MINERÍA DE DATOS

Yolanda Belinchón Monjas
Ingeniería de Telecomunicación
Universidad Carlos III de Madrid
Madrid, España

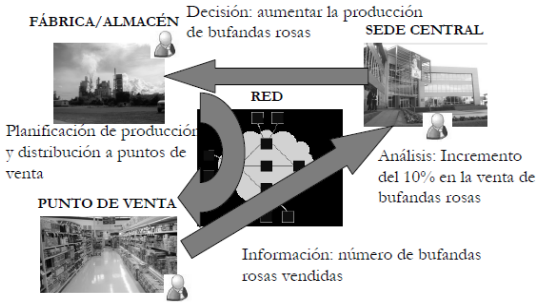
100060624@alumnos.uc3m.es

1. INTRODUCCIÓN

En la actualidad, nuestra sociedad es considerada la sociedad de la información, donde las tecnologías que ayudan a la creación, distribución y manipulación de información, facilitan las actividades sociales, culturales y económicas.

Alrededor del año 1999 se inició un cambio en las sociedades en cuanto a la manera de generar la riqueza, que se fue trasladando de los sectores industriales a los sectores de servicios. La mayor parte de los empleos estarán asociados a la generación, almacenamiento y procesamiento de todo tipo de información. Los sectores relacionados con las tecnologías de la información y la comunicación (TIC) desempeñan un papel particularmente importante dentro de esta sociedad.

Algunos sistemas que son sólo parcialmente conocidos, producen una cantidad inmensa de datos, datos que con frecuencia contienen información valiosa que puede resultar muy útil a ejecutivos de una empresa, a la hora de la toma de decisiones y de resolver problemas de negocio como:



- Generación de recomendaciones (¿Qué servicios se deberían ofrecer a los clientes?)
- Detección de anomalías (fraudes)
- Gestión de riesgos
- Segmentación de clientes
- Personalización de la publicidad
- Previsión
- …

Las dimensiones de las bases de datos grandes y sus velocidades de crecimiento, hacen muy difícil al ser humano su análisis y la extracción de alguna información importante. Aún con el uso de herramientas estadísticas clásicas esta tarea es casi imposible.

El descubrimiento de conocimiento en base de datos (KDD), que se esplli9cará con mayor exactitud en el siguiente punto, combina las técnicas tradicionales con numerosos recursos desarrollados en el área de la inteligencia artificial.

En estos casos habrá una parte del sistema que es conocida y habrá una parte aparentemente de naturaleza aleatoria. Bajo ciertas circunstancias, a partir de una gran cantidad de datos asociada con el sistema, existe la posibilidad de encontrar nuevos aspectos previamente desconocidos del modelo.

Por todo ello, en sistemas donde una parte es conocida y otra de naturaleza aleatoria con el fin de extraer conocimiento útil y comprensible, previamente desconocido, de grandes cantidades de datos almacenados en distintos formatos, aparece lo que conocemos como la Minería de Datos (DM, Data Mining).

2. DEFINICIÓN

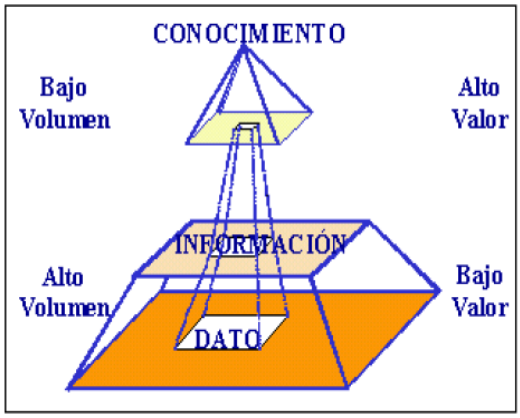
La minería de datos se encarga de preparar, sondear y explorar los datos para sacar la información oculta y útil en ellos. Si los datos son leídos y analizados, pueden proporcionar, en conjunto, un verdadero conocimiento (futuras tendencias y comportamientos) que ayude en la toma de decisiones, ya que para el responsable de un sistema, los datos en sí no son lo más relevante, sino la información que se encierra en sus relaciones, fluctuaciones y dependencias.

