Minería de datos
Unidad I y II

Profesor:

M. en C.C. José Hernández Santiago

Programa educativo:

Ingeniería en Computación

C.U UAEM-Texcoco

Av. Jardín Zumpango s/n Fraccionamiento
El Tejocote, 56259

“Patria Ciencia y Trabajo”

Unidad I.
Introducción a la Minería de Datos

Profesor:

M. en C.C. José Hernández Santiago

Programa educativo:

Ingeniería en Computación

C.U UAEM-Texcoco

Av. Jardín Zumpango s/n Fraccionamiento
El Tejocote, 56259

“Patria Ciencia y Trabajo”

Minería de Datos

Objetivo general:

El alumno estudiara y aplicara el proceso metodológico, las bases teóricas de las técnicas de minería de datos más usuales y las herramientas de software disponibles para adquirir la capacidad de aplicar minería de datos en un contexto real

Introducción a la minería de datos

Las empresas guardan muchos datos, almacenados en diversos medios:

- Encuestas impresas
- Formularios impresos
- Bitácoras
- Libros contables
- Archivos de texto plano
- Archivos multimedia (audio, fotos, video)
- Paginas Web
- Hojas de cálculo
- Bases de datos

Son ricos en datos pero pobres en información.

¿Como tomar decisiones?

"Quien olvida su historia está condenado a repetirla".

Jorge Agustín Nicolás Ruiz de Santayana y Borrás

Introducción a la minería de datos

Esquema de la jerarquía de una BDD

Definición

La minería de datos es la extracción de patrones o “conocimiento” útil a partir de grandes cantidades de datos almacenados en distintos formatos.

Requisitos para el conocimiento:

- No trivial
- Implícito en los datos
- Previamente desconocido
- Potencialmente útil

Tiene como objetivo la obtención de modelos o patrones

Definición

Es la extracción de información oculta y predecible a partir de grandes bases de datos.

Es una poderosa tecnología con gran potencial para ayudar a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse)

Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y dirigidas.

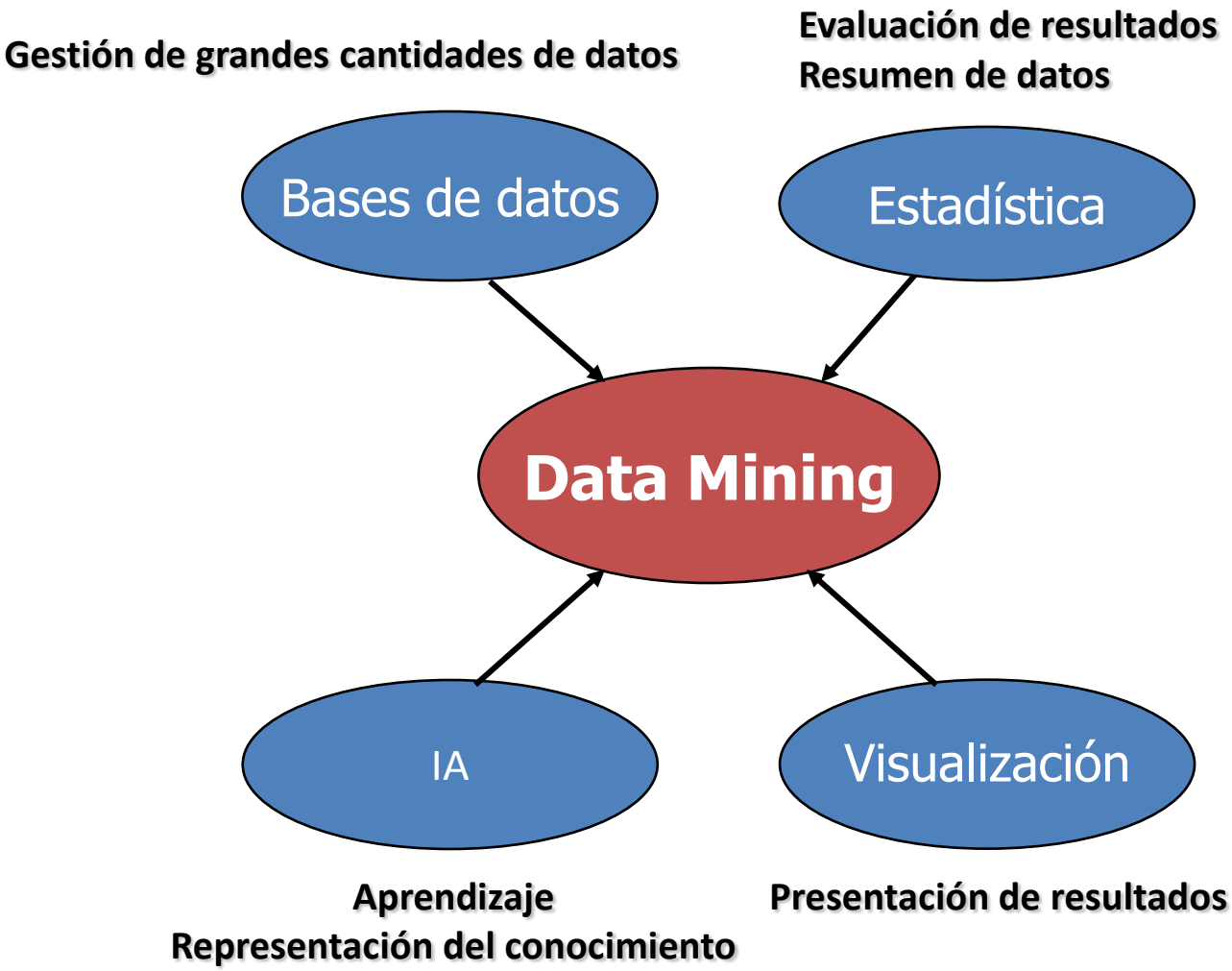
El uso de los patrones descubiertos deberían ayudar a tomar decisiones más seguras que reflejen, por tanto, algún beneficio a la organización.

