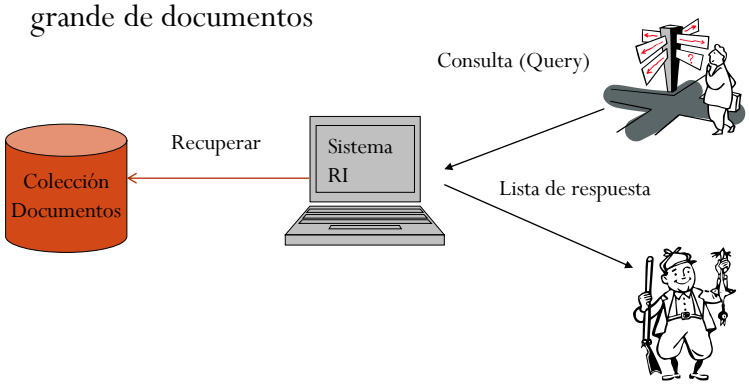
Introducción

Dra. Maya Carrillo Ruiz

El problema de Recuperación de Información

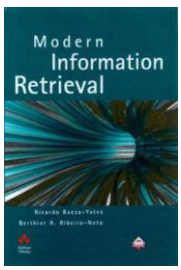
- Meta = encontrar documentos relevantes para una necesidad de información específica, a partir de un conjunto grande de documentos



Concepto de RI

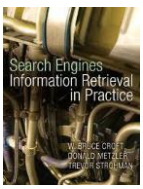
- ***Calvin N. Mooers*** *(1919-1994)*
  - *"Recuperación de Información (RI) es el ámbito que comprende los aspectos intelectuales de la descripción de la información y su especificación para buscar, así como cualquier sistema, técnica o máquina que se emplee para desarrollar la operación"*

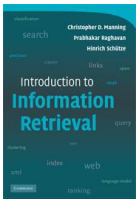
- ***Baeza & Ribeiro*** *(1999)*
  - *La recuperación de información trata con la representación, el almacenamiento, la organización y el acceso a elementos de información*



Concepto de RI

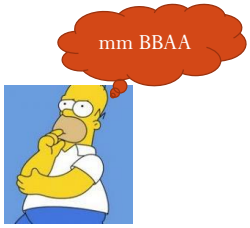
- **Van Rijsbergen**, 1979
  - *"Supongamos que existe un almacén de documentos y una persona formula una pregunta que tiene por respuesta un conjunto de documentos que satisfacen la necesidad de información expresada por esa pregunta"*
- **Croft, Metzler & Strohman**, 2010
  - *"Trata de modelar, diseñar e implementar sistemas capaces de proporcionar acceso basado en contenidos, rápido y efectivo. La meta de un sistema de recuperación es estimar la relevancia de elementos de información a la necesidad de información de un usuario, expresada como consulta"*
- **Manning, Raghavan & Schütze**, 2008
  - *"Trata de encontrar material de una naturaleza no estructurada (típicamente texto) que satisface una necesidad de información en una colección grande"*





Concepto de RI

- Ciertos elementos en común en las tres definiciones:
- El sistema de RI recibe una petición del usuario, una **consulta**
- Debe encontrar **información relacionada con la consulta**
- La información se busca dentro de un **repositorio**
- En este contexto los resultados se muestran ordenados (**ranking**)