Se conoce como minería de datos a todo un conjunto de técnicas encargas de la extracción de conocimiento procesable, implícito en las bases de datos (ayuda a comprender su contenido). Está fuertemente ligada con la supervisión de procesos industriales, pues resulta muy útil para aprovechar los datos almacenados en las bases de datos.

Las bases de la minería de datos se encuentran en la inteligencia artificial, el análisis estadístico, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo. Mediante la utilización de

técnicas de minería de datos se puede dar solución a problemas de predicción, clasificación y segmentación.

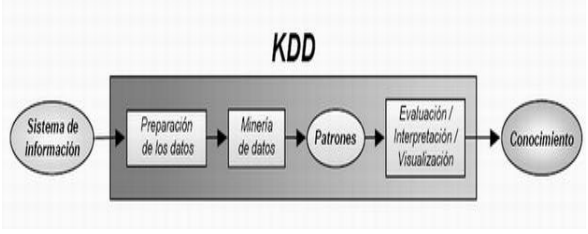
Los datos son la materia prima bruta y en el momento en el que el ser humano les atribuye un significado pasan a convertirse en información y cuando se encuentra un modelo que interpretando la información represente un valor agregado, se denomina conocimiento. En la siguiente figura se muestra la jerarquía que sigue una base de datos, entre datos y conocimiento:



Como podemos observar a medida que subimos de nivel el volumen de datos disminuye, puesto que cuanto más alto estemos en la pirámide, necesitaremos información más específica y procesada. El data mining trabaja en los niveles superiores buscando patrones, comportamientos, secuencias, tendencias o asociaciones que puedan generar algún modelo que nos permita comprender mejor el negocio, a través de una combinación de tareas como: Extracción de datos, limpieza de datos, selección de características, análisis de resultados,...

El término data mining se considera una etapa dentro de un proceso mayor llamado extracción o descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases o KDD). Aunque algunos autores usan los términos Minería de Datos y KDD indistintamente, como sinónimos, existen claras diferencias entre los dos. KDD como se ha comentado es un proceso que consta de un conjunto de fases, una de las cuales es la minería de datos, por lo tanto se denomina KDD al proceso completo que incluye pre-procesamiento, minería y post-procesamiento de los datos.

La siguiente figura muestra las fases del proceso de KDD:



Como puede observarse la minería de datos es una de las fases del proceso de KDD, como ya se ha comentado.

Algunas de las tareas más frecuentes en procesos de KDD son la clasificación y clustering, el reconocimiento de patrones, las predicciones y la detección de dependencias o relaciones entre los datos.

3. TAREAS DE MINERÍA DE DATOS

La principal fase del proceso de la minería de datos es el descubrimiento de reglas, las cuales mostrarán nuevas relaciones entre las variables o excepciones según el negocio que utilice este proceso. Puede ocurrir que algunas de las reglas descubiertas no puedan ser cambiadas, pero si modificadas para mejorar su desempeño.

Una vez descubiertas reglas importantes, pueden ser utilizadas para estimar algunas variables de salida. En esta técnica se complementan las técnicas estadísticas tradicionales con aquellas provenientes de la inteligencia artificial.

Conceptos adaptativos como los algoritmos genéticos y las redes neuronales, permiten realizar predicciones más acertadas, especialmente en casos de gran complejidad.

Entre las principales tareas de la minería de datos se encuentran:

1. **Tareas descriptivas:**

   Orientadas a describir un conjunto de datos.

   - **Clasificación:**

     Se asigna una categoría a cada caso. Cada caso tiene un conjunto de atributos, donde uno de ellos es el atributo clase.

     Se busca un modelo que describa el atributo clase como una función de los atributos de

salida. Existen principalmente dos tipos de clasificación:

- Clasificación basada en árboles de decisión.
- Clasificación neuronal.

- **Segmentación (agrupación):**

Esta tarea también es conocida como segmentación, y se encarga de identificar grupos naturales basándose en un conjunto de atributos. Existen diversas técnicas:

-Clustering: El número de segmentos se determina durante la ejecución del algoritmo. Procesa bien tanto las variables cualitativas como las cuantitativas.