Definición

Las herramientas de Data Mining exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas.

Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional porque las relaciones son demasiado complejas o porque hay demasiado datos, en cuyo caso puede hacer uso de prácticas estadísticas y algoritmos de búsqueda próximos a la Inteligencia Artificial.

Disciplinas involucradas



Aplicaciones de la minería de datos

En los negocios:

Administración empresarial basada en la relación con el cliente.

En lugar de contactar con el cliente de forma indiscriminada a través de un centro de llamadas o enviando cartas, sólo se contactará aquellos que tienen una mayor probabilidad de responder positivamente a una determinada oferta o promoción.

Determinar qué clientes van a ser rentables durante una ventana de tiempo (una quincena, un mes, año, etc) y sólo enviar las ofertas a las personas con alta probabilidad de ser rentables.

Aplicaciones de la minería de datos

Hábitos de compra en supermercados:

Es una aplicación clásica de la minería de datos.

Un estudio muy citado detectó que los viernes había una cantidad inusualmente elevada de clientes que adquirían a la vez pañales y cerveza.

Usando minería de datos se detectó que ambos productos se vendían principalmente los viernes en la tarde y eran comprados por hombres con edades entre los 25 y 35 años de edad.

Se detectó que se debía a que dicho día solían acudir al supermercado padres jóvenes cuya perspectiva para el fin de semana consistía en quedarse en casa cuidando de su hijo y viendo la televisión con una cerveza en la mano.

En otros casos las esposas, que en muchos casos hacen la compra de la casa, dejan los pañales para que el esposo los compre debido a que los paquetes son voluminosos y el esposo los compraba junto con las cervezas para el fin de semana.

El supermercado pudo incrementar sus ventas de cerveza colocándolas cerca de los pañales para fomentar las ventas compulsivas.

El resultado fue que los padres que normalmente llegaban a comprar los pañales y la cerveza, compraron más cervezas; mientras que los que antes no compraban cerveza, empezaron a comprarla por la proximidad de ésta con los pañales.

Aplicaciones de la minería de datos

Patrones de fuga:

En muchas industrias (como la banca, las telecomunicaciones y otras dedicadas a servicio), existe un comprensible interés en detectar cuanto antes aquellos clientes que puedan estar pensando en rescindir sus contratos para, posiblemente, pasarse a la competencia.

A estos clientes (en función de su valor) se les podrían hacer ofertas personalizadas, ofrecer promociones especiales, etc., con el objetivo de retenerlos.

La minería de datos puede ayudar a determinar qué clientes son los más proclives a darse de baja estudiando sus patrones de comportamiento y comparándolos con muestras de clientes que se dieron de baja en el pasado.

Aplicaciones de la minería de datos

Fraudes:

La detección de transacciones de lavado de dinero o de fraude en el uso de tarjetas de crédito, de servicios de telefonía móvil e, incluso, en la relación de los contribuyentes con el fisco.

Generalmente, estas operaciones fraudulentas 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.

En 2001, las instituciones financieras a escala mundial perdieron del orden de 2000 millones de dólares en fraudes cometidos con tarjetas de crédito.

El Falcon Fraud Manager es un sistema inteligente que examina transacciones, propietarios de tarjetas y datos financieros para intentar detectar y reducir el número de fraudes.

La solución de Falcon usa una sofisticada combinación de modelos de redes neuronales artificiales para analizar el pago mediante tarjeta y detectar los más remotos casos de fraude.

Ha sido usado durante más de 15 años y monitoriza alrededor de 450 millones de cuentas distribuidas en los 6 continentes.

Aplicaciones de la minería de datos

Otras áreas:

Recursos humanos. La identificación de las características de sus empleados de mayor éxito para crear un perfil y ayudar en la contratación de personal.

Comportamiento en Internet. Para ofrecer propaganda adaptada específicamente al perfil del usuario. O para, una vez que adquieren un determinado producto, inmediatamente ofrecerle otro teniendo en cuenta la información histórica disponible de los clientes que han comprado lo mismo.

Medicina y farmacia. Diagnóstico de enfermedades y análisis de la efectividad de tratamientos.

Astronomía. Identificación de nuevas galaxias y estrellas

Aplicaciones de la minería de datos

Otras áreas:

Geología, minería, agricultura y pesca. Identificación de áreas propicias para distintos cultivos , usadas para pesca o de explotación minería en bases de datos de imágenes satelitales.

Ciencias sociales. Estudios de los flujos de la opinión publica, identificar barrios con conflictos en función de valores socio demograficos.

Detección de terroristas.

Juegos.

Genética.

Unidad II.
La Minería de Datos en el proceso de KDD

Profesor:

M. en C.C. José Hernández Santiago

Programa educativo:

Ingeniería en Computación

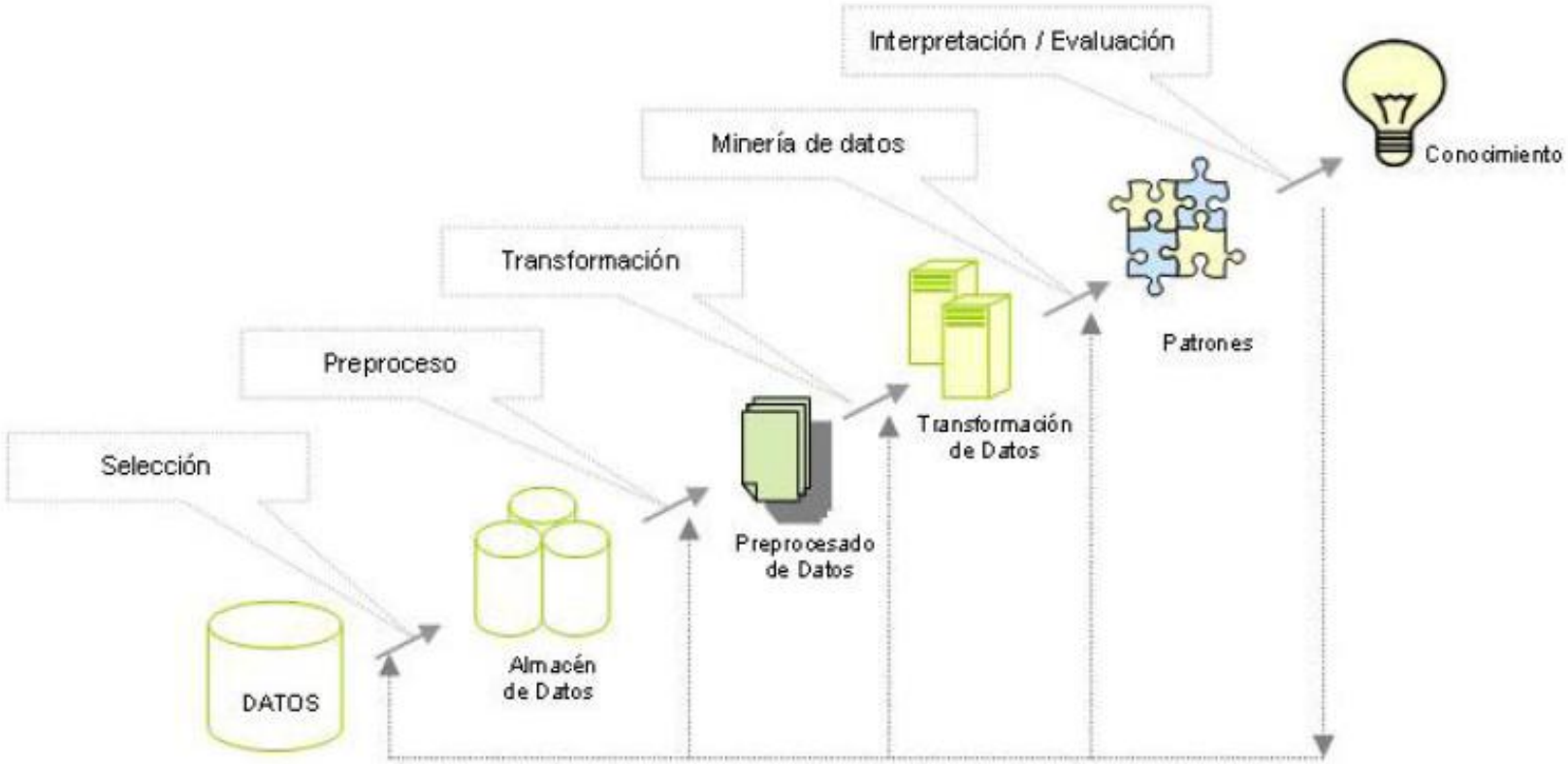
C.U UAEM-Texcoco

Av. Jardín Zumpango s/n Fraccionamiento
El Tejocote, 56259

“Patria Ciencia y Trabajo”

Knowledge Discovery in Databases

Fases del proceso de extracción del conocimiento (KDD) según Fayyad



KDD

Fayyad (2002):

Define el KDD como un proceso no trivial para identificar patrones validos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos.

El KDD utiliza la minería de datos como una de sus fases.

Es un proceso más complejo que la minería de datos.

Además de obtener modelos y patrones, incluye una evaluación e interpretación de los mismos.

