

# SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

C4: EVALUACIÓN Y RETROALIMENTACIÓN EN LOS SRI

5to Año Licenciatura en Ciencia de la Computación

Departamento de Programación  
Facultad de Matemática y Computación  
Universidad de La Habana, 2022

Prof. Carlos Fleitas Aparicio

[carlos.fleitas@matcom.uh.cu](mailto:carlos.fleitas@matcom.uh.cu)

Prof. Marcel E. Sánchez Aguilar

[marcel.sanchez@matcom.uh.cu](mailto:marcel.sanchez@matcom.uh.cu)

En la conferencia estudiaremos los mecanismos de evaluación y retroalimentación de los Sistemas de Recuperación de Información. Esto nos posibilitará determinar preliminarmente la eficacia de nuestro modelo y otras formas de mejorar la respuesta al usuario.

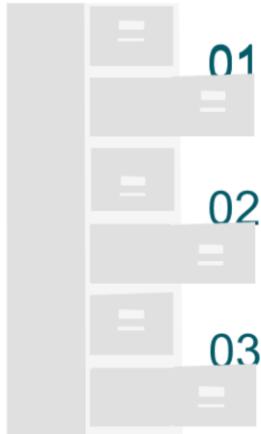


## BIBLIOGRAFÍA DE LA CLASE

1. Baeza-Yates, R. a.-N. (2002). *Modern Information Retrieval*, cap. 3, págs. 73-96, cap. 5, págs. 117-138.
2. Manning, C. D. (2009). *An Introduction to Information Retrieval*. Cambridge UP, cap. 8, págs. 151-176, cap. 9, págs. 177-194.

Recomendamos el estudio de la bibliografía. En ambos libros se encuentran bien delimitados y detallados los contenidos abordados en la conferencia.

# OBJETIVOS

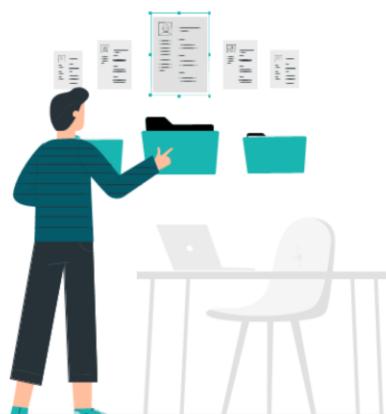


01 Dominar las principales medidas de evaluación de procesos de Recuperación de Información

02 Ejemplificar las variantes más utilizadas para lograr la retroalimentación en los Sistemas de Recuperación de Información

03 Valorar ideas sobre la expansión de consultas

# ¿QUÉ DATOS EMPLEAR PARA EVALUAR UN MODELO DE RECUPERACIÓN DE INFORMACIÓN?



## Colección de prueba

- Colección de documentos
- Conjunto de consultas
- Conjunto de documentos relevantes por cada ejemplo de consulta (juicios realizados por expertos sobre la relevancia o la pertinencia de cada documento a cada consulta)

4

Para evaluar un Modelo de Recuperación de Información es necesario contar con colecciones de prueba. Éstas no solo tienen la colección de documentos también el conjunto de consultas y la información de los documentos relevantes y no relevantes. De esta manera, podemos determinar el rendimiento de nuestro modelo. Estos documentos relevantes son seleccionados por expertos en el área.

En la práctica, se trabaja con varias colecciones de prueba. Con una realizamos el ajuste de los parámetros del modelo. Y con otra evaluamos su rendimiento, sin sesgo del conjunto de aprendizaje o entrenamiento del modelo.

## COLECCIONES DE PRUEBA DISPONIBLES

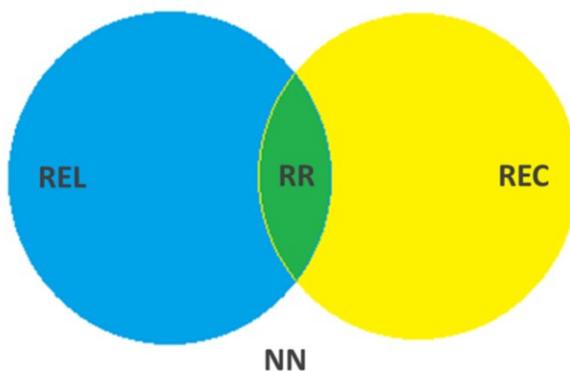
- *Cranfield*
- *Text Retrieval Conference (TREC)*
- *GOV2*
- *NII Test Collections for IR Systems (NTCIR)*
- *Cross Language Evaluation Forum (CLEF)*
- *Reuters-21578*
- *20 Newsgroups*

Le compartimos algunas colecciones de prueba de temáticas variadas que están disponibles en Internet. Más información pueden encontrarla en la bibliografía.

Manning, C. D. (2009). *An Introduction to Information Retrieval*. Cambridge UP, cap. 8, págs. 153-154.

Baeza-Yates, R. a.-N. (2002). *Modern Information Retrieval*, cap. 3, págs. 84-96.

# EVALUACIÓN



**REL** conjunto de documentos relevantes

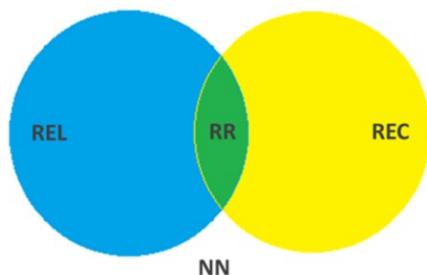
**REC** conjunto de documentos recuperados

**RR** conjunto de documentos relevante recuperados

**NN** conjunto de documentos no relevantes no recuperados

Para la evaluación partimos de estos cuatro conjuntos de documentos.

# EVALUACIÓN



	RELEVANTES	IRRELEVANTES
RECUPERADOS	RR	RI
NO RECUPERADOS	NR	NI

**RR:** conjunto de documentos recuperados relevantes ( $REL \cap REC$ )

**RI:** conjunto de documentos recuperados irrelevantes ( $REC - RR$ )

**NR:** conjunto de documentos no recuperados relevantes ( $REL - RR$ )

**NI:** conjunto de documentos no recuperados irrelevantes ( $NN$ )

7

En particular, se destacan cuatro conjuntos fundamentales. Éstos son los que intervienen en la mayoría de las medidas de evaluación. Noten que los documentos recuperados irrelevantes pueden ser considerados como falsos positivos, pues se recuperaron pero no eran interesantes para el usuario. De manera análoga, los documentos no recuperados relevantes son los falsos negativos.

# MEDIDAS DE EVALUACIÓN

- Precisión
- Recobrado (en inglés *Recall*)
- Medida F
- Medida F1

8

Estas cuatro medidas de evaluación no tienen en cuenta el *ranking*. Sin embargo, posteriormente veremos cómo pueden ser aplicadas a modelos que realizan ordenación por relevancia.

# PRECISIÓN

Fracción de los documentos recuperados que son relevantes.

$$P = \frac{|RR|}{|RR \cup RI|}$$

La Precisión usualmente tiende a decrecer cuando la cantidad de documentos recuperados aumenta.

	RELEVANTES	IRRELEVANTES
RECUPERADOS	RR	RI
NO RECUPERADOS	NR	NI

9

La precisión es una de las medidas fundamentales. En la práctica, ocurre que la cantidad de documentos relevantes a una consulta es significativamente menor que la cantidad de documentos presentes en la colección. Esto hace que, a medida que aumenta la cantidad de documentos recuperados, disminuya también la proporción de relevantes en la respuesta.

Esta medida es muy importante en la evaluación de modelos de inteligencia artificial, pues mide la calidad de la respuesta o la clasificación. Sin embargo, en el caso de la recuperación de información no solo interesa qué tan precisa es la respuesta, también que incluya a la mayor cantidad posible de documentos relevantes. Esto hace que la medida de precisión por sí sola no sea suficiente en el caso de un modelo de recuperación de información. Esta es una de las diferencias respecto a los problemas de inteligencia artificial. Pueden emplearse los mismos métodos en ambas áreas pero con enfoques, objetivos y consideraciones distintas.

# RECOBRADO

Fracción de los documentos relevantes que fueron recuperados.

$$R = \frac{|RR|}{|RR \cup NR|}$$

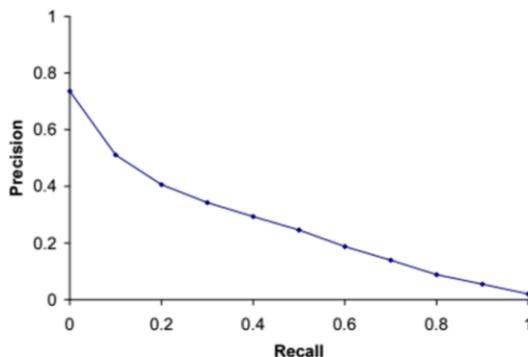
La medida de Recobrado es una función no decreciente para el número de documentos recuperados.

Recuperando todos los documentos para todas las consultas se puede conseguir un Recobrado alto pero la Precisión baja

	RELEVANTES	IRRELEVANTES
RECUPERADOS	RR	RI
NO RECUPERADOS	NR	NI

Por otro lado, tenemos el recobrado, que es fundamental en los procesos de recuperación de información. En contraposición a la precisión el recobrado aumenta a medida que incorporamos más documentos a la respuesta, pues es cada vez más probable que los elementos del conjunto de documentos relevantes estén contenidos en nuestra respuesta. Esto hace que siempre sea posible tener el mayor valor de recobrado, 1. Esto se lograría devolviendo la colección completa aunque, lógicamente, no es una solución factible.

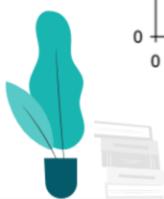
## PRECISIÓN VS. RECOBRADO



**P = 1:** todos los documentos recuperados son relevantes

**R = 1:** todos los documentos relevantes fueron recuperados

**P = 1, R = 1: Respuesta perfecta**  
todos los documentos recuperados son relevantes y todos los documentos relevantes fueron recuperados.



1  
1

En el gráfico se muestra cómo a medida que aumenta una de las dos métricas, la otra disminuye. Esto no significa que sean inversamente proporcionales, sino que se compensan. Y es precisamente eso lo que se espera de un buen modelo de recuperación de información.

## PRECISIÓN VS. RECOBRADO

- Las dos medidas se compensan (no son inversamente proporcionales).
- La Precisión usualmente tiende a decrecer cuando la cantidad de documentos recuperados aumenta.
- Siempre se puede obtener un Recobrado = 1 si se recuperan todos los documentos para todas las consultas, pero en ese caso, se obtiene una Precisión muy baja.
- Ideal: Lograr un Recobrado alto grado tolerando una pequeña cantidad de falsos positivos.

12

La solución ideal plantea un alto valor de recobrado tolerando la menor cantidad posible de falsos positivos, es decir, de documentos recuperados irrelevantes. La aparición de algunos documentos irrelevantes puede ser aceptada por parte del usuario siempre que se le proporcione la mayor cantidad posible de documentos relevantes.

## MEDIDA F

Permite enfatizar la Precisión sobre el Recobrado o viceversa

$$F = \frac{(1 + \beta^2)PR}{\beta^2P + R} = \frac{(1 + \beta^2)}{\frac{1}{P} + \frac{\beta^2}{R}}$$

$\beta = 1$ : Igual peso o énfasis para la Precisión y el Recobrado ( $F = F_1$ )

$\beta > 1$ : Mayor peso a la Precisión

$\beta < 1$ : Mayor peso al Recobrado

Con el objetivo de lograr una compensación entre la precisión y el recobrado se define la medida F. Se utiliza la media armónica en lugar de la media aritmética pues ofrece características convenientes a la hora de mejorar esta evaluación. El análisis de por qué es mejor está disponible en el libro y se propone como uno de los ejercicios de estudio individual de la conferencia.

Manning, C. D. (2009). *An Introduction to Information Retrieval*. Cambridge UP, cap. 8, págs. 156-157.

## MEDIDA $F_1$

Es una medida que armoniza Precisión y Recobrado teniéndolos en cuenta a ambos

$$F_1 = \frac{2PR}{P + R} = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

La medida  $F_1$  toma un alto valor cuando la Precisión y el Recobrado son altos, por lo que la determinación del valor máximo para  $F_1$  puede interpretarse como un esfuerzo por encontrar el mejor compromiso entre Precisión y Recobrado.

La medida  $F_1$  es un caso particular de la medida  $F$  en la que la precisión y el recobrado tienen igual importancia.

## EJEMPLO

	RELEVANTES	IRRELEVANTES
RECUPERADOS	20	40
NO RECUPERADOS	60	1 000 000

$$P = \frac{20}{(20+40)} = \frac{1}{3} = 0.33 \quad R = \frac{20}{(20+60)} = \frac{1}{4} = 0.25 \quad F_1 = \frac{\frac{2}{\frac{1}{3} + \frac{1}{4}}}{\frac{1}{3} + \frac{1}{4}} = \frac{2}{7} = 0.29$$

15

En estas medidas no se tienen en cuenta los documentos irrelevantes, por lo que podríamos no estar empleando todos los datos disponibles.

Por otro lado, noten los valores de precisión y recobrado y cómo el valor de F1 hace un balance de ambas, con una tendencia al mínimo de los dos.

## DIFICULTADES CON LA PRECISIÓN Y EL RECOBRADO

- El número de documentos irrelevantes no es tenido en cuenta.
- La Precisión es indefinida cuando no se recupera ningún documento.

$$P = \frac{|RR|}{|RR \cup RI|}$$

- El Recobrado es indefinido cuando no hay documentos relevantes en la colección.

$$R = \frac{|RR|}{|RR \cup NR|}$$

Estos son casos particulares y limitaciones que deben ser tenidas en cuenta a la hora de implementar un modelo y poder interpretar adecuadamente los valores de evaluación.

## DIFICULTADES CON LA PRECISIÓN Y EL RECOBRADO

- Se necesita del juicio de expertos para aplicar las medidas (las colecciones de pruebas son difíciles de crear).
- Para evaluar un SRI se necesita calcular con un gran conjunto de consultas (que el sistema obtenga un buen resultado en un conjunto pequeño de consultas no garantiza su rendimiento).
- La Precisión, el Recobrado y la Medida  $F$  no tienen en cuenta el ranking.
  - › Computar las medidas para cada prefijo: top 1, top 2, ..., top  $n$  documentos recuperados.

17

Las mayores dificultades en la evaluación se encuentran en la creación de las colecciones de prueba y el volumen que deben tener para garantizar un resultado satisfactorio.

Por otro lado, las medidas básicas de evaluación están diseñadas para modelos que no hacen *ranking* de los documentos en la respuesta. La alternativa es analizar una cantidad fija de documentos dada la ordenación. Por ejemplo, los  $n$  documentos de mayor valor de relevancia.

## R-PRECISIÓN

- Precisión en la posición R del *ranking* de documentos relevantes a una consulta dada para la cual existen *R documentos relevantes*.

$$P = \frac{|RR|}{|RR \cup RI|}$$

$$R - \text{Precisión}_6 = \frac{4}{6} = 0.67$$

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

De manera análoga se puede aplicar este criterio al recobrado y la medida F.

## PROPORCIÓN DE FALLO (FALLOUT)

- Tiene en cuenta la cantidad de documentos irrelevantes y el *ranking*

$$Fallout = \frac{|RI|}{|RI \cup NI|}$$

- Total de documentos irrelevantes = 100

$$Fallout_6 = \frac{2}{100} = 0.02$$

$$Fallout_9 = \frac{5}{100} = 0.05$$

	RELEVANTES	IRRELEVANTES
RECUPERADOS	RR	RI
NO RECUPERADOS	NR	NI

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Se introduce una nueva medida de evaluación que tiene en cuenta los documentos irrelevantes.

A pesar de todas estas medidas objetivas, ¿cuál sería realmente la mejor medida de evaluación? La satisfacción del usuario. Esto hace que también sea necesario considerar un conjunto de medidas subjetivas que influyen en la percepción y en la calidad de la presentación de los resultados a los usuarios.

## MEDIDAS SUBJETIVAS

- **Proporción de novedad:** La proporción entre los documentos recuperados relevantes *desconocidos* por el usuario y los documentos relevantes recuperados.
- **Proporción de cobertura:** La proporción entre la cantidad de documentos recuperados relevantes *conocidos* por el usuario y la cantidad de documentos relevantes *conocidos* por el usuario.

## OTROS ASPECTOS A CONSIDERAR

- **Esfuerzo del usuario:** Trabajo requerido al usuario en la formulación de consultas, guiado de la búsqueda y visualización del resultado.
- **Tiempo de respuesta:** Intervalo de tiempo entre la especificación de la consulta y la presentación de los resultados.
- **Forma de presentación:** Formato de presentación de los resultados que posibilite su utilización.

Estos otros aspectos pueden ser evaluados de manera objetiva o subjetiva, según sea el caso, lo que depende en gran medida del contexto en que se desarrolla la solución y los usuarios a la que va destinada.

## RETROALIMENTACIÓN

- Para los usuarios es difícil plantear las consultas de modo que expresen sus necesidades informativas.
- Muchas veces los SRI no logran dar respuesta a la necesidad de información del usuario.
- Idea: Involucrar al usuario en el proceso de recuperación de información para obtener mejores resultados.

22

En ocasiones resulta difícil satisfacer una determinada necesidad de información del usuario debido a la complejidad de la misma y a la manera en que se expresan como consultas. Es por esto que una idea interesante podría ser involucrar al usuario en un proceso de retroalimentación que permita ir refinando la respuesta.

Veamos cómo se incorporarían estas ideas en los modelos clásicos.

## ALGORITMO CLÁSICO DE RETROALIMENTACIÓN

1. El usuario plantea la consulta
2. El sistema devuelve un conjunto de documentos
3. El usuario selecciona de estos documentos los que considera relevantes o no
4. El sistema obtiene una mejor representación de las necesidades del usuario utilizando esta información
5. Se regresa al paso 2

23

En el caso del modelo clásico Booleano la correspondencia es exacta, por lo que no tenemos una manera de incorporar la información de la retroalimentación directamente en el motor de búsqueda. Para ello se emplea la expansión de consultas.

Sin embargo, en el caso del modelo vectorial se puede abordar la retroalimentación como un problema de *clustering* (agrupamiento).

## RETROALIMENTACIÓN EN EL MRI VECTORIAL

- Considera que los vectores de los documentos relevantes (a una consulta) son similares y que tienen una diferencia notable con los vectores de los documentos no relevantes.
- Idea: Encontrar un vector consulta  $\vec{q}$  que maximice la similitud con los documentos relevantes mientras minimice la similitud con los documentos no relevantes.

## RETROALIMENTACIÓN EN EL MRI VECTORIAL

Idea: Encontrar un vector consulta  $\vec{q}$  que maximice la similitud con los documentos relevantes mientras minimice la similitud con los documentos no relevantes.

Asumiendo que se conocen:

$C_r$ - conjunto de documentos relevantes

$C_{nr}$ - conjunto de documentos no relevantes

$$\vec{q}_{opt} = \arg \max [sim(\vec{q}, C_r) - sim(\vec{q}, C_{nr})]$$

Teóricamente esta función devuelve el vector ( $\vec{q}_{opt}$ ) que maximiza esta diferencia. Esto representa el centroide, es decir, el vector promedio del conjunto de documentos relevantes y no relevantes respecto a la consulta.

## RETROALIMENTACIÓN EN EL MRI VECTORIAL

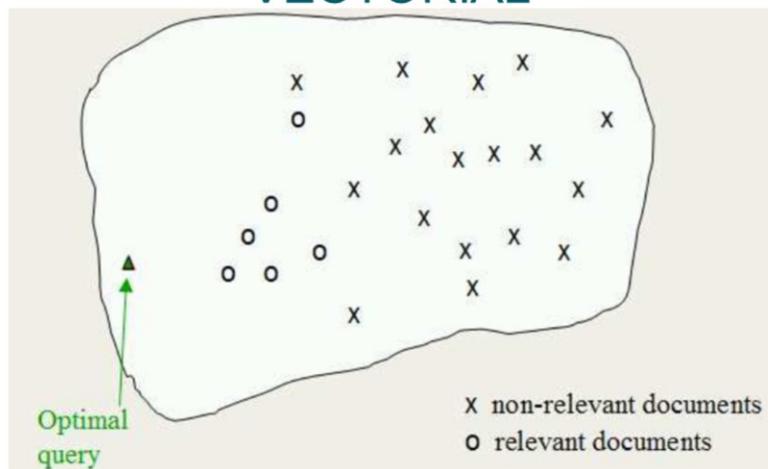
Utilizando el cálculo del coseno del ángulo entre los vectores de la consulta y los documentos como medida de similitud, resulta:

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\vec{d_j} \in C_r} \vec{d_j} - \frac{1}{|C_{nr}|} \sum_{\vec{d_j} \in C_{nr}} \vec{d_j}$$

La consulta óptima es el vector diferencia entre los centroides de los documentos relevantes y los no relevantes.

No conocemos con certeza cuáles son los conjuntos de documentos relevantes y no relevantes, por lo que debemos estimarlos a partir de la retroalimentación, con el supuesto de que podemos poco a poco acercarnos a la solución ideal.

## RETROALIMENTACIÓN EN EL MRI VECTORIAL



27

Esta es la representación gráfica de la ubicación del vector que representa a la consulta y los vectores que representan a los documentos, ya sean relevantes o no. La idea es que, a partir del conocimiento de algunos documentos relevantes y no relevantes, podamos obtener un nuevo vector consulta que maximice la diferencia entre los centroides de ambos conjuntos.

## ALGORITMO DE ROCCHIO (1971)

En un SRI real, se tiene una consulta y solo se conoce parcialmente el conjunto de los documentos relevantes y no relevantes:

$$\vec{q}_m = \alpha \vec{q}_0 + \frac{\beta}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

$D_r$  y  $D_{nr}$ : conjuntos conocidos de documentos relevantes y no relevantes respectivamente

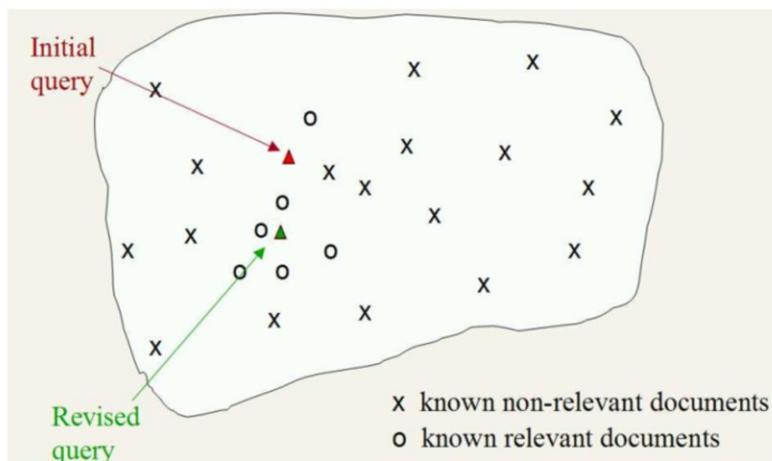
$\alpha, \beta, \gamma$ : pesos establecidos para cada término de la consulta

$D_r$  - conjunto de documentos relevantes identificados por el usuario a partir de los documentos recuperados

$D_n$ - conjunto de documentos no relevantes a partir de los documentos recuperados

El algoritmo de Rocchio ofrece una alternativa para el cálculo de esta consulta ideal a partir del conocimiento de algunos documentos relevantes y no relevantes. Además, considera la influencia de la consulta original y otorga un peso a cada componente de la fórmula. De esta manera podemos ajustar a conveniencia la importancia de cada uno.

## ALGORITMO DE ROCCHIO



29

Comenzando en la consulta inicial  $q_0$ , la nueva consulta se mueve cierta distancia hacia el centroide de los documentos relevantes y otra lejos del centroide de los no relevantes. Esta nueva consulta puede ser utilizada en el modelo vectorial.

## CONSIDERACIONES SOBRE EL ALGORITMO DE ROCCHIO

- Si tenemos muchos documentos,  $\beta$  y  $\gamma$  pueden tener valores grandes
- Se puede asumir que es más importante la información obtenida de los documentos relevantes que de los no relevantes, por lo que  $\beta$  puede ser mayor que  $\gamma$
- Valores comúnmente utilizados:  $\alpha = 1$ ,  $\beta = 0.75$ ,  $\gamma = 0.15$

30

Los valores de cada constante dependen del contexto de aplicación. Es común que se seleccione un coeficiente  $\beta$  (retroalimentación positiva) mayor que  $\gamma$  (retroalimentación negativa).

## RETROALIMENTACIÓN EN EL MRI PROBABILÍSTICO

- Tenemos un conjunto conocido de documentos relevantes y no relevantes a una consulta.
- Idea: construir un clasificador binario. Puede ser utilizado el modelo probabilístico de Naive-Bayes.
- Estimar la probabilidad de que un término  $t$  aparezca en un documento, teniendo en cuenta si es relevante o no.

En el caso del modelo de recuperación de información probabilístico vimos cómo incorporaba de manera natural la retroalimentación.

## RETROALIMENTACIÓN EN EL MRI PROBABILÍSTICO

Estimar la probabilidad de que un término  $t$  aparezca en un documento, teniendo en cuenta si es relevante o no.

$$P(x_t = 1|R = 1) = \frac{|VR_t|}{|VR|}$$
$$P(x_t = 1|R = 0) = \frac{(df_t - |VR_t|)}{(N - |VR|)}$$

$N$ : cantidad total de documentos

$df_t$ : cantidad de documentos que contienen el término  $t$

$VR$ : conjunto de documentos relevantes conocido

$VR_t$ : subconjunto de documentos  $VR$  que contienen el término  $t$

En este caso la retroalimentación que se propone es una alternativa en la que no interviene el usuario y la estimación de la cantidad de documentos relevantes y no relevantes se obtiene a partir de la selección de los primeros  $k$  documentos del *ranking* devuelto, tal y como estudiamos en la conferencia anterior.

## OTRO MÉTODOS DE RETROALIMENTACIÓN

- **Pseudo-Retroalimentación:** Se seleccionan como documentos relevantes los  $k$  primeros documentos del *ranking* dado como resultado por el sistema y luego se aplica la retroalimentación.
- **Retroalimentación Indirecta:** Se seleccionan como documentos relevantes aquellos en los que el usuario muestra interés y luego se aplica la retroalimentación. Dicho interés se estima a partir de la captura de información implícita (Ejemplo: cantidad de clics y tiempo de lectura).

# RETROALIMENTACIÓN

## VENTAJAS

- ✓ Aumenta la precisión y el recobrado.

## LIMITACIONES

- ✗ Los usuarios generalmente no pasan de la primera página de documentos y no es común que quieran participar en la retroalimentación.
- ✗ Es complicado que los usuarios comunes entiendan cómo hacer la retroalimentación y por qué funciona.
- ✗ Las consultas que se generan son muy grandes y son inefficientes para los algoritmos de recuperación.

Aunque la retroalimentación ha probado ser uno de los mecanismos más eficientes para mejorar los resultados de un SRI, tiene varias desventajas.

# EXPANSIÓN DE CONSULTAS

Una variante para lograr la retroalimentación del Sistema de Información es adecuar mejor la consulta al conjunto de documentos relevantes esperados.

## Alternativas para adecuar la consulta:

- Expandir el conjunto de términos en la consulta.
- Variar los pesos de los términos en la consulta

La expansión de consultas es una variante muy atractiva que permite mejorar las respuestas de modelos en los que no se realiza retroalimentación, como el Booleano. Además, se puede usar conjuntamente con la retroalimentación en otros modelos.

# EXPANSIÓN DE CONSULTAS

Web | [Images](#) | [Video](#) | [Local](#) | [Shopping](#) | [more](#)

Options

1 - 10 of about 534,000,000 for palm (About this page) - 0.11 sec.

Also try: [palm trees](#), [palm springs](#), [palm centro](#), [palm treo](#), [More...](#)

**Palm - AT&T**  
[att.com/wireless](#) - Go mobile effortlessly with the PALM Treo from AT&T (Cingular).

**Palm Handhelds**  
[Palm.com](#) - Organizer, Planner, WiFi, Music Bluetooth, Games, Photos & Video.

**Palm, Inc.**  
Maker of handheld PDA devices that allow mobile users to manage schedules, contacts, and other personal and business information.  
[www.palm.com](#) - Cached

**Palm, Inc. - Treo and Centro smartphones, handhelds, and accessories**  
Palm, Inc., innovator of easy-to-use mobile products including Palm® Treo\_ and Centro\_ smartphones, Palm handhelds, services, and accessories.  
[www.palm.com/us](#) - Cached

**SPONSOR RESULTS**

**Handhelds at Dell**  
Stay Connected with Handheld PCs & PDAs. Shop at Dell™ Official Site.  
[www.Dell.com](#)

**Buy Palm Centro Cases**  
Ultimate selection of cases and accessories for business devices.  
[www.Cases.com](#)

**Free Palm Treo**  
Get A Free Palm Treo 700W Phone. Participate Today.  
[EvaluationNation.com/treo](#)

En la imagen se muestra cómo un motor de búsqueda (Yahoo) ofrece alternativas a la consulta original expandiendo los términos con otros relacionados.

# EXPANSIÓN DE CONSULTAS

- Análisis Global
  - Utilizar fuentes externas como tesauros u otras bases de conocimientos.
  - Es la forma más común de expansión.
- Análisis Local
  - Analizar los documentos en la colección

Los métodos de expansión de consultas pueden ser clasificados atendiendo a si la información que emplean es local o externa al modelo.

## ANÁLISIS GLOBAL

- Por cada término  $t$  en la consulta, la consulta puede ser expandida automáticamente con sinónimos y palabras relacionadas con  $t$  utilizando tesauros.
- Combina el uso de tesauros con la ponderación de los términos.
- Incrementa el Recobrado.
- No requiere la intervención del usuario.

## ANÁLISIS LOCAL

- Los documentos recuperados para una consulta son analizados antes de ser devueltos al usuario para encontrar términos para expandir la consulta.
- Una alternativa es construir una matriz de correlación entre los términos (por ejemplo, la cantidad de documentos en los cuales dos términos coocurren).
- Para la expansión de la consulta, se toman los términos más relacionados con los términos de la consulta.

## REFORMULACIÓN DE CONSULTAS

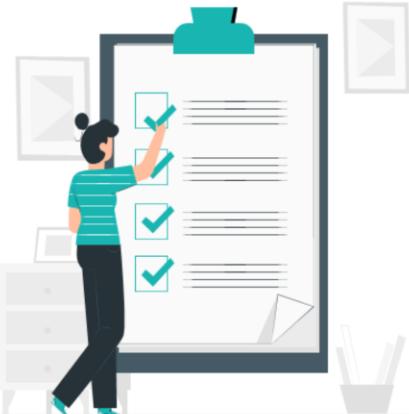
- Mostrar las palabras que fueron omitidas en la búsqueda (*stopwords*).
- Mostrar las palabras a las que se le aplicó *stemming*.
- Permitir navegar en los términos indexados para seleccionar de ellos algunos para su búsqueda.
- Sugerir nuevos términos semejantes a los de la consulta utilizando un diccionario de sinónimos.

40

Éstos pueden ser algunos de los métodos a aplicar para expandir las consultas. Por otro lado, esto no implica que siempre se puedan mejorar los resultados. Depende de las características de la colección de documentos, de la consulta y de la cantidad de documentos relevantes o no relevantes que se encuentran en la colección.

El proceso de *stemming* hace referencia a la sustitución de palabras por sus raíces gramaticales.

## CONCLUSIONES



- Las medidas de evaluación permiten determinar la efectividad de un modelo a partir de diversos criterios. El uso de esquemas de varias métricas contribuye a la calidad del ajuste del modelo.
- No existe una métrica de evaluación superior a las otras. Es necesario identificar cuál o cuáles son oportunas en cada contexto.
- La retroalimentación es un mecanismo muy útil en la obtención de mejores respuestas a las consultas pero supone algunas limitaciones que ocasionalmente impiden su uso.
- La expansión de consultas puede emplearse como una alternativa factible a la retroalimentación.

41

En la próxima conferencia estudiaremos la evolución de la Web y su papel en la Recuperación de Información.

## ESTUDIO INDIVIDUAL

Del libro, Manning, C. D. (2009). *An Introduction to Information Retrieval*. Cambridge UP., responda:

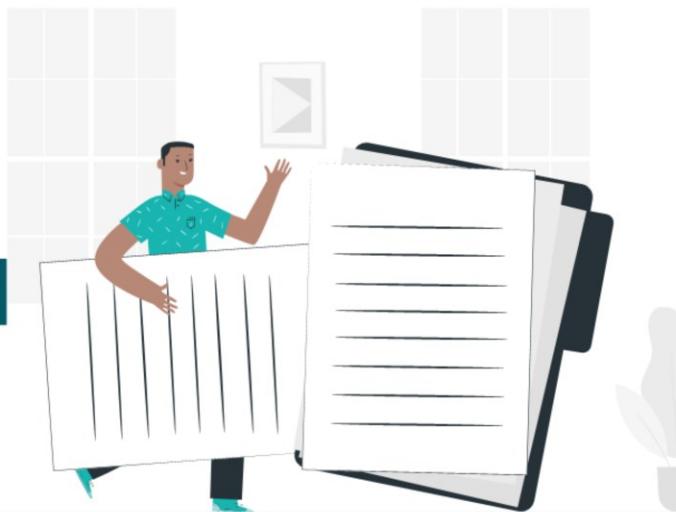
1. Ejs. 8.1 y 8.2, pág. 157.
2. Ej. 9.6, pág. 189.



Les dejamos propuestos algunos ejercicios interesantes relacionados con los contenidos de la clase.

# GRACIAS POR LA ATENCIÓN

CREDITS: This presentation template was created by [Slidesgo](#),  
including icons by [Flaticon](#), and infographics & images by [Freepik](#)



# SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

C4: EVALUACIÓN Y RETROALIMENTACIÓN EN LOS SRI

5to Año Licenciatura en Ciencia de la Computación

Departamento de Programación  
Facultad de Matemática y Computación  
Universidad de La Habana, 2022

Prof. Carlos Fleitas Aparicio

[carlos.fleitas@matcom.uh.cu](mailto:carlos.fleitas@matcom.uh.cu)

Prof. Marcel E. Sánchez Aguilar

[marcel.sanchez@matcom.uh.cu](mailto:marcel.sanchez@matcom.uh.cu)

En la conferencia estudiaremos los mecanismos de evaluación y retroalimentación de los Sistemas de Recuperación de Información. Esto nos posibilitará determinar preliminarmente la eficacia de nuestro modelo y otras formas de mejorar la respuesta al usuario.