Resumen-sistemas-de-informacion.pdf

user_2320516

Sistemas de Información

3º Grado en Ingeniería Informática

Escuela de Ingeniería y Arquitectura
Universidad de Zaragoza

RESUMEN SISTEMAS DE INFORMACIÓN

CURSO 2023 - 2024

1. INTRODUCCIÓN

1.1 INTRODUCCIÓN

Datos: valores en crudo que representan hechos.
Información: colección de datos organizados de forma que proporcionen valor añadido.
Sobrecarga de información (information overload) vs infobesidad (consumo de información innecesaria). Recientemente: infodemia (pandemia de información).
Conocimiento: conciencia o familiaridad adquirida por la experiencia de hechos o situaciones a través del aprendizaje, observación o introspección.

El origen de los datos se denomina provenance y la trazabilidad o data lineage indica los procesos a los que han sido sometidos.

Proceso: conjunto de tareas lógicamente relacionadas que a partir de datos de entrada proporciona resultados (datos de salida).
Algoritmo: lista ordenada de pasos o especificación de instrucciones para llevar a cabo una determinada tarea.
Sistema: conjunto de elementos (procesos, algoritmos, conocimiento, información y datos) que interactúan para lograr un objetivo.

Los componentes de cualquier tipo de sistema son:

Entradas
Mecanismos de procesado: elementos físicos y protocolos
Salidas
Retroalimentación (feedback)

Métricas de cualquier tipo de sistema:

- **Feedback**
  Efectividad: mide hasta qué punto se ha alcanzado el objetivo del sistema
  Eficiencia: mide el beneficio con respecto al consumo realizado para obtenerlo (optimización de recursos)
- **Medidas de rendimiento estándar** específicas del sistema

1.2 SISTEMA DE INFORMACIÓN

Es un **sistema** que está compuesto por un **conjunto de elementos interrelacionados** que **recogen** (entrada), **manipulan** (proceso), **almacenan** información y **diseminan** (salida) datos y además, proporcionan **mecanismos correctores** (feedback) para alcanzar determinado objetivo.

Componentes de un sistema de información:

Hardware
Software
Datos
Personas
Procedimientos y protocolos

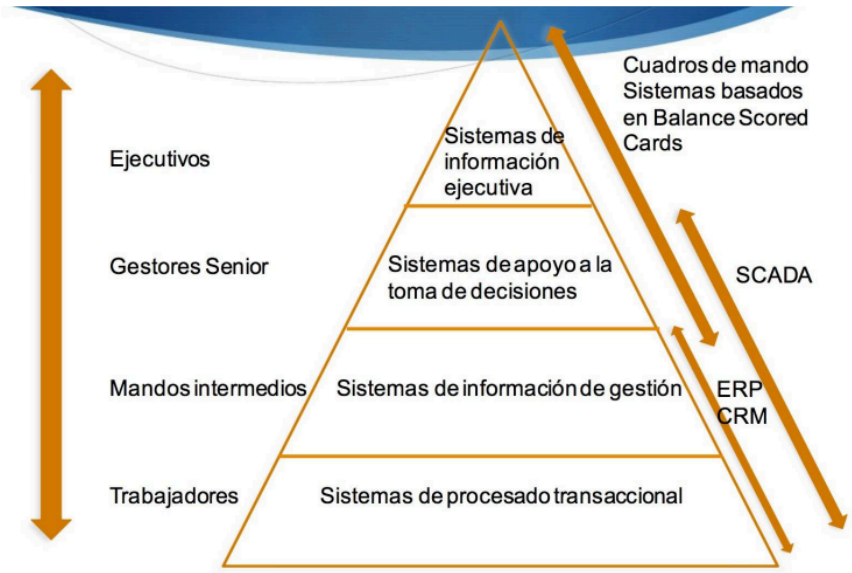
Clasificación de los sistemas de información:

Sistemas de procesado transaccional (TPS): gestionan la información referente a las transacciones producidas diariamente en una organización como la compra de materiales, ventas de productos.

Sistemas de gestión (MIS): orientados a los responsables técnicos de las diferentes áreas de la organización para la definición de procedimientos rutinarios como la generación de nóminas, facturación.

Sistemas de apoyo a la toma de decisiones (DSS): dan soporte a la toma de decisiones para un problema específico complejo. Además, en general, la información a considerar para analizar el problema no está definida. Ejemplo: software destinado a la creación de data warehouses.

Sistemas de información para ejecutivos (EIS): DSS para altos directivos. Orientados a conseguir los objetivos estratégicos de la empresa. Su principal uso es informativo.



1.3 APLICACIONES EMPRESARIALES

Características:

Almacenan y manipulan datos: bases de datos y ficheros xml

Realizan transacciones: propiedades ACID (Atomicity-Consistency-Isolation-Durability)

Escalables: más carga de trabajo sin necesidad de modificar el software

Disponibles: no dejar de prestar servicio

Seguras: permisos acceso a datos y funcionalidades

Integración: diferentes tecnologías

Tipos de interfaces: **texto, entorno de ventanas** (cliente de escritorio)**, web, apps móviles**

Arquitecturas software:

Aplicaciones monocapa:

Modelo de datos: dependiente de la aplicación en concreto (no tiene en cuenta la integración con el resto del sistema/aplicaciones)

Persistencia: ficheros.

Ventajas: rápidas, útiles para aplicaciones de propósito específico.

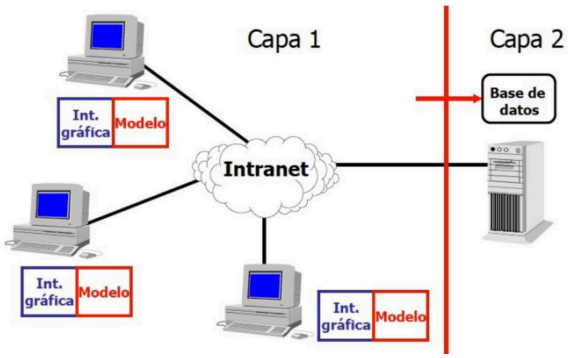
Inconvenientes: Necesaria instalación y re-compilación en todas las máquinas. Datos duplicados y necesidad de procesos ETL.

Aplicaciones de dos capas:

Separación entre interfaz y modelo.

Ventajas: cada capa puede ser desarrollada por personal con perfiles específicos. Reuso de la capa modelo para diferentes dispositivos.

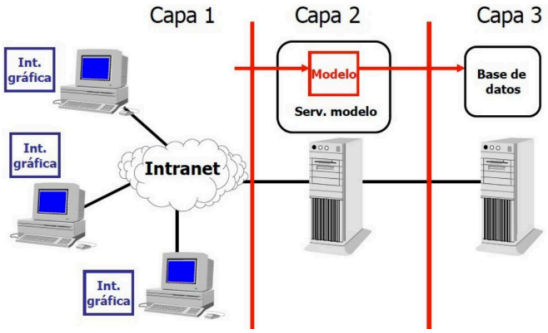
Inconvenientes: cambios en el modelo implican la re-compilación e instalación en todas las máquinas clientes.



