TEMA 105. SISTEMAS DE RECUPERACIÓN DE LA INFORMACIÓN. POLÍTICAS, PROCEDIMIENTOS Y MÉTODOS PARA LA CONSERVACIÓN DE LA INFORMACIÓN

Actualizado a 16/01/2018

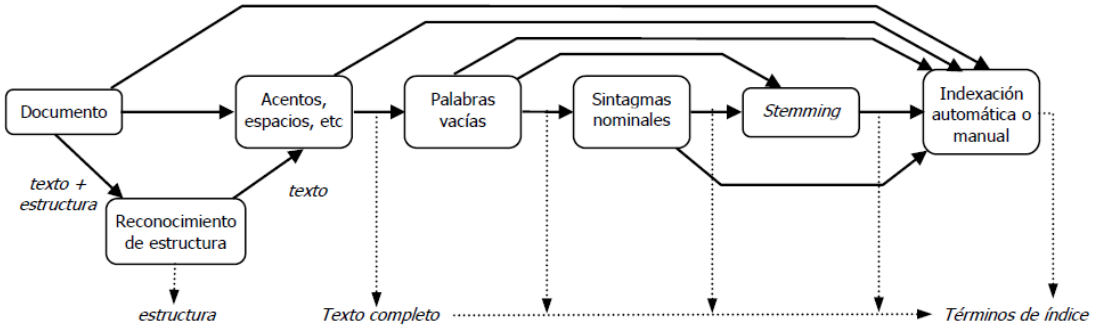
1. CONTEXTO NORMATIVO CONSERVACION INFORMACIÓN

NORMA	OBSERVACIONES
LEY 39/2015 DEL PROCEDIMIENTO ADMINISTRATIVO COMÚN DE LAS ADMINISTRACIONES PÚBLICAS	- Artículo 17. Archivo de documentos - Artículo 26. Emisión de documentos por las Administraciones Públicas. -Artículo 27. Validez y eficacia de las copias realizadas por las Administraciones Públicas -Artículo 70. Expediente administrativo
REAL DECRETO 1671/2009, DE 6 DE NOVIEMBRE, POR EL QUE SE DESARROLLA PARCIALMENTE LA LEY 11/2007	Artículo 41. Características del documento electrónico Artículo 42. Adición de metadatos a los documentos electrónicos Artículo 43. Copias electrónicas de los documentos electrónicos realizadas por la Administración General del Estado y sus organismos públicos Artículo 51. Archivo electrónico de documentos Artículo 52. Conservación de documentos electrónicos
LA LEY ORGÁNICA 15/1999, DE PROTECCIÓN DE DATOS PERSONALES Y SU RD DECRETO DE DESARROLLO 1720/2007 OJO NUEVO REGLAMENTO	Conservación datos de carácter personal
EL ESQUEMA NACIONAL DE INTEROPERABILIDAD, REAL DECRETO 4/2010	Persigue la creación de las condiciones necesarias para garantizar el adecuado nivel de interoperabilidad técnica, semántica y organizativa de los sistemas y aplicaciones empleados por las Administraciones Públicas Un aspecto clave relacionado con este tema como es la recuperación y conservación del documento electrónico
EL ESQUEMA NACIONAL DE SEGURIDAD, REAL DECRETO 3/2010	Tiene como objeto establecer la política de seguridad en la utilización de medios electrónicos y está constituido por principios básicos y requisitos mínimos que permitan una protección adecuada de la información.
LEY 25/2007, DE 18 DE OCTUBRE, DE CONSERVACIÓN DE DATOS RELATIVOS A LAS COMUNICACIONES ELEC-TRÓNICAS Y A LAS REDES PÚBLICAS DE COMUNICACIONES	En el capítulo II de conservación y cesión de datos se pone de manifiesto la obligación a los operadores que presten servicios de comunicaciones a conservar los datos de comunicaciones realizadas (art. 4)
LEY 25/2013, DE 27 DE DICIEMBRE, DE IMPULSO DE LA FACTURA ELECTRÓNICA Y CREACIÓN DEL REGISTRO CONTABLE DE FACTURAS EN EL SECTOR PÚBLICO	El artículo 7 habla sobre el archivo y custodia de la información de facturas
ESTÁNDARES	
UNE-ISO/TR 15801:2008	recomendaciones para la veracidad y fiabilidad de información almacenada electrónicamente
UNE-ISO/TR 18492:2008	Conservación a largo plazo de la información

	basada en documentos.
ISO/TR 26102	Requisitos para la conservación a largo plazo de documentos electrónicos
UNE-ISO 19005-1:2008	Formato de archivo de documentos electrónicos para conservación a largo plazo

