Ezequiel Palomeque Bernal
Francisco Moreno Sánchez

Resumen rec inf

Objetivo de los RI

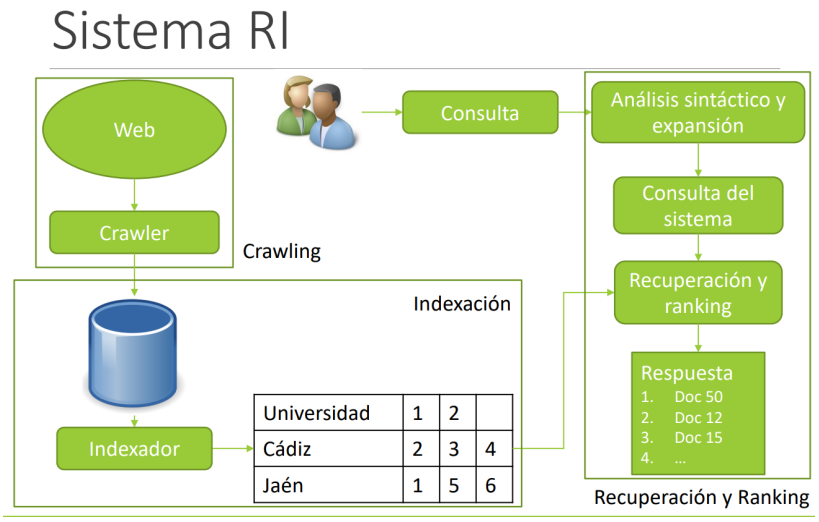
El objetivo principal de un sistema de recuperación de datos de la información es recuperar todos los documentos que sean relevantes para la consulta dada por el usuario, minimizando a su vez el número de documentos no relevantes recuperados.

(No confundir una base de datos o un sistema de recuperación de datos con el nuestro, que es de la información, pues el nuestro recoge lenguaje natural y no datos y no tiene una estructura definida, a diferencia de los de datos)

Recuperación de la Información

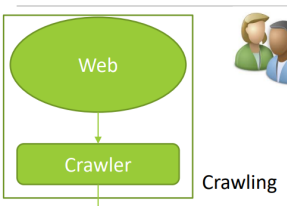
Es un área dentro de la informática que se dedica a la construcción de índices, procesamiento de consultas, etc, con el fin de facilitar a los usuarios el acceso a la información.

Otra forma de definir un modelo RI es diciendo que es un 4-tupla [D, Q, F, R(qi, dj)], siendo D el conjunto de de los documentos (corpus); Q la consulta; F el modelo de representación (lógico, vectorial o probabilístico) y R(qi,dj) la función de ranking con qi perteneciente a Q y dj a D.



Crawlers

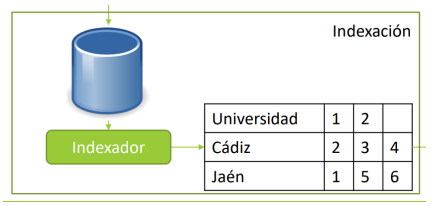
Un crawler es un programa que dada una semilla, accede a esta, la descarga y busca links en ella para repetir este proceso hasta una condición de parada, normalmente con el fin de luego indexarlas. Estos programas se usan debido a la cantidad de webs que hay hoy día.



Indexador

La colección principal tiene que pre-procesarse:

- Eliminación de palabras vacías (stopwords)
- Lematización (stemming)
- Selección de los términos de indexación
- Utilizados como representación de un documento
- Deben ser más pequeños que los documentos en sí.



La colección de documentos del repositorio central se tiene que indexar para permitir:

- Búsquedas rápidas
- Ordenación y ranking

Índice Invertido

Es una estructura para ordenar los los términos y documentos que aparecen en ellos.
Su funcionamiento se basa en indexar los términos y decir en cada caso, a qué documentos pertenece. Esto agiliza mucho las búsquedas.