Aplicaciones de tres capas:

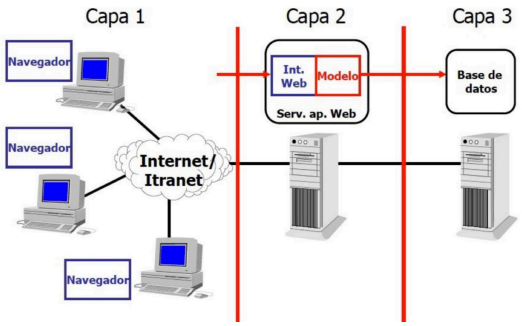
Ventajas: cambios en el modelo sólo afectan al servidor de la aplicación. Clientes ligeros que necesitan poca capacidad de procesamiento.

Inconvenientes: cambios en la interfaz gráfica implican la re-compilación e instalación de la aplicación cliente.



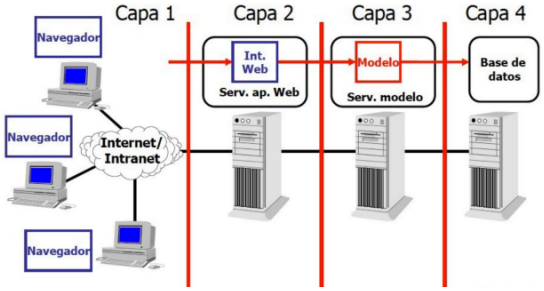
Aplicaciones de tres capas con interfaz Web:

Ventajas: cambios en la interfaz gráfica sólo implican la re-compilación e instalación de la capa interfaz del servidor de aplicaciones Web. Los servidores de aplicaciones Web suelen tener soporte para gestionar la escalabilidad y disponibilidad.



Aplicaciones de cuatro capas:

Esta arquitectura suele emplearse cuando la interfaz gráfica web y la capa de modelo están construidas con tecnologías diferentes. Se requiere una aplicación web para facilitar el acceso a la aplicación.



1.4 NUEVAS TENDENCIAS: CLOUD COMPUTING

Es un paradigma que permite el acceso ubicuo bajo demanda a servicios TIC a través de Internet. Surge de la externalización del servicio TIC y para ahorrar costes.
Cloud computing no es una nueva arquitectura de aplicaciones empresariales, sino una nueva forma de proveer servicios tecnológicos para soportar las aplicaciones empresariales.
Tipos de cloud:
Según su **funcionalidad**:

- **IaaS (Infrastructure as a Service)** -> AWS de Amazon
  · Ofrece recursos de computación: almacenamiento, red, procesadores, …
- **PaaS (Platform as a Service)** ->Microsoft Azure
  · Se ofrecen entornos de desarrollo cooperativo y despliegue de aplicaciones rápido.
  · Es decir PaaS = IaaS + middleware, herramientas de desarrollo, servicios de inteligencia empresarial (BI), sistemas de administración de bases de datos.
  · Diseñado para sustentar el ciclo de vida completo de las aplicaciones web: compilación, pruebas, implementación, administración y actualización.
- **SaaS (Software as a Service)** -> Gmail, Dropbox
  · Se ofrecen aplicaciones en concreto sin controlar el cliente ni la infraestructura hardware ni su configuración.
  · SaaS = PaaS + despliegue de aplicaciones

Según la **compartición:**

- **Público:** cloud destinado a público general o empresas que deseen contratarlo. Ejemplo: AWS de Amazon.
- **Privado:** para uso exclusivo de una organización. Puede ser propio o alquilado. En las máquinas donde se ejecutan los sistemas de la empresa que alquila no se ejecutan sistemas de otras empresas. Ejemplo: Inditex.
- **Híbrido:** los picos se gestionan mediante un cloud público de forma totalmente transparente al usuario. Ejemplo: AWS de Amazon.
- **Comunitario:** cloud privado de una comunidad de organizaciones.

2. EVOLUCIÓN DE LA WEB

2.1 TIPOS DE FUENTES DE DATOS

- **Fuentes de datos no estructuradas:**
  · Documentos o audio en lenguaje natural cuyo contenido es interpretado por personas.
  · Búsqueda de información en repositorios de documentos cerrados, la Web o interfaces basadas en palabras clave
- **Fuentes de datos estructuradas:**
  · Datos que tienen asociados un conjunto de metadatos.
  · Búsqueda de información localizando la fuente de datos y analizando su estructura (metadatos) o en interfaces basadas en SQL.
- **Fuentes de datos híbridas** (La Web)

2.2 WEB ESTÁTICA O TRADICIONAL (WEB 1.0)

Internet: Nodos interconectados (red) y conjunto de protocolos (TCP/IP) que permite comunicar y compartir datos e información entre los diferentes nodos y compartir recursos (impresora, memoria, etc). El origen fue DARPA en la guerra fría y ARPANET.

Web: su objetivo es facilitar la compartición e intercambio de documentos entre científicos involucrados en un proyecto.

URI: Uniform Resource Identifier (Identificador universal de recursos). Es un identificador que se utiliza para identificar de manera única un recurso. Un recurso puede ser cualquier cosa que tenga una identidad, como un documento, una imagen, un servicio, un libro. Ejemplo: urn:isbn:0451450523 es un URI que identifica de manera única un libro por su ISBN.

URL: Uniform Resource Locator (Localizador universal de recursos). Es un tipo de URI que proporciona la ubicación específica de un recurso en la web, es decir, la forma de acceder a él. En otras palabras, una URL es un tipo de URI. Ejemplo: https://www.ejemplo.com/pagina es una URL que se refiere a una página web específica.
En resumen, todas las URL son URI, pero no todas las URI son URL. Mientras que una URL específicamente proporciona la ubicación de un recurso en la web, una URI es un identificador más general que puede referirse a cualquier recurso, ya sea en la web o no.

Documento HTML: Fichero de texto escrito en HTML (Hyper Text Markup Language) que los navegadores interpretan y muestran.

HTTP: es un protocolo sin estado (no guarda información sobre las peticiones anteriores) que sigue un esquema petición-recurso y sirve para comunicar clientes y servidores web. Los recursos se identifican mediante URL. Las peticiones pueden ser de tipo GET (los parámetros se codifican en la URL), POST (los parámetros se codifican en el cuerpo del mensaje), HEAD, PUT o DELETE. Las respuestas del servidor incluyen un código de estado: 1xx (información), 2xx (éxito), 3xx (redirección), 4xx (error cliente) o 5xx (error servidor).