2. RECUPERACIÓN DE LA INFORMACIÓN

los documentos de una colección se representan frecuentemente a través de un conjunto de términos de índice o palabras clave (keywords). Dichas palabras clave podrían ser extraídas directamente del texto o ser especificadas por un operador humano, y proporcionan la visión lógica de los documentos desde la óptica del sistema de recuperación de información



2.1. OPERACIONES SOBRE TEXTOS

- Eliminación de palabras vacías (stop words), como artículos o conjunciones.
- Identificación de la raíz gramatical (stemming)
- Identificación de sintagmas nominales
- Compresión de los conjuntos de palabras obtenidos.

Al aplicar las operaciones sobre textos a los documentos originales, se obtiene la visión lógica de los mismos. A partir de la misma, el gestor de bases de datos realiza la indexación automática de los textos. Los índices generados en este proceso son estructuras de datos críticas que permiten realizar búsquedas rápidas sobre volúmenes extensos de información. Aunque existen múltiples estructuras de indexación, la más popular es la conocida como archivo invertido (inverted file).

2.2. MODELOS Y TÉCNICAS DE RECUPERACIÓN

MODELO BOOLEANO

Las consultas se formulan como una combinación booleana de términos. Los operadores utilizados habitualmente son Y, O y NO.

El peso relativo de un término dentro de un documento sólo será 1 (si está presente en el mismo) ó 0 (si está ausente).

Respecto a una consulta dada, un documento será relevante si cumple todas las condiciones establecidas en la consulta, o no relevante si no las satisface: no se permiten distintos grados de relevancia. El proceso de consulta basado en el modelo booleano se caracteriza por ser del tipo "prueba con reintento" con alto grado de retroalimentación hasta obtener una versión definitiva de la consulta que suministre una cantidad razonable de documentos.

MODELO VECTORIAL

El modelo vectorial, por tanto, se basa en representar cada documento según un vector formado por los pesos de sus términos de índices, y que permite ver a un documento como un punto en el espacio formado por todos los términos de índice.

El cálculo de los pesos se puede realizar de varias maneras, pero uno de los esquemas más empleado es el cono-cido como term frequency x inverse document frequency, **tf*idf**

MODELO PROBABILÍSTICO

El modelo probabilístico, también conocido como modelo de recuperación de independencia binaria, (Binary Independence Retrieval, BIR)

Existe un resultado ideal

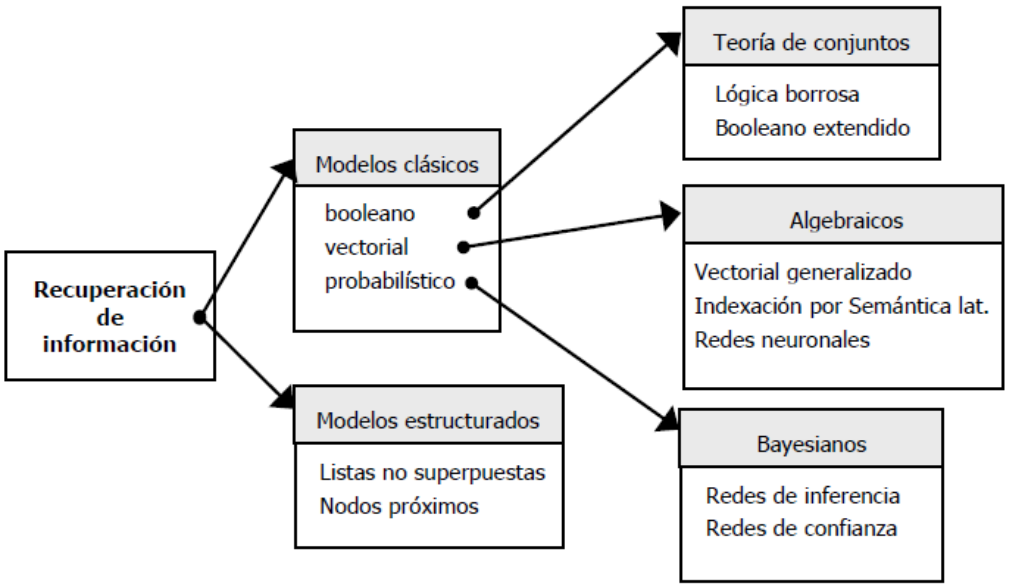
Considera exclusivamente de la presencia o ausencia de los términos en los documentos de la colección. Se trata pues, de un modelo binario, como el modelo booleano.

