SISTEMAS KDD (Knowledge Discovery in Databases)

***Abstract*— Knowledge Discovery implica la evaluación e interpretación de patrones y modelos para tomar decisiones con respecto a lo que constituye conocimiento y lo que no lo es. Por lo tanto, el KDD requiere de un amplio y profundo conocimiento sobre tu área de estudio. Por otra parte, la Minería de Datos, exploración de datos o Data Mining, no requiere tanto conocimiento sobre el área de estudio, sino más conocimiento técnico. Como mencionamos anteriormente, la Minería de Datos es un paso que forma parte del KDD e implica el análisis de grandes cantidades de datos observacionales, para encontrar relaciones insospechadas. https://mnrva.io/kdd-platform.html.**

I. INTRODUCCION

En este texto se estudia uno de los campos que más se están estudiando en estos días: La extracción de conocimiento a partir de fuentes masivas de datos. Para ello se emplean las denominadas técnicas de minería de datos, que son algoritmos capaces de obtener relaciones entre distintos atributos o conceptos para ayudar, por ejemplo, a la toma de decisiones. Además de las técnicas estadísticas se estudian las técnicas de Minería de Datos [Data Mining] basadas en técnicas de aprendizaje automático que se implementan en una herramienta de minería de datos de libre distribución: WEKA. Esta herramienta permite, a partir de ficheros de texto en un formato determinado, utilizar distintos tipos de técnicas para extraer información. A continuación se definen los conceptos fundamentales empleados en el texto: KDD y, sobretodo, minería de datos, así como sus principales características. Posteriormente se comenta la estructura del proyecto.

II. 1.1. KDD y Minería de Datos

KDD [Knowledge Discovery in Databases] [PSF91] es el proceso completo de extracción de información, que se encarga además de la preparación de los datos y de la interpretación de los resultados obtenidos. KDD se ha definido como "el proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles" [FAYY96]. Se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones. Para conseguirlo harán falta técnicas de aprendizaje automático [Machine Learning] [MBK98], estadística [MIT97, DEGR86], bases de datos [CODD70], técnicas de representación del conocimiento, razonamiento basado en casos [CBR, Case Based Reasoning], razonamiento aproximado, adquisición de conocimiento, redes de neuronas y visualización de datos. Tareas comunes en KDD son la inducción de reglas, los problemas de clasificación y clustering, el reconocimiento de patrones, el modelado predictivo, la detección de dependencias, etc.

Los datos recogen un conjunto de hechos (una base de datos) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto). KDD involucra un proceso iterativo e interactivo de búsqueda de modelos, patrones o parámetros. Los patrones descubiertos han de ser válidos, novedosos para el sistema (para el usuario siempre que sea posible) y potencialmente útiles.

Se han de definir medidas cuantitativas para los patrones obtenidos (precisión, utilidad, beneficio obtenido...). Se debe establecer alguna medida de interés [interestingness] que considere la validez, utilidad y simplicidad de los patrones obtenidos mediante alguna de las técnicas de Minería de Datos. El objetivo final de todo esto es incorporar el conocimiento obtenido en algún sistema real, tomar decisiones a partir de los resultados alcanzados o, simplemente, registrar la información conseguida y suministrársela a quien esté interesado.

Ha llegado un momento en el que disponemos de tanta información que nos vemos incapaces de sacarle provecho. Los datos tal cual se almacenan [raw data] no suelen proporcionar beneficios directos. Su valor real reside en la información que podamos extraer de ellos: información que nos ayude a tomar decisiones o a mejorar nuestra comprensión de los fenómenos que nos rodean.

Se requiere de grandes cantidades de datos que proporcionen información suficiente para derivar un conocimiento adicional. Dado que se requieren grandes cantidades de datos, es esencial el proceso de la eficiencia.

La exactitud es requerida para asegurar que el descubrimiento del conocimiento es válido. Los resultados deberán ser presentados de una manera entendible para el ser humano. Una de las premisas mayores de KDD es que el conocimiento es descubierto usando técnicas de aprendizaje inteligente que van examinando los datos a través de procesos automatizados.

Para que una técnica sea considerada útil para el descubrimiento del conocimiento, éste debe ser interesante; es decir, debe tener un valor potencial para el usuario.

