Introducción a la Recuperación de Información

UCR – ECCI

CI-2414 Recuperación de Información

Prof. Kryscia Daviana Ramírez Benavides

Características de la Web

- Gigantesco volumen de texto.
- Texto altamente volátil.
- Información heterogénea (datos, formatos, idiomas, alfabetos).
- Información distribuida y conectada por una red de calidad variable.
- Información mal estructurada y redundante.
- Información de mala calidad, sin revisión editorial de forma ni contenido.

Características de la Web (cont.)

- No cuenta con una buena organización.
- Consultado por usuarios inexpertos.
- Nadie se hace responsable de los contenidos.
- No es fácil de buscar ni indexar.
- Documentos no uniformes, websites no uniformes, en promedio 8 links de una página.
  - 70% Inglés
  - 3% Español

Recuperación de Datos versus Recuperación de Información

Recuperar Datos (RD)

- Lenguaje de consulta
- Estructurado
- Determinístico
- Datos específicos
- No permite errores - Exacto
- Importante: Eficiencia (velocidad y espacio)
- Registros

Recuperar Información (RI)

- Lenguaje natural
- No estructurado
- No determinístico
- Necesidad de información
- Permite errores - Inexacto
- Importante: Calidad de la respuesta
- Documentos

Recuperación de Datos versus Recuperación de Información (cont.)

- Recuperar Datos (RD) versus Recuperar Información (RI):
  - SELECT name,position,salary

    FROM Employees

    WHERE company = “Coca-Cola” AND salary >= 3000

    versus
  - Quiero información sobre las consecuencias de la crisis de los misiles cubanos en el desarrollo de la guerra fría

Recuperación de Datos versus Recuperación de Información (cont.)

- Los datos *se pueden estructurar* en tablas, árboles, etc. para recuperar exactamente lo que se quiere. El texto *no tiene estructura clara y no es fácil crearla*.
- En RD se sabe exactamente lo que se quiere, en RI no existe la respuesta correcta.
- En RD se hace *búsqueda de datos específicos*, en RI se hace una *búsqueda ante una necesidad de información*.

Recuperación de Datos versus Recuperación de Información (cont.)

- En RD se dan *datos exactos* (no se permiten errores), en RI cada *documento puede ser más o menos relevante* y esto puede cambiar según el usuario y la situación (se permiten errores).
- En RD sólo importa la *eficiencia* (básicamente velocidad y espacio), mientras que en RI importa también (o más) la *calidad* de la respuesta.
- En RD se recuperan *registros*, en RI se recuperan *documentos*.

¿Qué es Recuperación de Información?

- Dado que comprender cabalmente el significado de un texto en forma automática es imposible en la práctica, RI busca una *aproximación* a responder lo que el usuario busca.
- El problema de RI se puede definir como:

  Dada una necesidad de información (consulta + perfil del usuario + ...) y un conjunto de documentos, ordenar los documentos de más a menos relevantes para esa necesidad y presentar un subconjunto de lo más relevantes.

¿Qué es Recuperación de Información? (cont.)

- En general, RI no congenia bien con el modelo relacional. Las extensiones hechas a los RDBMS para incorporar texto son limitadas y no todo lo eficientes que se necesita.
- Dos grandes etapas para abordar el problema:
  - Elegir un *modelo* que permita calcular la relevancia de un documento frente a una consulta.
  - Diseñar *algoritmos y estructuras de datos* que lo implementen eficientemente (*índices*).

¿Qué es Recuperación de Información? (cont.)

- Cuán bien hecha está la primera etapa se mide comparando las respuestas del sistema contra las que un conjunto de expertos consideran relevantes.
- Cuán bien hecha está la segunda etapa se mide considerando el tiempo de respuesta del sistema, espacio extra de los índices, tiempo de construcción y actualización del índice, etc.
- RI requiere modelos y algoritmos especializados.