El modelo probabilístico actúa sobre los términos que configuran la consulta del usuario, asignándoles un peso de manera que se cuanto mejor permita discernir los documentos relevantes de los irrelevantes mayor será, y menor en caso contrario.

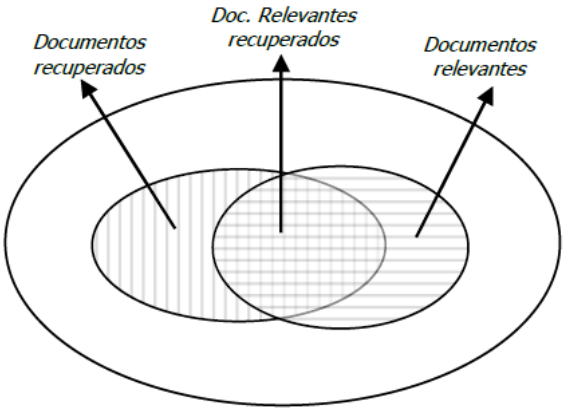
Para depurar los resultados, el modelo probabilístico requiere de un aprendizaje en el que haya una intensa inter-acción con el usuario, lo que se denomina **realimentación por relevancia**

COMPARACION DE MODELOS

- El modelo booleano puro se considera el más débil al no proporcionar coincidencias parciales ni, por tanto, ordenación por relevancia.
- Los experimentos realizados apuntan a que, en general, el modelo vectorial tiene un comportamiento superior al probabilístico cuando se utiliza contra colecciones genéricas de documentos.



2.3. MEDIDAS DE EVALUACIÓN



MEDIDAS TRADICIONALES

Precisión = Factor de pertinencia = Ratio de aceptación

$$\textit{Precisión} = \frac{\textit{Documentos relevantes recuperados}}{\textit{Documentos recuperados}}$$

Exhaustividad (recall) = índice de retorno = sensibilidad

$$\textbf{Exhaustividad} = \frac{\textit{Documentos relevantes recuperados}}{\textit{Documentos relevantes}}$$

Ruido

$$\textbf{Factor de ruido} = \frac{\textit{Documentos no relevantes recuperados}}{\textit{Documentos recuperados}}$$

MEDIDAS DE EVALUACIÓN DESDE EL PUNTO DE VISTA DEL USUARIO

Ratio de cobertura

$$\textbf{Cobertura} = \frac{\textit{Docs. relevantes conocidos recuperados}}{\textit{Documentos relevantes conocidos}}$$

Ratio de novedad

$$\textbf{Novedad} = \frac{\textit{Docs. relevantes desconocidos recuperados}}{\textit{Documentos relevantes recuperados}}$$

Exhaustividad relativa, que indica la proporción de documentos relevantes recuperados que han sido examinados por el usuario, respecto del número total de documentos que el usuario quiere examinar.

Esfuerzo de exhaustividad, que indica la proporción de documentos relevantes deseados partido por el número de documentos que ha de examinarse para localizar los anteriores

2.4. INDEXACIÓN Y RECUPERACIÓN AUTOMÁTICAS

- Indices invertidos. Hay una fila por cada término de índice contemplado y una columna por documento de la colección. Existen diversas estructuras utilizadas para recoger el resultado de la indexación, de forma que se agilicen las tareas de recuperación. Pese a las diferencias entre ellas, la gran mayoría se basan en el concepto de fichero **índice invertido, o fichero inverso**

	D1	D2	.....	Dn
T1	1	0	.....	1
T2	0	1	.....	0
.....	.....	.....	.....	.....
Tt	0	1	.....	1