Características Web 1.0:

- Web de solo lectura, grandes corporaciones difunden información (comunicación unidireccional)
- Pocos emisores vs muchos consumidores
- No existe interacción entre los interlocutores
- Problema: información desactualizada
- Coste actualización alto

2.3 WEB DINÁMICA O DE TRANSICIÓN (WEB 1.5)

Cambio tecnológico respecto a la Web 1.0: Generación dinámica de documentos HTML cuando se solicitan, a partir del acceso a información almacenada en bases de datos y ficheros plantilla, para evitar información y datos obsoletos. Se diferencia la ejecución en el cliente y en el servidor.

Web de solo lectura.

Sesión: período de tiempo durante el cual un determinado usuario interactúa con un sitio Web. La debe implementar cada aplicación. Diferentes alternativas: incluir el id de sesión en cookies , en todas las URLs o en campos ocultos en los formularios.

La búsqueda de información se hace mediante buscadores Web o motores de búsqueda.

2.4 WEB SOCIAL (WEB 2.0)

Supone un cambio de filosofía y uso: se usan formularios para recoger datos e información generada por los usuarios, que se convierte en contenido del propio sitio web. Las aplicaciones se centran en usuarios que participan activamente.

Web de lectura / escritura.

Wikipedia: fuente de información no estructurada

La búsqueda de información pasa a ser más compleja por la cantidad de sitios web y páginas, por esto surge la necesidad de escalar los sistemas de búsqueda. También hay buscadores específicos como Google Scholar.

Web oculta (deep web)

2.5 WEB SEMÁNTICA (WEB 3.0)

Web orientada a máquinas y personas y no solo a personas como la Web tradicional y social. Los agentes SW pueden interactuar de forma automática con distintos sitios web.

Web de lectura, escritura y ejecución.

Supone un cambio tecnológico incremental basado en tres pilares fundamentales:
anotación, ontologías y uso de reglas y razonadores.

ANOTACIÓN
Se basa en la existencia de metadatos que identifican los datos.

Tecnologías para la **anotación** de recursos Web:

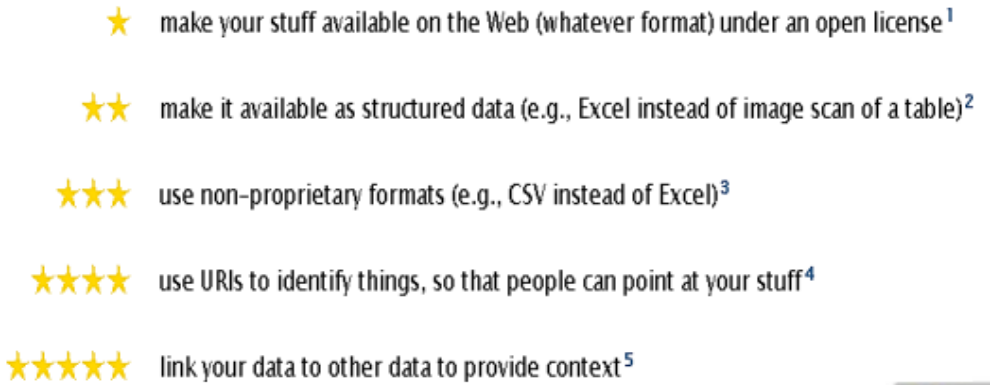
- Para integrar información semántica en las páginas web y realizar anotaciones: **RDFa**
- Para recuperar información semántica: **SPARQL**

ONTOLOGÍAS
Se crea un vocabulario estandarizado ccomún.

Tecnologías para la definición de **ontologías**:
- **RDF** (Resource Description framework)
- **RDFS** (RDF Schema)
- **OWL** (Ontology Web Language)

REGLAS Y RAZONADORES
Existen motores de inferencia automáticos y motores de razonamiento (razonadores)

LA WEB DE DATOS ENLAZADOS
Para que los datos enlazados sean óptimos, hay que cumplir los siguientes principios:



La aplicación que demostró la viabilidad de la Web Semántica: DBpedia, cuyo objetivo era hacer disponible a los agentes software la información disponible en las cajas de información de Wikipedia.Integran información de diversos sitios web.

Las tres reglas de la web de datos enlazados:
- Definir una **URI** para cada **entidad**
- **Formato estándar** para mostrar información de la URI mediante http
- **Relacionar** ese dato con otros

2.6 WEB 4.0

Web en la que no haga falta preguntar a buscadores de información. En base a información del contexto que detecten las necesidades de los usuarios y les ofrezcan recomendaciones. Web donde la información se encuentre geo-localizada.

4. INTRODUCCIÓN A SERVLETS Y JSPs

1. INTRODUCCIÓN

Aplicación Web: es una aplicación que se ejecuta en un servidor Web y a la que el usuario accede desde un cliente de propósito general. En ellas hay contenido estático y dinámico.

Servlet: es un programa escrito en Java que normalmente se ejecuta en respuesta a una petición HTTP. Puede recibir peticiones HTTP. Procesa datos y genera una respuesta acorde con los parámetros generando una página HTML dinámica. Cada servlet puede estar asociado a una o más URLs.

Métodos públicos de **HttpServletRequest**: **getParameter** (atributos monovaluados) y **getParameterValues** (atributos multivaluados y monovaluados).

Método público de **HttpServletResponse: setContentType** establece el tipo de contenido de la respuesta.

Método **sendRedirect** le indica al cliente (como respuesta a la petición http que ha hecho) que genere una nueva petición a la URL que se le indica como parámetro. Es decir, se realiza una redirección con una nueva petición. También se puede realizar una redirección reenviando la petición actual sin que el cliente sea consciente de ello, empleando el método **forward.**

JSPs (JAVA SERVER PAGES)
Una página JSP es un tipo especial de Servlet orientado a crear una interfaz gráfica ya que tiene aspecto de una página HTML y puede incluir scriptlets (scripts) que se escriben en Java.

Cuando se accede a una página JSP:

- Si es **primera vez**, el servidor de aplicaciones web implementa (genera a partir de la página jsp) un servlet asociado a dicha página, lo compila y carga en memoria.
- Si **no es primera vez**, le pasa la petición al servlet.
- Si la página jsp **se ha modificado,** se genera un nuevo servlet.

JUNTANDO SERVLETS Y JSP
JSP genera la vista de la aplicación (visualización de formularios, mensajes de error y resultados de una operación).

Servlets para el procesamiento de los formularios y las llamadas a la capa modelo de la aplicación (si parámetros válidos: realiza la correspondiente operación y pasa el control a la página JSP para mostrar resultados; si no, detecta errores y pasa el control a la página JSP que permita al usuario subsanar errores).

5. TECNOLOGÍA SEMÁNTICA EN LA WEB DE DATOS

RDF (Resource Description Framework) es un modelo de datos estándar para el intercambio de datos en la Web "entendible" por computadoras y para describir relaciones entre los diferentes "recursos" (que se presentan en forma de grafo).