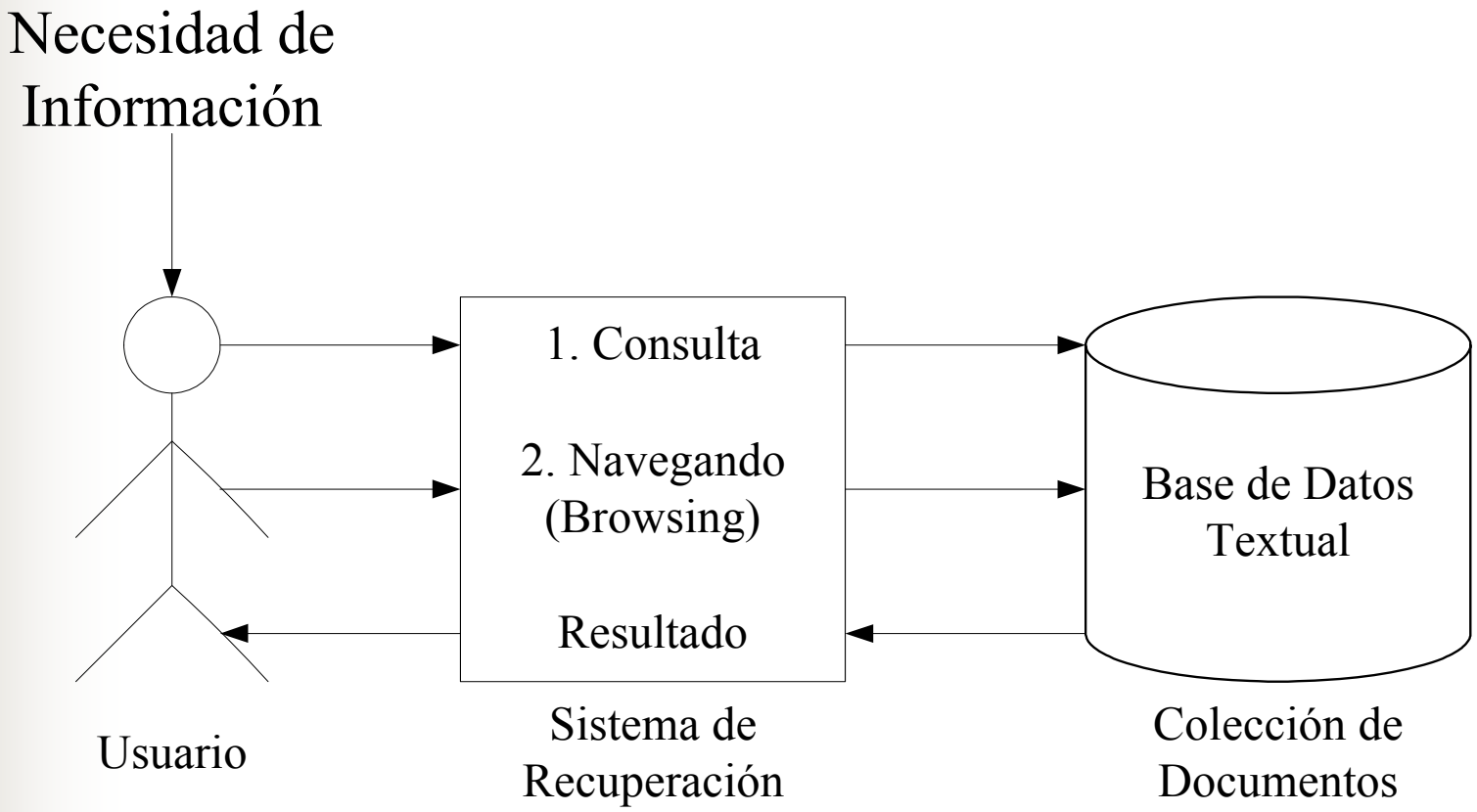
¿Qué es Recuperación de Información? (cont.)

- RI se basa en la utilización de términos índice para indexar y recuperar documentos.
- Indexar un documento puede consistir en sustituir su contenido por un conjunto de términos índices que lo representan.
- Recuperar puede consistir en especificar un conjunto de términos que deben hallarse entre los índices de un documento, estableciendo un ranking de relevancia.