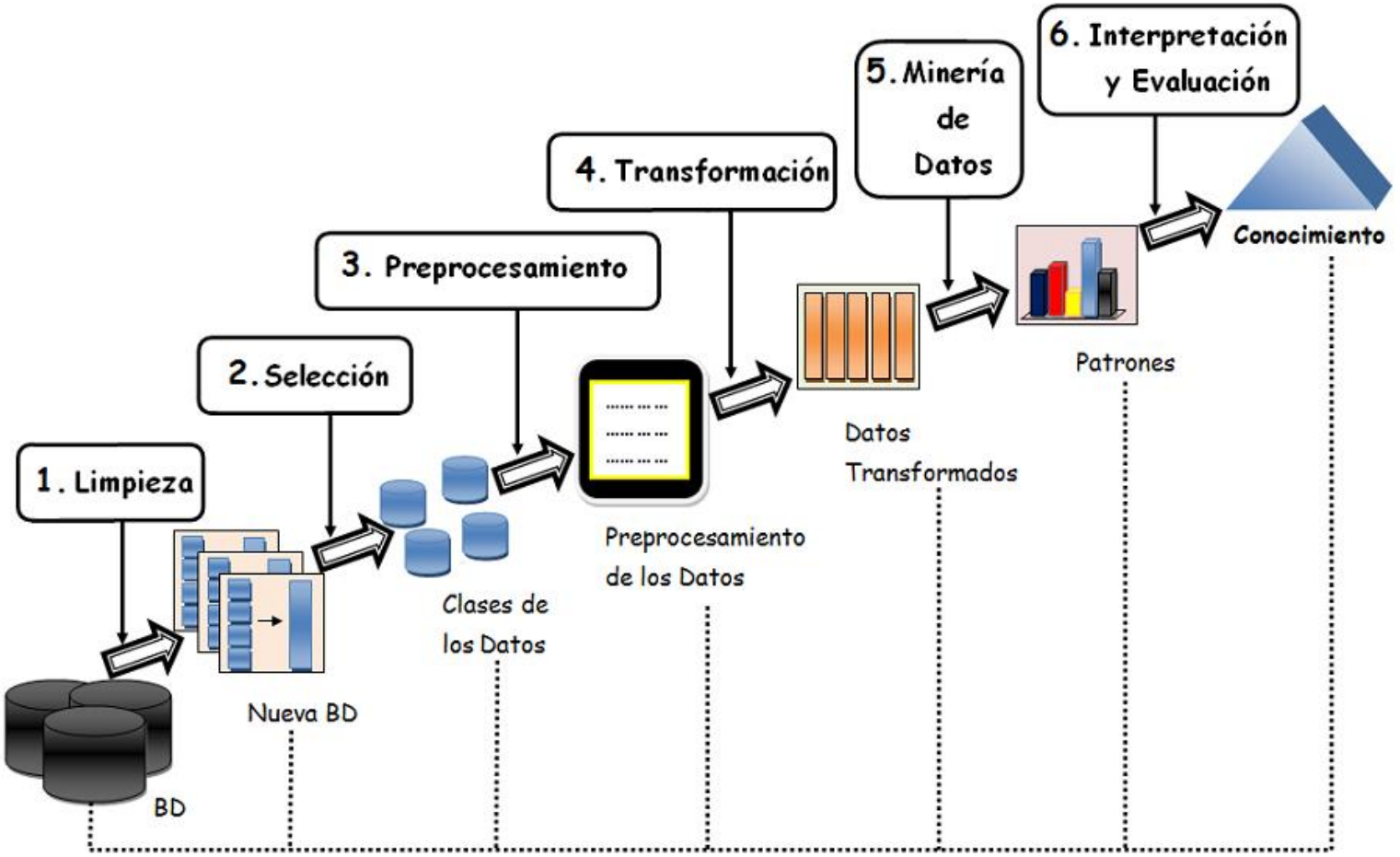
Rhodes (2002) hace la distinción entre

KDD como el proceso de **descubrir conocimiento**

mientras que la minera de los datos es el método de extraer patrones de los datos.

Knowledge Discovery in Databases

KDD: Extracción de conocimiento en BDD



KDD: 1. Limpieza

Al iniciar el proceso de KDD, usualmente se cuenta con una BD, que contiene todo lo almacenado por una organización sobre algún contexto.

Podemos pensar en una BD que nos ofrece algún banco, con la información de todos sus clientes que tienen una tarjeta de crédito.

Se denominan atributos a cada una de las características. Por ejemplo para un cliente, sus características esenciales serían el identificador del cliente, nombre, apellidos, edad, RFC(Registro Federal de Contribuyentes), domicilio, teléfono, empresa donde labora, ingresos, etc.

Cada uno de los atributos tiene un conjunto de valores que distinguen si un atributo es discreto (número finito de valores) o continuo (número infinito de valores posibles).

KDD: 1. Limpieza

Una base de datos relacional es una colección de tablas con un nombre único.

Cada tabla tiene un conjunto de atributos (campos o columnas), almacenando los valores en conjuntos de tuplas (registros o filas).

En una tabla relacional, cada tupla representa un objeto identificado por una clave única, permitiendo crear un modelo semántico tal como el modelo de datos Entidad-Relación.

Ejemplo de una base de datos relacional:

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	...
...	...	...	...	...	...	...	...

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
...	...	...	...	...	...	...	...	...

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
...	...	...

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
...	...	...	...	...	...

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
...	...	...	...	...	...	...

items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
...	...	...

KDD: 1. Limpieza

Cada uno de estos atributos presenta un diferente grado de beneficio para el contexto que se esta manejando.

Algunos de los valores asignados para cada ejemplo en la BD puede presentar errores o inconsistencias.

Esta fase dentro del proceso de KDD es la encargada de tratar con este estado inicial de los datos en las BD, **eliminando el ruido e inconsistencias sobre los datos**, encontrando los atributos que aporten mayor beneficio al contexto.

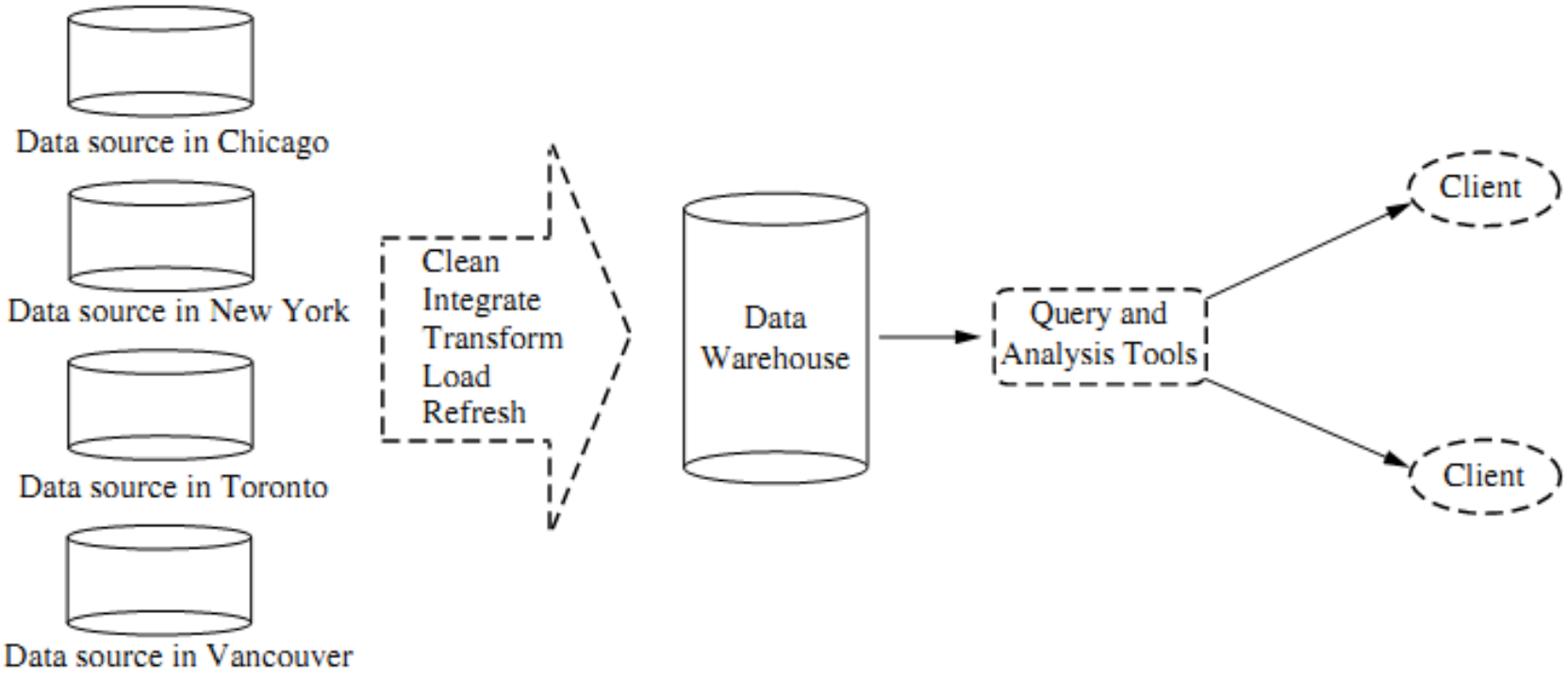
También se deberán combinar o integrar múltiples fuentes de datos:

Bases de datos relacionales, data warehouses, bases de datos transaccionales, bases de datos orientadas a objetos, bases de datos temporales, bases de datos secuenciales, series de tiempo, datos espaciales (mapas), diseños, hojas de datos sobre circuitos, hipertextos, multimedia, stream data, archivos de texto.

KDD: 1. Limpieza

Data warehouse: Es un repositorio de información integrada a partir de múltiples fuentes, almacenado bajo un esquema unificado y que usualmente reside en un solo sitio. Es construido a través del proceso de limpieza de datos, integración, transformación, carga y actualización de los datos.

Ejemplo de data warehouse:



KDD: 2.Selección

Una vez que tenemos una BD más refinada, gracias a la fase de limpieza

La fase de selección es aplicada para encontrar los datos relevantes del problema, las “clases” o grupos a las que corresponden cada uno de los registros, patrones o muestras.

En caso de no existir dichas clases, se debe aplicar técnicas de agrupamiento para asignarle a los patrones a algún tipo de etiqueta, identificando las clases y finalmente la unión de varias fuentes de datos que puedan enriquecer el tamaño de los datos a ser utilizados.

El estado que tendrá la BD será más completa que la que se obtuvo en la fase de limpieza, ya que se incluyen las clases que serán de suma importancia para la fase de Minería de Datos.

KDD: 3.Preprocesamiento y 4.Transformación

Una vez que los datos han pasado por la fase de limpieza y selección, aun tienen que redefinirse, usando un formato que permita aplicar los métodos de Minería de Datos.

Dentro de las herramientas que permiten aplicar Minería de Datos sobre BD grandes, se utiliza un formato especial para el manejo de las BD, algunas manejan el formato relacional de SQL o DB2 por ejemplo.

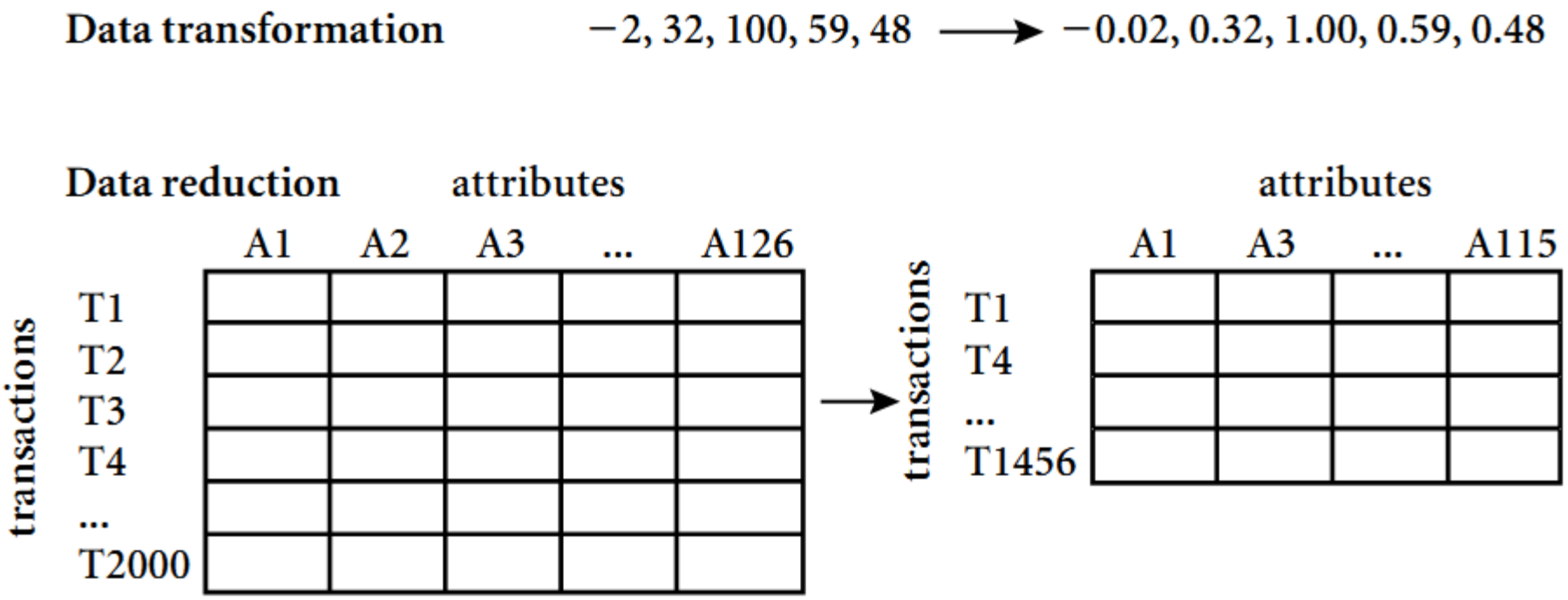
Hay que decidir primero qué herramientas se van a usar para aplicar Minería de Datos, para saber de antemano qué formato deberá llevar la(s) BD