Un recurso es cualquier cosa (concepto) del entorno digital o de otro entorno (página web, un dato de una pág web, libro, artículo, persona…)

RDF está basado en **tripletas** (S, P, O) que se pueden representar en forma de grafo (nodos -> recursos y aristas -> relaciones entre recursos).
Sujeto y objeto representan recursos y predicado representa una propiedad.
Sujeto puede ser URI o nodo en blanco.
Predicado puede ser URI.
Objeto puede ser URI, nodo en blanco o Literal (valor en concreto de un tipo de dato).

RDFS es un lenguaje basado en RDF para definir vocabularios para RDF (RDF Schema); es un lenguaje para definir los metadatos o la estructura de fuentes de datos RDF.

SPARQL es un lenguaje declarativo para extraer información de grafos RDF. Su funcionamiento se basa en el emparejamiento de patrones de la pregunta contra la Base de Conocimientos que estamos interrogando.

	Modelos relacionales	Modelos basados en tripletas
Componente base	Tablas o relaciones	Tripletas (S, P, O)
Definición de metadatos	Sentencias SQL CREATE TABLE	Tripletas RDF considerando RDFS
Definición de datos o instancias	Sentencias SQL INSERT	Sentencias SPARQL CONSTRUCT o definición de Tripletas RDF
Lenguaje de consulta	Sentencias SQL SELECT	Sentencias SPARQL SELECT y ASK

Consulta expresada en SPARQL:

```
SELECT ?ganador
WHERE {
        ?ganador dcterms:subject dbpediaCat:Nobel_laureates_in_Literature
        }
```

6. RECUPERACIÓN DE INFORMACIÓN

6.1 BÚSQUEDA DE INFORMACIÓN EN CORPUS

El objetivo de un sistema de RI es que dada una colección de documentos (corpus documental) y una necesidad de información de un determinado usuario expresada en forma de pregunta (query) recuperar (extraer o listar) los documentos para resolver la necesidad de información del usuario.

Componentes básicos de un sistema de RI:

- Formalismo para representar cada uno de los documentos
- Formalismo para representar las consultas
- Una medida de similitud entre un documento y una consulta

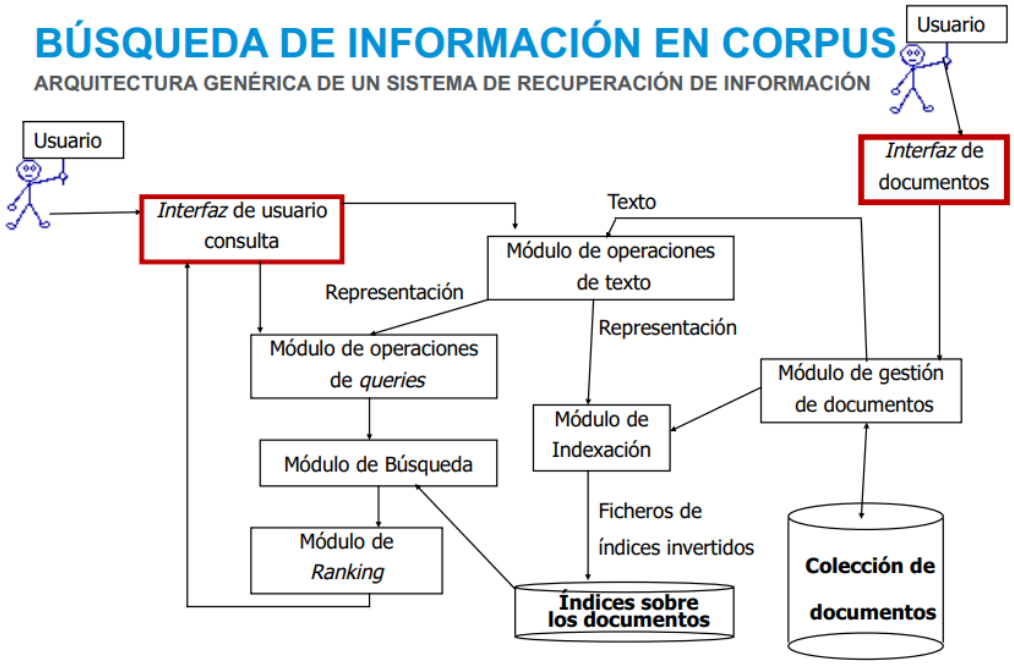
Hay dos posibles soluciones: matching sintáctico (cuando se trabaja con pocos documentos, conlleva problemas como polisemia de palabras y diferentes representaciones del mismo concepto) y recorrido secuencial (se recorre todo el corpus documental comparando el contenido de cada documento con las palabras de la consulta, conlleva problemas para corpus >200Mb pq requiere demasiado tiempo, soluciones: índices invertidos).

Medidas de similitud entre queries y documentos (cómo evaluar un determinado sistema de RI):

- Medidas estándar de **efectividad**:
  - **Precision (P):** nº doc relevantes recuperados / nº doc recuperados
  - **Recall (R):** nº doc relevantes recuperados / nº doc relevantes
  - **F measurement:** medida armónica de P y R:
    { (1 + B^2) P * R } / {B^2 * (P + R) }
  - **Corpus TREC (1992)**

- Medidas de **eficiencia**: menor cantidad de recursos empleados

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN



Módulo de gestión de documentos: gestionar los documentos del corpus y los metadatos que tienen asociados.
Módulo de operaciones de texto: transformar el documento/consulta original en una representación del mismo/de la misma (vista lógica).
Módulo de indexación: gestionar los índices sobre los documentos.

Módulo de operaciones de consulta: uso de recursos lingüísticos y ontologías para enriquecer la representación lógica de la consulta.
Módulo de operaciones de búsqueda: dada la representación lógica de una query realiza consultas en los índices para determinar cuáles son los documentos relevantes para dicha consulta.
Módulo de ranking: ordena los documentos recuperados de acuerdo con la relevancia respecto a la consulta que se está resolviendo.

MÓDULO DE BÚSQUEDA Y RANKING
Modelo booleano: cada documento se representa por una lista de bits (0 ó 1) que indican si en ese documento aparece determinado término del vocabulario del corpus o no.
Ventajas: sencillo de implementar y rápido.
Inconvenientes: recupera muy pocos o demasiados documentos y es difícil representar consultas booleanas para los usuarios.

Modelo booleano extendido: se considera el número de veces que aparece un término en un documento.
Ventajas: se puede hacer ranking de los documentos recuperados.
Dio lugar al modelo vectorial.

Modelo vectorial: las queries y los documentos se representan mediante vectores cuya dimensión es la cardinalidad del vocabulario (conjunto de términos considerados). Medida de similitud entre dos vectores: coseno del ángulo que forman.
Ventajas: sencillo de implementar, poco coste computacional, alto rendimiento.
Inconvenientes: asume independencia de términos, no tiene en cuenta el tamaño de los documentos.