Preprocesamiento

El preprocesamiento tiene 5 operaciones o transformaciones:

Análisis del texto

Aquí es donde reconocen los espacios los dígitos y los símbolos extra que no sean necesarios. También se detectan los acentos y las mayúsculas o minúsculas.

Eliminación de las Palabras Vacías

Las también llamadas "Stopwords" son aquellas palabras con un valor muy bajo para el sistema RI.
Una buena forma de detectarlas suele ser si aparece en más del 80% de los documentos.
(Hay listas en internet de palabras vacías frecuentes para muchos idiomas)

Stemming (Lematización)

Es el proceso de eliminar los afijos a las palabras para así reducir la cantidad de palabras a indexar. Es parte del preprocesamiento del texto.

Selección de términos a indexar

Suelen elegirse dependiendo de la naturaleza sintáctica de la palabra, por ejemplo, los sustantivos se suelen elegir de forma preferente a las demás.

Tesauros

def diapos: Construcción de categorías de términos (tesauros). Extracción de la estructura contenida en los documentos para permitir la expansión de la consulta.

def wikipedia: Un tesauro es una lista de palabras o términos empleados para representar conceptos.

Motor de búsqueda (recuperación y ranking)

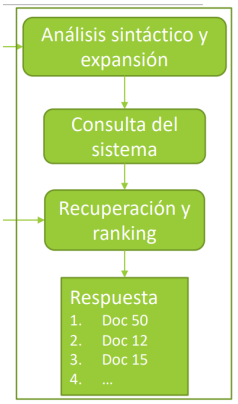
En primer lugar, el usuario tiene que especificar una consulta que refleje sus necesidades de información.

El sistema aplica un procesamiento similar al realizado sobre los documentos:

- Eliminación de palabras vacías
- Corrección de errores
- Stemming (Lematización)

El sistema puede expandir la consulta:

- Sugerencias hechas por el sistema y confirmadas por el usuario



Recuperación y Ranking

Proceso de recuperación: la consulta del sistema se lanza contra el motor de recuperación para obtener el conjunto de documentos que contienen las palabras de la consulta.

El motor de búsqueda debe ordenar los documentos de forma que los más relevantes sean los primeros que vea el usuario, esto se realiza gracias al ranking.

Ranking

Es una función que puntúa a los documentos para poder ordenarlos fácilmente. Este proceso se realiza a la hora de consultar en el motor de búsqueda.
Esta puntuación dependerá del modelo implementado. (Ej. en el modelo vectorial usamos el coseno)
El ranking puede utilizar la información proveniente del usuario para mejorar su modelo de ranking, a esto lo llamamos feedback. Ejemplos:

- Clics.
- Contexto geográfico (IP, lenguaje)
- Contexto tecnológico (sistema operativo, navegador, dispositivo móvil)
- Contexto temporal: histórico de consultas (es necesario el uso de cookies o el registro por parte del usuario)

Modelo Booleano

Es un modelo basado en teoría de conjuntos y álgebra booleana, es sencillo y con semántica clara. Los primeros sistemas de RI se basaban en este modelo.

Características: Los términos de indexación están presentes o ausentes, es decir, la matriz términos-documentos es binaria y la consulta es una expresión booleana (NOT, AND, OR).

Ventajas:
- Definición formal clara
- Simplicidad (peso binario de los términos)

Desventajas:
- No existe ranking
- La formulación de consultas booleanas es muy compleja

Modelo Vectorial

El modelo booleano es demasiado limitado.
- No se permite una combinación parcial.
- Los documentos o son relevante o no lo son.
- No se puede establecer un ranking.
- Es difícil traducir una necesidad de información en una expresión booleana.

El modelo vectorial tiene en cada pareja termino-documento (ki-dj) un peso asociado Wij, el cual es positivo y no binario. Este modelo asume la independencia de términos y por tanto cada término es representado como un vector unitario t-dimensional.

La consulta se trata como un documento, pues así es más sencillo encontrar las similitudes.