-Segmentación neuronal: Es necesario definir antes de la ejecución del algoritmo el número de segmentos y su distribución bidimensional. Procesa mejor las variables cuantitativas que las cualitativas

- **Asociación:**

Organizar según relaciones entre atributos (Análisis de la cesta de la compra).

Expresa las afinidades entre elementos siguiendo el modelo de las reglas de asociación X→Y, facilitando una serie de métricas como el soporte y confianza.

-Regresión:

Tarea muy similar a la de clasificación pero con el objetivo de buscar patrones para determinar su valor único.

2. **Tareas Predictivas:**

Orientadas a estimar valores de salida.

-Previsión:

A partir de la entrada, conjunto de valores obtenidos a lo largo de un tiempo determinado de los que se extrae un comportamiento futuro. Para la estimación de variables cuantitativas, los métodos más usados son:

-Funciones de base radial: Tienen la capacidad de poder procesar variables cualitativas y cuantitativas a la vez.

-Predicción neuronal.

-Análisis de secuencia:

Se encarga de la búsqueda de patrones en una serie de eventos denominados secuencias o transacciones, lo que permite optimizar las ventas a lo largo del tiempo

-Análisis de desviaciones:

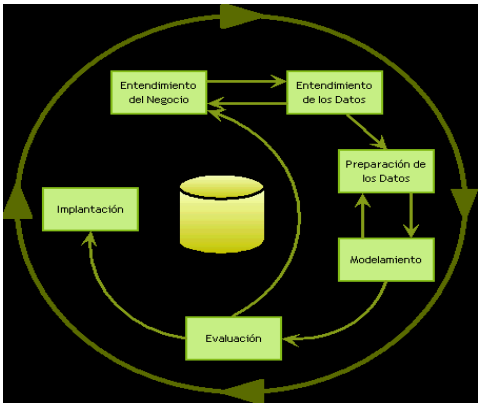
Busca datos distintos, raros, diferentes en comparación con el resto de los datos obtenidos.

-Análisis de similitud en series temporales:
Detecta todas las ocurrencias de secuencias similares en una colección de series temporales.

4. CICLO DE UN PROYECTO DE MINERÍA DE DATOS

Los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento.

A la hora de implantar la técnica de minería de datos en un determinado proyecto, hay seguir el siguiente ciclo:



1. **Entendimiento del negocio:**

   Formulación del problema de negocio (uno de los ya antes mencionados: previsión, gestión de riesgos, segmentación de clientes,…).

2. **Entendimiento de los datos:**

   Recolección de datos.

3. **Preparación de los datos:**

   - Transformación de datos: Generalmente, el formato de los datos contenidos en las fuentes de datos no es el idóneo, y la mayoría de las veces no es posible aplicar algún algoritmo de minería sobre los datos iniciales sin que requieran algún cambio (Por ejemplo, transformaciones numéricas).
   - Limpieza o filtrado de datos: En esta fase se filtran los datos con el objetivo de eliminar valores erróneos o desconocidos, según las necesidades y el algoritmo a utilizar.

   - Preprocesado**:** Se analizan las propiedades de los datos, en especial los histogramas,

diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos) y se obtienen muestras de los datos en busca de mayor velocidad y eficiencia de los algoritmos o se reduce el número de valores posibles, mediante tareas como:

- Redondeo
- Agrupación
- Agregación

4. **Modelado:**

Creación del modelo.

- Selección de variables: Después de haber sido preprocesados y realizar la limpieza de datos, se sigue teniendo una cantidad enorme de variables o atributos.
La selección de características reduce el tamaño de los datos, eligiendo las variables más influyentes del problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.
Los métodos para la selección de los atributos que más influencia tienen en el problema son básicamente dos:

- Aquellos basados en la elección de los mejores atributos del problema.
- Aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

- Extracción de Conocimiento**:** La extracción del conocimiento es la esencia de la Minería de Datos donde mediante una técnica, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. Los modelos que se generan son expresados de diversas formas:

- Reglas
- Árboles
- Redes neuronales

También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre-procesado diferente de los datos. Normalmente se suele seguir el procedimiento de prueba y error.