- Lenguaje Natural (interesante ver Agenda Digital Plan de Impulso de las Tecnologías del Lenguaje)

2.5. RECUPERACIÓN DE INFORMACIÓN MULTIMEDIA

Los metadatos necesarios para categorizar colecciones de objetos multimedia se pueden dividir en tres tipos

- Metadatos semánticos, que tratan de caracterizar el asunto o material de la que trata el documento.
- Metadatos de contexto, que describen relaciones del documento con objetos externos, como autor, intérprete, etc.
- Metadatos estructurales, que describe la estructura interna y modo de presentación del documento.

2.6. EXTRACCIÓN INFORMACIÓN EN LA WEB

Aunque los diversos motores de búsqueda existentes en la Web tienen cada uno su arquitectura y estructura específica, en todos ellos es posible encontrar los siguientes subsistemas:

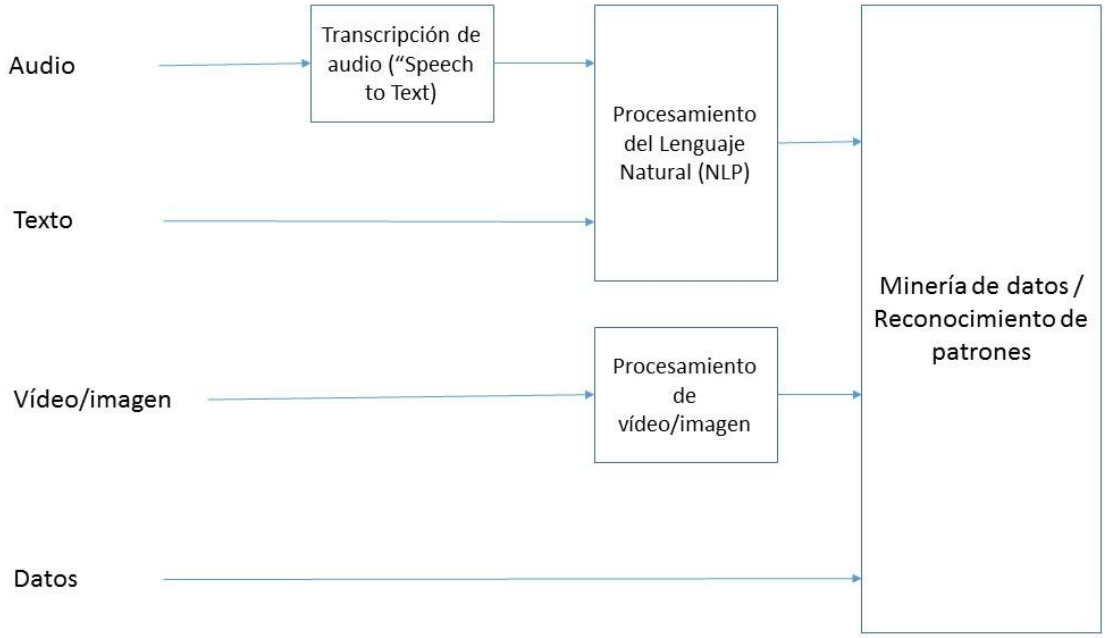
- Recolector (crawler), que visita una serie de páginas Web con el objeto de incorporarlas a la colección de páginas conocidas (y por tanto recuperables) por el motor de búsqueda;
- Indexador (indexer), que convierte la colección en una estructura más manejable y pequeña, llamada Índice;
- Buscador (search engine), encargado propiamente de recuperar ciertas páginas del índice a partir de la consulta realizada por el usuario

Siempre que se necesite extraer información de algún sitio de Internet, ya sean redes sociales, blogs, o webs en general, es necesesario que usar un crawler, que se encarga de recuperar esa información. El crawler puede volcar directamente el texto html sobre una base de datos (cuando se opere sobre webs estáticas, por ejemplo) o puede utilizar APIs (todas las redes sociales tienen las suyas, Twitter, Facebook, Youtube, Google+, etc)

Para el contenido textual, se utiliza un módulo de minería de textos que incluye la parte de procesamiento del lenguaje natural (NLP) y la parte de reconocimiento de patrones y machine-learning.

