	sí	sí	no	no	no	sí	no	sí	...
6	sí	no	no	sí	sí	sí	sí	no	...
7	no	no	no	no	no	no	no	no	...
8	sí	sí	sí	sí	sí	sí	sí	no	...
...	...	...	...	...	...	...	...	...	...

Tabla 1.2. Datos de las cestas de la compra.

Analizando estos datos el supermercado podría encontrar, por ejemplo, que el 100 por cien de las veces que se compran pañales también se compra leche, que el 50 por ciento de las veces que se compran huevos también se compra aceite o que el 33 por ciento de las veces que se compra vino y salmón entonces se compran lechugas. También se puede analizar cuáles de estas asociaciones son frecuentes, porque una asociación muy estrecha entre dos productos puede ser poco frecuente y, por tanto, poco útil.

1.2.3 Ejemplo 3: determinar las ventas de un producto

Una gran cadena de tiendas de electrodomésticos desea optimizar el funcionamiento de su almacén manteniendo un *stock* de cada producto suficiente para poder servir rápidamente el material adquirido por sus clientes. Para ello, la empresa dispone de las ventas efectuadas cada mes del último año de cada producto, tal y como se refleja en la Tabla 1.3.

Producto	mes–12	...	mes–4	mes–3	mes–2	mes–1
televisor plano 30' Phlipis	20	...	52	14	139	74
video-dvd-recorder Miesens	11	...	43	32	26	59
discman mp3 LJ	50	...	61	14	5	28
frigorífico no frost Jazzussi	3	...	21	27	1	49
microondas con grill Sanson	14	...	27	2	25	12
...	...	...	...	...	...	...

Tabla 1.3. Ventas mensuales durante el último año.

Esta información permite a la empresa generar un modelo para predecir cuáles van a ser las ventas de cada producto en el siguiente mes en función de las ventas realizadas en los meses anteriores, y efectuar así los pedidos necesarios a sus proveedores para disponer del *stock* necesario para hacer frente a esas ventas.

1.2.4 Ejemplo 4: determinar grupos diferenciados de empleados

El departamento de recursos humanos de una gran empresa desea categorizar a sus empleados en distintos grupos con el objetivo de entender mejor su comportamiento y tratarlos de manera adecuada. Para ello dispone en sus bases de datos de información sobre los mismos (sueldo, estado civil, si tiene coche, número de hijos, si su casa es propia o de alquiler, si está sindicado, número de bajas al año, antigüedad y sexo). La Tabla 1.4 muestra algunos de los registros de su base de datos.

Id	Sueldo	Casado	Coche	Hijos	Alq/prop	Sindicado	Bajas/año	Antigüedad	Sexo
1	1.000	Sí	No	0	Alquiler	No	7	15	H
2	2.000	No	Sí	1	Alquiler	Sí	3	3	M
3	1.500	Sí	Sí	2	Prop	Sí	5	10	H
4	3.000	Sí	Sí	1	Alquiler	No	15	7	M
5	1.000	Sí	Sí	0	Prop	Sí	1	6	H
6	4.000	No	Sí	0	Alquiler	Sí	3	16	M
7	2.500	No	No	0	Alquiler	Sí	0	8	H
8	2.000	No	Sí	0	Prop	Sí	2	6	M
9	2.000	Sí	Sí	3	Prop	No	7	5	H
10	3.000	Sí	Sí	2	Prop	No	1	20	H
11	5.000	No	No	0	Alquiler	No	2	12	M
12	800	Sí	Sí	2	Prop	No	3	1	H
13	2.000	No	No	0	Alquiler	No	27	5	M
14	1.000	No	Sí	0	Alquiler	Sí	0	7	H
15	8 00	No	Sí	0	Alquiler	No	3	2	H
...	...	...	...	...	...	...	...	...	...

Tabla 1.4. Datos de los empleados.