Por ejemplo transformar los valores booleanos true-false a valores numéricos 0-1 ya que la mayoría de las técnicas solo trabaja con números.

Otro ejemplo es normalizar los valores, usando un valor real o flotante entre 0 y 1 en vez de números grandes o negativos.

KDD: 3.Preprocesamiento y 4.Transformación

Ejemplo de transformación y reducción de datos:



KDD: 5.Minería de datos

Es en esta fase donde se aplican métodos de extracción de patrones .

También se aplican técnicas para evaluar el rendimiento de los métodos anteriores.

En esta fase se plantean una serie de metas que pueden realizarse al aplicar Minería de Datos:

Predicción: Descubrir la manera de cómo ciertos atributos, dentro de los datos, se comportarán en el futuro.

Identificación: Identificar la existencia de objetos, eventos y actividades dentro de los datos.

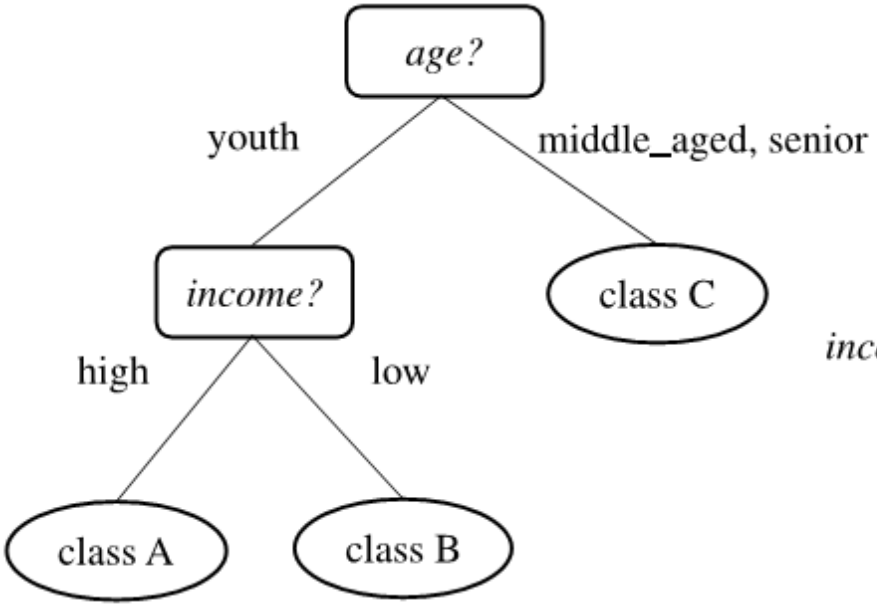
KDD: 5.Minería de datos

Clasificación: Particionar los datos de acuerdo a las clases o etiquetas que sean asignadas a cada ejemplo de muestra.

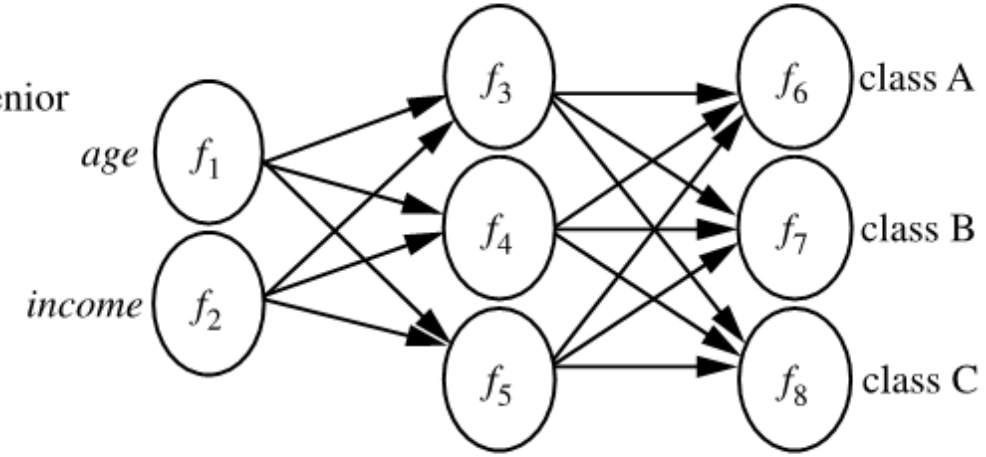
Ejemplo de clasificación usando reglas IF-THEN:

age(X, "youth") AND income(X, "high") → class(X, "A")
age(X, "youth") AND income(X, "low") → class(X, "B")
age(X, "middle_aged") → class(X, "C")
age(X, "senior") → class(X, "C")

