Extracción de conocimiento en bases de datos

Christopher Expósito Izquierdo
Airam Expósito Márquez
Israel López Plata
Belén Melián Batista
J. Marcos Moreno Vega

{**cexposit, aexposim, ilopezpl, mbmelian, jmmoreno**}**@ull.edu.es**

Departamento de Ingeniería Informática y de Sistemas
Universidad de La Laguna

Índice

INTRODUCCIÓN

Magnitud

Información generada

- *There was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing.*

Eric Schmidt
CEO, Google
(2001-2011)

Tipo y variedad de datos

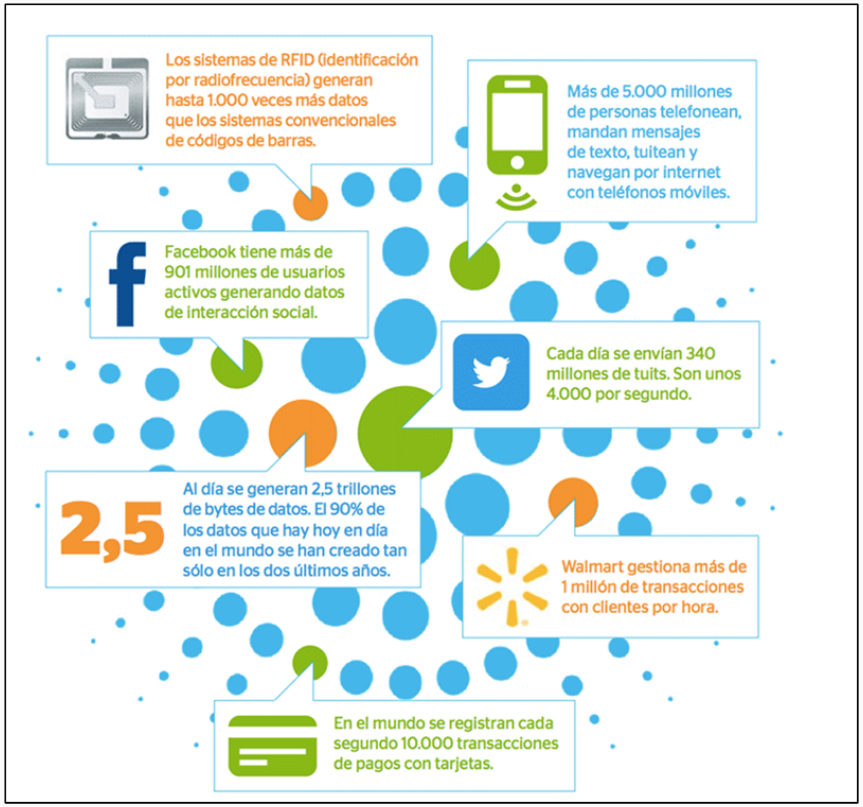


Figura 1: centrodeinnovacionbbva.com

Qué pasa en un día en internet

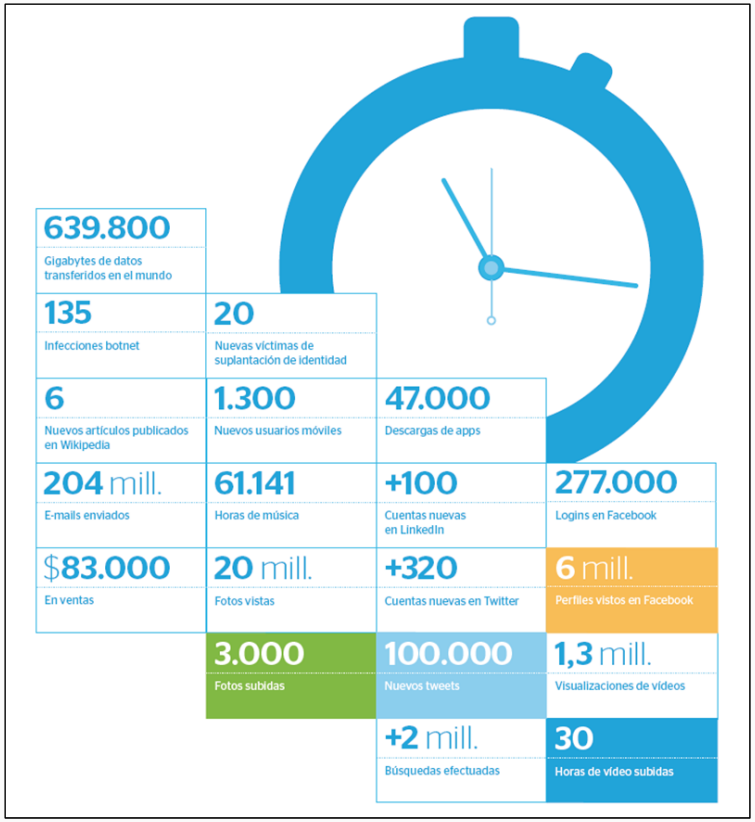


Figura 2: centrodeinnovacionbbva.com

Aplicaciones

El Big Data echa una mano al campo

- Combinando los datos que suministran los sensores que instala sobre el terreno con otros procedentes de plataformas abiertas o públicas (AEMET, Google ...), la empresa española bynse recomienda a los agricultores cuándo regar, cuándo plantar o por dónde empezar un tratamiento fitosanotario.

Aplicaciones

Minería de datos para identificar las causas de los accidentes de tráfico

- El grupo de investigación Transporte y Seguridad (TRYSE) del departamento de Ingeniería Civil de la Universidad de Granada, analizando 3229 accidentes y 18 variables relacionadas con las carreteras, los vehículos, los conductores y el entorno en que se produjeron los accidentes, ha identificado relaciones interesantes entre las variables que podrían usarse para mejorar la seguridad vial.

Durante el análisis emplearon técnicas de agrupamiento (clustering) con las que segmentaron los accidentes en cuatro subgrupos bien definidos. Estos subgrupos revelaron las relaciones existentes entre la severidad del accidente y las condiciones de la carretera, las características de los conductores o las franjas horarias.

Aplicaciones

Revolución en el trabajo policial

- PREDPOL es un sistema software que, analizando el listado histórico de delitos cometidos, calcula la distribución y frecuencia con que estos se producirán en cada una de las zonas geográficas previamente establecidas.

La información suministrada por el sistema se emplea para planificar el patrullaje de la ciudad.