Un sistema de minería de datos podría obtener tres grupos con la siguiente descripción:

Grupo 1:
Sueldo: 1.535,2€
Casado: No -> 0,777
Sí -> 0,223
Coche: No -> 0,82
Sí -> 0,18
Hijos: 0,05
Alq/Prop: Alquiler -> 0,99
Propia -> 0,01
Sindic.: No -> 0,8
Sí -> 0,2
Bajas/Año: 8,3
Antigüedad: 8,7
Sexo: H -> 0,61
M -> 0,39

Grupo 2:
Sueldo: 1.428,7€
Casado: No -> 0,98
Sí -> 0,02
Coche: No -> 0,01
Sí -> 0,99
Hijos: 0,3
Alq/Prop: Alquiler -> 0,75
Propia -> 0,25
Sindic.: Sí -> 1,0
Bajas/Año: 2,3
Antigüedad: 8
Sexo: H -> 0,25
M -> 0,75

Grupo 3:
Sueldo: 1.233,8€
Casado: Sí -> 1,0
Coche: No -> 0,05
Sí -> 0,95
Hijos: 2,3
Alq/Prop: Alquiler -> 0,17
Propia -> 0,83
Sindic.: No -> 0,67
Sí -> 0,33
Bajas/Año: 5,1
Antigüedad: 8,1
Sexo: H -> 0,83
M -> 0,17

Estos grupos podrían ser interpretados por el departamento de recursos humanos de la siguiente manera:

- Grupo 1: sin hijos y con vivienda de alquiler. Poco sindicados. Muchas bajas.
- Grupo 2: sin hijos y con coche. Muy sindicados. Pocas bajas. Normalmente son mujeres y viven en casas de alquiler.
- Grupo 3: con hijos, casados y con coche. Mayoritariamente hombres propietarios de su vivienda. Poco sindicados.

1.3 Tipos de datos

Llegados a este punto surge una pregunta obligada, ¿a qué tipo de datos puede aplicarse la minería de datos? En principio, ésta puede aplicarse a cualquier tipo de información, siendo las técnicas de minería diferentes para cada una de ellas. En esta sección damos una breve introducción a algunos de estos tipos. En concreto, vamos a diferenciar entre datos estructurados provenientes de bases de datos relacionales, otros tipos de datos estructurados en bases de datos (espaciales, temporales, textuales y multimedia) y datos no estructurados provenientes de la web o de otros tipos de repositorios de documentos.

1.3.1 Bases de datos relacionales

Una base de datos relacional es una colección de relaciones (tablas). Cada tabla consta de un conjunto de atributos (columnas o campos) y puede contener un gran número de tuplas (registros o filas). Cada tupla representa un objeto, el cual se describe a través de los valores de sus atributos y se caracteriza por poseer una clave única o primaria que lo identifica. Por ejemplo, la Figura 1.1 ilustra una base de datos con dos relaciones: *empleado* y *departamento*. La relación *empleado* tiene seis atributos: el identificador o clave primaria (*IdE*), el nombre del empleado (*Enombre*), su sueldo (*Sueldo*), su edad (*Edad*), su sexo (*Sexo*) y el departamento en el que trabaja (*IdD*), y la relación *departamento* tiene tres atributos: su identificador o clave primaria (*IdD*), el nombre (*Dnombre*) y su director (*Director*). Una relación puede además tener claves ajenas, es decir, atributos que hagan referencia a otra relación, como por ejemplo el sexto atributo de la relación *empleado*, *IdD*, que hace referencia (por valor) al *IdD* de *departamento*.

empleado					
IdE	*Enombre*	*Sueldo*	*Edad*	*Sexo*	*IdD*
1	Juan	2.100	45	H	Ge
2	Elena	2.400	40	M	Ma
3	Maria	?	53	M	Ge
4	Pedro	1.000	20	H	Ge
5	Lucía	3.500	35	M	Ma

departamento		
IdD	*Dnombre*	*Director*
Ge	Gestión	Rubio
Ve	Ventas	Burriel
Ma	Márketing	Torrubia

Figura 1.1. Base de datos relacional.

Una de las principales características de las bases de datos relacionales es la existencia de un esquema asociado, es decir, los datos deben seguir una estructura y son, por tanto, estructurados. Así, el esquema de la base de datos del ejemplo indica que las tuplas de la relación *empleado* tienen un valor para cada uno de sus seis atributos y las de la relación *departamento* constan de tres valores, además de indicar los tipos de datos (numérico, cadena de caracteres, etc.).