III. 1.1.2. El proceso de KDD

El proceso de KDD se inicia con la identificación de los datos. Para ello hay que imaginar qué datos se necesitan, dónde se pueden encontrar y cómo conseguirlos. Una vez que se dispone de datos, se deben seleccionar aquellos que sean útiles para los objetivos propuestos. Se preparan, poniéndolos en un formato adecuado. Una vez se tienen los datos adecuados se procede a la minería de datos, proceso en el que se seleccionarán las herramientas y técnicas adecuadas

para lograr los objetivos pretendidos. Y tras este proceso llega el análisis de resultados, con lo que se obtiene el conocimiento pretendido.

KDD es un proceso interactivo e iterativo, que involucra numerosos pasos e incluye muchas decisiones que deben ser tomadas por el usuario, y se estructura en las siguientes etapas [FAYY96]:

• Comprensión del dominio de la aplicación, del conocimiento relevante y de los objetivos del usuario final.

• Creación del conjunto de datos: consiste en la selección del conjunto de datos, o del subconjunto de variables o muestra de datos, sobre los cuales se va a realizar el descubrimiento.

• Limpieza y preprocesamiento de los datos:

Se compone de las operaciones, tales como: recolección de la información necesaria sobre la cual se va a realizar el proceso, decidir las estrategias sobre la forma en que se van a manejar los campos de los datos no disponibles, estimación del tiempo de la información y sus posibles cambios.

• Reducción de los datos y proyección:

Encontrar las características más significativas para representar los datos, dependiendo del objetivo del proceso. En este paso se pueden utilizar métodos de transformación para reducir el número efectivo de variables a ser consideradas o para encontrar otras representaciones de los datos.

• Elegir la tarea de Minería de Datos:

Decidir si el objetivo del proceso de KDD es:

Regresión, Clasificación, Agrupamiento, etc.

• Elección del algoritmo(s) de Minería de Datos:

Selección del método(s) a ser utilizado para buscar los patrones en los datos. Incluye además la decisión sobre que modelos y parámetros pueden ser los más apropiados.

• Minería de Datos:

Consiste en la búsqueda de los patrones de interés en una determinada forma de representación o sobre un conjunto de Técnicas de Análisis de Datos.

IV. 1.1.3. Mineria de datos

Minería de Datos es usado comúnmente por los estadísticos, analistas de datos, y por la comunidad de administradores de sistemas informáticos como todo el proceso del descubrimiento, mientras que el término KDD es utilizado más por los especialistas en Inteligencia Artificial.

El análisis de la información recopilada (por ejemplo, en un experimento científico) es habitual que sea un proceso completamente manual (basado por lo general en técnicas estadísticas). Sin embargo, cuando la cantidad de datos de los que disponemos aumenta la resolución manual del problema se hace intratable. Aquí es donde entra en juego el conjunto de técnicas de análisis automático al que nos referimos al hablar de Minería de Datos o KDD.

Hasta ahora, los mayores éxitos en Minería de Datos se pueden atribuir directa o indirectamente a avances en bases de datos (un campo en el que los ordenadores superan a los humanos). No obstante, muchos problemas de representación del conocimiento y de reducción de la complejidad de la búsqueda necesaria (usando conocimiento a priori) están aún por resolver. Ahí reside el interés que ha despertado el tema entre investigadores de todo el mundo.

A continuación se presentan varias definiciones de Minería de Datos (MD):

• “MD es la extracción no trivial de información implícita, desconocida previamente, y potencialmente útil desde los datos” [PSF91].

• “MD es el proceso de extracción y refinamiento de conocimiento útil desde grandes bases de datos” [SLK96].

• “MD es el proceso de extracción de información previamente desconocida, válida y procesable desde grandes bases de datos para luego ser utilizada en la toma de decisiones” [CHSVZ].

• ”MD es la exploración y análisis, a través de medios automáticos y semiautomáticos, de grandes cantidades de datos con el fin de descubrir patrones y reglas significativos” [BERR97].

• ”MD es el proceso de planteamiento de distintas consultas y extracción de información útil, patrones y tendencias previamente desconocidas desde grandes cantidades de datos posiblemente almacenados en bases de datos” [THUR99].

• “MD es el proceso de descubrir modelos en los datos” [WF00].

References

V. Minerva

https://mnrva.io/kdd-platform.html