Importancia

El Big Data creará 4,4 millones de empleos en los próximos dos años

... Big Data también ofrece un nicho laboral importante a tenor de las expectativas puestas en este sector de la industria tecnológica. De hecho, según la consultora Gartner, se estima que para los próximos dos años se habrán generado 4,4 millones de empleos en todo el mundo (1,2 millones sólo en Europa Occidental) relacionados con el sector de las Tecnologías de la Información bajo el tirón tecnológico del Big Data.

Miguel A. Pérez
Blog Think Big
31/05/2013

Importancia

Los titulados en ciencias e ingeniería que transformarán el mercado

El futuro de las empresas pasa por saber lo que quiere el consumidor antes de que él lo sepa. Es lo que se llama comportamiento predictivo, una metodología que a partir de los algoritmos de las redes sociales rastrea las preferencias del usuario e identifica lo que puede necesitar. La migración de las empresas a este nuevo escenario de la economíaa digital requiere de los llamados STEM (siglas en inglés de Science, Technology, Engineering and Mathematics), titulados universitarios en ciencias, tecnología, ingeniería y matemáticas.

Ana Torres Menarguez
Periódico El Pas
01/12/2014

Importancia

El análisis de Big Data se convierte en la profesión más atractiva

En un mercado sujeto a booms pasajeros, la revista Harvard Business Review ha calificado el análisis de Big Data como la profesión más atractiva del siglo XXI, debido a la creciente demanda y a la escasez de expertos preparados para afrontar esta tarea.

A grandes rasgos, el cometido de un analista de Big data consiste en recopilar grandes cantidades de datos en diferentes formatos (documentos de texto, páginas web, archivos de imagen y video, contenidos en redes sociales, dispositivos móviles, apps, sensores, etc) y traducirlos en información relevante y útil para la empresa.

Redacción
Periódico La Vanguardia
28/01/2015

Problemática

Big Data: ¿Vidas privadas al alcance de todos?

Alejandro Giménez, CTO (Chief Technology Officer) de la compañía Dell EMC en España, relativiza el riesgo para nuestra privacidad del Big Data al compararlo con los beneficios que obtendremos al usarlo.