La integridad de los datos se expresa a través de las restricciones de integridad. Éstas pueden ser de dominio (restringen el valor que puede tomar un atributo respecto a su dominio y si puede tomar valores nulos o no), de identidad (por ejemplo la clave primaria tiene que ser única) y referencial (los valores de las claves ajenas se deben corresponder con uno y sólo un valor de la tabla referenciada).

Como se ha comentado en la Sección 1.1, la obtención de información desde una base de datos relacional se ha resuelto tradicionalmente a través de lenguajes de consulta especialmente diseñados para ello, como SQL. La Figura 1.2 muestra una consulta típica SQL sobre una relación *empleado* que lista la media de edad de todos los empleados de una empresa cuyo sueldo es mayor de 2.000 euros, agrupado por departamento.

```
SELECT D.IdD, D.DNombre, AVG(E.edad)
FROM empleado E JOIN departamento D ON E.IdD=D.IdD
WHERE E.sueldo > 2.000
GROUP BY D.IdD, D.DNombre
```

Resultado de la consulta		
IdD	Dnombre	AVG_edad
Ge	Gestión	45,0
Ma	Marketing	37,5

Figura 1.2. Consulta SQL.

Aunque las bases de datos relacionales (recogidas o no en un almacén de datos, normalizadas o estructuradas de una manera multidimensional) son la fuente de datos para la mayoría de aplicaciones de minería de datos, muchas técnicas de minería de datos no son capaces de trabajar con *toda* la base de datos, sino que sólo son capaces de tratar con una sola tabla a la vez. Lógicamente, mediante una consulta (por ejemplo en SQL, en una base de datos relacional tradicional, o con herramientas y operadores más potentes, en los almacenes de datos) podemos combinar en una sola tabla o *vista minable* aquella información de varias tablas que requiramos para cada tarea concreta de minería de datos. Por tanto, la presentación tabular, también llamada atributo-valor, es la más utilizada por las técnicas de minería de datos.

En esta presentación tabular, es importante conocer los tipos de los atributos y, aunque en bases de datos existen muchos tipos de datos (enteros, reales, fechas, cadenas de texto, etc.), desde el punto de vista de las técnicas de minería de datos más habituales nos interesa distinguir sólo entre dos tipos, numéricos y categóricos. Para tratar otras representaciones más complejas, como las cadenas de caracteres, los tipos textuales "memo", los vectores, y otras muchas, harán falta técnicas específicas.

- Los atributos *numéricos* contienen valores enteros o reales. Por ejemplo, atributos como el *salario* o la *edad* son numéricos.
- Los atributos *categóricos* o *nominales* toman valores en un conjunto finito y preestablecido de categorías. Por ejemplo, atributos como el *sexo* (H, M), el nombre del departamento (Gestión, Márketing, Ventas) son categóricos.

Incluso sólo considerando estos dos tipos de datos, no todas las técnicas de minería de datos son capaces de trabajar con ambos tipos. Este hecho puede requerir la aplicación de un proceso previo de transformación/preparación de los datos del que hablaremos en el Capítulo 4.

1.3.2 Otros tipos de bases de datos

Aunque las bases de datos relacionales son, con gran diferencia, las más utilizadas hoy en día, existen aplicaciones que requieren otros tipos de organización de la información. Otros tipos de bases de datos que contienen datos complejos son:

- Las bases de datos espaciales contienen información relacionada con el espacio físico en un sentido amplio (una ciudad, una región montañosa, un atlas cerebral...). Estas bases de datos incluyen datos geográficos, imágenes médicas, redes de transporte o información de tráfico, etc., donde las relaciones espaciales son muy relevantes. La minería de datos sobre estas bases de datos permite encontrar patrones entre los datos, como por ejemplo las características de las casas en una zona montañosa, la planificación de nuevas líneas de metro en función de la distancia de las distintas áreas a las líneas existentes, etc.
- Las bases de datos temporales almacenan datos que incluyen muchos atributos relacionados con el tiempo o en el que éste es muy relevante. Estos atributos pueden referirse a distintos instantes o intervalos temporales. En este tipo de bases de datos las técnicas de minería de datos pueden utilizarse para encontrar las características de la evolución o las tendencias del cambio de distintas medidas o valores de la base de datos.
- Las bases de datos documentales contienen descripciones para los objetos (documentos de texto) que pueden ir desde las simples palabras clave a los resúmenes. Estas bases de datos pueden contener documentos no estructurados (como una biblioteca digital de novelas), semi-estructurados (si se puede extraer la información por partes, con índices, etc.) o estructurados (como una base de datos de fichas bibliográficas). Las técnicas de minería de datos pueden utilizarse para obtener asociaciones entre los contenidos, agrupar o clasificar objetos textuales. Para ello, los métodos de minería se integran con otras técnicas de recuperación de información y con la construcción o uso de jerarquías específicas para datos textuales, como los diccionarios y los tesauros.
- Las bases de datos multimedia almacenan imágenes, audio y vídeo. Soportan objetos de gran tamaño ya que, por ejemplo, los vídeos pueden necesitar varios gigabytes de capacidad para su almacenamiento. Para la minería de estas bases de datos también es necesario integrar los métodos de minería con técnicas de búsqueda y almacenamiento.

Las bases de datos objetuales y las objeto-relacionales son aproximaciones generales a la gestión de la información y, por tanto, pueden utilizarse para los mismos usos que las relacionales o para algunas de las bases de datos especiales que acabamos de ver.

1.3.3 La *World Wide Web*

La World Wide Web es el repositorio de información más grande y diverso de los existentes en la actualidad. Por ello, hay gran cantidad de datos en la web de los que se puede extraer conocimiento relevante y útil. Éste es precisamente el reto al que se enfrenta la "minería web". Minar la web no es un problema sencillo, debido a que muchos de los datos son no estructurados o semi-estructurados, a que muchas páginas web contienen datos multimedia (texto, imágenes, vídeo y/o audio), y a que estos datos pueden residir en diversos servidores o en archivos (como los que contienen los *logs*). Otros aspectos que dificultan la minería web son cómo determinar a qué páginas debemos acceder y cómo seleccionar la información que va a ser útil para extraer conocimiento. Toda esta diversidad hace que la minería web se organice en torno a tres categorías:

- minería del contenido, para encontrar patrones de los datos de las páginas web.
- minería de la estructura, entendiendo por estructura los hipervínculos y URLs.
- minería del uso que hace el usuario de las páginas web (navegación).

La mayoría de las técnicas de la Parte III de este libro se enfocan a bases de datos tradicionales. La Parte V se dedica a presentar técnicas específicas o adaptar las tradicionales para aboradar los otros tipos de datos comentados: espaciales, temporales, documentales y textuales, multimedia y la web.

1.4 Tipos de modelos

La minería de datos tiene como objetivo analizar los datos para extraer conocimiento. Este conocimiento puede ser en forma de relaciones, patrones o reglas inferidos de los datos y (previamente) desconocidos, o bien en forma de una descripción más concisa (es decir, un resumen de los mismos). Estas relaciones o resúmenes constituyen el modelo de los datos analizados. Existen muchas formas diferentes de representar los modelos y cada una de ellas determina el tipo de técnica que puede usarse para inferirlos.

En la práctica, los modelos pueden ser de dos tipos: *predictivos* y *descriptivos*. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que denominamos *variables objetivo* o *dependientes*, usando otras variables o campos de la base de datos, a las que nos referiremos como *variables independientes* o *predictivas*. Por ejemplo, un modelo predictivo sería aquel que permite estimar la demanda de un nuevo producto en función del gasto en publicidad.

Los modelos descriptivos, en cambio, identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Por ejemplo, una agencia de viaje desea identificar grupos de personas con unos mismos gustos, con el objeto de organizar diferentes ofertas para cada grupo y poder así remitirles esta información; para ello analiza los viajes que han realizado sus clientes e infiere un modelo descriptivo que caracteriza estos grupos.

