Modelos de RI

UCR – ECCI

CI-2414 Recuperación de Información

Prof. Kryscia Daviana Ramírez Benavides

¿Qué es un Modelo?

- Es la primera etapa para abordar el tema de la RI.
- Representación matemática para resolver el problema de recuperación de información.
- Un buen modelo debe de “adivinar” lo que el usuario realmente quiso preguntar.
- Los modelos son utilizados para el cálculo de la relevancia.
- Para construir un modelo:
  - Analizar las representaciones de documentos y consultas.
  - Concebir el marco en el que pueden ser representados.
  - Construcción de función de ranking.

Caracterización Formal de los Modelos de RI

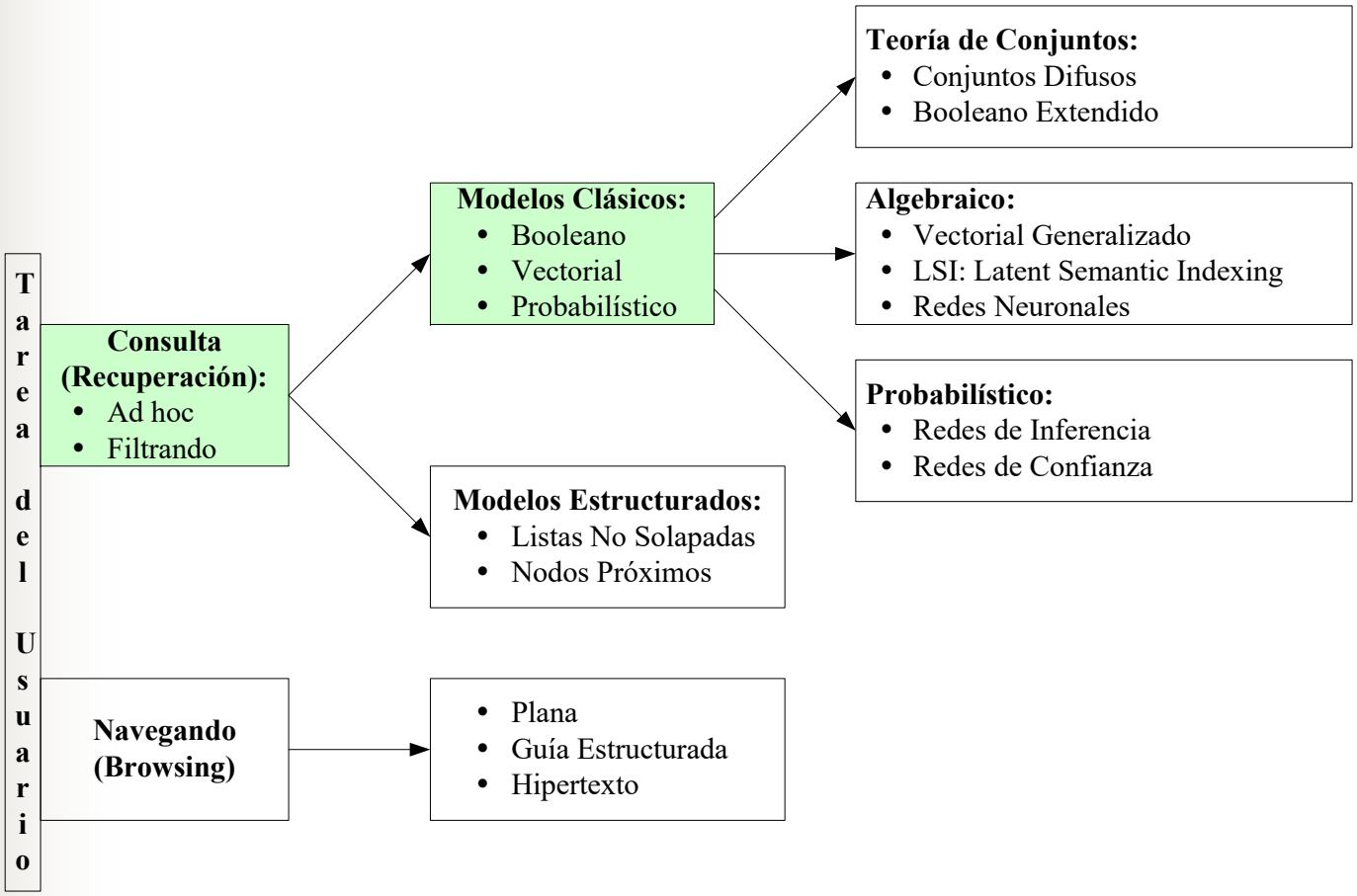
- *Un modelo de recuperación de información es una cuádrupla:*

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

donde:

  - **D** *es el conjunto de vistas lógicas (o representaciones) de los documentos de la colección.*
  - **Q** *es el conjunto de vistas lógicas (o representaciones) de las necesidades de información del usuario (llamadas consultas).*
  - ***F*** *es el marco de trabajo (framework) de las representaciones de documentos modeladas, consultas y sus relaciones.*
  - $\boldsymbol{R(q_i,d_j)}$ *es la función de ordenamiento (ranking), número real que relaciona la consulta* $\boldsymbol{q_i} \in \mathbf{Q}$ *con la representación de documento* $\boldsymbol{d_j} \in \mathbf{D}$. *Tal ranking define un orden en el que el documento satisface la consulta* $\boldsymbol{q_i}$.