La similitud entre un documento y una consulta puede calcularse mediante el coseno del ángulo entre los vectores que representan al documento dj y a la consulta q.

$$\cos\theta = sim(d_j, q) = \frac{\overrightarrow{d_j} \cdot \vec{q}}{|\overrightarrow{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

Modelo probabilístico

No es necesario aprenderlo, solo saber que existe.

Distancias de Hamming y Levenshtein

Hamming: Dadas dos cadenas de caracteres de la misma longitud, la distancia es el número de posiciones que tienen caracteres diferentes.

Levenshtein: Mínimo número de inserciones, eliminación o sustitución de caracteres para que las cadenas sean iguales.

Ponderación de términos (TF - IDF)

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{n_i} & Si\ f_{i,j} > 0 \\ 0 & en\ otro\ caso \end{cases}$$

TF: $1 + \log_2$ frecuencia
IDF: $\log_2$ (numTotalDocumentos / numDocsApareceElTermino)

Características del IDF:

- **Exhaustividad** (exhaustivity) es una propiedad de la descripción de un documento. Debe interpretarse como la importancia que tiene dicha descripción para el tema general del documento.
- **Particularidad** (specificity) es una propiedad de los términos de indexación. Debe interpretarse como la bondad de los términos para describir el tema del documento.

Relación entre exhaustividad y particularidad:

- Si la descripción de un documento crece, la particularidad de los términos decrece.
- Si un término aparece en todos los documentos, su particularidad es mínima, por lo que el término no es útil para la recuperación

Vector normal

Considerando que cada término va asociado con el vector unitario ortonormal ki en un espacio d-dimensional (con t=num Términos), en este espacio, los documentos son vectores de términos ponderados.
Es decir, el término ki del documento dj, se asocia con el documento Wij * ki y representa la contribución del término al documento.
La representación del documento es, básicamente, el vector formado por todos los términos (componentes vectoriales).
La longitud del documento se calcula con el sumatorio entre el IDF y el TF de cada término.

Evaluación de sistemas RI

Precision y Recall

Supongamos la petición de información I.
Sea R el conjunto de documentos relevantes para la petición de información I.

- Sea $|R|$ el número de documentos del conjunto R

Tras un proceso de recuperación, el sistema genera el conjunto de respuestas A.

- Sea $|A|$ el número de documentos del conjunto A

Sea $|R \cap A|$ el número de documentos de la intersección de los conjuntos A y R

Precisión: documentos recuperados (conjunto A) que son relevantes.

$$Precisión = p = \frac{|R \cap A|}{|A|}$$

Recall: documentos relevantes (conjunto R) que han sido recuperados.

$$Recall = r = \frac{|R \cap A|}{|R|}$$

Ejemplo de ejercicio:
Al llegar al al primer documento relevante tenemos:

$|A| = 3, |R| = 3, |R \cap A| = 1$

Por tanto aplicando las fórmulas obtenemos los datos en la tabla del medio:

Precision = 33.33%, Recall = 33.33%

Al llegar al al segundo documento relevante tenemos:

$|A| = 8, |R| = 3, |R \cap A| = 2$

Por tanto aplicando las fórmulas obtenemos los datos en la tabla del medio:

Precision = 25%, Recall = 66.66%

Al llegar al al tercero documento relevante tenemos:

$|A| = 15, |R| = 3, |R \cap A| = 3$

Por tanto aplicando las fórmulas obtenemos los datos en la tabla del medio:

Precision = 20%, Recall = 100%

R	d3	d56	d129
A2	d425	d615	d193
	d87	d512	d715
	d56	**d129**	d810
	d32	d4	d5
	d124	d130	**d3**

recall	precision
0	33,33%
10	33,33%
20	33,33%
30	33,33%
40	25%
50	25%
60	25%
70	20%
80	20%
90	20%
100	20%

	precision	recall
primer	33,33%	33,33
segundo	25%	66,66
tercero	20%	100