Ejemplo de clasificación usando un árbol de decisión

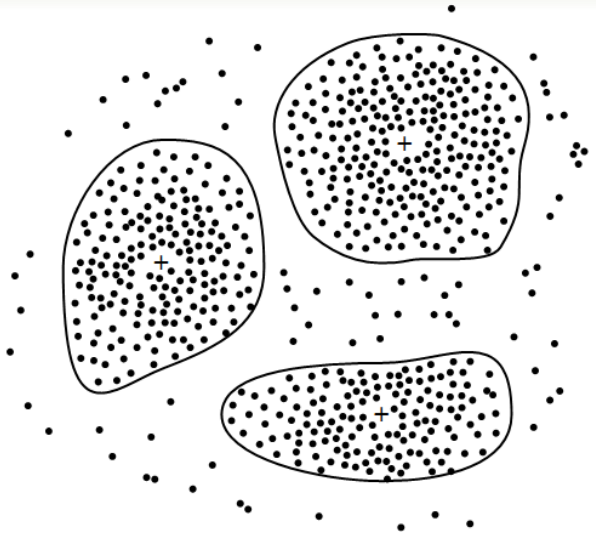


Ejemplo de clasificación usando una red neuronal artificial



KDD: 5.Minería de datos

Agrupamiento: Permite maximizar las similaridades y minimizar las diferencias entre los objetos, mediante algún criterio de agrupamiento.



Asociación: Las reglas de asociación intentan descubrir cuales son las conexiones que se pueden tener entre los objetos identificados.

Ejemplos de reglas de asociación multidimensionales:

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}software\text{”}) \quad [support = 1\%, confidence = 50\%]$$

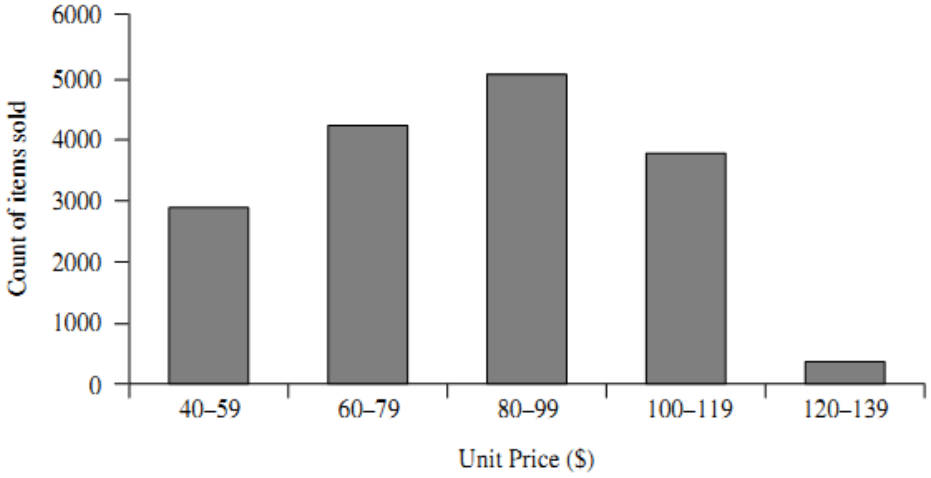
$$age(X, \text{“}20...29\text{”}) \wedge income(X, \text{“}20K...29K\text{”}) \Rightarrow buys(X, \text{“}CD\ player\text{”})$$
$$[support = 2\%, confidence = 60\%]$$

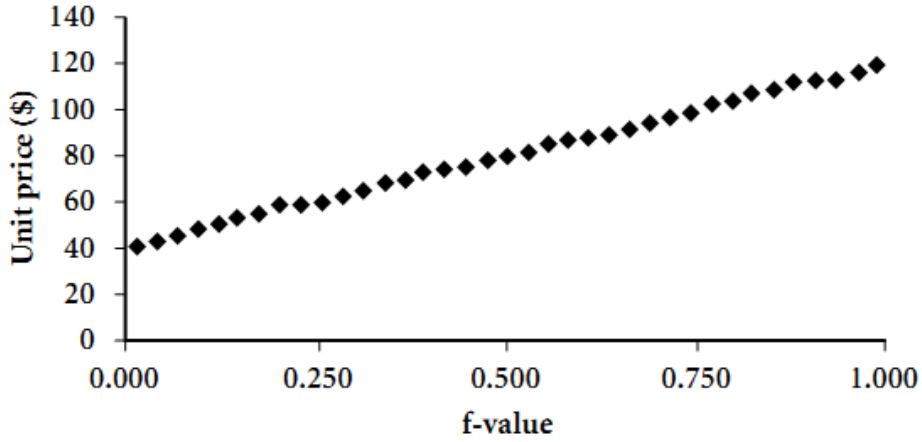
KDD: 6.Interpretación y evaluación

Se realiza un análisis de los patrones obtenidos hasta la fase de Minería de Datos.

Se usan técnicas de visualización y de representación del conocimiento, para obtener la solución a la problemática implícita dentro de la BD.

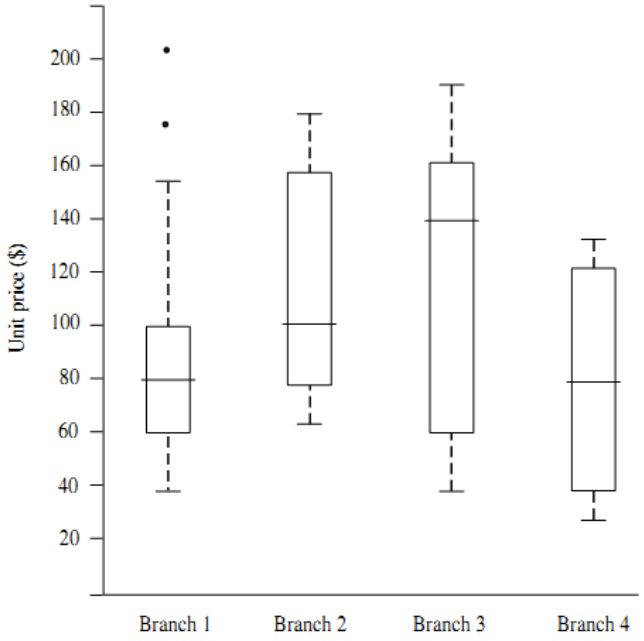
El objetivo principal es encontrar el conocimiento, un valor más significativo a partir de los datos.

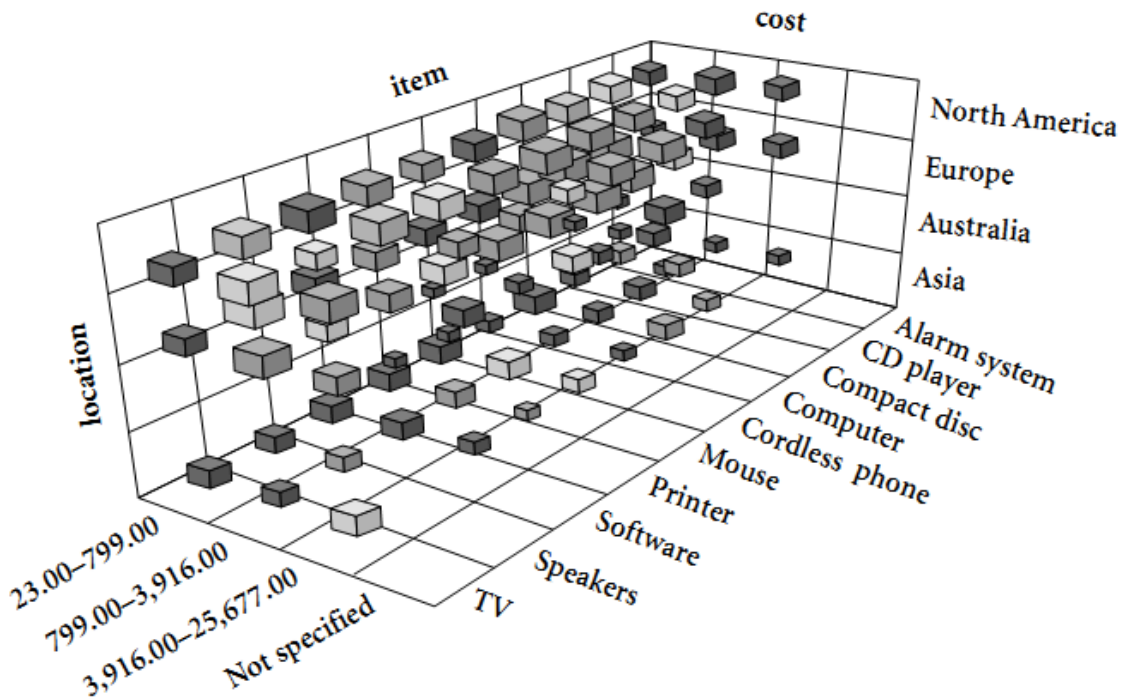




KDD: 6.Interpretación y evaluación

Las técnicas de visualización y representación del conocimiento también permitirán que el usuario pueda identificar de forma clara el conocimiento obtenido.





Bibliografía

Número	Autor	Título	Año de Publicación	Edición	Editorial
1	Jose Hernandez Orallo	Introducción a la mineria de datos	2004	1era	Pearson
2	Jose Tomas Palma Mendez. Roque Marin Morales	Inteligencia artificial: métodos, técnicas y aplicaciones	2008	1era	McGraw-Hill
3	Han, D. J.	Principles of Data Mining	2007		MIT Press
4	Maimon, O. Z. and L. Rokach	Data mining and knowledge discovery handbook	2005		Springer
5	Pérez López, C. and D. Santín Gonzalez	Data Mining-Soluciones Con Enterprise Miner	2006		Alfaomega, Ra-Ma
6	Sumathi, S. and S. N. Sivanandam	Introduction to data mining and its applications	2006		Springer-Verlag
7	Tan, P. N., M. Steinbach, et al.	Introduction to data mining	2005		Addison-Wesley Longman Publishing Co., Inc.

Bibliografía

Número	Autor	Título	Año de Publicación	Edición	Editorial	ISBN
8	Samira ElAtia, Donald Ipperciel, Osmar R. Zaiane.	Data mining and learning analytics : applications in educational research	2017		John Wiley & Sons, Inc.	ISBN: 9781118998205 libro electrónico,1118998200
9	ALFREDO DAZA VERGARAY	DATA MINING. MINERIA DE DATOS	2017		S.A. MARCOMBO	ISBN: 9788426724588
10	Jordi Gironés Roig, Jordi Casas Roma, Julia Minguillon Alfonso, Ramon Caihuelas Quiles	MINERÍA DE DATOS: MODELOS Y ALGORITMOS	2017		UOC (UNIVERSITAT OBERTA DE CATALUNYA)	ISBN: 9788491169031
11	Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar	Introduction to Data Mining	2005	2nd	Pearson	ISBN-13: 9780134545943
12	Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman	Mining of Massive Datasets	2014		Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman	ISBN-13: 978-1107015357 ISBN-10: 1107015359

Bibliografía

Número	Autor	Título	Año de Publicación	Edición	Editorial	ISBN
13	Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze	An Introduction to Information Retrieval	2009		Cambridge University Press	ISBN-10: 0521865719; ISBN-13: 978-0521865715
14	H. Garcia-Molina, J.D. Ullman, and J. Widom	Database Systems: The Complete Book Second Edition	2009		Prentice-Hall	ISBN-10: 0131873253; ISBN-13: 978-0131873254
15	J.D. Ullman and J. Widom	A First Course in Database Systems	2008	3rd	Prentice-Hall	ISBN-10: 013600637X; ISBN-13: 978-0136006374
16	Han, J. and M. Kamber	Data mining: concepts and techniques	2006		Morgan Kaufmann	ISBN-13: 978-9380931913 ISBN-10: 0123814790