Como veremos en el Capítulo 2 y, con más detalle, en el Capítulo 6, algunas tareas de minería de datos que producen modelos predictivos son la clasificación y la regresión, y las que dan lugar a modelos descriptivos son el agrupamiento, las reglas de asociación y el análisis correlacional. Por ejemplo, los problemas planteados en los ejemplos 1 y 3 de la Sección 1.2 se resolverían mediante modelos predictivos, mientras que los de los ejemplos 2 y 4 usarían modelos descriptivos.

Cada tarea puede ser realizada usando distintas técnicas. Por ejemplo, los modelos inferidos por los árboles de decisión y las redes neuronales (por citar dos técnicas de las más conocidas y utilizadas) pueden inferir modelos predictivos. Igualmente, para una misma técnica se han desarrollado diferentes algoritmos que difieren en la forma y criterios concretos con los que se construye el modelo. Todas estas cuestiones serán abordadas en capítulos sucesivos. Así, el Capítulo 2 presenta una relación de las técnicas de generación de modelos más significativas. Asimismo, en la Parte III del libro se hace un estudio más profundo de algunas de estas técnicas y de los diferentes algoritmos desarrollados en ellas.

1.5 La minería de datos y el proceso de descubrimiento de conocimiento en bases de datos

Existen términos que se utilizan frecuentemente como sinónimos de la minería de datos. Uno de ellos se conoce como "análisis (inteligente) de datos" (véase por ejemplo [Berthold & Hand 2003]), que suele hacer un mayor hincapié en las técnicas de análisis estadístico. Otro término muy utilizado, y el más relacionado con la minería de datos, es la extracción o "descubrimiento de conocimiento en bases de datos" (*Knowledge Discovery in Databases, KDD*). De hecho, en muchas ocasiones ambos términos se han utilizado indistintamente, aunque existen claras diferencias entre los dos. Así, últimamente se ha usado el término KDD para referirse a un proceso que consta de una serie de fases, mientras que la minería de datos es sólo una de estas fases.

En [Fayyad et al. 1996a] se define el KDD como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos". En esta definición se resumen cuáles deben ser las propiedades deseables del conocimiento extraído:

- válido: hace referencia a que los patrones deben seguir siendo precisos para datos nuevos (con un cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención.
- novedoso: que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario.
- potencialmente útil: la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.
- comprensible: la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento (al menos desde el punto de vista de su utilidad).

Como se deduce de la anterior definición, el KDD es un proceso complejo que incluye no sólo la obtención de los modelos o patrones (el objetivo de la minería de datos), sino también la evaluación y posible interpretación de los mismos, tal y como se refleja en la Figura 1.3.

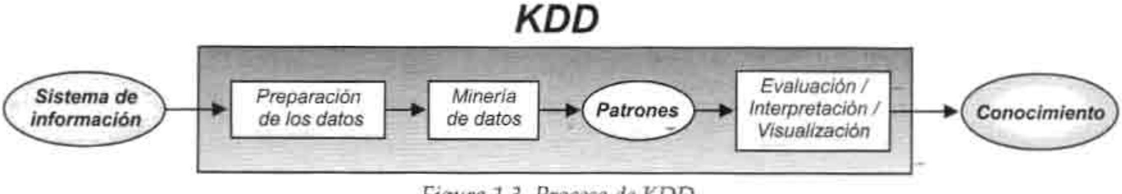

Figura 1.3. Proceso de KDD.

Así, los sistemas de KDD permiten la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento; consolidar el conocimiento resolviendo posibles conflictos con conocimiento previamente extraído; y hacer el conocimiento disponible para su uso. Esta definición del proceso clarifica la relación entre el KDD y la minería de datos: el KDD es el proceso global de descubrir conocimiento útil desde las bases de datos mientras que la minería de datos se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos. Al ser la fase de generación de modelos, comúnmente se asimila KDD con minería de datos. Además, las connotaciones de aventura y de dinero fácil del término "minería de datos" han hecho que éste se use como identificador del área, especialmente en el ámbito empresarial.

1.6 Relación con otras disciplinas

La minería de datos es un campo multidisciplinar que se ha desarrollado en paralelo o como prolongación de otras tecnologías. Por ello, la investigación y los avances en la minería de datos se nutren de los que se producen en estas áreas relacionadas.

