¿Qué elementos tiene un sistema REC-INF?

1. Un conjunto de documentos o unidades de información (corpus)
2. Un índice, que permite localizar rápidamente los documentos que contienen un término o conjunto de términos
3. Una consulta o módulo de búsqueda, que permite al usuario especificar los términos de búsqueda y recibir una lista de documentos relevantes
4. Un módulo de recuperación, que se encarga de obtener los documentos relevantes del repositorio de documentos
5. Un módulo de visualización, que muestra al usuario los resultados de la búsqueda

Otros módulos opcionales incluyen pero no se limitan a: filtración, procesamiento de lenguaje y personalización

Definición de REC-INF

Un área dentro de la informática que toca elementos informáticos y humanísticos, se centra en el proceso de obtener información relevante a una consulta a partir de una colección de fuentes de información. Tiene como objetivo facilitar a los usuarios el acceso a la información.

¿Qué es el stemming?

Del inglés stem(tallo) es un proceso de procesamiento del lenguaje natural que consiste en reducir las palabras a su raíz o tronco léxico. Por ejemplo : Caminar - Caminando - Caminó ~ “Camin”

Fases de REC-INF

1. **Indexación**
    a. Preprocesado de los documentos
    b. Creación de un índice para localizar rápidamente los documentos
2. **Consulta**
    a. El usuario especifica los términos de la búsqueda
    b. Preprocesado de los términos de la búsqueda
    c. Uso del índice para encontrar los documentos relevantes
    d. Presentación al usuario

Crawler

Programa informático que se utiliza para recorrer y examinar las páginas web con el objetivo de indexar su contenido. Se utilizan a menudo en los motores de búsqueda para actualizar y mantener sus índices de páginas web. Se encarga de analizar el contenido de una página y siguiendo los enlaces que tiene, se mueve a otras páginas y las analiza.

Exhaustividad y particularidad

1. **Exhaustividad:** Es una propiedad de la descripción de un documento, debe interpretarse como la importancia que tiene dicha descripción para el tema general del documento (aumenta si añadimos nuevos términos al documento)
2. **Particularidad:** Es una propiedad de los términos de indexación. Debe interpretarse como la bondad de los términos para describir el tema del documento

Exhaustividad óptima: El número medio de términos de indexación por documento debería de optimizarse para que la probabilidad de ser relevante sea maximizada.
Exhaustividad y particularidad estática: La exhaustividad puede cuantificarse como el número de términos de indexación que contiene el documento. Consecuentemente, la particularidad de un documento puede cuantificarse como la función inversa del número de documentos en los que aparece.
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.

TF/IDF
Es el esquema de ponderación de términos más comúnmente utilizados
Viene de la unión de la frecuencia de los términos (TF) y la frecuencia inversa de los documentos (IDF)

1. **TF**

Conjetura de Luhn: El valor o peso de un término que aparece en un documento es proporcional a su frecuencia, es decir, cuanto mayor sea el número de veces que la palabra aparece en el documento, mayor será el peso.

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & Si f_{i,j} > 0 \\ 0 & en\ otro\ caso \end{cases}$$

2. **IDF**

A partir de los conceptos de la exhaustividad, la particularidad y la Ley de Zipf (la frecuencia de un término en la colección puede modelarse mediante una función exponencial de su rango) obtenemos la frecuencia inversa del término, el IDF. El IDF puede calcularse como:
$IDF_i = \log (N/n_i)$
Donde N es el número de elementos en el corpus y n_i la frecuencia de el término en el corpus.

3. **TF-IDF**

Combinando la frecuencia inversa y la frecuencia en el documento podemos definir el IDF como

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

TF x IDF

Modelo vectorial

El modelo vectorial permite combinaciones parciales, hacer un ranking de relevancia de los documentos y expresar más fácilmente la necesidad de información.

Para esto se asignan pesos no binarios a los términos de las consultas y los documentos que serán utilizados para medir el grado de similitud entre los documentos del corpus y la consulta dada por el usuario.

Se asigna el peso $w_{i,j}$ a una pareja término-documento (k_i , d_j) positivo y no binario. Los términos de indexación pueden representarse como vectores unitarios en un espacio t-dimensional donde t es el número total de términos de indexación. En este sentido, la representación de un documento d_j y una consulta q puede exprsesarse de la siguiente forma: $d_j = (w_{i,j}, w_{2,j}, \dots, w_{t,j})$ y $q = (w_{i,q}, w_{2,q}, \dots, w_{t,q})$

donde $w_{i,q}$ es el peso asociado con la pareja término-consulta (k_i, q) con $w_{i,q} >= 0$

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

Para la consulta, su peso se reduce al IDF ya que su frecuencia será 1

#	Término	IDF	d_1	d_2	d_3	d_4
1	to	1	3	2		
2	do	0,415	0,830		1,073	1,073
3	is	2	4			
4	be	0				
5	or	2		2		
6	not	2		2		
7	I	1		2	2	
8	am	1		2	1	
9	what	2		2		
10	think	2			2	
11	therefore	2			2	
12	da	2				5,170
13	let	2				4
14	it	2				4