Taxonomía de los Modelos de RI



Clasificación de Modelos de RI

Vista Lógica de Documentos

Tarea del Usuario		**Términos Índice**	**Full Text**	**Full Text + Estructura**
	Consulta (Recuperación)	Modelos Clásicos Teoría de Conjuntos Algebraico Probabilístico	Modelos Clásicos Teoría de Conjuntos Algebraico Probabilístico	Modelos Estructurados
	Navegación (Browsing)	Plana	Plana Hipertexto	Guía Estructurada Hipertexto

Modelos Clásicos

- Los documentos se describen a través de un conjunto de términos representativos llamados índices o términos índice.
- Se pueden considerar todos los términos como importantes en una aproximación llamada *full text*.
- No todos los términos son igualmente importantes. El proceso de decidir la importancia de un término se puede realizar a través de la asignación de *pesos*.
- Todos los modelos clásicos tienen ciertas falencias comunes, la más notoria es la incapacidad para capturar las relaciones entre términos, ya que los términos son independientes.

Modelos Alternativos

- Se extiende de los modelos clásicos.
- Son en general bastante costosos de implementar, y no siempre dan grandes resultados.
- Por ello, son en general poco populares, con la excepción del LSI y, las Redes Neuronales y Bayesianas.

Modelos Estructurados

- Tratan de combinar la información del contenido del texto con la estructura del texto.
- Se pierde la noción de relevancia, y estamos ante una recuperación de datos.
- Ejemplo:
  - Un usuario tiene mucha memoria visual. Recuerda un documento donde aparece 'holocausto atómico' en cursiva, cerca de una imagen que tiene en la etiqueta la palabra 'tierra'.

    same-page(near('holocausto atómico', Figure(etiqueta('tierra'))))
  - Se recuperarán aquellos documentos que satisfagan exactamente la consulta, por tanto no hay orden de relevancia en los resultados.

Modelos Estructurados (cont.)

- Aunque no se proporciona escala de relevancia, este es un tema de investigación hoy en día y se pude conseguir una relevancia parcial.
- Cuanto más expresivo es el lenguaje de consulta, más ineficiente resulta.
- Componentes:
  - ***Match point*:** Posición inicial en el texto de una secuencia que satisface la condición.
  - **Región:** Porción de texto.
  - **Nodo:** Componente estructural del texto, sección, capítulo, etc.
- Los documentos se estructuran en nodos (secciones), los cuales son regiones con propiedades predefinidas que son conocidas tanto por el autor como por el usuario que busca.

Modelos de Navegación

- No se basan en consultas del usuario, sino que el usuario navega a través de una jerarquía hasta encontrar los documentos relacionados a lo que anda buscando.
- Son una guía jerárquica de directorios que va de los temas más generales a los más particulares. Listan lugares (URLs) y los clasifican en categorías, además de añadir comentarios identificativos sobre ellos.
- Su objetivo es encontrar los documentos que pertenezcan al área temática seleccionada.

Modelos de Navegación (cont.)

- Están compuestos por dos partes:
  - La BD que es construida por los URLs remitidos.
  - Una estructura jerárquica que facilita la consulta a la base.
- Al conectar con algún buscador nos encontraremos con una página que contiene una estructura jerárquica de temas, es decir, hay un grupo de temas generales, al seleccionar uno nos sale otro grupo de temas dependiente (cada vez mas específico) del que nos llevó allí, y podemos seguir así hasta que localicemos el tema de nuestro interés o se acaben las categorías creadas por el autor del buscador.

Modelos de Navegación (cont.)

- No es muy bueno, porque distrae al usuario antes de que localice lo que se había propuesto encontrar, puede que el usuario al andar navegando se encuentre y se entretenga con documentos que pierden el objetivo de la búsqueda original. Tal vez no era muy atractivo lo que buscábamos.
- No suele estar muy actualizado, ya que se hacen a mano.
- Es lento para encontrar lo deseado, pues exige varios pasos previos.
- Existen ítems de difícil categorización.
- Se pierde la noción de relevancia, y estamos ante una búsqueda entre carpetas.

Referencias Bibliográficas

- La información fue tomada de:
  - Libro de texto del curso.