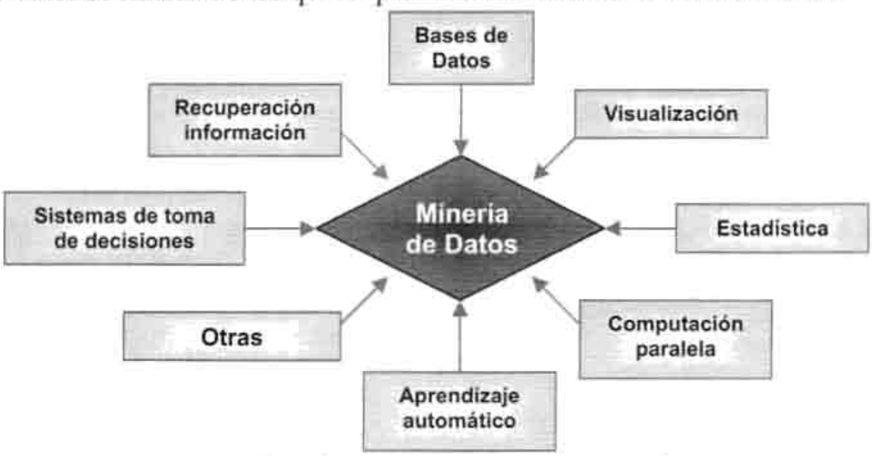

Figura 1.4. Disciplinas que contribuyen a la minería de datos.

Podemos destacar como disciplinas más influyentes las siguientes (Figura 1.4):

- **las bases de datos**: conceptos como los almacenes de datos y el procesamiento analítico en línea (OLAP) tienen una gran relación con la minería de datos, aunque en este último caso no se trata de obtener informes avanzados a base de agregar los datos de cierta manera compleja pero predefinida (como incluyen muchas herramientas de *business intelligence*, presentes en sistemas de gestión de bases de datos comerciales), sino de extraer conocimiento novedoso y comprensible. Las técnicas de indización[1] y de acceso eficiente a los datos son muy relevantes para el diseño de algoritmos eficientes de minería de datos.
- **la recuperación de información** (*information retrieval, IR*): consiste en obtener información desde datos textuales, por lo que su desarrollo histórico se ha basado en el uso efectivo de bibliotecas (recientemente digitales) y en la búsqueda por Internet. Una tarea típica es encontrar documentos a partir de palabras claves, lo cual

[1] Es el término correcto para "*indexing*", aunque frecuentemente se utiliza el término "indexación".

puede verse como un proceso de clasificación de los documentos en función de estas palabras clave. Para ello se usan medidas de similitud entre los documentos y la consulta. Muchas de estas medidas se han empleado en aplicaciones más generales de minería de datos.