Inma Zamora
Periódico ABC
28/10/2013

Problemática

Mi préstamo para ti, tus datos para mí

Evgeny MOrozov, profesor visitante en la Universidad de Stanford y profesor en la New America Foundation, enumera algunas de las compañías que, analizando datos provenientes de diversas fuentes, identifican a potenciales clientes para un préstamo. También describe algunas de sus técnicas y alerta sobre los peligros que ello conlleva.

Evgeny Morozov
Periódico El País
23/02/2013

EXTRACCIÓN DE CONOCIMIENTO

Extracción de conocimiento

Definición

El proceso de extracción de conocimiento en bases de datos (Knowledge Discovery in Databases, KDD) es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de datos.

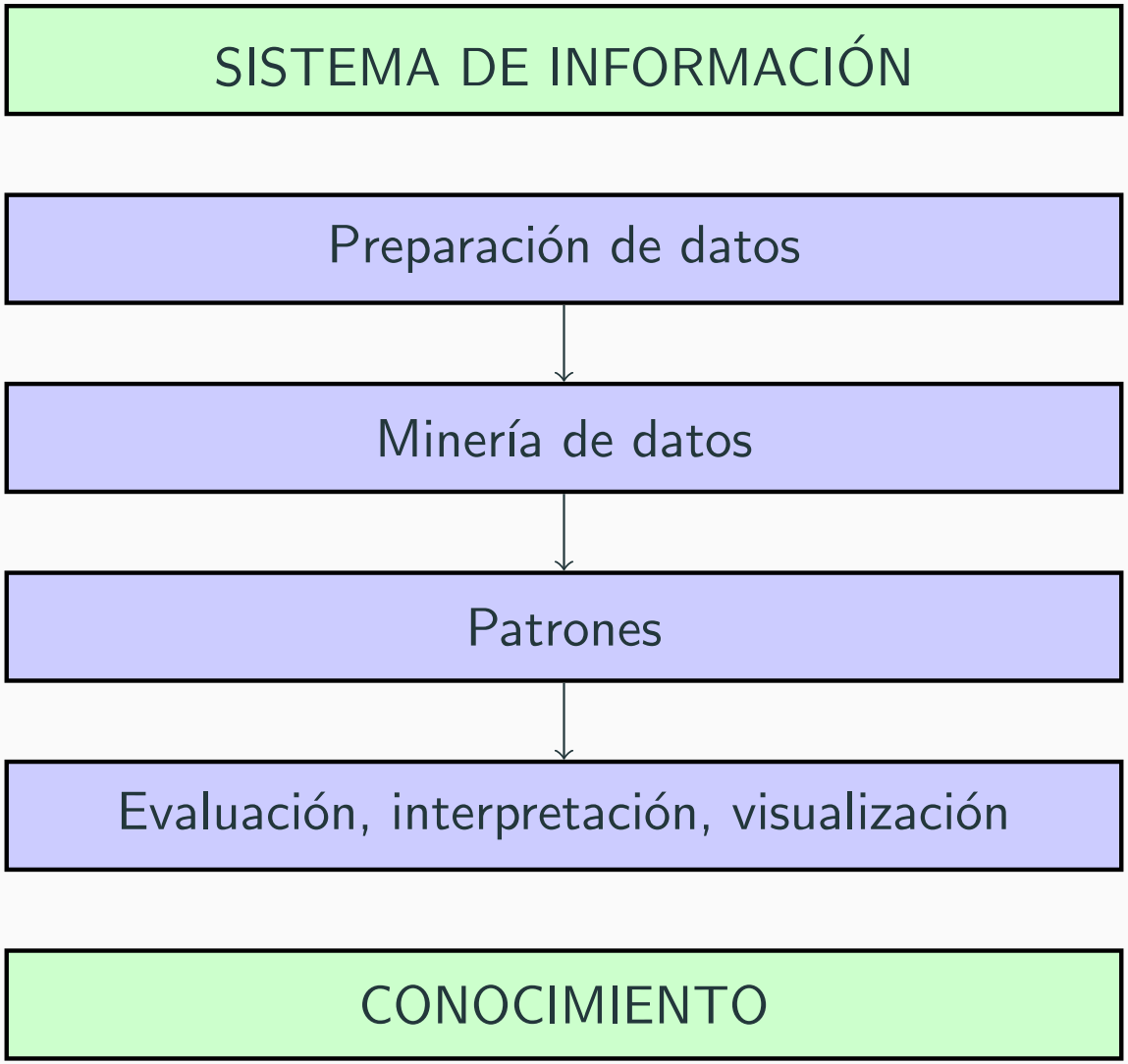
Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. 1996 [1].