Longitud	d_1	d_2	d_3	d_4
Vector normal	5,068	4,899	3,762	7,738

$$d_1 = \frac{1 \times 3 + 0{,}415 \times 0{,}830}{5{,}068} = 0{,}660$$

$$d_2 = \frac{1 \times 2 + 0{,}415 \times 0}{4{,}899} = 0{,}408$$

$$d_3 = \frac{1 \times 0 + 0{,}415 \times 1{,}073}{3{,}762} = 0{,}118$$

$$d_4 = \frac{1 \times 0 + 0{,}415 \times 1{,}073}{7{,}738} = 0{,}058$$

Ventajas:

- El esquema de ponderación de términos mejora la calidad de la recuperación
- La estrategia de búsqueda parcial permite recuperar documentos que coinciden con los términos de búsqueda de forma parcial
- La normalización por el tamaño del documento es inherente al modelo

Desventajas:

- Se asume la independencia entre los términos

Nombrar otros modelos

1. Booleano, expresa si un término está o no en un documento
2. Probabilístico, dada una consulta estima la posibilidad de que el usuario encuentre un documento relevante
3. Alternativos, de conjuntos, algebraicos y probabilísticos

Elaboración del proyecto

1. **Indexación**
    a. Procesado de caracteres por cada documento a través de filtros
        i. Mayúsculas / Minúsculas
        ii. Caracteres especiales
        iii. Números
    b. División del documento en lista de términos
    c. Procesamiento de los términos (palabras vacías)
    d. Cálculo del TF-IDF y creación del índice invertido
        i. Cálculo del TF a medida que proceso los términos (1a iteración)
        ii. Creo el índice invertido : Término 1 -> [IDF, [Doc1, peso] [Doc2, peso]]
        iii. Calculo la longitud de los documentos
        iv. Calculo el IDF
2. **Consulta**
    a. Procesamiento de caracteres
    b. División de consulta en lista de términos
    c. Recuperación de documentos
    d. Ordenación de documentos
    e. Mostrar n primeros resultados

Precisión, Recall y cómo evaluar un sistema de REC-INF

Sea:
R: El conjunto de documentos relevantes para una petición y |R| el número de documentos
A: Conjunto de respuestas tras el proceso de recuperación y |A| el número de documentos
|R ∩ A|: el número de documentos de la intersección
Precisión: Documentos recuperados (conjunto A) que son relevantes

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

Recall: Documentos relevantes (conjunto R) que han sido recuperados

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

Los documentos del conjunto A se devuelven ordenados, el usuario los examinará en ese orden y la precisión y el recall varían a lo largo del proceso. Por ejemplo:
R = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}
A = {***d*123**, *d*84, ***d*56**, *d*6, *d*8, ***d*9**, *d*511, *d*129, *d*187, ***d*25**, *d*38, *d*48, *d*250, *d*113, ***d*3**}

Recall	Precisión	
0	**100**	Siempre
19	100	1 de 1
20	66,7	2 de 3
30	50	3 de 6
40	40	4 de 10
50	33,3	5 de 15
60	0	
70	0	
80	0	
90	0	

Otras formas de evaluar un sistema de REC-INF son:
Valores únicos:

1. Precisión media en n: P@n : Mide la precisión cuando el usuario ha visto n docs
2. Precisión media (MAP): Media cada vez que se encuentra un doc relevante
3. Precisión - R : Precisión en la posición r-ésima del ranking
4. Histogramas de precisión: Comparación de la precisión-R de dos algoritmos
5. Ranking recíproco medio (MRR): Muy relacionado con la precisión media
6. Medida - E: Combina en una sola métrica las medidas de precisión y recall
7. Medida - F (media armónica): “”

Orientadas al usuario:

1. Cobertura : Docs conocidos por el user y relevantes en el conjunto de respuestas
2. Novedad: Docs relevantes en el conjunto de respuestas no conocidos por el user
3. Recall relativo: Relaciona los docs relev encontrados y los esperados por el user
4. Esfuerzo de recall: Relaciona los docs relev esperados y el nº de docs examinados

Evaluación por humanos:

1. Características de la interfaz
2. Facilidad de uso
3. Paneles paralelos: Dos paneles cada uno con una opción y el usuario decide cual es más relevante
4. Test A/B: Los resultados se muestran con pequeñas variaciones y analizan cómo los users responden al cambio
5. Clic: Analiza el número de veces que los usuarios hacen click en un doc

Concepto de retroalimentación:
Se refiere al proceso iterativo en el que los documentos que se saben que son relevantes para una consulta q son utilizados en la consulta modificada q'.
Para facilitar la colaboración de los usuarios, el concepto de retroalimentación por relevancia (usuario da explícitamente la información sobre los documentos relevantes) se ha relajado, observando los documentos en los que el usuario ha hecho clic y analizando los términos más frecuentes en los documentos que están en las primeras posiciones.

PageRank
Algoritmo que usa google, simula a un usuario navegando al azar por la web

- El usuario se encuentra con una página p
- A continuación, se mueve a una de las páginas enlazadas desde p
- Después repite el proceso
- Tras un número de movimientos, se calcula la probabilidad de visitar cada página

¿Qué es importante en una interfaz de usuario?

- Que sea sencilla pero con capacidad de hacer búsquedas de diversa índole
- Que tenga en cuenta *cómo* buscan los usuarios
- Que tenga un cuadro de búsqueda definido donde situar palabras claves o, en caso de que no, que tenga un buen conjunto de enlaces útiles y jerarquizados
- Que la relación entre el formulario y la longitud de una consulta tengan sentido
- Que los formularios muestren sugerencias
- Que ayuden al usuario a reformular su consulta