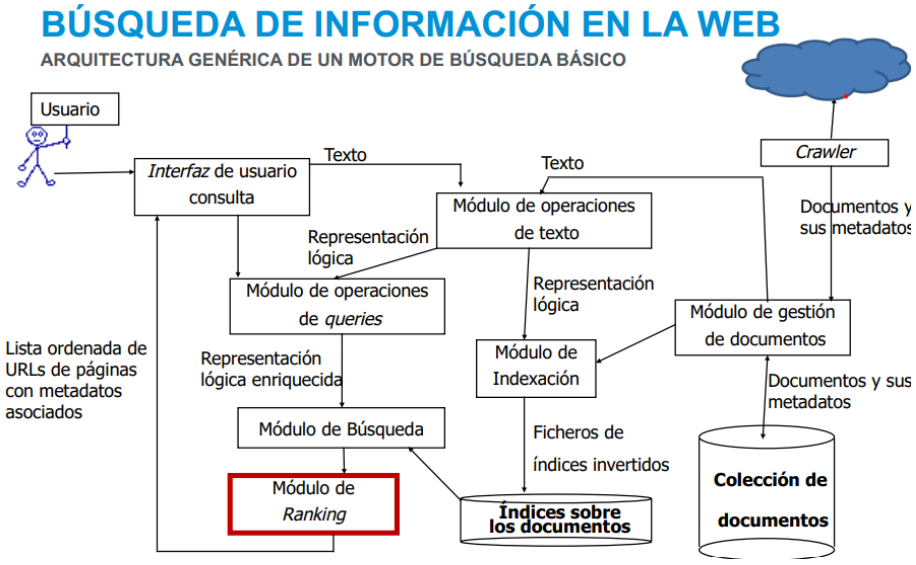
Modelo probabilístico: estrategia adaptativa basada en probabilidades condicionadas y el teorema de Bayes.

6.2 BÚSQUEDA DE INFORMACIÓN EN LA WEB

La diferencia entre Corpus y Web se basa en la necesidad de localizar documentos con los que se va a trabajar. Se usan 3 técnicas: localización manual (dirs de búsqueda), automática (crawlers) e híbrida (crawlers localizan y usuarios clasifican).

Los **directorios de búsqueda** se basan en la organización manual de las páginas en categorías (Yahoo en sus inicios) aunque tiene inconvenientes: escalabilidad y definición de la estructura.

Los **motores de búsqueda** se basan en la adaptación de RI en grandes corpus a la Web mediante el uso de crawlers generalmente con interfaces basadas en palabras clave (Google). Consisten en la construcción de un gran índice de palabras sobre todos los documentos del web estático.



META BUSCADORES

Son buscadores que buscan en buscadores y luego integran sus resultados en tiempo real (ej dogpile). Mejoran la relevancia en buscadores en Internet.

ALGORITMO DE PAGE-RANK

Los enlaces son considerados como “citas” de otros documentos. Se asumen como más relevantes los documentos más citados pero también importa quién es el que te cita. Por tanto, una página tiene un page-rank alto si tiene muchas páginas que la apuntan o la apuntan pocas páginas con un PageRannk alto (“Hubs”).

ALGORITMO HITS

La relevancia se basa en hiperenlaces; primero se realiza una búsqueda previa sobre un índice preconstruido y luego se aplica un algoritmo iterativo sobre los enlaces entre documentos. Hubs son páginas que enlazan muchas páginas buenas (autoridades) y autoridades son páginasa enlazadas desde muchos referentes buenos (hubs).

El algoritmo se basa en que cada página (nodo del grafo) comienza con un ‘peso de hub’ y un ‘peso de autoridad’. En cada iteración el ‘peso de autoridad’ de un nodo se calcula como la suma del ‘peso de hub’ de la iteración anterior de los nodos que lo apuntan. El ‘peso de hub’ se calcula como la suma del ‘peso de autoridad’ de los nodos a los que apunta. Está demostrado que el algoritmo converge.

DEEP WEB

Contenido de la Web no indexado por motores de búsqueda y por tanto, difícilmente localizable a no ser que se conozca su existencia. Los servidores que los alojan se encuentran aislados. Las páginas se generan dinámicamente en función de las peticiones del usuario. Tiene 65% más de tráfico.

7. BASES DE DATOS DISTRIBUIDAS

7.1 INTRODUCCIÓN

Base de datos centralizada: guarda toda la información en la misma base de datos.
Bases de datos distribuidas: buscan la integración de los datos en lugar de la centralización.

7.2 BASES DE DATOS DISTRIBUIDOS (TOP-DOWN)

Es una colección de varias bbdd que se encuentran lógicamente interrelacionadas y desplegadas sobre una red de ordenadores. Un sistema gestor de bbdd distribuidas es el software que permite el manejo de sistemas de bbdd distribuidas y que hace dicha distribución transparente al usuario.

NO son bbdd distribuidas:

- Un sistema de multiprocesado (bd paralelas)
- Un sistema de bbdd que reside en uno de los nodos de una red. Esto es una bd centralizada accesible a través de la red.
- Un conjunto de bbdd que pueden comunicarse unas con otras donde no existe un esquema global.

DISEÑO DE BBDD DISTRIBUIDAS
El diseño top-down se basa en el diseño de un esquema global (E/R, relacional) que posteriormente se fragmenta y finalmente, se elaboran esquemas locales y se asignan los fragmentos a los esquemas locales.

Un **fragmento** es la unidad a distribuir y puede ser parte de una tabla, una tabla entera o un conjunto de tablas.
Tipos de fragmentación:

- **Horizontal:** basada en encontrar condiciones de selección
- **Vertical:** basada en encontrar conjuntos de atributos a proyectar.
- **Híbrida:** primero horizontal y luego vertical.

La fragmentación debe ser:

- **Completa:** todo elemento de la relación debe estar en alguno de los fragmentos.
- **Reconstruible:** la relación inicial debe poder reconstruirse aplicando operadores (JOIN, UNION, …) sobre los factores.
- **Con intersección vacía (disjointness):** a excepción de las claves.

Hay 3 tipos de **asignación:**

- **Sin replicación:** todo fragmento reside en un único nodo (fácil actualizaciones, difícil consultas).
- **Replicación completa:** todos los fragmentos residen en todos los nodos (difícil actualizaciones, fácil consultas).
- **Replicación parcial:** compromiso entre actualizaciones y consultas.

	REPLICACIÓN COMPLETA	REPLICACIÓN PARCIAL	SIN REPLICACIÓN
PROCESAMIENTO DE CONSULTAS	Más fácil	Más difícil	Más difícil
CONTROL DE CONCURRENCIA	Difícil	Más difícil	Más fácil
DISPONIBILIDAD DE LOS DATOS	Muy alta	Alta	Baja

7.3 BASES DE DATOS FEDERADAS (BOTTOM-UP)

Formadas por bases de datos autónomas. Proporcionan un esquema global que se obtiene de abajo a arriba. Los esquemas locales son pre-existentes y se integran en un esquema global. **No hay que fragmentar**, y la redundancia probablemente ya existe.