Extracción de conocimiento

Propiedades deseables del conocimiento extraído

- **Válido.** El modelo o patrón encontrado debe mostrar un adecuado nivel de precisión cuando se usa con nuevos datos.
- **Novedoso.** El conocimiento extraído debía ser desconocido antes de su obtención.
- **Útil.** El conocimiento obtenido debe permitir mejorar el sistema o adoptar decisiones que aporten algún beneficio al usuario.
- **Comprensible.** En otro caso se dificulta la interpretación, validación y uso del mismo.

Extracción de conocimiento



Relación con otras disciplinas

- Bases de datos
- Aprendizaje automático
- Estadística
- Computación paralela
- Visualización
- Sistemas de soporte a la decisión

Extracción de conocimiento

Aplicaciones

- Banca y finanzas
- Comercio y mercado
- Educación y deporte
- Medicina, biologa y bioingenieía
- Recursos humanos

Extracción de conocimiento

Fases del proceso de extracción de conocimiento

- Integración y recopilación de datos
- Selección, limpieza y transformación
- Minería de datos.
- Evaluación e interpretación
- Difusión y uso

Integración y recopilación de datos

- Las organizaciones complejas se estructuran en departamentos (nóminas, recursos humanos, ventas ...), cada uno de los cuáles genera, almacena y utiliza sus propios datos.
- Sin embargo, para adoptar decisiones estratégicas que implican analizar, planificar o predecir a largo plazo, es necesario recopilar e integrar datos desde diferentes fuentes (internas y externas, públicas o privadas, estructuradas o no).
- Los datos recopilados de múltiples fuentes de datos son integrados en un almacén de datos siguiendo un esquema unificado.

Selección, limpieza y transformación

- La calidad del conocimiento extraído de un conjunto de datos depende del algoritmo de extracción empleado y de la calidad de los datos.
- Para mejorar la calidad de los datos se debe seleccionar y preparar un subconjunto del conjunto total de datos para formar lo que se conoce como vista minable.
- Algunas tareas que pueden o deben realizarse en esta etapa son:
  - Identificar los datos irrelevantes.
  - Identificar los *outliers* y lo valores perdidos (*missing values*).
  - Reducir la dimensionalidad (seleccionar los atributos relevantes, seleccionar una muestra apropiada de datos)
  - Construir nuevos atributos (discretización, transformación ...)

Minería de datos (i)

- En esta fase se obtiene nuevo conocimiento a partir de los datos.
- Para ello se construye un modelo basado en datos.
- El modelo describe los patrones y relaciones encontrados entre los datos y se usa para entender mejor los datos, explicar situaciones pasadas o predecir situaciones futuras.

Minería de datos (ii)

- Identificar la tarea (clasificación, agrupamiento, detección de outlier, regla de asociación) que es más apropiada para la problemática que se tiene.
- Elegir el modelo (árbol de decisión, clasificador bayesiano, agrupamiento jerárquico ...) que resuelve la tarea.
- Seleccionar el algoritmo que construye el modelo escogido en el paso anterior.

Minería de datos (iii)

- **Objetivo:** predecir cuáles de nuestos clientes dejarán de serlo.
- **Alternativa 1**
  - **Tarea:** clasificación
  - **Modelo:** árbol de decisión
  - **Algoritmo:** ID3
- **Alternativa 2**
  - **Tarea:** clasificación
  - **Modelo:** árbol de decisión
  - **Algoritmo:** C4.5

Evaluación e interpretación (i)

- Antes de su uso, todo modelo obtenido desde los datos debe ser evaluado para medir su calidad.
- Es una tarea no trivial que puede depender de varios criterios, algunos de ellos bastante subjetivos y que pueden estar enfrentados entre sí.
- Debe recordarse que los patrones descubiertos deben tener tres cualidades: debe ser válidos (precisos), interesantes (útiles y novedosos) y comprensibles.