Importante, la minería de textos está orientada a textos escritos; si se mezclan textos con datos en bruto de cualquier otro tipo es necesario hacer uso de minería de datos

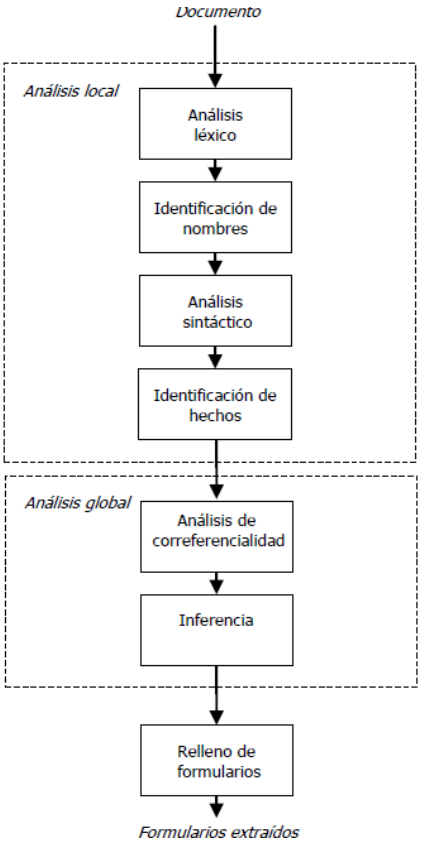
Y una vez que se tienen los datos para dispuestos, limpios y estructurados, se aplican de técnicas de aprendizaje supervisado o no supervisado.



2.7. TECNICAS SRI

EXTRACCIÓN DE INFORMACIÓN (INFORMATION EXTRACTION)

El proceso de extracción de información comprende dos fases principales. En primer lugar, el sistema identifica "hechos" aislados del texto de un documento a partir del análisis de cada frase (análisis local, local text analysis). En segundo lugar, se integran estos hechos en un análisis global (discourse analysis) para obtener hechos más generales o nuevos hechos, mediante correferencialidad e inferencia



MINERÍA DE TEXTOS (TEXT MINING)

La minería de textos es un área multidisciplinaria basada en la recuperación de información, minería de datos, aprendizaje automático, estadísticas y la lingüística computacional. El proceso de minería de textos habitualmente requiere la ejecución de los siguientes pasos:

1. Estructuración del texto de entrada: se aplican técnicas de análisis gramatical, junto con la adición y elimina-ción de ciertas características lingüísticas para posteriormente insertar el resultado en una base de datos.

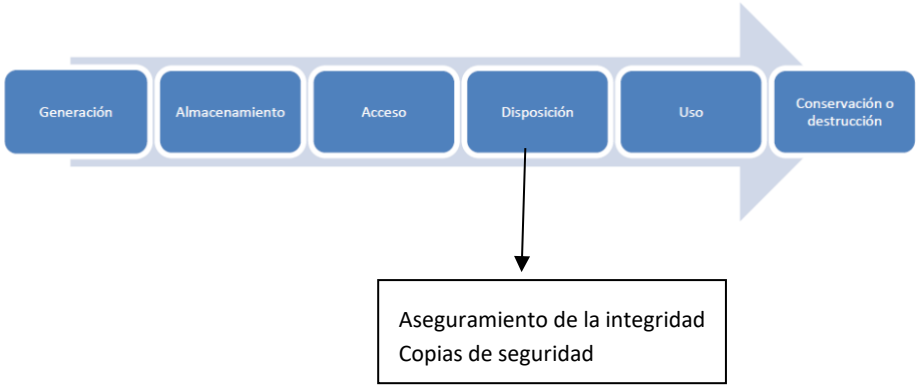
2. Extracción de patrones

3. Evaluación e interpretación del resultado

Para llevar a cabo estas tareas, la minería de textos se apoya en técnicas derivadas de la Recuperación de Información, Procesamiento de Lenguaje Natural, Extracción de Información y finalmente Minería de Datos.

