Information Retrieval – IR

Franco Giustozzi

Bases de Datos Avanzadas

AGENDA

- Hacia una definición de Recuperación de Información.
- Sistemas de Recuperación de Información.
- Modelos de Recuperacíon de Información.
- Evaluación de la Recuperación de Información.

HACIA UNA DEFINICIÓN DE RECUPERACIÓN DE INFORMACIÓN

¿Qué se entiende por Recuperación de Información?

- Según **Ricardo Baeza-Yates** *“la recuperación de infromación trata con la representación, el almacenamiento, la organización y el acceso a items de informacíon”*.

- **Croft** estima que la recuperación de información es *“el conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de informacion que son pertienentes para la resolución del problema planteado. En estas tareas desempeñan un papel fundamental los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental, etc.”*.

- Por otro lado, **Korfhage** definió la IR como *“la localización y presentación a un usuario de informacíon relevante a una necesidad de información expresada como una pregunta”*.

¿Qué se entiende por Recuperación de Información?

. Grossman y **Frieder** indican que recuperar información es *“encontrar documentos relevantes, no encontrar simples correspondencias a unos patrones de bits”.*

***. Karen Sparck Jones** y **Peter Willet** indican que: “la recuperacíon de información es considerada como sinónimo de recuperación de documentos, y en la actualidad, como recuperación de texto, e implica dos actividades relacionadas, aunque diferentes: indización, referida a la representación de los documentos y de la petición de información, y búsqueda”.*

¿Qué se entiende por Recuperación de Información?

Por lo tanto, podemos plantear que la Recuperación de la Información intenta resolver el problema de:

“encontrar y rankear documentos relevantes que satisfagan la necesidad de información de un usuario, expresada en un determinado lenguaje de consulta”.

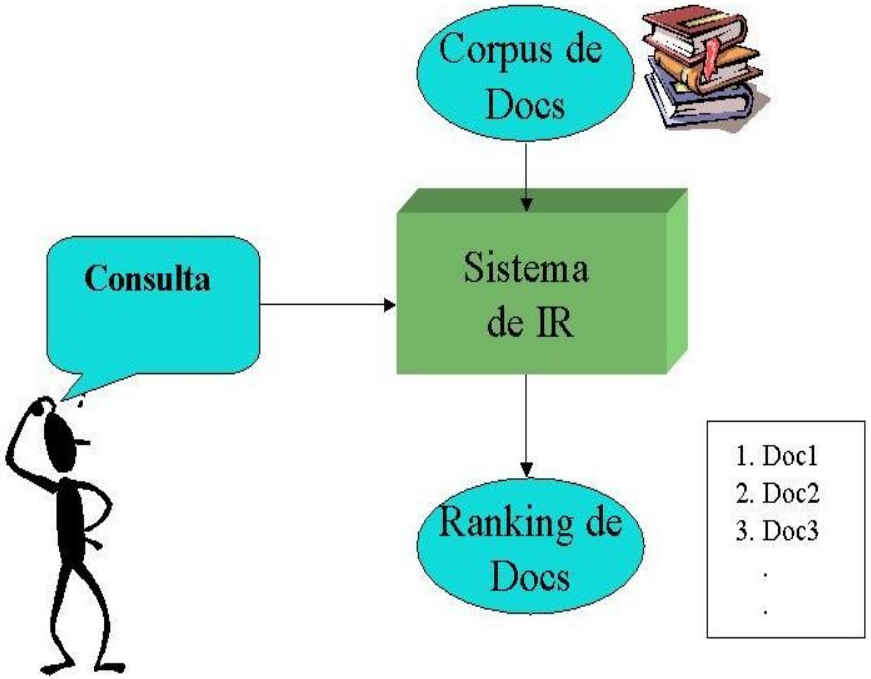
Diferencia entre Data Retrival e Information Retrival