Evaluación e interpretación (ii)

- **Etapas de entrenamiento y validación.** Para medir la calidad de los modelos predictivos deben definirse apropiadamente las etapas de entrenamiento y validación. De esta manera se asegura que las predicciones sean precisas y robustas.
  - **Datos de entrenamiento** (training dataset): porción del conjunto de datos que se emplea para construir el modelo.
  - **Datos de validación** (test dataset): porción del conjunto de datos que se usa para validar el modelo obtenido en la etapa anterior.

Evaluación e interpretación (iii)

- **Técnicas de evaluación (i)**

  • **Validación simple:** se divide aletoriamente el conjunto de datos en dos subconjuntos: uno para el entrenamiento y otro para la validación. Con el primero se construye el modelo que luego es validado con los datos del segundo. El conjunto de validación suele constar del 5% al 50% de los datos iniciales.

Evaluación e interpretación (iv)

- **Técnicas de evaluación (ii)**
  - **Validación cruzada:** se divide aleatoriamente el conjunto de datos en dos subconjuntos.
    - Se construye el modelo con el primer subconjunto y se valida con el segundo. Se obtiene así un primer ratio de error (o precisión).
    - Se construye el modelo con el segundo subconjunto y se valida con el primero, obteniéndose un segundo ratio de error.
    - Finalmente, se construye el modelo con todos los datos y se valida sobre el conjunto total.
    - La precisión del clasificador se obtiene como la media de las precisiones anteriores.

Evaluación e interpretación (v)

- **Técnicas de evaluación (iii)**
  - **Validación cruzada con k pliegues:** similar a la validación cruzada, pero dividiendo el conjunto de datos iniciales en k subconjuntos.
    - Se toma el primer subconjunto para validar y se construye el modelo con la unión del resto de subconjuntos.
    - Se repite el paso anterior con cada uno de los $k-1$ subconjuntos restantes.
    - Finalmente, se construye el modelo con todos los datos y se valida sobre el conjunto total.
    - La precisión del clasificador se obtiene como la media de las precisiones anteriores.

Evaluación e interpretación (vi)

- **Medidas de evaluación.** La medida de evaluación empleada depende, entre otros factores, del contexto de aplicación o de la tarea a evaluar.
  - **Clasificación.** Normalmente, se emplea como medida la precisión, que se obtiene como la proporción de instancias clasificadas correctamente.
  - **Regresión.** La medida más empleada es el *error cuadrático medio* entre el valor real y el obtenido con el modelo de regresión (valor predicho).

Evaluación e interpretación (vi)

- **Interpretación.** En ocasiones las medidas empleadas pueden suministrar valores que deben ser interpretados en el contexto de aplicación o teniendo en cuenta la estructura de los datos.
  - En clasificación, la adopción de la precisión del clasificador como medida de calidad debe tomarse con cierta cautela en la presencia de clases no balanceadas (muchas instancias en algunas clases y pocas o ninguna en otras). En esta situación, la precisión puede ser globalmente aceptable, a pesar de no serlo en algunas clases (en algunas aplicaciones estas últimas clases son las de interés).
  - La matriz de confusión que muestra, para cada clase, el número de casos predichos y actuales, suministra mayor información para evaluar el modelo.

Difusión, uso y monitorización

- El conocimiento extraído de los datos debe difundirse entre los miembros de la organización para su uso.
- Este conocimiento entra a formar parte del *know-how* de la organización y, como tal, es uno de sus bienes intangibles más preciados.
- El modelo obtenido y el conocimiento que de él se deriva deben ser monitorizados para identificar posibles desviaciones en su comportamiento que aconsejen su rediseo.

BIBLIOGRAFÍA

Bibliografía

Fayyad, U. M., Piatetsky-Shapiro, G., , and Smyth, P.
Advances in Knowledge Discovery and Data Mining.
AAI Press, Menlo Park, California, 1996, ch. From Data Mining to Knowledge Discovery: An Overview, pp. 1–30.

Hernández Orallo, J., Ramírez Quintana, M., and Ferri Ramírez, C.
Introducción a la Minería de datos.
Pearson Prentice Hall, 2004.

Licencia