- **la estadística**: esta disciplina ha proporcionado muchos de los conceptos, algoritmos y técnicas que se utilizan en minería de datos, como por ejemplo, la media, la varianza, las distribuciones, el análisis univariante y multivariante, la regresión lineal y no lineal, la teoría del muestreo, la validación cruzada, la modelización paramétrica y no paramétrica, las técnicas bayesianas, y un largo etcétera. De hecho, algunos paquetes de análisis estadístico se comercializan como herramientas de minería de datos.
- **el aprendizaje automático**: ésta es el área de la inteligencia artificial que se ocupa de desarrollar algoritmos (y programas) capaces de aprender, y constituye, junto con la estadística, el corazón del análisis inteligente de los datos. Los principios seguidos en el aprendizaje automático y en la minería de datos son los mismos: la máquina aprende un modelo a partir de ejemplos y lo usa para resolver el problema.
- **los sistemas para la toma de decisión**: son herramientas y sistemas informatizados que asisten a los directivos en la resolución de problemas y en la toma de decisiones. El objetivo es proporcionar la información necesaria para realizar decisiones efectivas en el ámbito empresarial o en tareas de diagnóstico (por ejemplo en medicina). Herramientas como el análisis ROC (ver Capítulo 17.) o los mismos árboles de decisión provienen de esta área.
- **la visualización de datos**: el uso de técnicas de visualización permite al usuario descubrir, intuir o entender patrones que serían más difíciles de "ver" a partir de descripciones matemáticas o textuales de los resultados. Existen técnicas de visualización, como, por ejemplo, las gráficas (diagramas de barras, gráficas de dispersión, histogramas, etc.), las icónicas (basadas en figuras, colores, etc.), las basadas en píxeles (cada dato se representa como un único píxel), las jerárquicas (dividiendo el área de representación en regiones dependiendo de los datos) y muchas otras.
- **la computación paralela y distribuida**: actualmente, muchos sistemas de bases de datos comerciales incluyen tecnologías de procesamiento paralelo, distribuido o de computación en *grid*. En estos sistemas el coste computacional de las tareas más complejas de minería de datos se reparte entre diferentes procesadores o computadores. Su éxito se debe en parte a la explosión de los almacenes de datos (su adaptación distribuida) y de la minería de datos, en los que las prestaciones de los algoritmos de consulta son críticas. Una de las principales ventajas del procesamiento paralelo es precisamente la escalabilidad de los algoritmos, lo que lo hace idóneo para estas aplicaciones.
- **otras disciplinas**: dependiendo del tipo de datos a ser minados o del tipo de aplicación, la minería de datos usa también técnicas de otras disciplinas como el lenguaje natural, el análisis de imágenes, el procesamiento de señales, los gráficos por computadora, etc.

1.7 Aplicaciones

La integración de las técnicas de minería de datos en las actividades del día a día se está convirtiendo en algo habitual. Los negocios de la distribución y la publicidad dirigida han sido tradicionalmente las áreas en las que más se han empleado los métodos de minería, ya que han permitido reducir costes o aumentar la receptividad de ofertas. Pero éstas no son las únicas áreas a las que se pueden aplicar. De hecho, podemos encontrar ejemplos en todo tipo de aplicaciones: financieras, seguros, científicas (medicina, farmacia, astronomía, psicología, etc.), políticas económicas, sanitarias o demográficas, educación, policiales, procesos industriales y un largo etcétera.

Siendo un poco más concretos, a continuación incluimos una lista (que en modo alguno pretendemos que sea exhaustiva) de ejemplos en algunas de las áreas antes mencionadas para ilustrar en qué ámbitos se puede usar la minería de datos.

- Aplicaciones financieras y banca:
  - Obtención de patrones de uso fraudulento de tarjetas de crédito.
  - Determinación del gasto en tarjeta de crédito por grupos.
  - Cálculo de correlaciones entre indicadores financieros.
  - Identificación de reglas de mercado de valores a partir de históricos.
  - Análisis de riesgos en créditos.
- Análisis de mercado, distribución y, en general, comercio:
  - Análisis de la cesta de la compra (compras conjuntas, secuenciales, ventas cruzadas, señuelos, etc.).
  - Evaluación de campañas publicitarias.
  - Análisis de la fidelidad de los clientes. Reducción de fuga.
  - Segmentación de clientes.
  - Estimación de *stocks*, de costes, de ventas, etc.
- Seguros y salud privada:
  - Determinación de los clientes que podrían ser potencialmente caros.
  - Análisis de procedimientos médicos solicitados conjuntamente.
  - Predicción de qué clientes contratan nuevas pólizas.
  - Identificación de patrones de comportamiento para clientes con riesgo.
  - Identificación de comportamiento fraudulento.
  - Predicción de los clientes que podrían ampliar su póliza para incluir procedimientos extras (dentales, ópticos...).
- Educación:
  - Selección o captación de estudiantes.
  - Detección de abandonos y de fracaso.
  - Estimación del tiempo de estancia en la institución.
- Procesos industriales:
  - Extracción de modelos sobre comportamiento de compuestos.
  - Detección de piezas con trabas. Modelos de calidad.