El problema de obtener un esquema global a partir de N esquemas locales se divide en dos: **traducción** (cada esquema local se traduce a un modelo canónico) e **integración** (los esquemas locales se integran en uno solo).

DISEÑO DE BBDD FEDERADAS

El modelo de datos (canónico) utilizado para expresar el esquema global es muy importante: las bbdd locales pueden ser heterogéneas (distintos modelos de datos) y se utilizan modelos más ricos semánticamente que el relacional (OO, modelos funcionales, …).

Supongamos que los esquemas locales son relacionales y se usa como modelo canónico el modelo E/R extendido de Chen (?):

- **Traducción:** cada esquema local se traduce a un modelo canónico (relacional).
- **Integración:** los esquemas locales se integran en uno solo.

Las consultas realizadas sobre el esquema global deben responderse sobre los esquemas locales. Para ello hay que preservar la **relación de enlace**, es decir, la relación entre los elementos del esquema global y los de los esquemas locales.

7.4 BASES DE DATOS INTEROPERANTES

Están formadas por bbdd autónomas. No proporcionan esquema global sino lenguajes de acceso a bbdd. El usuario es consciente de que trabaja con varias bbdd.

8. ALMACENES DE DATOS

8.1 INTRODUCCIÓN

Problema: las organizaciones manejan enormes cantidades de datos en distintos formatos, que residen en distintas bases de datos y organizados utilizando distintos tipos de gestores de bases de datos. Así, resulta difícil acceder y utilizar todos los datos en aplicaciones de análisis (las cuales requieren extraer, preparar e integrar los datos).
Un **data warehouse** es un repositorio de datos estructurados a nivel de empresa, con datos históricos y actuales, que facilita la toma de decisiones. => Business Intelligence

Tipos de sistemas de información:

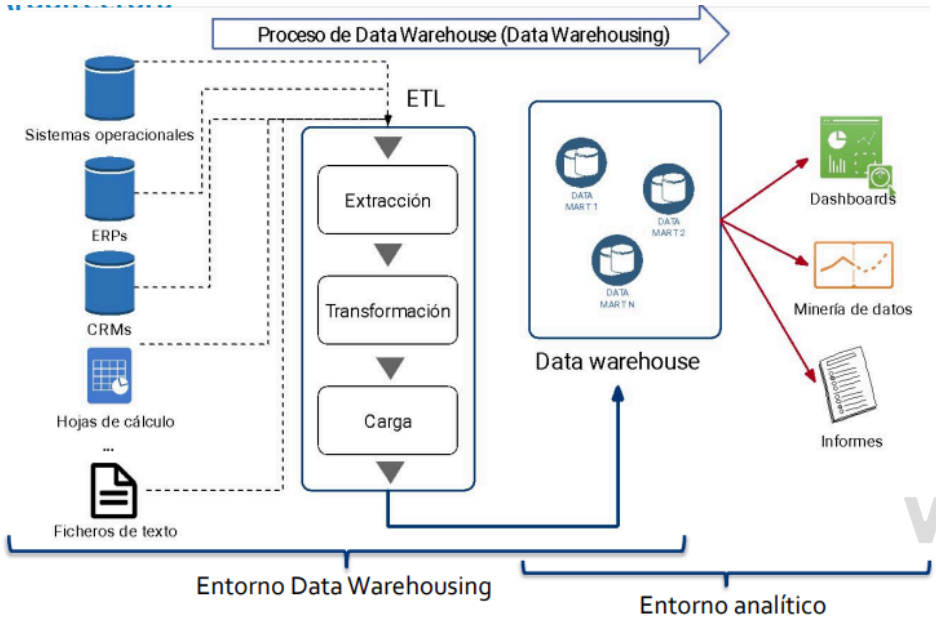
Transaccionales (OLTP)	Analíticos (OLAP)
Datos operacionales	Datos consolidados (suelen provenir de distintas BD OLTP)
Muchas transacciones (INSERT, UPDATE, DELETE)	Pocas transacciones
Datos actuales	Datos actuales e históricos
Información detallada	Información detallada y resumida (integrada) (Consultas complejas – agregaciones □ Data mining)
Los datos cambian continuamente (volátiles)	Datos con mayor estabilidad y menos cambios (no volátiles)

Las características de los almacenes de datos son:

- **Orientados a un aspecto concreto:** la información se guarda en base a un tema de interés para los directivos de la entidad.
- **Integrados:** el almacén de datos suele contener, entre otros, todos los datos de los sistemas operacionales de la empresa. Estos deben ser consistentes.
- **No volátiles:** una vez que los datos han sido incorporados al sistema (registrados) estos no se borran ni actualizan. Además, están pensados para un horizonte de tiempo mucho mayor que los datos operacionales.

8.2 CONSTRUCCIÓN DE DATA WAREHOUSES

ARQUITECTURA



PROCESOS ETL

Primero se realiza una **extracción** de datos de fuentes de datos heterogéneas (bbdd relacionales con datos estructurados, semi-estructurados o no estructurados). Puede llevarse a cabo para realizar una imagen inicial o para actualizar una imagen ya existente. Es muy costoso en tiempo por lo que puede afectar al rendimiento de los sistemas de fuentes de datos.

Por último, se realiza la **carga** de los datos de la transformación. Hay 2 métodos: **carga completa** (imagen inicial) y **carga incremental** (carga en intervalos de tiempo regulares y planificados; se puede hacer en streaming (volúmenes pequeños de datos) o por lotes (volúmenes grandes); se realiza un mantenimiento de históricos).

El **staging area** facilita los procesos de extracción y transformación de los datos antes de ser incluidos en el data warehouse.

MODELOS MULTIDIMENSIONALES

Para organizar datos de un DW se usan **cubos n-dimensionales o hipercubos** (estructura que se emplea para organizar los datos en el Data Mart. Tiene múltiples dimensiones). En cada una de las dimensiones hay un nivel de detalle diferente.

Un **slice** (loncha) es el subconjunto de datos multidimensionales definidos por seleccionar valores específicos de cada uno de los atributos que definen las dimensiones.