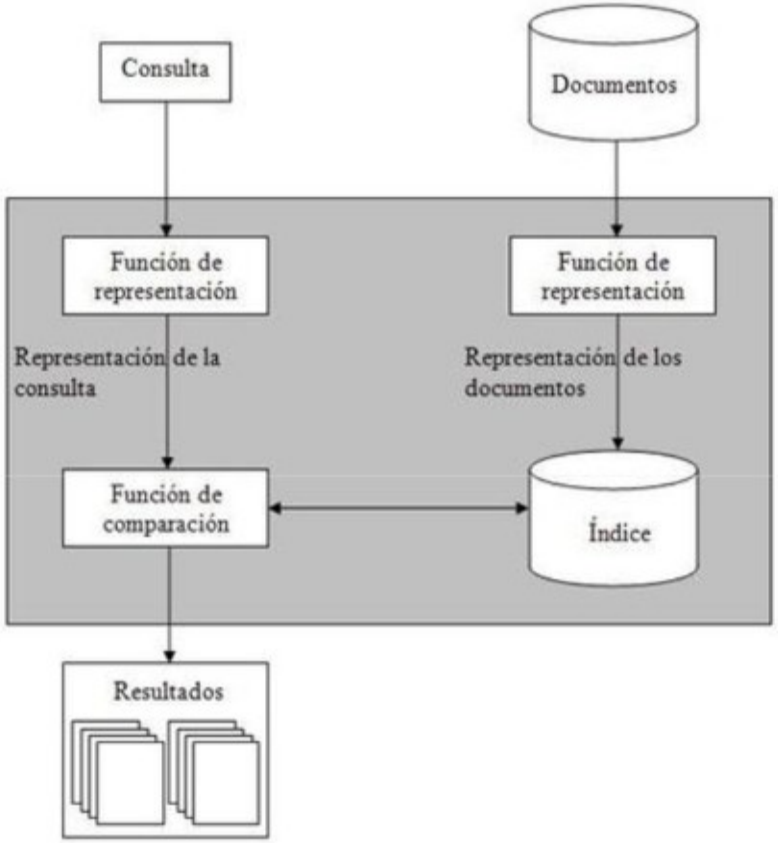
	Data Retrieval (DR)	Information Retrieval (IR)
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Query language	Artificial	Natural
Items wanted	Matching	Relevant

SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

Sistema de Recuperación de la Información - SRI

Para **Salton**, *"Cualquier SRI puede ser descripto como un conjunto de ítems de información (DOCS), un conjunto de peticiones (REQS) y algún mecanismo que determine qué ítems satisfacen las necesidades de información expresadas por el usuario en la petición"*





Sistema de Recuperación de la Información - SRI

Funciones principales en un SRI (Chowdhury, 1999)

1. Identificar las fuentes de información relevantes a las áreas de interés de las solicitudes de los usuarios.

2. Analizar los contenidos de los documentos.

3. Representar los contenidos de las fuentes analizadas de manera adecuada para compararlas con las preguntas de los usuarios.

4. Analizar las preguntas de los usuarios y representarlas de forma adecuada para compararlas con las representaciones de los documentos de la base de datos.

5. Realizar la correspondencia entre la representación de la búsqueda y los documentos almacenados en la base de datos.

6. Recuperar la información relevante.

7. Realizar los ajustes necesarios en el sistema basados en la retroalimentación con los usuarios.

Evolución de los SRI

- **Desarrollos iniciales**
- **Recuperación de información en las bibliotecas**
- **La WORLD WIDE WEB**

Evolución de los SRI

- **Desarrollos iniciales**
- **Recuperación de información en las bibliotecas**
- **La WORLD WIDE WEB**

ESTA EVOLUCIÓN NO ES UN PROCESO FINALIZADO, SINO MAS BIEN UN PROCESO EN REALIZACIÓN.

MODELOS DE RECUPERACIÓN DE INFORMACIÓN

Modelos de Recuperación de informacíon

Un **modelo de recuperacion de informacion** es una idealización o abstracción del proceso real de recuperación.