5. **Evaluación:**

Evaluación de la integridad del modelo en el negocio.

Una vez obtenido el modelo, se procede a su validación; comprobando que las conclusiones obtenidas son válidas y satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos para buscar el que mejor se ajuste al problema.

Si ninguno de los modelos alcanza los resultados esperados, debe modificarse alguna de las fases anteriores para generar nuevos modelos. Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un modelo válido.

Una vez validado el modelo, si resulta ser aceptable (proporciona salidas adecuadas y/o con márgenes de error admisibles) éste ya está listo para su explotación e implantación.

6. **Implantación:**

Integración en aplicaciones para solucionar el problema de negocio expuesto.

5. TÉCNICAS

Como ya se ha comentado, las técnicas de la minería de datos provienen de la Inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados.

Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en:

- **Forecasting (Predicción)** :

  Dada una tendencia de los datos se busca cuál será su previsión.

- **Supervisados (o predictivos)**:

  Predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.

- **No supervisados (o del descubrimiento del conocimiento)**:

  Se descubren patrones y tendencias en los datos.

Las técnicas más representativas son:

- **Redes neuronales**:

  Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales, es decir, un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.

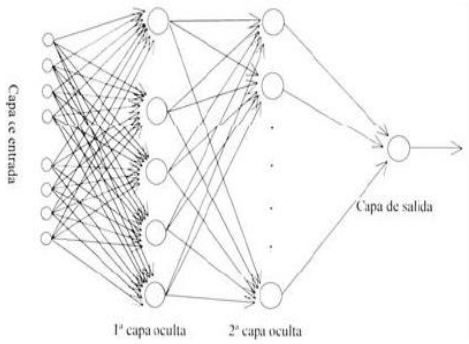
Esta tecnología puede ser desarrollada tanto en software como en hardware y con ella se pueden construir sistemas capaces de aprender, de adaptarse a condiciones variantes, o inclusive si se dispone de una colección suficiente grande de datos, predecir el estado futuro de algunos modelos. Estas técnicas son adecuadas para enfrentar problemas que hasta ahora eran resueltos sólo por el cerebro humano y resultaban difíciles o imposibles para las máquinas lógicas secuenciales. Un procesamiento paralelo realizado por un gran número de elementos altamente interconectados, es la clave de su funcionamiento.
Algunos ejemplos de red neuronal son:

-Los Mapas Auto-organizados, también conocidos como redes de Kohonen.

-El Perceptrón.

-El Perceptrón multicapa:

Este tipo de red neuronal se organiza generalmente en capas, como puede observarse en la siguiente figura:



1. Capa de entrada
2. Capa(s) oculta(s)
3. Capa de salida

- **Árboles de decisión**:

  Algoritmo de aprendizaje por inducción supervisada que pretende modelar los datos de ejemplo mediante un árbol.

  Un árbol de decisión se describe como un modelo de predicción utilizado en el ámbito de la inteligencia artificial, ya que dada una base de datos se construyen diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.

  En este tipo de árbol, los nodos intermedios son los atributos de entrada de los ejemplos presentados, las ramas representan valores de dichos atributos y los nodos finales son los valores de la clase.

  Para elegir qué atributos y en qué orden aparecen en el árbol, se utiliza una función de evaluación: ganancia de información.

  Ejemplos:

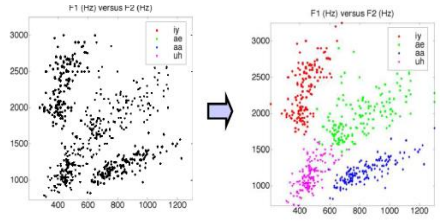
  -Algoritmo ID3 (Inicialmente sólo datos nominales).
  -Algoritmo C4.5 (Funciona con datos numéricos).

- **Modelos estadísticos**:

  Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.

- **Agrupamiento o Clustering :**

  Se trata de un algoritmo no supervisado, cuyo objetivo es detectar agrupamientos o estructuras intrínsecas en el conjunto de datos, identificando grupos homogéneos de individuos parecidos.