3. CONSERVACIÓN DE LA INFORMACIÓN

CICLO VIDA DE LA INFORMACIÓN



REPLICACIÓN DE INFORMACIÓN

- **Replicación continua (copia en tiempo real)** cada vez que un dato es almacenado o modificado en un sistema de almacenamiento, este mismo dato es transmitido al dispositivo remoto, en el que se realiza la misma operación que en el original, manteniendo en todo momento una réplica de los datos en dos sistemas de almacenamiento distintos.
    - **Copia síncrona** el servidor inicia una escritura en el sistema de almacenamiento y no recibe la confirmación de la escritura hasta que esta no se ha realizado tanto en el sistema principal como en el de respaldo→ PROBLEMA LIMITE DISTANCIA.
    - **copia asíncrona**. En este tipo de copia el servidor recibe la confirmación de la escritura en cuanto el sistema principal la ha realizado. La información se manda al sistema principal y al de respaldo, pero no se espera la respuesta del sistema de respaldo, por lo que la latencia de la línea no afecta al rendimiento de la aplicación
        - no se tendrá la certeza de que las últimas escrituras realizadas hubiesen llegado al sistema de respaldo
        - Para evitar la pérdida de datos, los sistemas de copia asíncrona son capaces de situar el sistema remoto en un punto de consistencia anterior al fallo del sistema principal.
    - No es "retroceder" en el tiempo hasta un punto que se sepa consistente.
    - Es necesario tener un gran ancho de banda

- **Replicación discreta (backups)**
    - se conserva una imagen del estado de los datos en un momento determinado en un soporte distinto del que mantiene los datos que usan las aplicaciones.
    - es posible extraer los soportes y trasladarlos a centros remotos para poder hacer frente a desastres locales sin necesidad de disponer de líneas de comunicaciones de gran ancho de banda entre los centros principales y los centros remotos de respaldo.
    - **NDMP (Network Data Management Protocol)** → que permite separación de los caminos de control y de datos, de forma que el backup puede ser gestionado desde un servidor central de backup mientras que los datos viajan directamente del servidor de ficheros a las unidades de cinta, pudiendo pasar directamente por la SAN sin tener que hacer uso de la LAN.
    - **Copia de seguridad normal/completa**: copia de seguridad total de todos los archivos y directorios seleccionados en Copia de Seguridad de Windows. El programa borra el bit de modificado de cada archivo. Es la base para futuras tareas que solo realizan copias de seguridad de los archivos modificados
    - **Copia de seguridad incremental**: el programa examina el bit de modificado y hace una copia de seguridad solo de los archivos que han cambiado desde la última copia de seguridad

incremental o normal. Esta tarea borra el bit de modificado de cada archivo que copia. Utilizan la mínima cantidad de cinta y ahorran tiempo, sin embargo, realizar una restauración es un inconveniente

- **Copia de seguridad diferencial:** es lo mismo que una copia de seguridad incremental exceptuando que el programa no elimina el bit de modificación. Requiere más espacio en cinta y tiempo que las incrementales pero su ventaja radica en que cuando se realiza una restauración se necesitan solo las cintas que contengan la copia de seguridad normal y la más reciente diferencial.
- **Copia de seguridad intermedia**: equivalente a una copia de seguridad normal, excepto que el programa no desactiva el bit de modificado.
- **Un esquema de rotación de medios** dicta cuantas cintas se usan para realizar las copias de seguridad. Un esquema popular es el método del abuelo-padre-hijo, utiliza tres generaciones de cintas que representan copias de seguridad mensual, semanal y diaria

REPLICACIÓN DE CPDS

- **Activo-Activo**: Todos los nodos del cluster reciben una parte de la misma carga.
- Activo-Pasivo: La carga la asume uno de los nodos y en caso de caída se traspasa a otro nodo
- La *pecera* de un centro de respaldo recibe estas denominaciones en función de su equipamiento:
  - Sala blanca: cuando el equipamiento es exactamente igual al existente en el CPD principal.
  - Sala de back-up: cuando el equipamiento es similar pero no exactamente igual.

ILM (INFORMATION LIFECYCLE MANAGEMENT)

Es una estrategia de almacenamiento de grandes volúmenes de información en empresas que alinea la infraestructura IT con los requisitos de negocio basada en el valor cambiante de la información en el tiempo

POLITICAS CONSERVACIÓN INFORMACIÓN

- Política de copias de seguridad
- Políticas de borrado seguro de la información
- Plan de contingencias