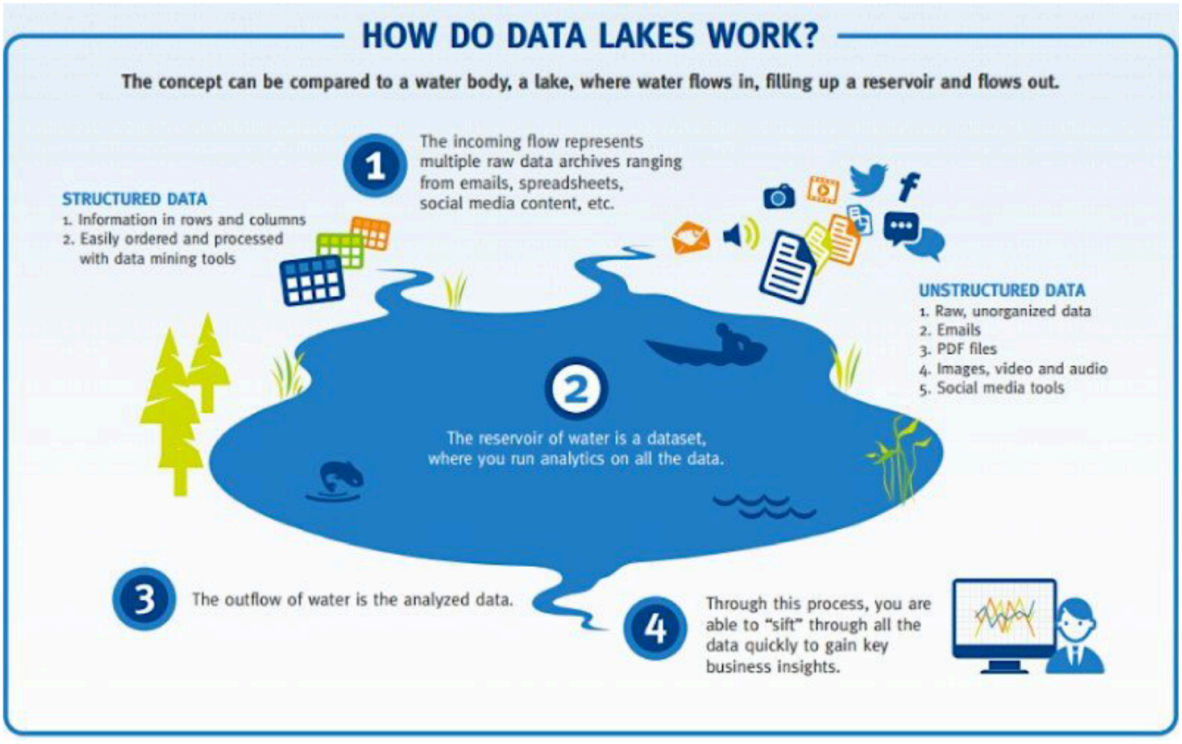
Hay 2 formas de implementar los cubos: **virtual** (opción más simple: una sola tabla con múltiples columnas que representan o bien las dimensiones que se consideran o bien los datos de interés para el análisis, pe nº ventas de un producto; esquema en estrella) o **física** (bbdd multidimensionales, matriz n-dimensional almacenando los valores).

La **arquitectura en estrella** se basa en una tabla central que contiene la información de los hechos que se desea analizar conectada a las diferentes tablas que representan las diferentes dimensiones.

Se pueden generar **informes configurables**. Hay 2 operadores definidos sobre los informes: **drill down** detalla los resultados obtenidos añadiendo un campo y **roll up** agrega los resultados obtenidos eliminando un campo.

TEMAS RELACIONADOS: DATA LAKES

Data warehouse	vs	Data lake
estructurados, preprocesados	DATOS	estructurados, semi-estructurados , no estructurados
esquema al escribir	PROCESAMIENTO	esquema al leer
costoso para grandes volúmenes	ALMACENAMIENTO	Diseñado para bajo coste
menos ágil, configuración fija	AGILIDAD	muy ágil, configuración bajo demanda
madura	SEGURIDAD	en proceso
directivos	USUARIOS	analistas de datos (entre otros)



9. MINERÍA DE DATOS Y TEXTOS

9.1 INTRODUCCIÓN

La **minería de datos** se basa en descubrir, a partir de los datos, conocimiento interesante.

Según el **objetivo general** la minería de datos puede ser predictiva o descriptiva.

Los **pasos** en el proceso de la minería de datos son:

1. Aprender sobre el dominio de la aplicación
2. Seleccionar los datos a analizar
3. Preparar los datos (cleaning -> fiabilidad)
4. Reducir y transformar los datos
5. Escoger el objetivo de la minería de datos (agrupamiento, clasificación, asociación…)
6. Escoger un algoritmo de minería de datos apropiado
7. Analizar los datos
8. Explotación del conocimiento descubierto

9.2 REGRESIÓN

La regresión lineal trata de obtener la línea que mejor se ajusta a los datos. La regresión no lineal trata de obtener el polinomio que mejor se ajusta a los datos.

9.3 MINERÍA DE PATRONES Y REGLAS DE ASOCIACIÓN

Conjuntos de ítems:
X = {x1,…,xk} //conjunto de elementos del antecedente
Y = {y1,…,yl} // conjunto de elementos del consecuente
A = {evento | evento ⊂ X} B = {evento | evento ⊂ Y} A ∩ B = {evento | evento ⊂ (X ∪ Y)}

Reglas X -> Y
Si ocurre X entonces también ocurre Y. A ∩ B es el conjunto de eventos que cumple la regla $X \rightarrowtail Y$, es decir, que contienen el conjunto de items X ∪ Y = {x1,…,xk, y1,…,yi}.

Tres **medidas** importantes:

- **Soporte (support)**: probabilidad de que una transacción tenga X∪Y.

$$\text{soporte } (X \rightarrowtail Y) = \frac{N_{X \cup Y}}{N}$$

N = Número total de instancias

$N_{X \cup Y}$ = Número de instancias que contienen X e Y

Valores entre 0 y 1 (0 = ningún soporte, 1 = soporte total)

- **Confianza (confidence):** probabilidad condicional de que una transacción que contenga X también contenga Y.

$$\text{confianza } (X \rightarrowtail Y) = \frac{N_{X \cup Y}}{N_X} = \frac{soporte(X \cup Y)}{soporte(X)}$$

NXUY=número de instancias que contienen X e Y

Nx=número de instancias que contienen X
Valores entre 0 y 1

- **Elevación (lift):** tasa (ratio) entre el soporte y el producto de las probabilidades de cada conjunto de ítems por separado.

$$lift(X \to Y) = \frac{N_{X \cup Y}/N}{(N_X/N)*(N_Y/N)} = \frac{soporte(X \to Y)}{(N_X/N)*(N_Y/N)} = \frac{confianza(X \to Y)}{N_Y/N}$$

La elevación indica el incremento de la probabilidad de que ocurra el consecuente de la regla si se da el antecedente.
Si es >1 → correlación positiva (si se da X es más probable que se de Y).
Si es =1 → sucesos independientes. Da igual si se da X o no para que se de Y.
Si es <1 → correlación negativa (si se da X, es menos probable que se de Y).
Cuanto más se aleja de 1, mayor es la evidencia de que la regla no se debe a un artefacto aleatorio.

