Recuperación de Información en el Contexto de la Ciencia de la Computación

Edgar Casasola Murillo
Universidad de Costa Rica
Escuela de Ciencias de la Computación
edgar.casasola@ecci.ucr.ac.cr

Temas tratados

- Introducción a la Recuperación de Información
- Recuperación en el WWW
- Modelos formales
- Investigación en R.I.
- Conclusiones

¿Qué es Recuperación de Información?

- *R.I.* tiene como objetivo que el usuario logre satisfacer su *necesidad de información*, permitiéndole recuperar aquellos documentos que considera *relevantes* entre una colección o masa de información.
- Se centra en la estructuración, almacenamiento y recuperación de información.

Historia

- Índice es el término utilizado por los Romanos para referirse a las tiras que colgaba de los papiros (técnica utilizada por Egipcios, Griegos y Romanos)
- Vannevar Bush 1938 plantea en un ensayo la necesidad de automatizar el proceso utilizado para administrar y acceder a la información.
- Menciona aspectos como la compresión, reducción del tamaño de los dispositivos de almacenamiento, y el hipertexto. Visualiza la compresión de la Enciclopedia Británica en un dispositivo no mayor que una caja de fósforos.

Recuperación Automática de Información

- Gerarld Salton [1962] y Cleverdon [1966] se centran en la publicación de artículos científicos relacionados con la recuperación automática de documentos indexados, y proponen nuevos modelos, algoritmos y métricas para evaluación de efectividad y rendimiento.
- Surgen conceptos como Precisión y Recall, ampliamente utilizados en la actualidad

Recuperación de Datos versus Recuperación de Información

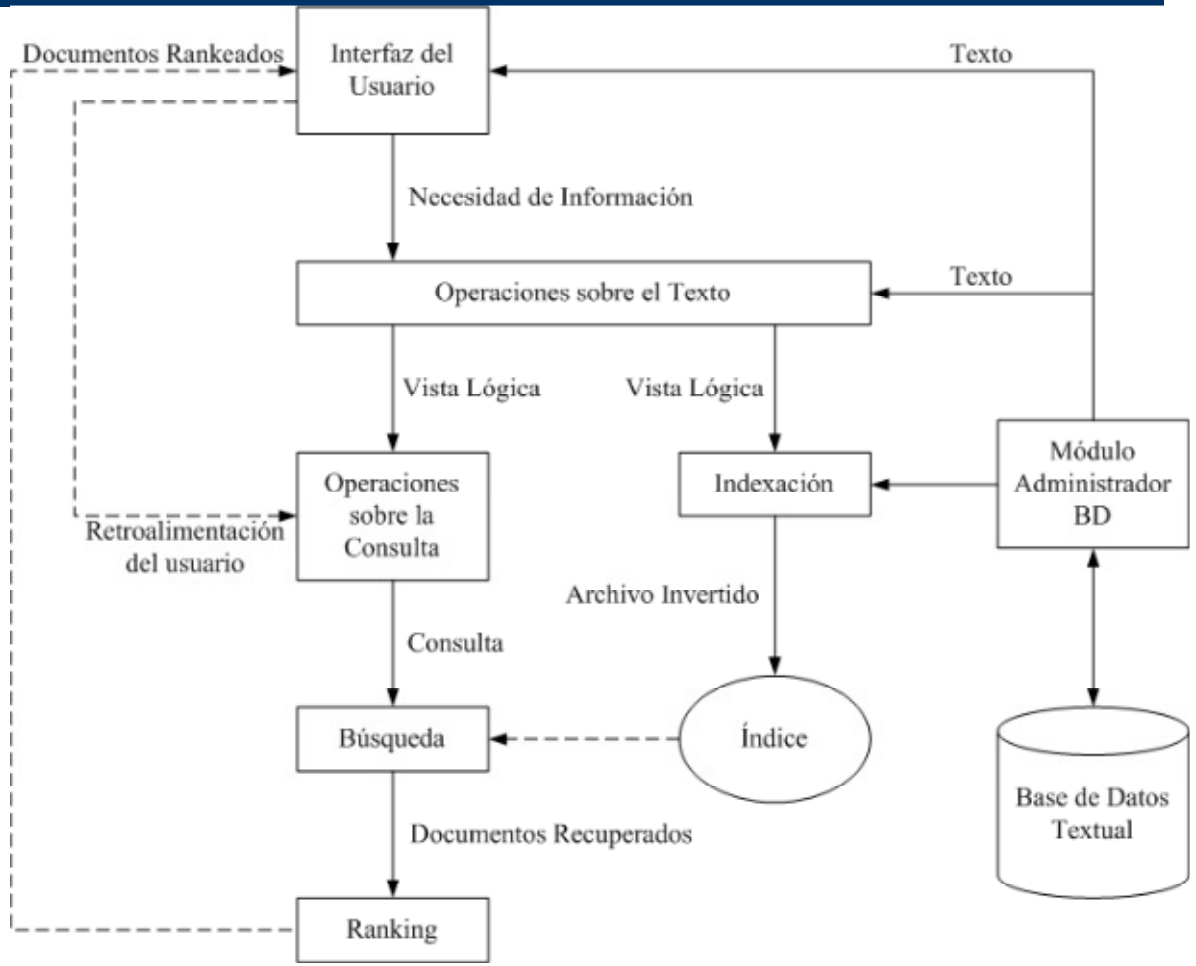
Recuperar Datos (RD)