Dentro de un modelo se pueden diferenciar tres partes:

I. Modelo para la representacion de documentos.

II. Modelo para representar las consultas de los usuarios.

III. Una funcion de ranking, que asocia un número real con una consulta *q* y un documento *d.*
Ese numero, representa la probabilidad de que *d* resulte relevante para *q*.

OBS:

- En un SRI no se trabaja con los documentos propiamente dichos, sino con una representación más manejable de los mismos.

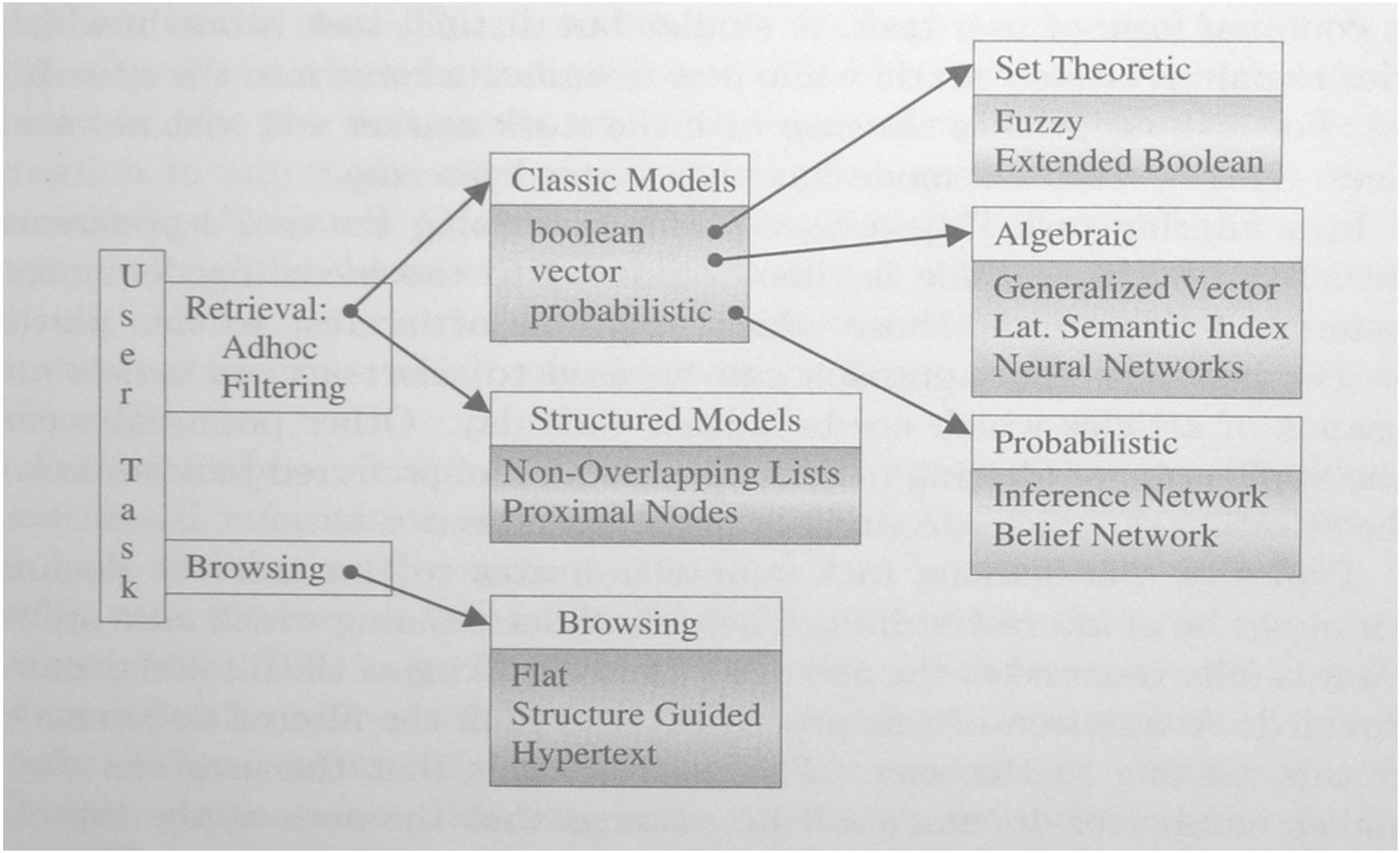
Modelos de IR

Definición formal:

Un modelo de IR es una cuadrupla (*D, Q, F, R(q,d)*), donde:

- *D* es una representación de la colección de documentos.
- *Q* es una representación de la información que necesita el usuario.
- *F* es el entorno de trabajo para modelar la colección de documentos, las consultas y las relaciones que hay entre ellos.
- *R (q,d)* es una función que devuelve un numero real que permite asociar la consulta *q (q ϵ Q)* y la representación de la colección de documentos *d (d ϵ D)*.

Modelos de IR



Modelos Clásicos

- **Modelo Booleano.**
- **Modelo Espacio Vectorial.**
- **Modelo Probabilístico.**

<u>Algunas observaciones de los modelos clásicos:</u>

- Consideran que los documentos son descritos por una serie de términos $K= \{k_1 ... k_n\}$.
- Los términos tienen asociados pesos, que varían según el documento que describan. W_{1j} = peso término 1 en documento j
- Un documento es entonces descrito por: $d_j = \{w_{1j} ... w_{nj}\}$

Modelo Booleano

- Es un modelo simple, basado en la teoría de conjuntos y el algebra de Bool.
- Los documentos se representan por el conjunto de terminos contenidos en ellos.
- Las consultas se especifican como expresiones booleanas entre los términos. Tienen una semántica concreta.

  Por ejemplo, sean T1 y T2 dos términos. Una consulta podría ser:

  T1 AND T2 = Conj. de documentos cuyas representaciones contienen al término T1 y al T2.

- Los pesos pertencen al conjunto {0,1}, siendo el peso igual a 1 si el término pertenece al documento y 0 en caso contrario.
- La similitud de un documento *d* y una consulta *q* valdrá 1 si los términos contenidos en la representación del documento *d* hacen verdadera a la expresión de la consulta *q*, y valdrá 0 en caso contrario.

Modelo Booleano

Ventajas:

- Posee un formalismo muy simple y una semantica clara y concisa para la formulación de las consultas.

Desventajas:

- De difícil uso por los usuarios. Se requieren cierto tipo de conocimientos y habilidades, para el manejo de expresiones lógicas.
- Escaso control sobre el volumen del resultado producido por una petición concreta. Se hacen necesarias reformulaciones de la pregunta para lograr un volumen aceptable de resultados.
- Todos los registros recuperados son supuestamente de la misma utilidad para el usuario. Se entregan de manera aleatoria. No existen mecanismos que permitan ordenarlos en función de su relevancia.
- No permite reflejar la importancia relativa de los diferentes componentes de la pregunta todos los términos tienen un peso 1 o 0, dependiendo de si están o no presentes en la pregunta.

Extensiones:

- Booleano extendido.
- Conjuntos difusos.

Modelo Booleano

Ejemplo:

Documento 1: “los **coches** tienen **ruedas** y circulan por cualquier **vía**”
Documento 2: “por la **autopista** pueden circular **coches**, **motos**...”

- Los términos son:

K={coches, ruedas, vía, autopista, motos}

- Los documentos quedan descriptos por (son los pesos):

D1 = (1,1,1,0,0)
D2 = (1,0,0,1,1)

- Algunas consultas:

Q1 : ruedas AND (autopista OR coches) = D1
Q2 : coches AND motos = D2

Modelo Espacio Vectorial

- Después del booleano, es el modelo de mayor influencia.
- Los términos de indización son considerados como coordenadas en un espacio informativo multidimensional.
- Documentos y preguntas son representados como vectores. Con tantas dimensiones como términos en K. Cada componente del vector representa al término de indización correspondiente.

$$d_j = (w_{1j}, w_{2j}, \ldots , w_{|K| j})$$

- La similaridad entre un documento y una pregunta se calcula mediante la comparación entre sus vectores. La similitud se entiende como afinidad entre el significado del documento y el tema de la pregunta → **Relevancia.**
- Para el cálculo del coeficiente de similitud se utilizan varias funciones.

