UAEMéx
Universidad Autónoma
del Estado de México

Universidad Autónoma del Estado de México

Material didáctico multimedia

Sólo visión

La minería de datos en el proceso de KDD (Knowledge Discovery and Data Mining)

Unidad de Aprendizaje Minería de Datos
Licenciatura de Ingeniería en Computación
Facultad de Ingeniería

Elaborado por M en I Sara Vera Noguez
Durante el período intensivo verano 2016

Minería de datos

Unidad 2. El proceso KDD

M en I Sara Vera Noguez

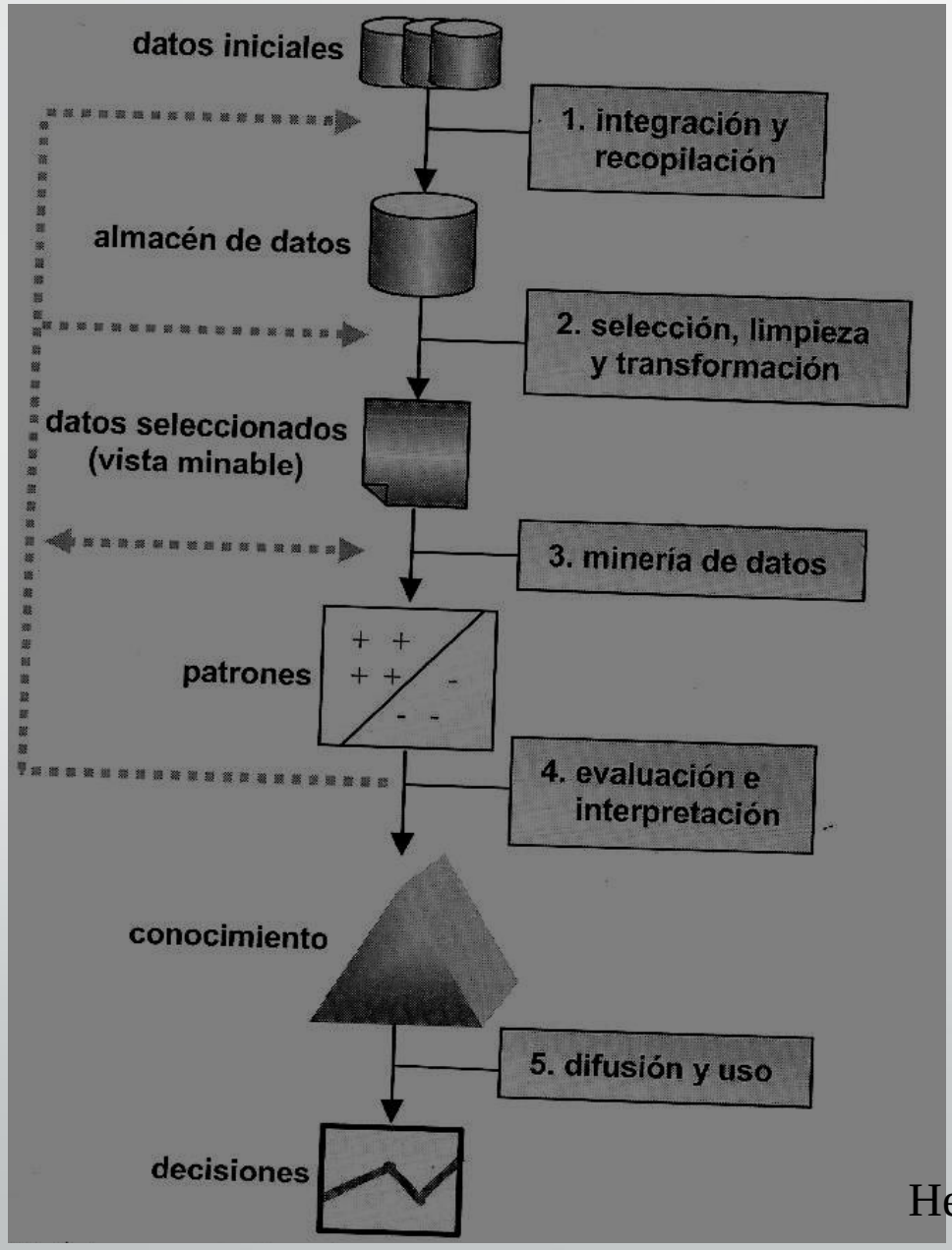
2. La minería de datos en el proceso de KDD