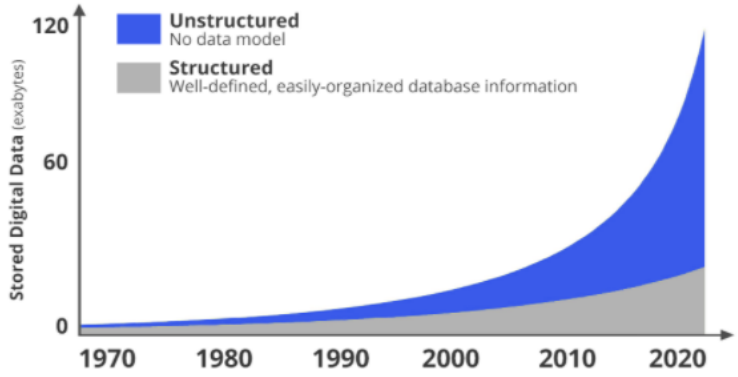


Problema: **SUBJETIVIDAD**

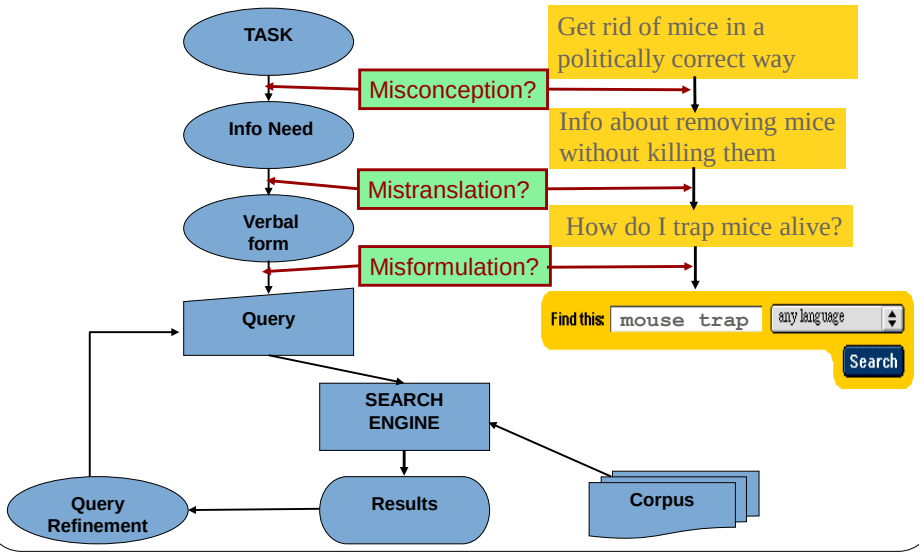
Concepto de RI

- **Información no estructurada** son datos con estructura semántica arbitraria
- **Documento** es cualquier información que se presenta al usuario (texto, página web, vídeo, música, imágenes,…)
- **Colección** es un repositorio de documentos
- La consulta debe traducirse para que el SRI la procese y recupere la información relevante (**palabras clave**)
- El usuario debe **encontrar fácilmente** los documentos de su interés en la salida del SRI (y navegar, filtrar resultados)

¿Porqué estudiar RI?



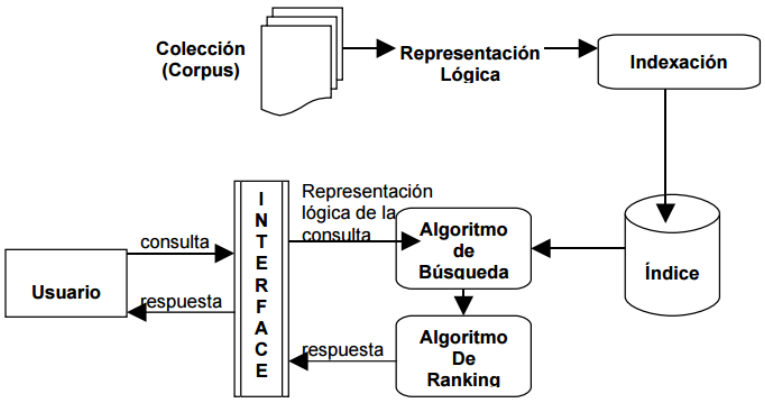
Modelo de búsqueda clásico



Tareas básica de un Sistema de Recuperación de Información (SRI)

- Representación lógica de los documentos y opcionalmente almacenamiento del original
- Representación delas necesidades de información del usuario en forma de consulta
- Evaluación de los documentos respecto de una consulta para establecer la relevancia de cada uno
- Ranking de los documentos considerados relevantes para formar el "conjunto solución" o respuesta
- Presentación de la respuesta al usuario
- Retroalimentación o refinamiento de las consultas (para aumentar la calidad de la respuesta)

Arquitectura de un SRI



RI abarca tópicos como

- Modelos de Recuperación: La tarea de la recuperación puede ser modelada desde distintos enfoques, por ejemplo la estadística, el álgebra de boole, el álgebra de vectores, la lógica difusa, el procesamiento del lenguaje natural y demás.
- Filtrado y Ruteo: Es un área que permite la definición de perfiles de necesidades de información por parte de usuarios y ante el ingreso de nuevos documentos al SRI, se los analiza y se lo reenvía para quienes se estima que van a ser relevantes.
- Clasificación: Aquí se realiza la rotulación automática de documentos de un corpus en base a clases previamente definidas

RI abarca tópicos como

- Agrupamiento (Clustering): Es una tarea similar a la clasificación pero no existen clases predefinidas. El proceso automáticamente determinará cuáles son las particiones.
- Sumarización: Área que entiende sobre técnicas de extracción de aquellas partes (palabras, frases, oraciones, párrafos) que contienen la semántica que determina la esencia de un documento.
- Detección de novedades (Novelty Detection): Se basa en la determinación de la introducción de nuevos tópicos o temas a un SRI.

RI abarca tópicos como

- Respuestas a Preguntas (Question Answering): Consiste en hallar aquellas porciones de texto de un documento que satisfacen expresamente a una consulta, es decir, la respuesta concreta a una pregunta dada.
- Extracción de Información: Extraer aquellas porciones de texto con una alta carga semántica y establecer relaciones entre los términos o pasajes extraídos.
- Recuperación cross-language: Hallar documentos escritos en cualquier lenguaje que son relevantes a una consulta expresada en otro lenguaje (búsqueda multilingual).

RI abarca tópicos como

- Búsquedas Web: Se refiere a los SRI que operan sobre un corpus web privado (intranet) o público (Internet). La web ha planteado nuevos desafíos al área de RI, debido a sus características particulares como – por ejemplo – dinámica y tamaño.
- Recuperación de Información Distribuida: A diferencia de los SRI clásicos donde el corpus y las estructuras de datos que auxilian a la búsqueda están centralizadas, aquí se plantea la tarea sobre los mismos elementos pero distribuidos sobre una red de computadoras.
- Modelado de Usuarios: Esta área – a partir de la interacción de los usuarios con un SRI – estudia como se generan de forma automática perfiles que definan las necesidades de información de éstos.

RI abarca tópicos como

- Recuperación de Información Multimedia: Más allá de que los SRI tradicionales operan sobre corpus de documentos textuales, la recuperación de información tiene que tratar con otras formas alternativas de representación como imágenes, registro de conversaciones y video.

- Desarrollo de Conjuntos (data-sets) de Prueba: A los efectos de evaluar SRI completos o nuevos métodos y técnicas es necesario disponer de juegos de prueba normalizados (corpus con preguntas y respuestas predefinidas, corpus clasificados, etc.). Esta área tiene que ver con la producción de tales conjuntos, a partir de diferentes estrategias que permitan reducir la complejidad de la tarea, manejando la dificultad inherente a la carga de subjetividad existente.

Sistemas de Recuperación de Datos (SRD)

- Recuperación de datos
  - Objetos
    - Son estructuras de datos conocidas
    - Su representación en formato previamente definido y significado implícito (sintaxis y semántica no ambigua) para cada elemento
  - Consultas
    - Estructuras bien definidas
    - No son ambiguas
    - Consisten en un conjunto de condiciones que deben cumplir los ítems a evaluar para que la misma se satisfaga
    - Utilizando lenguajes como SQL (Structured Query Language) cuya semántica es precisa

```
SELECT *
FROM Clientes Chivilcoy
WHERE Localidad = "Chivilcoy"
AND Saldo_Cuenta > 10000
```

    - Conjunto completo de resultados

SRI

- RI
  - Documentos que contengan información biográfica de los deportistas de México que ganaron más medallas olímpicas en los últimos 10 años"
  - Objetos
    - documentos de texto sin estructura
  - Consultas
    - Construir una expresión que refleje exactamente las necesidad de información del usuario
    - concepto de relevancia
      - la salida (respuesta) se encuentra confeccionada de acuerdo con algún criterio que evalúa la "similitud" que existe entre la consulta y cada documento.
    - El resultado es un ranking (que no es sinónimo de "orden", tal como se lo entiende habitualmente en RD), donde la primera posición corresponde al documento más relevante a la consulta y así decrece sucesivamente.
    - Conjunto de respuesta no es exacto.

Diferencias entre un SRD y SRI

	SGBD	SRI
Estructura	Información estructurada con semántica bien definida.	Información semi o no estructurada.
Recuperación	Determinística. Todo el conjunto solución es relevante para el usuario	Probabilística. Una porción de los documentos recuperados puede no ser relevante.
Consulta y Lenguaje	Especificación precisa (no hay ambigüedad). Lenguaje formal, preciso y estructurado.	Hay imprecisión en su formulación. Lenguaje natural, ambiguo y no estructurado.
Resultados	Aciertos exactos	Aciertos parciales

SRI de acuerdo a la interacción con el usuario

- 1) Recuperación inmediata: El usuario plantea su necesidad de información y – a continuación – obtiene referencias a los documentos que el sistema evalúa como relevantes. Existen dos modalidades:
  - a) Búsqueda (propiamente dicha) o recuperación "ad-hoc", donde el usuario formula una consulta en un lenguaje y el sistema la evalúa y responde. En este caso, el usuario tiene suficiente comprensión de su necesidad y sabe cómo expresar una consulta al sistema. Un ejemplo clásico son los buscadores de Internet como Google ,Yahoo o MSN Search.
  - b) Navegación o browsing: En este caso, el usuario utiliza un enfoque diferente al anterior. El sistema ofrece una interfaz de temas donde el usuario "navega" por dicha estructura y obtiene referencias a documentos a relacionados. Esto facilita la búsqueda a usuarios que no pueden definir claramente cómo comenzar con su consulta e – inclusive – van definiendo su necesidad a medida que observan diferentes documentos.

SRI de acuerdo a la interacción con el usuario

- Recuperación diferida: El usuario especifica sus necesidades y el sistema entregará de forma continua los nuevos documentos que le lleguen y concuerden con ésta. Esta modalidad recibe el nombre de filtrado y ruteo y la necesidad del usuario – generalmente – define un "perfil" (profile) de los documentos buscados. Nótese que un "perfil" es – de alguna forma – un query y puede ser tratado como tal. Cada vez que un nuevo documento arriba al sistema se compara con el perfil y – si es relevante – se envía al usuario. Un ejemplo, es el servicio provisto por GoogleAlert.

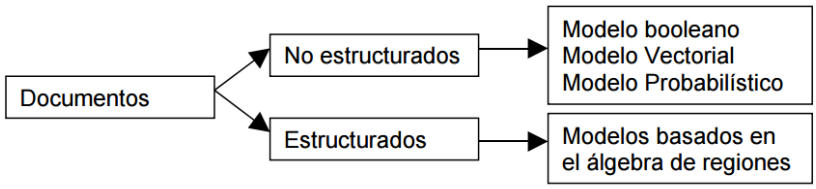
Relevancia

- La recuperación de información intenta resolver el problema de encontrar documentos relevantes que satisfagan la necesidad de información de un usuario
- Imposibilidad de expresar exactamente tal necesidad
- La noción de relevancia es un juicio subjetivo y depende de diferentes factores relacionados con el usuario
- La relevancia de un documento respecto de un query se refiere a cuánto el primero responde al segundo.

Relevancia

- Se plantea la relevancia como similitud de manera de poder comparar documentos con consultas y – bajo ciertos criterios – definir una medida de distancia entre ambos
- Un documento es relevante a una consulta si son similares
- La medida de similitud puede estar basada en diferentes criterios (coincidencias de términos, significado de éstos, frecuencia de aparición de términos, distribución del vocabulario, entre otros)

Modelos de RI



Concepto de RI

Information Retrieval		Data Retrieval
Probabilístico	Modelos de recuperación	Determinista
Se deriva de los contenidos	Indexación	Todos los items
Mejor coincidencia	Devuelve	Coincidencia exacta
Lenguaje Natural	Tipo de las consultas	Estructuradas
Relevancia	Criterio de resultados	Todos los resultados
Ranking	Orden de los resultados	No