- Lenguaje de consulta
- Estructurado
- Determinístico
- Resultados específicos
- Eficiencia se mide en términos de velocidad, consistencia y reducción del espacio de almacenamiento
- Registros y tuplas

Recuperar Información (RI)

- Lenguaje natural
- No estructurado
- No determinístico, y centrado en la necesidad de información del usuario
- Resultados ordenados de mayor a menor relevancia en forma de ranking
- Eficiencia se mide en términos de la calidad de los resultados
- Documentos

Proceso de Recuperación de Información



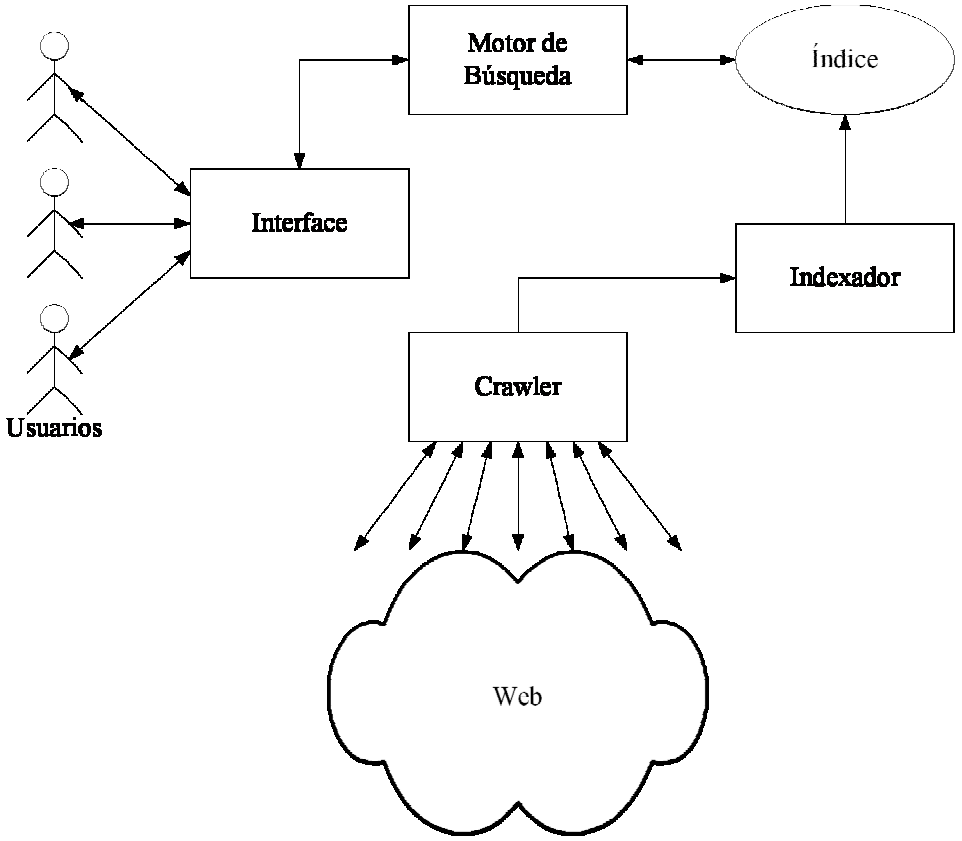
La Web reactivó la investigación en R.I.

- Almacenamiento no se estructura de manera uniforme.
- Consultado por todo tipo de usuarios (expertos e inexpertos).
- Sin regulación de contenidos ni formatos.
- Contiene información redundante.

Características de la Web

- Gigantesco volumen de texto.
- Crecimiento exponencial
- Alta volatilidad.
- Información heterogénea a nivel de: datos, formatos, idiomas y conjuntos de caracteres.
- Información distribuida y conectada por una red de enlaces constantemente variables.