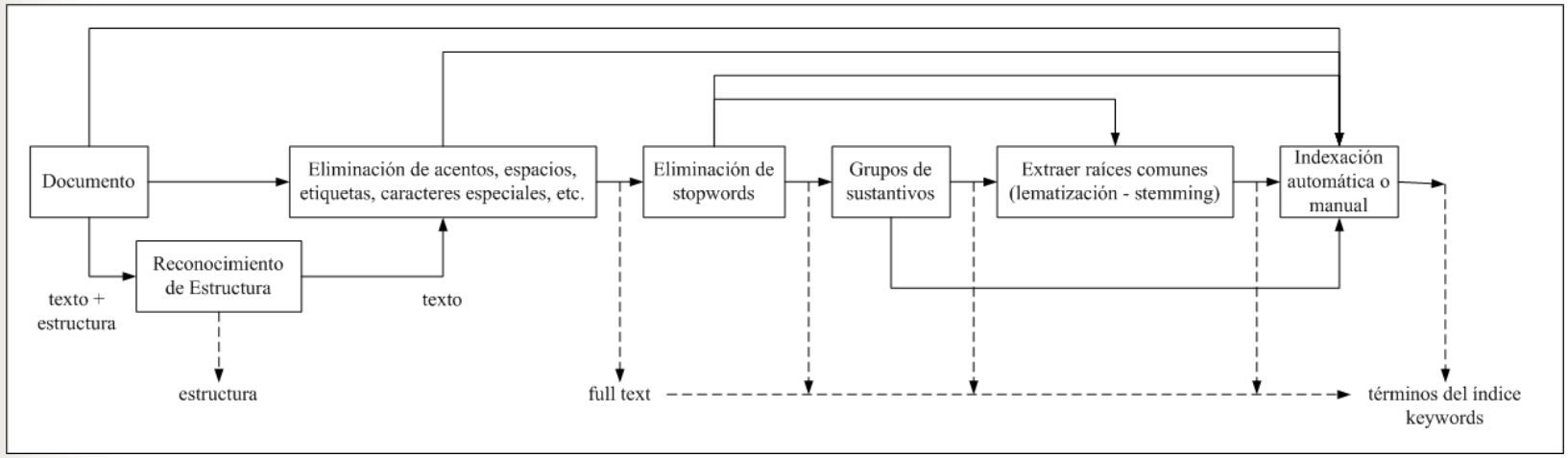
¿Qué es Recuperación de Información? (cont.)

- El problema de RI es la manera de predecir la relevancia de los documentos y su grado de relevancia (ranking).

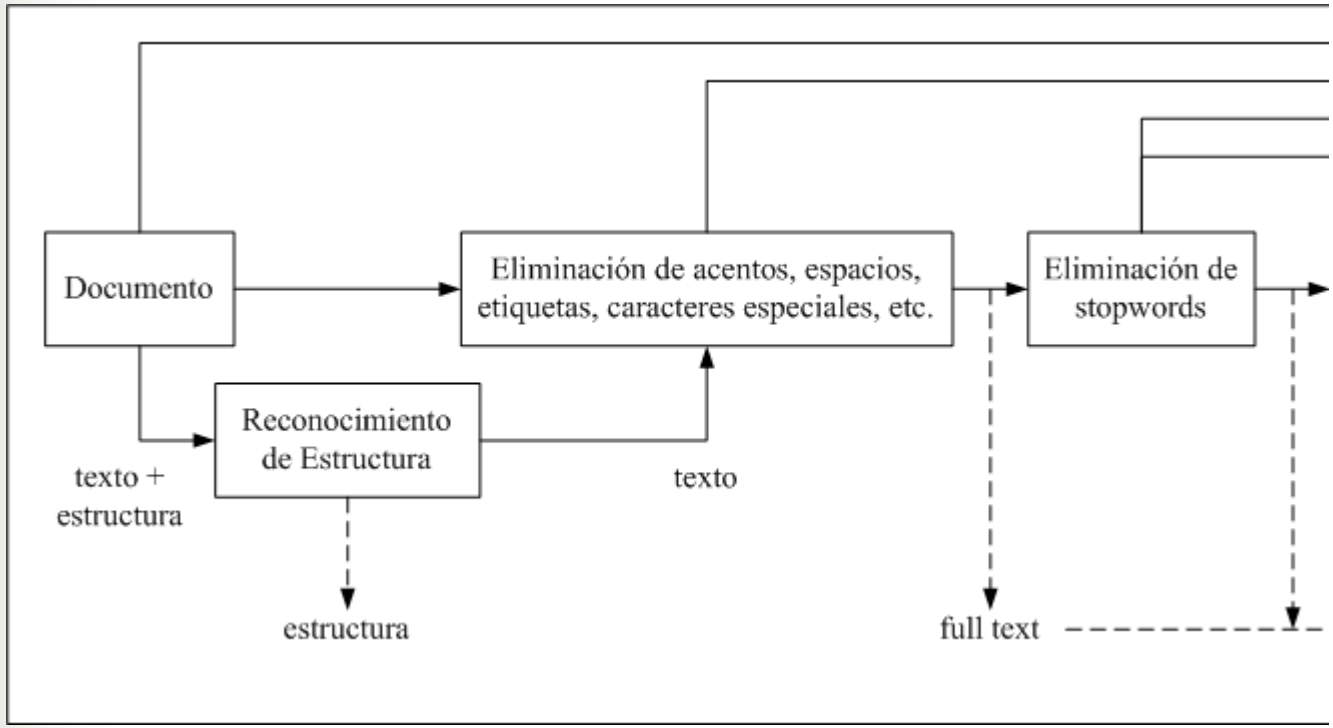
Interacción del Usuario con el Sistema de Recuperación