- Etapas de proceso de KDD:

1) Integración y recopilación
2) Selección, Limpieza (también llamada preprocesamiento),Transformación
3) **Minería de Datos**
4) Evaluación e Interpretación
5) Difusión y uso.

Algunas fuentes no lo contemplan

El proceso KDD



Hernández, 2008 p.20

KDD

- “Es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de datos” [Fayyad, 1996]

Propiedades del conocimiento

- Válido: debe ser preciso para datos nuevos y para aquello que se han usado para obtenerlo
- Novedoso: que aporte algo antes desconocido en el sistema o para el usuario
- Potencialmente útil
- Comprensible: ya que información incomprensible no proporciona conocimiento

¿De donde viene los datos?

- Bases de datos relacionales
- Bases de datos espaciales
- Bases de datos temporales
- Bases de datos documentales
- Bases de datos multimedia

La BD relacionales

- Cuentan con un esquema que describe las relaciones
- Integridad:
  - Dominio (valores que puede tomar)
  - Identidad (llave primaria)
  - Referencial

Aplicamos minería sobre toda la base?

- Se aplican técnicas de acuerdo al tipo de datos. Se distinguen diferentes tipos de campos:
- Numérico
- Categóricos
- Textos
- Vectores

El conjunto *minable*

- Las técnicas de minería se aplican a una tabla o una vista de la DB o data *warehouse*.

Las 5 fases centrales

- **Selección de datos:** En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos.
- **Limpieza (Preprocesamiento):** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

Las 5 fases centrales

- **Transformación:** Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.
- **Data Mining:** Es la fase de modelamiento propiamente tal, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u "ocultos" en los datos.

Las 5 fases centrales

- **Interpretación y Evaluación:** Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

Bibliografía básica

- Hernández Orallo, J., M. J. Ramírez Quintana, et al. (2004). Introducción a la Mineria de Datos. España, Pearson Educación SA.
- Han, D. J. (2007). Principles of Data Mining, MIT Press.
- Maimon, O. Z. and L. Rokach (2005). Data mining and knowledge discovery handbook. USA, Springer.
- Pérez López, C. and D. Santín Gonzalez (2006). Data Mining-Soluciones Con Enterprise Miner. México, Alfaomega, Ra-Ma.
- Sumathi, S. and S. N. Sivanandam (2006). Introduction to data mining and its applications. Berlín, Germany, Springer-Verlag New York Inc.
- Tan, P. N., M. Steinbach, et al. (2005). Introduction to data mining, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Bibliografía complementaria

- Everitt, B.S. (1994). A Handbook of Statistical Analyses using S-Plus. Chapman and Hall.
- Inmon, W.H. (1996). Building the Datawarehouse. J.Wiley & Sons.
- Han, J. and M. Kamber (2006). Data mining: concepts and techniques, Morgan Kaufmann.
- Kimball, R (1996). The Data Warehouse Toolkit. John Wiley & Sons.
- Hastie, T., R. Tibshirani, et al. (2005). The elements of statistical learning: data mining, inference and prediction, Springer
- Dunham. H. Margaret (2003). Data Mining. Introductory and Advanced Topics, Prentice Hall.
- Pyle, D. (1999), "Data Preparation for Data Mining", Morgan Kaufmann, San Francisco, CA.
- Hand, David; Mannila, Heikki; Smyth, Padhraic (2001), Principles of Data Mining, A Bradford Book. The MIT Press.
- Ian Witten and Eibe Frank (2002), Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers.