Unidad 7

Fases del Proceso de Descubrimiento de Conocimiento

Introducción

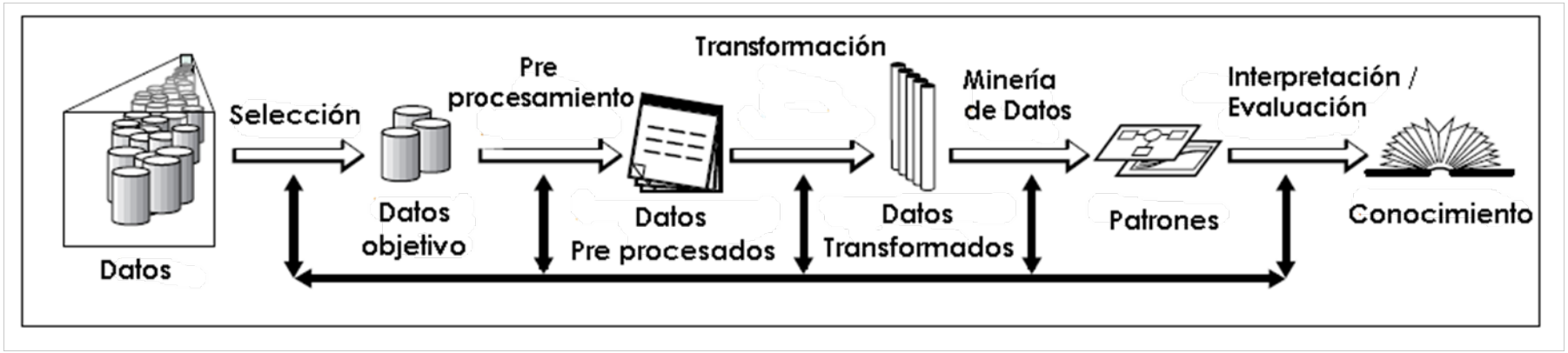
Descubrimiento de Conocimiento

- Permite a partir de ciertos ejemplos y patrones identificados, poder realizar clasificaciones y predicciones de otros casos similares

KDD

- La Extracción del Conocimiento en Base de Datos es conocido como KDD
- *Knowledge Discover in Databases*
- Busca descubrir información útil dentro de los datos almacenados en bases de datos
- Busca encontrar patrones y determinar relaciones

Etapas de KDD



Etapas de KDD

- Se han clasificado diversas etapas para KDD, las más generales son:
  - Identificar Objetivos
  - Seleccionar Datos e Información
  - Pre procesamiento de los Datos
  - Transformación
  - Técnicas de Minería de Datos
  - Interpretación de los Resultados o Modelos
  - Integración al Negocio

Planteamiento de Objetivos

1. Identificación de Objetivos

- Consiste en definir los objetivos del procesamiento de información
- Considerar el tipo de información que se tiene disponible y las relaciones que se desean encontrar entre ellas

El Manejo de los Datos

La Materia Prima

- Los datos se consideran como la materia prima de la Minería de Datos
- La recopilación de los mismos debe ir acompañada de una limpieza e integración para su mejor análisis

Tipos de Datos

- A pesar de la gran cantidad de tipos de datos, se pueden agrupar en las siguientes categorías:
  - Numéricos
  - Nominales sin orden {lógicos, booleanos}
  - Nominales ordenados {bajo, medio, alto}

2. Selección de Datos e Información

- Ya definidos los objetivos, se debe seleccionar la información a utilizar
- Un análisis más profundo de los datos disponibles para determinar su utilidad
- Seleccionar un conjunto de datos o sub datos para ser procesados
- Buscar elementos en común entre los datos para su relación

Integración de los Datos

- Se realiza normalmente durante la recopilación de los datos
- Se busca que los datos sobre el mismo objeto se unifiquen y los datos de diferentes objetos permanezcan separados

Problemas con la Integración

- El principal problema de integrar distintas fuentes de información viene dada por la inconsistencia
- También es posible que aparezcan datos mezclados sin relación unos con otros

Ejemplo

ID	Edad	CP	Estado	Años Socio
123	35	31678	Casado	10
456	23	45677	Soltero	3

ID	Nacimiento	Ciudad	Casado	Credencial
765	1992	México	SI	A-45
123	1980	Toluca	NO	B-68

ID	Edad	CP	Estado	Casado	Nac	Ciudad	Años	Cred
123	35	31678	Casado	NO	1980	Toluca	10	B-68
456	23	45677	Soltero	...	...	...	3	
765	...	...	...	SI	1992	México	...	A-45

Reconocimiento

- Una vez integrados los datos, se realiza un resumen de sus características
- Se generan tablas con las características generales de los atributos
  - Medias
  - Mínimos
  - Máximos
  - Posibles valores
- Se puede distinguir entre valores numéricos y nominales

3. Pre procesamiento de los Datos

- Una vez definidas las fuentes de datos e información se procede a una etapa de limpieza
  - Manejo de Datos Faltantes
  - Manejo de Datos Redundantes
  - Manejo de Datos Incompletos
  - Manejo de Datos Inconsistentes
- Se busca obtener una estructura bien definida para el procesamiento posterior

Preparación de los Datos

- También conocida como *Datacooking*
- Su principal objetivo es eliminar la mayor cantidad de datos erróneos o inconsistente para presentarlos de la mejor manera

Valores Faltantes

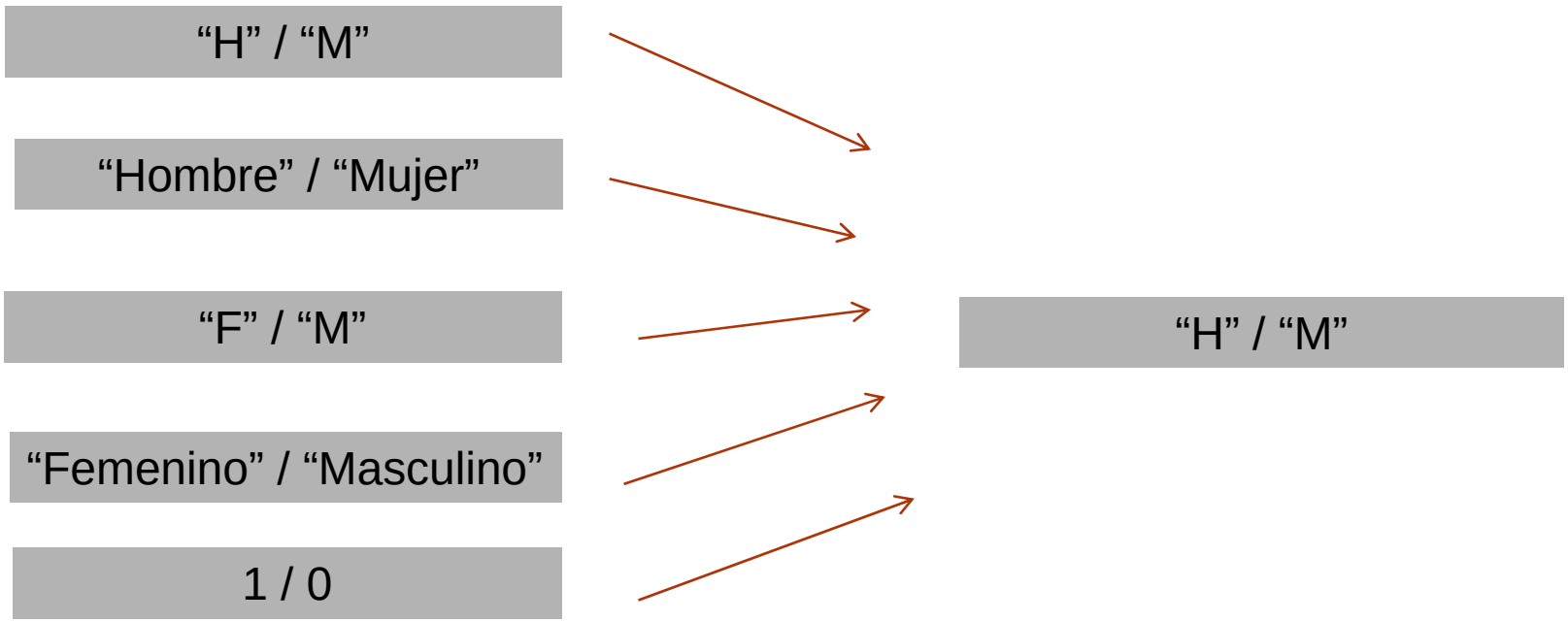
- En el caso de los datos inconsistentes, se recomienda dejar los datos como faltantes
- Los valores faltantes pueden tratarse de diferentes maneras
  - Ignorarlos. Solo cuando se usan ciertos algoritmos
  - Eliminar la columna. Cuando son muchos los faltantes
  - Filtrar la fila. Puede ocasionar un sesgo en la información
  - Reemplazar el valor. Con la media, varianza o moda
  - Predecir el valor. En base a otras filas similares

4. Transformación

- Se aplican diversas técnicas para la modificación o generación de nuevas variables a partir de las que ya se tienen disponibles
- Se eliminan variables no deseadas
- Entre las técnicas más comunes se tiene:
  - Selección de Características
  - Reducción
  - Transformación de Atributos
  - Discretización de valores numéricos

Manejo de Atributos

- Cuando sea posible, se recomienda fusionar los datos o generar solamente una categoría



5. Técnicas de Minería de Datos

- Se utilizan diversas técnicas para obtener patrones de interés a partir de la información oculta en los datos
  - Minería de Datos
  - Minería de Textos
  - Análisis de Series de Tiempos
- Elegir los algoritmos de Minería de Datos más adecuados al tipo de análisis que se quiere realizar

6. Interpretación de Resultados o Modelos

- Se interpretan los modelos o patrones que se obtuvieron a partir de ciertas medidas de evaluación
- Se utiliza el conocimiento y se incorpora a un sistema para futuras acciones

7. Integración al Negocio

- Se realizan los reportes necesarios
- A pesar de que se liga mucho la Minería de Datos con el negocio, los modelos obtenidos también se utilizan para la toma de decisiones en otras áreas