  - Predicción de fallos y accidentes.
  - Estimación de composiciones óptimas en mezclas.
  - Extracción de modelos de coste.
  - Extracción de modelos de producción.
- Medicina:
  - Identificación de patologías. Diagnóstico de enfermedades.
  - Detección de pacientes con riesgo de sufrir una patología concreta.
  - Gestión hospitalaria y asistencial. Predicciones temporales de los centros asistenciales para el mejor uso de recursos, consultas, salas y habitaciones.
  - Recomendación priorizada de fármacos para una misma patología.
- Biología, bioingeniería y otras ciencias:
  - Análisis de secuencias de genes.
  - Análisis de secuencias de proteínas.
  - Predecir si un compuesto químico causa cáncer.
  - Clasificación de cuerpos celestes.
  - Predicción de recorrido y distribución de inundaciones.
  - Modelos de calidad de aguas, indicadores ecológicos.
- Telecomunicaciones:
  - Establecimiento de patrones de llamadas.
  - Modelos de carga en redes.
  - Detección de fraude.
- Otras áreas
  - Correo electrónico y agendas personales: clasificación y distribución automática de correo, detección de correo *spam*, gestión de avisos, análisis del empleo del tiempo.
  - Recursos Humanos: selección de empleados.
  - Web: análisis del comportamiento de los usuarios, detección de fraude en el comercio electrónico, análisis de los *logs* de un servidor web.
  - Turismo: determinar las características socioeconómicas de los turistas en un determinado destino o paquete turístico, identificar patrones de reservas, etc.
  - Tráfico: modelos de tráfico a partir de fuentes diversas: cámaras, GPS...
  - Hacienda: detección de evasión fiscal.
  - Policiales: identificación de posibles terroristas en un aeropuerto.
  - Deportes: estudio de la influencia de jugadores y de cambios. Planificación de eventos.
  - Política: diseño de campañas políticas, estudios de tendencias de grupos, etc.

Todos estos ejemplos muestran la gran variedad de aplicaciones donde el uso de la minería de datos puede ayudar a entender mejor el entorno donde se desenvuelve la organización y, en definitiva, mejorar la toma de decisiones en dicho entorno.

1.8 Sistemas y herramientas de minería de datos

La diversidad de disciplinas que contribuyen a la minería de datos está dando lugar a una gran variedad de sistemas de minería de datos. Cada uno de ellos posee unas características apropiadas para realizar determinadas tareas o para analizar cierto tipo de datos. En esta sección presentamos varias clasificaciones de los sistemas y herramientas atendiendo a varios criterios (el modelo de datos que generan, el tipo de datos que minan, el tipo de técnica o el tipo de aplicación al que se pueden aplicar):

- Tipo de base de datos minada: teniendo en cuenta los diferentes modelos de datos podemos hablar de sistemas de minería de datos relacionales, multidimensionales, orientados a objetos, etc. Asimismo, atendiendo al tipo de datos manejados, hablamos de sistemas textuales, multimedia, espaciales o web.
- Tipo de conocimiento minado: también pueden tenerse en cuenta los niveles de abstracción del conocimiento minado: conocimiento generalizado (alto nivel de abstracción), a nivel primitivo (a nivel de filas de datos), o conocimiento a múltiples niveles (de abstracción). Por último, podemos igualmente distinguir entre los sistemas que buscan regularidades en los datos (patrones) frente a los que analizan las irregularidades (excepciones).
- Tipo de funcionalidad y de técnica: los sistemas de minería de datos se pueden clasificar basándose en su funcionalidad (clasificación, agrupamiento, etc.) o por los métodos de análisis de los datos empleados (técnicas estadísticas, redes neuronales, etc.).
- Tipo de aplicación: podemos distinguir dos clases de sistemas según la aplicación para la que se usan: los sistemas de propósito general y los sistemas específicos (como los usados en aplicaciones financieras, web, e-mail, análisis de *stocks*, etc.).

A finales de la década de los 90 aparecen las *suites*, que son capaces de trabajar con distintos formatos, de incorporar distintas técnicas, de obtener distintos tipos de conocimiento y de aplicarse a un gran abanico de áreas. A esto nos referimos cuando hablamos de sistemas integrados o paquetes de minería de datos.

En el Apéndice A se ha incluido una descripción de algunos de los sistemas y herramientas que consideramos más importantes y representativos.