Tabla de contenido

INTRODUCCIÓN AL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS (KDD)

1.1 Proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD

) El término Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases, o KDD para abreviar) empezó a utilizarse en 1989 para referirse al amplio proceso de búsqueda de conocimiento en bases de datos, y para enfatizar la aplicación a "alto nivel" de métodos específicos de minería de datos. En general, el descubrimiento es un tipo de inducción de conocimiento, no supervisado, que implica dos procesos: • Búsqueda de regularidades interesantes entre los datos de partida, • Formulación de leyes que las describan. Entre la literatura dedicada al tema, se pueden encontrar varias definiciones para descubrimiento: ⎫ Descubrimiento implica observar, recoger datos, formar hipótesis para explicar las observaciones, diseñar experimentos, comprobar la corrección de las hipótesis, comparar nuestros hallazgos con los de otros investigadores y repetir el ciclo. Las computadoras son capaces de observar y recoger datos, a veces mejor que los observadores humanos; los programas estadísticos pueden generar agrupaciones de forma automática entre los datos recogidos; también hay programas con cierta capacidad para diseñar experimentos; y algunos sistemas robóticos realizan las manipulaciones necesarias en ciertos experimentos. Pero ninguna computadora reúne todas estas habilidades ni es capaz de adaptarse para aplicarlas a nuevos problemas; en este sentido, las computadoras no serían capaces de descubrir. Sin embargo, el descubrimiento no requiere realizar simultáneamente todas estas tareas. De igual modo que un investigador puede descubrir nuevo conocimiento a través del análisis de sus datos, una computadora puede examinar los datos disponibles o recogidos por otras computadoras y encontrar relaciones y explicaciones previamente desconocidas, realizando así descubrimiento en un sentido más restringido. La capacidad de las computadoras para realizar búsquedas exhaustivas de forma incansable entre grandes cantidades de datos ofrece buenas expectativas para obtener descubrimiento de forma automática.

1.3 Fases del KDD

Hay distintas técnicas de las diferentes disciplinas que se utilizan en las diversas fases. Si bien los términos Minería de datos y descubrimiento de conocimiento en bases de datos son usados como sinónimos, el término KDD describe el proceso completo de extracción de conocimiento a partir de los datos. Mientras que Data Mining, se refiere exclusivamente al estadio de descubrimiento de un proceso general KDD. En este contexto, descubrimiento de conocimiento significa la identificación de relaciones y patrones existenciales en los datos. Un proceso KDD consiste en la extracción no trivial de conocimiento previamente desconocido y potencialmente útil a partir de un conjunto de datos. En el proceso KDD es posible definir al menos 6 estados: Recolección de datos, Selección, Limpieza y Transformación de datos, Minería de datos, Evaluación y Validación, Interpretación y Difusión, Actualización y Monitorización. 1.3.1 Recolección de Datos Las primeras fases del KDD determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original. Generalmente, la información que se requiere investigar sobre un cierto dominio de la organización se encuentra: ⎪ En bases de datos y otras fuentes muy diversas. ⎪ Tanto internas como externas. ⎪ Muchas de estas fuentes son las

1.3.3 Minería de Datos.

Características Especiales de los Datos: Aparte del gran volumen, ¿por qué las técnicas de aprendizaje automático y estadísticas no son directamente aplicables? ⎪ Los datos residen en el disco. No se pueden escanear múltiples veces. ⎪ Algunas técnicas de muestreo no son compatibles con algoritmos no incrementales. ⎪ Muy alta dimensionalidad (muchos campos). ⎪ Evidencia Positiva. ⎪ Datos Imperfectos Aunque algunos se aplican casi directamente, el interés en la investigación en minería de datos está en su adaptación. Patrones a descubrir: ⎪ Una vez recolectados los datos de interés, un explorador puede decidir qué tipos de patrón quiere descubrir. ⎪ El tipo de conocimiento que se desea extraer va a marcar claramente la técnica de minería de datos a utilizar. ⎪ Según como sea la búsqueda del conocimiento se puede distinguir entre: ♣ Directed data mining: se sabe claramente lo que se busca, generalmente predecir unos ciertos datos o clases. ♣ Undirected data mining: no se sabe lo que se busca, se trabaja con los datos. En el primer caso, los propios sistemas de minería de datos se encargan generalmente de elegir el algoritmo más idóneo entre los disponibles para un determinado tipo de patrón a busca