  Consiste en un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes.

  Ejemplos:

  -Algoritmo K-means (Distancia Euclídea).

  -Algoritmo K-medoids.

- **Algoritmos genéticos :**

  Los Algoritmos Genéticos, ni inductivos ni deductivos, ó en general, los métodos basados en la evolución nos suministran nuevas maneras de trabajar con cierto tipo de problemas. Estos se inspiran en las leyes sobre la evolución de Darwin y en los conceptos básicos de la biología genética.

  No es necesario conocer el comportamiento interno del sistema con el que se va a

trabajar. Sin embargo, se debe poseer el conocimiento suficiente de las salidas del sistema y sus efectos en el medio circundante de manera que se puedan evaluar las posibles soluciones.

El conjunto de reglas se considera como una población de "pseudo-organismos", los cuales al igual que los biológicos, pueden:

-Cruzarse unos con otros (operadores de cruce sexuales).

-Mutar aleatoriamente (operadores de mutación).

A partir de estos procedimientos se crean nuevos individuos de la población y se produce la selección natural: sólo "sobreviven" los mejores individuos (operadores de selección).

En este tipo de algoritmos existe una función de evaluación de la población, de tal forma que el proceso se "congela" si se alcanza el objetivo y se llega a la solución.

6. APLICACIONES

En este apartado se describirán diferentes aplicaciones de la minería de datos que facilitan los problemas de negocio y la toma de decisiones:

Detección de fraudes:

Detección de transacciones de blanqueo de dinero o de fraude en el uso de tarjetas de crédito o de servicios de telefonía móvil, donde estas operaciones fraudulentas o ilegales suelen seguir patrones característicos que permiten, con cierto grado de probabilidad, distinguirlas de las legítimas y desarrollar así mecanismos para tomar medidas rápidas frente a ellas. Por todo ello, el algoritmo, puede ser considerado como una técnica de clasificación, que analiza una gran cantidad de transacciones, tratando de categorizar aquellas que sean ilegítimas mediante la identificación de ciertas características que estas últimas tengan en común.

Recursos humanos:

La minería de datos también puede ser de gran utilidad en los departamentos de recursos humanos de cualquier empresa, en la identificación de las características y capacidades de sus mejores empleados.

La información obtenida mediante estas técnicas puede ayudar al personal de recursos humanos a la hora de la contratación de personal, centrándose en los esfuerzos de sus empleados y los resultados obtenidos por éstos. Además dicha ayuda ofrecida por la minería de dato (conocimiento), se traduce en la obtención de ventajas a nivel corporativo, como mejoras en las decisiones corporativas: desarrollo de planes de producción o gestión de mano de obra

Terrorismo:

La minería de datos es la técnica por la cual la unidad de Able Danger del ejército de los EE.UU. había identificado al líder de la banda terrorista autora de los atentados del 11 de septiembre de 2001, Mohammed Atta, y a otros tres secuestradores, como posibles miembros de una célula de Al Qaeda que operaba en los EE.UU. un año antes del ataque.

Juegos:

A comienzos de la década de 1960, se disponía de oráculos para determinados juegos combinacionales, se ha abierto un nuevo camino en la minería de datos que consiste en la extracción de estrategias utilizadas por personas para la implantación en dichos oráculos.

Los planteamientos actuales sobre reconocimiento de patrones, no parecen poder aplicarse con éxito al funcionamiento de estos oráculos.

Genética:

En el estudio de la genética humana, el objetivo principal es entender la relación cartográfica entre las partes y la variación individual en las secuencias del ADN del ser humano y los cambios que puedan producirse en la susceptibilidad a las enfermedades. Es decir, como los cambios en la secuencia del ADN de un individuo afectan al riesgo de desarrollar enfermedades comunes (como por ejemplo el cáncer).

La minería de datos puede ayudar a mejor de esta forma el diagnóstico, prevención y tratamiento de enfermedades. Generalmente la técnica de minería de datos que se utiliza en este tipo de aplicaciones se conoce como "reducción de dimensionalidad multifactorial".

Ingeniería eléctrica:

Las técnicas de minería de datos en este ámbito han sido utilizadas principalmente para monitorizar las condiciones de las instalaciones de alta tensión.