8.4 AGRUPAMIENTO (CLUSTERING)

Un **cluster** o agrupamiento es un conjunto de entidades tales que hay cohesión (similitud dentro del cluster) y diferenciación (disimilitud entre agrupaciones diferentes). Interesa maximizar tanto la cohesión como la diferenciación.

8.5 CLASIFICACIÓN

La **clasificación** se basa en asociar cada elemento del conjunto de datos a una serie de categorías definidas previamente.

Hay 2 tipos de **variables:**
- Una variable **respuesta** o dependiente: etiqueta cada instancia con la categoría correspondiente.
- El resto son variables **predictoras** o independientes

El objetivo es explicar la variable dependiente en términos de las variables independientes.

MATRIZ DE CONFUSIÓN

Caso binario

		Clase detectada	
		Positiva	Negativa
Clase real	Positiva	tp	fn
	Negativa	fp	tn

Caso multiclase

		Clase detectada		
		A	B	C
Clase real	A	tp_A	e_{AB}	e_{AC}
	B	e_{BA}	tp_B	e_{BC}
	C	e_{CA}	e_{CB}	tp_C

En ambos casos la **diagonal** representa los datos bien clasificados, mientras que los otros elementos son los errores cometidos.

MÉTRICAS TÍPICAS

La **precisión (recall)** representa la capacidad para no identificar incorrectamente instancias como pertenecientes a una clase. Valor entre 0 y 1.
Precision = tp / (tp + fp)
Para la clase A: Precision A = tpA / (tpA + eBA + eCA)

El **recall (exhaustividad)** o sensitivity representa la capacidad para no dejarse instancias de una clase sin identificar correctamente como pertenecientes a la misma. Valor entre 0 y 1.
Recall = Sensitivity = tp / (tp + fn)
Para la clase A: Recall A = tpA / (tpA + eAB + eAC)

El **F-measure** es una medida armónica de la precision y el recall.
F-measure = (2 * precision * recall) / (precision + recall)

La **accuracy** es la medida de la corrección global del modelo.
Accuracy = nº clasificaciones correctas / nº total clasificaciones realizadas

MEDIDAS TÍPICAS CLASIFICACIÓN BINARIA

El **FPR** (false positive rate = false alarm rate = fallout) representa cuántos resultados se detectan como positivos de forma incorrecta entre todas las muestras negativas.
FPR = fp / (fp + tn)

El **TPR** (true positive rate = sensitivity = recall) representa cuántos resultados se detectan como positivos de forma correcta de entre todas las muestras positivas.
TPR = tp / (tp + fn)

El **FNR** (false negative rate) representa cuántos resultados se detectan como negativos de forma incorrecta de entre todas las muestras positivas.
FNR = fn / (tp + fn) = 1 - TPR

El **TNR** (true negative rate = specificity) representa cuántos resultados se detectan como negativos de forma correcta entre todas las muestras negativas.
TNR = SPC = tn / (fp + tn) = 1 - FPR

EVALUACIÓN

Para evaluar hay que separar los datos en conjunto de entrenamiento y conjunto de test (conjunto de validación). Nunca hay que evaluar sobre el mismo conjunto de datos utilizados para entrenar.

La **evaluación cruzada de k vías**(k-fold cross-validation) vías se realiza de la siguiente manera:

- Se particiona la muestra inicial en k muestras de igual tamaño.
- 1 de las muestras se utiliza como test y las k-1 restantes para entrenar.
- Se repite k veces, de forma que cada una de las k muestras se utilizan y vez como test.
- Se combinan los resultados de las k evaluaciones (promedios)
- Si k=n(número de muestras) => leave-one out cross-validation

9.6 MINERÍA DE TEXTOS

La **minería de datos textuales** es el proceso de derivar información de alta calidad a partir de fuentes de texto (no estructuradas o mínimamente estructuradas). Para ello se estructura la entrada (pre-procesamiento del texto) y se aplica minería de datos sobre los datos estructurados.

Las tareas típicas son:

- Clasificación de textos (categorización).
- Agrupación de textos (extracción automática de temas).
- Extracción de información (entidades y sus relaciones, datos de interés, correferencias)
- Análisis del sentimiento / minería de opiniones.
- Generación de resúmenes.

TÉCNICAS

En cuanto a las tareas de pre-procesamiento de textos destacan las siguientes técnicas:

- Reconocimiento de caracteres (OCR)
- Tokenización
- Eliminación de stopwords
- Lematización
- Stemming
- Etiquetado gramatical
- Análisis sintáctico (parsing)
- Desambiguación

Los documentos se representan de forma vectorial, es decir, se representan en un espacio vectorial multidimensional => bolsas de palabras. Los términos son las dimensiones del espacio y los documentos son puntos o vectores en este espacio. El valor de cada componente del vector se determina a partir de la frecuencia del término en el documento y su frecuencia inversa. La similitud entre los documentos puede calcularse midiendo el ángulo formado por sus vectores. Para mejorar el rendimiento, podrían seleccionarse únicamente las palabras más frecuentes.

11. CONTEXTO NORMATIVO

Principales derechos del ciudadano:

- **Derecho de acceso:** se puede ejercitar para conocer si el responsable está tratando o no los datos de carácter personal del solicitante, y en caso de que se esté realizando dicho tratamiento, obtener información como: una copia, los fines del tratamiento, los destinatarios, el plazo previsto de conservación, …
- **Derecho de rectificación:** se puede ejercitar si los datos personales incluidos en una actividad de tratamiento son inexactos, y deben ser rectificados sin dilación indebida del responsable. Teniendo en cuenta los fines del tratamiento, mediante este derecho se puede solicitar que se completen los datos personales que sean incompletos, inclusive mediante una declaración adicional.

- **Derecho de oposición:** se puede ejercitar para oponerse a que el responsable realice un tratamiento de los datos personales en los siguientes supuestos: cuando sean objeto de tratamiento basado en una misión de interés público o en el interés legítimo o cuando el tratamiento tenga como finalidad la mercadotecnia directa.
- **Derecho de supresión ("al olvido"):** Es el derecho de un individuo a solicitar la eliminación de sus datos personales cuando estos ya no sean necesarios para la finalidad, cuando se retira el consentimiento del tratamiento de estos (haciendo uso de su derecho a oposición), o cuando se hayan tratado los datos personales de forma ilícita.

MAGERIT

El análisis de riesgos considera los siguientes elementos:

- **Activos:** son los elementos del sistema de información que soportan la misión de la organización.
    - **Esenciales:** información y servicio.
    - **Secundarios:** datos, servicios auxiliares, aplicaciones informáticas, …
- **Amenazas:** eventos que pueden ocurrir a los activos y que pueden causar un perjuicio a la organización.
- **Salvaguardas:** medidas de protección desplegadas para reducir el potencial daño que producen las amenazas.