Modelo Espacio Vectorial

¿Cómo calcular los pesos?

Diferentes métodos, uno de los más utilizados el de frecuencias **tf-idf.**

- **tf (term frequency):** frecuencia de un término i en un documento j.
  Número de veces que un término aparece en un documento normalizado por la máxima frecuencia de cualquier término en ese documento.

$$tf_{i,j} = freq_{i,j} / max\ freq_{l,j}$$

- **idf (inverse document frequency):** frecuencia de un término i en el resto de la colección. Número de documentos de la colección (N) en los que aparece el término (n_i).

$$idf_i = \log (N / n_i)$$

- Peso de un término i en un documento j según tf-idf:

$$w_{i,j} = tf_{i,j} \cdot idf_i = (freq_{i,j} / \max freq_{l,j}) \cdot \log (N / n_i)$$

Modelo Espacio Vectorial

¿Como calcular el coeficiente de similitud?

1-

Mediante el producto escalar entre los vectores que representan al documento y a la consulta.

$$R(d,q) = \sum (d * q)$$

donde d y q son vectores que contienen los pesos de los terminos correspondientes en esos documentos.

OBS:

- Con pesos binarios. Equivale a contar el número de términos coincidentes entre el documento y la consulta.
- Usando el metodo **tf-idf**, la importancia de cada término influirá en el ranking.

Modelo Espacio Vectorial

Ejemplo (pesos binarios):

Si nuestro SRI contiene los siguientes cuatro documentos:

D1: el río Danubio pasa por Viena, su color es azul

D2: el caudal de un río asciende en Invierno

D3: el río Rhin y el río Danubio tienen mucho caudal

D4: si un río es navegable, es porque tiene mucho caudal

Su matriz correspondiente dentro del modelo del Espacio Vectorial podría ser la siguiente:

	Río	Danubio	Viena	color	azul	caudal	invierno	Rhin	navegable
D1	1	1	1	1	1	0	0	0	0
D2	1	0	0	0	0	1	1	0	0
D3	2	1	0	0	0	1	0	1	0
D4	1	0	0	0	0	1	0	0	1

- si la pregunta fuera “¿cuál es el caudal del río Danubio?” su vector de términos sería Q = (1,1,0,0,0,1,0,0,0)

Modelo Espacio Vectorial

Ejemplo (tf-idf):

Matriz ***tf-idf.***

	Río	Danubio	Viena	color	azul	Caudal	invierno	Rhin	navegable
D1	0	0.301	0.602	0.602	0.602	0	0	0	0
D2	0	0	0	0	0	0.124	0.602	0	0
D3	0	0.301	0	0	0	0.124	0	0.602	0
D4	0	0	0	0	0	0.124	0	0	0.602
Q	0	0.301	0	0	0	0.124	0	0	0

Tabla 1.5 Ejemplo de Matriz de términos y documentos en el Espacio Vectorial con los pesos calculados. Fuente: elaboración propia.

Ahora corresponde calcular las similitudes existentes entre los distintos documentos (D1, D2, D3 y D4) y el vector Q de la pregunta. Hay que multiplicar componente a componente de los vectores y sumar los resultados. El modo más sencillo de obtener la similitud es por medio del producto escalar de los vectores (es decir, multiplicando los componentes de cada vector y sumando los resultados).

Cálculo de similitudes

Sim (D1,Q) = 0*0 + 0.301*0.301 + 0.602*0 + 0.602*0 + 0.602*0 + 0*0.124+ 0*0 + 0*0 +0*0 = **0.09**

Sim (D2,Q) = 0*0 + 0*0.301 + 0*0 + 0*0 + 0*0 + 0.124*0.124 + 0.602*0 + 0*0 + 0*0 = **0.01**

Sim (D3,Q) = 0*0 + 0.301*0.301 + 0*0 + 0*0 + 0*0 + 0.124*0.124 + 0*0 + 0.602*0 + 0*0 = **0.10**

Sim (D4,Q) = 0*0 + 0*0.301 + 0*0 + 0*0 + 0*0 + 0.124*0.124 + 0*0 + 0*0 + 0.602*0 = **0.01**

Con estos valores de similitud, se obtiene la siguiente respuesta:

{D3,D1, D2, D4}

Modelo Espacio Vectorial

¿Como calcular el coeficiente de similitud?

2-

- Es el producto escalar de ambos vectores normalizado por la longitud de los mismos, es decir es el coseno del angulo formado por los vectores representativos de cada documento y consulta. Es un número entre 0 (ninguna coincidencia) y 1 (mayor coincidencia). Esta entre 0 y 1 porque es el rango del coseno.

$$\mathrm{CosSim}(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{\left|\vec{d_j}\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^{2} \cdot \sum_{i=1}^{t} w_{iq}^{2}}}$$

Modelo Espacio Vectorial

¿Como calcular el coeficiente de similitud?

- Hay otras formas de calcular el coeficiente de similitud, por ejemplo: Coeficiente de Dice, Coeficiente de Jaccard, etc.

Modelo Espacio Vectorial

Ventajas:

- Mejores puntuaciones en experimentos, sobre todo con grandes colecciones.
- Tiene en cuenta tf/idf.
- Grado de relevancia y matching parcial.

Desventajas:

- Necesita que consulta y documento compartan igual terminología
- No tiene en cuenta información de contexto, sintáctica, términos ambiguos.

Extensiones:

- Vectorial generalizado.
- Latent semantic indexing.
- Redes neuronales.

Modelo Probabilístico

- Esta basado en la Teoría de Probabilidades.
- Los documentos se representan por el conjunto de términos que contienen.
- Las consultas se expresan como una enumeración de términos.
- Los pesos de los términos son binarios ({0,1}).
- Suposición: dada una consulta, existe exactamente un conjunto de documentos, y no otro, que satisface dicha consulta. Este conjunto se llama **conjunto ideal**, y no se conoce de antemano, se necesitan hacer algunas supociones sobre las propiedades de este conjunto, que se refinan consulta tras consulta.

Modelo Probabilístico

- La similitud entre un documento y una consulta se define así:

$$sim(d_j, q) = \frac{P(R|d_j)}{P(R'|d_j)} = \frac{\frac{P(d_j|R) \times P(R)}{P(d_j)}}{\frac{P(d_j|R') \times P(R')}{P(d_j)}} = \frac{P(d_j|R)}{P(d_j|R')}$$

donde:

- P(R|dj) es la probabilidad de que el documento dj sea relevante a la consulta q.
- P(R'|dj) es la probabilidad de que el documento dj no sea relevante a la consulta q.
- P(dj|R) es la probabilidad de seleccionar al documento dj de entre los relevantes
- P(R) es la probabilidad de que seleccionando algún documento aleatoriamente de la colección, sea relevante.
- P(dj) es la probabilidad de obtener el documento dj aleatoriamente seleccionando uno de entre toda la colección.
- P(R'|dj), P(dj |R'), P(R') son los análogos, aplicados a la no relevancia

- Entonces, un documento dj será considerado como relevante si:

P(R|dj) > P(R'|dj) o P(dj|R) > P(dj|R')

Modelo Probabilístico

Ejemplo:

- **I1:** La **distribución de términos** en documentos relevantes es **independiente**, y en **todos los documentos** también.
- **I2:** La **distribución de términos** en documentos relevantes es **independiente**, y en **no relevantes** también.
- **O1:** La **probabilidad de relevancia** se basa sólo en la **presencia** de los términos de la consulta en el documento.
- **O2:** La **probabilidad de relevancia** se basa en la **presencia** de los términos de la consulta en el documento y en su **ausencia**.

N=número de documentos
R=número de relevantes para la consulta
n=número de documentos con el término
r=número de relevantes con el término

	I1	I2
O1	$\log\left(\frac{(r+0.5)/(R+1)}{(n+1)/(N+2)}\right)$	$\log\left(\frac{(r+0.5)/(R+1)}{(n-r+0.5)/(N-R+1)}\right)$
O2	$\log\left(\frac{(r+0.5)/(R-r+0.5)}{(n+1)/(N-n+1)}\right)$	$\log\left(\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/[(N-n)-(R-r)+0.5]}\right)$

Modelo Probabilístico

Ejemplo:

Q: “oro plata camión”

D1: “envío de **oro** dañado en incendio”

D2: “entrega de **plata** en un **camión** de **plata**”

D3: “envío de **oro** en un **camión**”

D2 y D3 son considerados relevantes

	oro	plata	camión
N	3	3	3
n	2	1	2
R	2	2	2
r	1	1	2

	I1-O1	I2-O1	I1-O2	I2-O2
oro	-0.079	-0.176	-0.176	-0.477
plata	0.097	0.301	0.176	0.477
camión	0.143	0.523	0.523	1.176

	I1-O1	I2-O1	I1-O2	I2-O2
D1	-0.079	-0.176	-0.176	-0.477
D2	0.240	0.824	0.699	1.653
D3	0.064	0.347	0.347	0.699

Modelo Probabilístico

Ventajas:

- Ordena los resultados por relevancia
- Sigue un razonamiento matemático basado en probabilidades, lo que permite que tenga extensiones populares

- **Desventajas:**
- Pesos binarios a los index terms.
- Poco intuitivo.
- Independencia de términos
- No es posible conocer al principio el conjunto de documentos relevantes

Extensiones:

- Considerar la frecuencia de términos en los documentos
- Redes bayesianas
- Redes de inferencia bayesianas

EVALUACIÓN DE LA RECUPERACIÓN DE INFORMACIÓN

Evaluación de la IR

Baeza-Yates señala 3 criterios de evaluación de los SRI

- Eficacia en la ejecución: medida del tiempo que tarda un SRI en realizar una operación
- Eficiencia del almacenamiento: medida del espacio que se precisa para almacenar los datos
- Efectividad en la recuperación de la información: medida del éxito en satisfacer la demanda de información de los usuarios → basada en la relevancia

Evaluación de la IR

Evaluación del rendimiento de los SRI

- Se parte del concepto de relevancia → un documento recuperado es relevante cuando el contenido del mismo responde a la necesidad de información del usuario (pregunta)

  Subjetividad: dificultad de determinar el grado de relevancia del documento (un mismo documento puede ser considerado relevante o no por dos personas distintas (motivos de la búsqueda, grado de conocimiento), incluso recibir distinta evaluación por el mismo usuario en dos momentos distintos.

  Existen distintos grados de relevancia (relevancia parcial), no puede medirse en términos binarios (relevante no relevante)

- Los juicios de relevancia son realizados por los usuarios, en función de la utilidad del contenido de los documentos recuperados y no tienen por qué coincidir con los juicios de valor de los expertos sobre el contenido de los mismos, por eso parece más apropiado utilizar el término pertinencia.

Evaluación de la IR

Relevancia:

Relación existente entre los contenidos de un documento con una temática determinada.

Pertinencia:

Relación de utilidad entre un documento recuperado y una necesidad de información individual.

Evaluación de la IR

Medidas más usadas: RECALL y PRECISION.

Dada una Consulta I y su conjunto de documentos relevantes R.
Dada una estrategia de recuperación, que ejecuta I y genera el conjunto de documentos A.