La finalidad de esta aplicación es obtener información valiosa sobre el estado de aislamiento de los equipos, vigilar las vibraciones producidas o por ejemplo para analizar los cambios de carga en los transformadores. Generalmente se usan técnicas encargadas de detectar condiciones anormales (Análisis de anomalías).

Previsiones de fuga:

En muchas industrias (banca, telecomunicaciones,…) existe un interés comprensible en detectar e identificar cuanto antes a aquellos clientes que puedan estar pensando en rescindir sus contratos, para muy probablemente pasarse a la competencia.

Con la ayuda de la minería de datos identificaríamos qué clientes son los más proclives a darse de baja estudiando sus patrones de comportamiento y comparándolos con clientes que ya han rescindido su contrato con la empresa, de esta forma se podría actuar realizando ofertas personalizadas y ofreciendo promociones con el objetivo de retener a dichos clientes.

Detección de hábitos de compra en supermercados

Un ejemplo clásico de aplicación de minería de datos, es la detección de hábitos de los clientes, a la hora de comprar en los supermercados. Un estudio muy conocido detectó que los viernes se compraban una cantidad inusual de pañales y cerveza, debido principalmente a que los viernes solían acudir a comparar padres jóvenes cuya perspectiva para el fin de semana era quedarse en casa cuidando de los niños y viendo la televisión tomándose una cerveza. Con este tipo de información, muy valiosa para el supermercado, se pudieron poner en práctica tácticas para incrementar por ejemplo la ventas de las cervezas colocándolas cercanas a los pañales y así fomentar las ventas compulsivas. O para, una vez que adquieren un determinado producto, saber inmediatamente qué otro ofrecerle teniendo en cuéntala información histórica disponible acerca de los clientes que han comprado primero.

Bioinformática

La bioinformática se encuentra en la intersección entre las ciencias de la vida y de la información, proporciona las herramientas y recursos necesarios para favorecer la investigación biomédica. Como campo interdisciplinario, comprende la investigación y el desarrollo de sistemas útiles para entender el flujo de información desde los genes a las estructuras moleculares, su función bioquímica, su conducta biológica y, finalmente, su influencia en las enfermedades y en la salud.

Los principales estímulos para su desarrollo son:

- El enorme volumen de datos generados por los distintos proyectos denominados genoma (humano y de otros organismos).
- Los nuevos enfoques experimentales, que permiten obtener datos genéticos a gran velocidad, bien de genomas individuales (mutaciones, polimorfismos) de enfoques celulares (expresión génica).

Uno de los retos de la bioinformática es el desarrollo de métodos que permitan integrar los datos genómicos para explicar el comportamiento global de la célula viva, minimizando la intervención humana. Dicha integración, sin embargo, no puede producirse sin considerar el conocimiento acumulado durante años, producto de la investigación de miles de científicos.

La bioinformática es un área del espacio que representa la biología molecular computacional, que incluye la aplicación de las computadoras y de las ciencias de la información en áreas como la geonómica, el mapeo, la secuencia y determinación de las secuencias y estructuras por métodos clásicos. Las metas fundamentales de la bioinformática son la predicción de la estructura tridimensional de las proteínas a partir de su secuencia, la predicción de las funciones biológicas y biofísicas a partir de la secuencia o la estructura, así como simular el metabolismo y otros procesos biológicos basados en esas funciones.

Web Mining:

Una aplicación especial de la minería de datos es la minería web (o minería de uso de la web, *web mining*) que consiste en extraer información y conocimiento útil específicamente de la actividad de un sitio web: análisis de tráfico (visitas y visitantes), contenidos más accedidos, procedencia, tipo de usuarios, navegadores y sistemas operativos, reglas de asociación entre páginas (tasa de conversión)...

El Web mining se considera una metodología de recuperación de la información que usa herramientas de la minería de datos para extraer información tanto del contenido de las páginas, de su estructura de relaciones (enlaces) y de los registro de navegación de los usuarios.