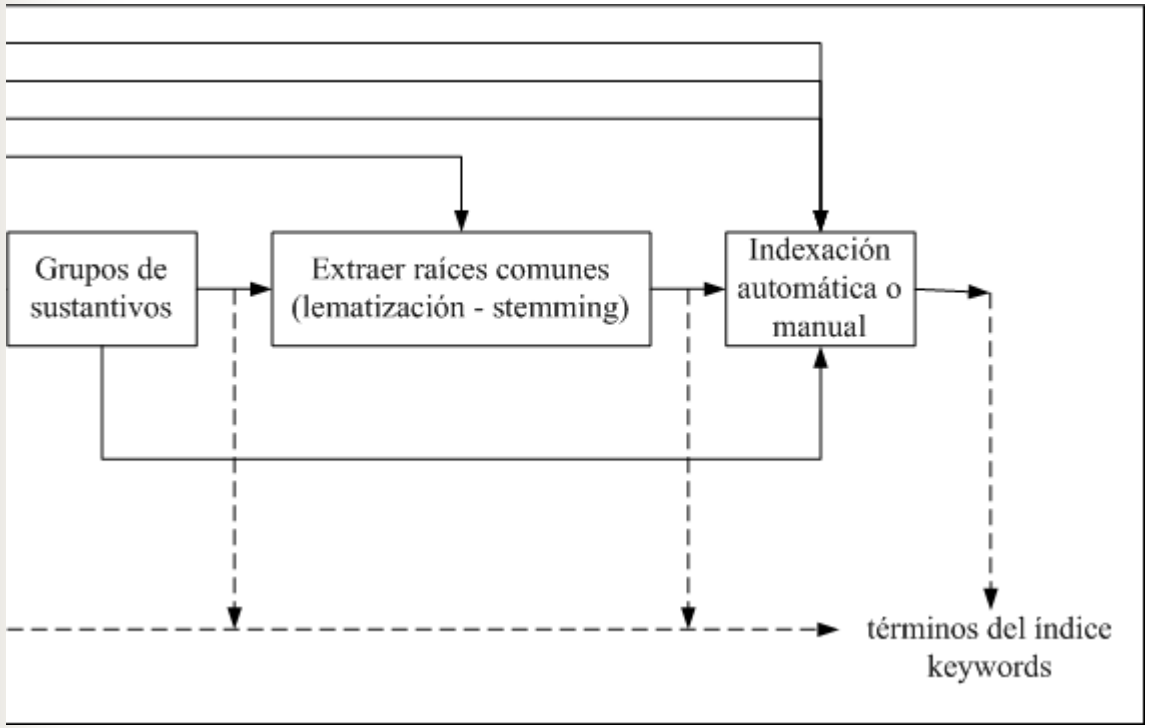
Vista Lógica de un Documento



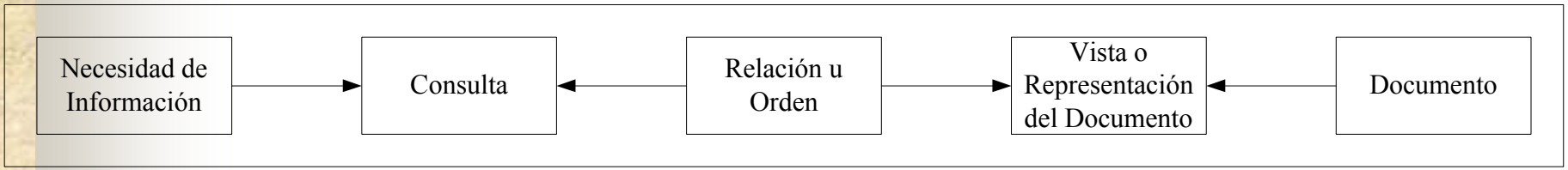
Vista Lógica de un Documento (cont.)



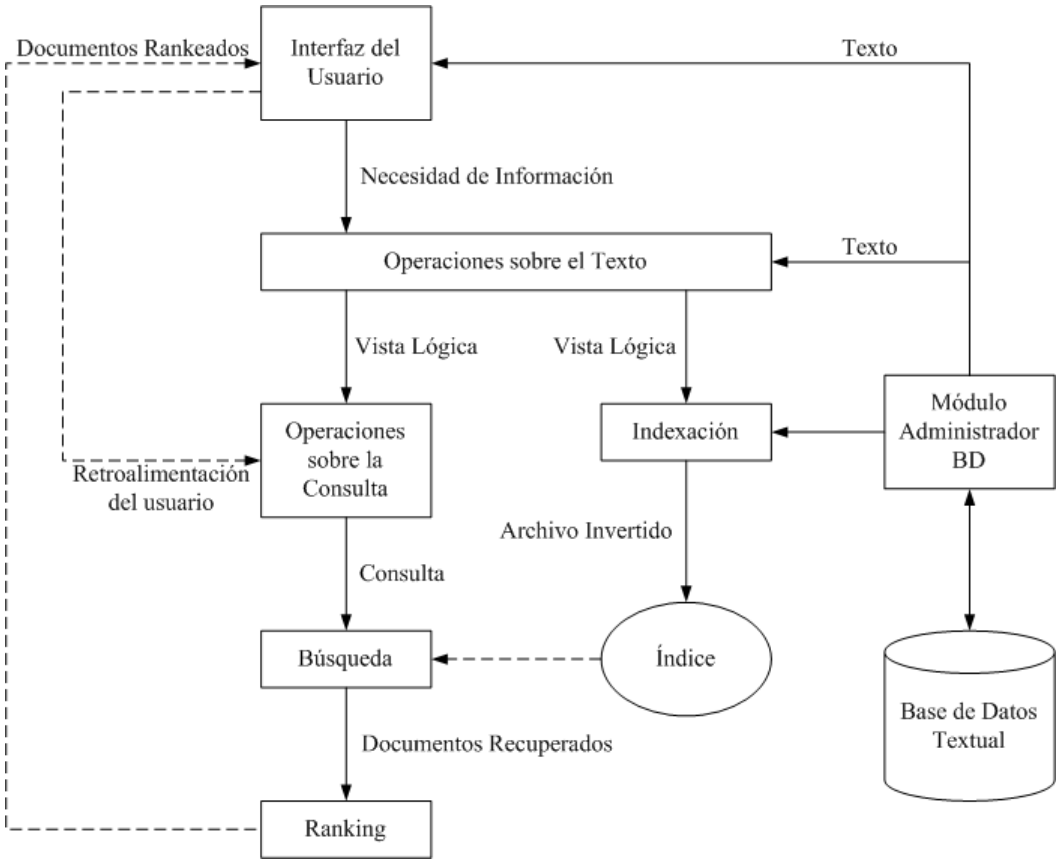
Vista Lógica de un Documento (cont.)



Vista de un Documento



Proceso de Recuperación de Información



Referencias Bibliográficas

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