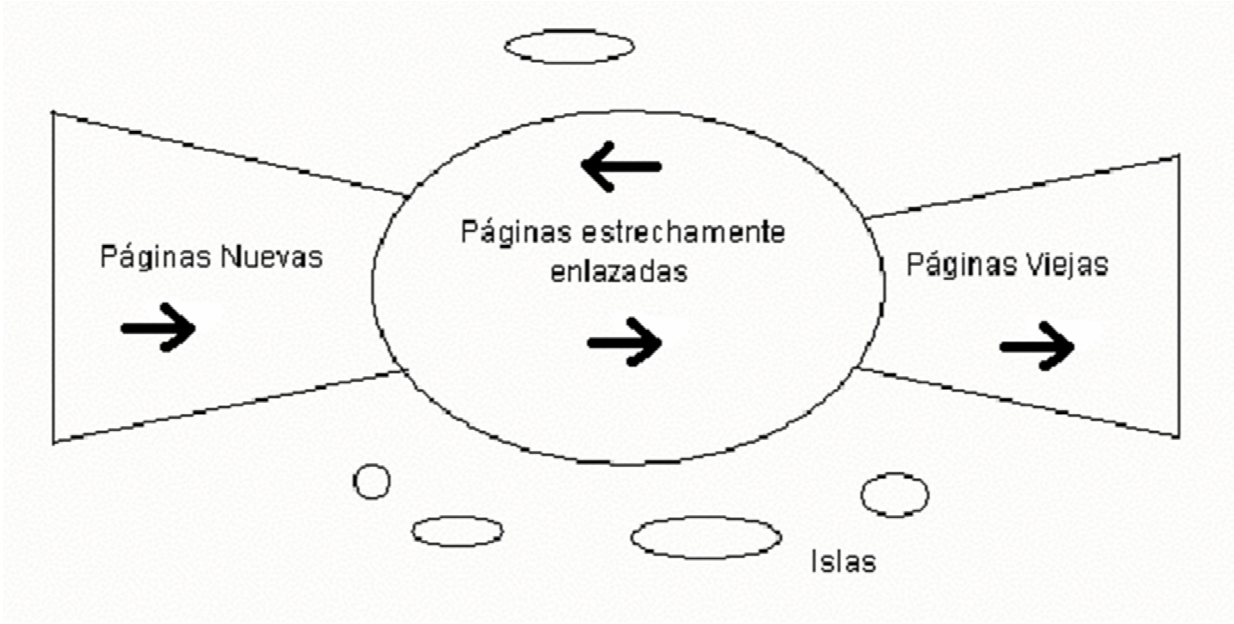
Funcionamiento de un Motor de Búsqueda WWW



No pueden indexar el 100% de la información en un momento dado

- Estimaciones indican que aproximadamente lo que se indexa es alrededor de 12% del total de páginas
- 23 % de ellas cambian constantemente
- Su vida promedio es de 10 días [Lau99]. Adicionalmente, dada la gran cantidad de páginas se hace prácticamente imposible obtener una copia local del Web [Kah97] [Cho98].

Estructura de la Web



Motores de búsqueda más utilizados en U.S. (Julio 2006)

Motor de Búsqueda		%
Google	Google	49.20
Yahoo Search	YAHOO!	23.80
MSN Search	msn	9.60
AOL Search	AOL Search	6.30
Ask.com	Ask.com	2.60
My Way	myway	2.30
Earth Link	EarthLink	0.60
iWon	iwon Search	0.60
Netscape Search	Netscape Search	0.50
Dogpile	DOGPILE	0.40

Los motores se diseñan siguiendo modelos matemáticos

Existen varios modelos:

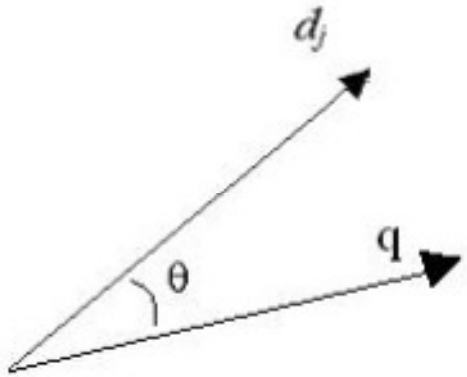
- Probabilístico
- Booleano
- Vectorial
- Vectorial generalizado
- Conjuntos difusos

Caracterización Formal del Modelo Vectorial

$$[\ D, Q, F, R(q_i, d_j)\ ]$$

- **D** = Vector de documentos, formado por los pesos de los términos de la colección en el documento.
- **Q** = Vector de consultas, formado por los pesos de los términos de la colección en la consulta.
- ***F*** = Álgebra Vectorial.
- $\boldsymbol{R(q_i, d_j)}$ = Distancia Coseno (Coseno del Ángulo entre los Vectores)

Modelo Vectorial



Similaridad entre documento y consulta en el modelo vectorial

$$sim(d_j,q) = \frac{\vec{d_j} \bullet \vec{q}}{\| \vec{d_j} \| \times \| \vec{q} \|}$$

$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} (w_{i,j})^2} \times \sqrt{\sum_{i=1}^{t} (w_{i,q})^2}}$$

Cada término tiene una rareza determinada en una colección

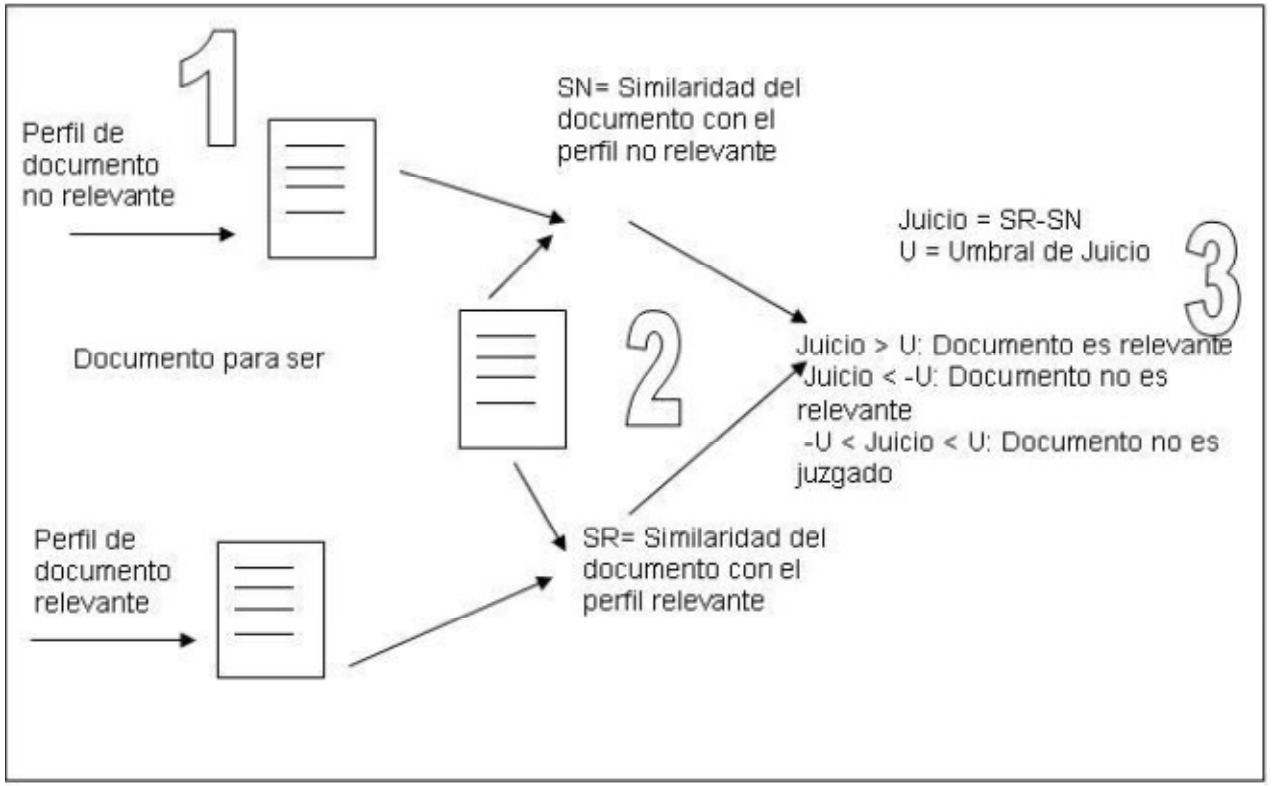
$$idf_i = \log \frac{N}{n_i}$$

Los pesos de un términos están dados por la fórmula idf*tf

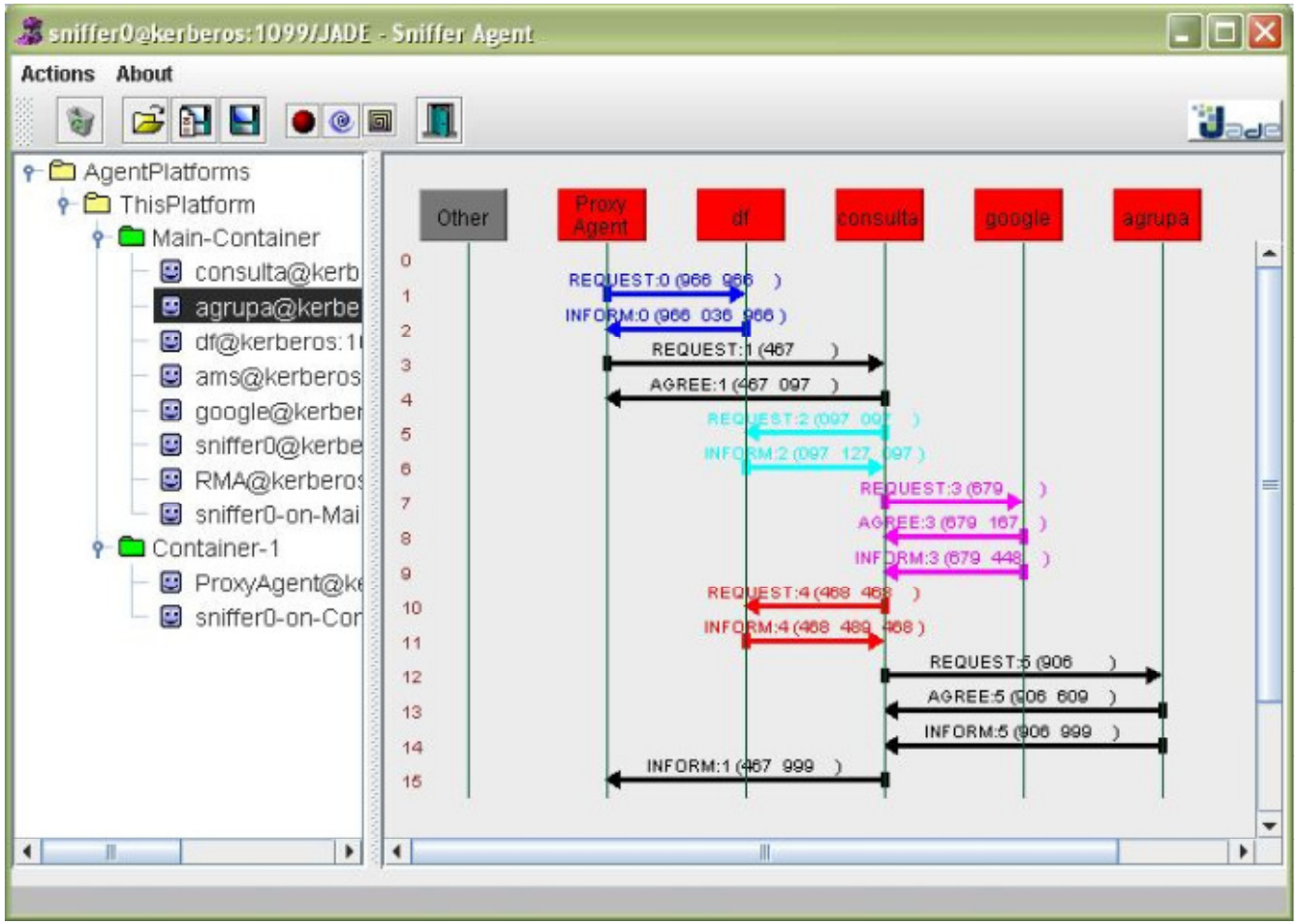
$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i}$$

$$w_{i,q} = \left(0.5 + \frac{0.5\, freq_{i,q}}{max_l freql_{,q}} \right) \times \log \frac{N}{n_i}$$

El filtrado y la clasificación pueden ser vistos como problemas de R.I.



Nuevas tecnologías y tendencias: Web Semántico, Agentes Inteligentes, Minería de Web, Minería de Uso



Conclusiones

- Con el WWW surgió una nueva era de investigación en R.I.
- Motores de búsqueda como Google no corresponden aún a una solución a todos los problemas de R.I.
- Los algoritmos clásicos y técnicas desarrollados desde hace más de cuarenta años continúan siendo los fundamentos primordiales del trabajo realizado en el presente.
- El objetivo de la R.I. es satisfacer la necesidad de información del usuario.
- El uso de Agentes Inteligentes, el Web Semántico, la minería de usuario, y se presentan como nuevas alternativas para .

¡Gracias!