En este sentido podemos definir tres variantes del el Web mining:

- Mineración del contenido de la Web, o Web Content Mining;
- Mineración de la estructura de la Web, o Web Structure Mining;
- Mineración de los registro de navegación en la Web, o Web Usage Mining.

Text Mining:

La minería de textos (text mining) es una disciplina englobada dentro de las técnicas de acceso, recuperación y organización de información y consiste en un conjunto de técnicas que nos permiten extraer información relevante y desconocida de manera automática dentro de grandes volúmenes de información textual, normalmente en lenguaje natural y generalmente no estructurada.

La minería de textos permite el descubrimiento de patrones interesantes y nuevos conocimientos en un conjunto de textos, es decir, su objetivo consiste en descubrir tendencias, desviaciones y asociaciones entre una gran cantidad de información textual. Esto nos permite encontrar conocimiento significativo a partir de datos textuales sin estructurar.

La minería de textos extrae información nueva por lo que es algo totalmente distinto a la búsqueda web, en la cual se busca información ya conocida, no se extrae información nueva.

Una de las principales características de la minería de textos consiste en que la información no está estructurada, al contrario de lo que ocurre en la minería de datos en la que la información suele extraerse de una base de datos, por lo que sí está estructurada. Esto hace que la extracción de información de una base de datos sea más sencilla, ya que las bases de datos están diseñadas para que sea posible el tratamiento automático de la información.

Las principales áreas de aplicación de las tecnologías de minería de textos cubren dos aspectos:

- El descubrimiento de conocimiento
- La extracción de información

La minería de textos constituye una herramienta de gran utilidad ya que alrededor de un 80% de la información de las organizaciones está almacenada en forma de texto no estructurado.

5. CONCLUSIONES

Generalmente, el conocimiento se ha venido obteniendo por el clásico método hipotético-deductivo de la ciencia. En él es fundamental el paso inductivo inicial: a partir de un conjunto de observaciones y de unos conocimientos previos, la intuición conduce a formular la hipótesis.

Las técnicas de análisis estadístico, permiten obtener ciertas informaciones útiles, pero no inducir relaciones cualitativas generales, o leyes, previamente desconocidas; para esto se requieren otras técnicas de análisis inteligente que están enfocadas a la inducción de conocimiento en bases de datos: la Minería de Datos (data mining), que pone al alcance del individuo lo que necesita en el momento preciso para que su actividad se haga efectiva.

Tradicionalmente, las técnicas de minería de datos se aplicaban sobre información contenida en almacenes de datos. De hecho, muchas grandes empresas e instituciones han creado bases de datos especialmente diseñadas para proyectos de minería de datos en las que centralizan información potencialmente útil de todas sus áreas de negocio. No obstante, actualmente está cobrando una importancia cada vez mayor la minería de datos desestructurados como información contenida en ficheros de texto, en Internet,…, ya que gran parte de la información es desestructurada.

Por tanto, la Minería de Datos surge a partir de sistemas de aprendizaje inductivo en ordenadores, al ser aplicados a bases de datos, y su importancia crece de forma masiva.

Los modelos obtenidos por técnicas de minería de datos se aplican incorporándolos en los sistemas de análisis de información de las organizaciones, e incluso, en los sistemas transaccionales. En este sentido cabe destacar los esfuerzos del Data Mining Group, que está estandarizando el lenguaje PMML (Predictive Model Markup Language), de manera que los modelos de minería de datos sean interoperables en distintas plataformas, con independencia del sistema con el que han sido construidos. Los principales fabricantes de sistemas de bases de datos y programas de análisis de la información hacen uso de este estándar.

6. REFERENCIAS

[1] "INTRODUCCIÓN A LA MINERÍA DE DATOS"
José Hernández Orallo, M.José Ramírez Quintana, Cèsar Ferri Ramírez**.**
Editorial Pearson, 2004. ISBN: 84 205 4091

[2] Artículo: “Data mining: torturando a los datos hasta que confiesen”.
Luis Carlos Molina Felix.
Universitat Politécnica de Catalunya.

[3] Artículo: “Web Mining: Fundamentos Básicos”
Francisco Manuel De Gyves Camacho
Doctorado en informática y automática
Universidad de Salamanca