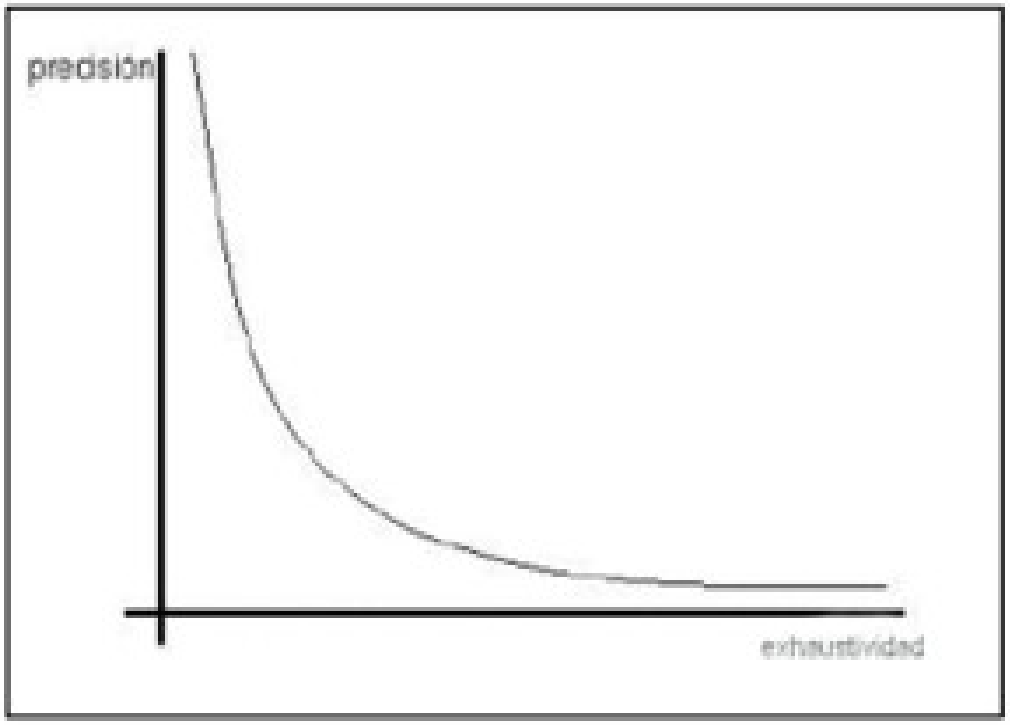
Recall: es la fracción de documentos relevantes (de R) que han sido recuperados.

$$R = \frac{C_{DRR}}{C_{TDR}}$$

Precision: es la fracción de documentos recuperados (de A) que son relevantes.

$$P = \frac{C_{DRR}}{C_{DR}}$$

Evaluación de la IR



Evaluación de la IR

Ejemplo:

Búsquedas específicas obtienen resultados muy precisos, pero habrán perdido documentos por ese alto nivel de especificación. Se reduceel Recall.
B1:“contaminación de agua en los ríos”
B2: “contaminación en los ríos”

Búsquedas generales recuperan la mayoría de los documentos relevantes, con el tema, pero también otros que no lo son. Se reduce la Precisión.
b1:“contaminación”
b2: “contaminación en los ríos”

Evaluación de la IR

Falsos positivos: un documento es un falso positivo cuando se recupera pero no es relevante.

Falsos negativos: un documento es un falso negativo cuando no se recupera aunque sea relevante.

Factores que generan los falsos positivos y negativos:

- Indización deficiente del documento (descriptores inadecuados)
- Indización deficiente de la necesidad de información
- Grado insuficiente de especificidad del lenguaje documental
- Algoritmo de relevancia deficiente: documentos relevantes en últimas posiciones de la lista de resultados o no relevantes en las primeras.

Bibliografía

- Bender, C. M., Deco, C., Tópicos avanzados de Bases de datos.
- Baeza-Yates, R., Ribeiro-Neto, B- (eds.), Modern Information Retrieval. New York. ACM Press, 1999.
- Salton, G. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.
- Martínez Méndez, Francisco Javier. Recuperación de información: modelos, sistemas y evaluación. Murcia: KIOSKO JMC, 2004.
- Van Rijsbergen, C. J. Information Retrieval. Butterworths, 1979.