Bases de Datos Textuales, Una Herramienta en el Trabajo de la Recuperación de Información

UCR – ECCI

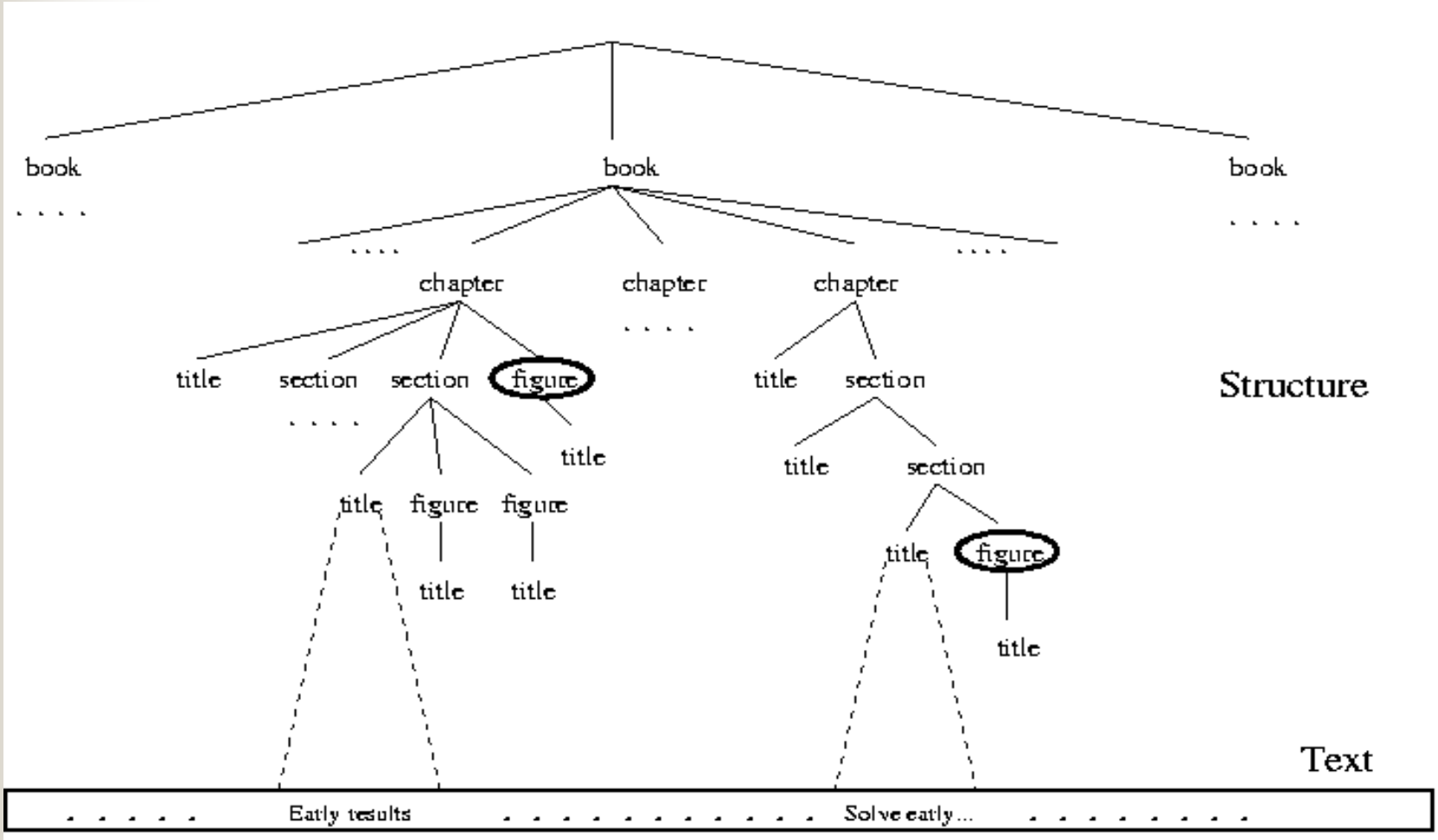
CI-2414 Recuperación de Información

Prof. Kryscia Daviana Ramírez Benavides

Un Acercamiento a las Bases de Datos Textuales

- Múltiples aplicaciones: bibliotecas, ingeniería de software, automatización de oficinas, diccionarios automatizados.
- Diferentes disciplinas: medicina, biología, historia, geografía, derecho, bibliotecología, lingüística.
- Compuestas de dos partes: contenido y estructura.
- El objetivo, como un sistema de recuperación de información, es ayudar a los usuarios a encontrar lo que buscan.

Partes en que se Divide una Base de Datos Textual



Recuperar Datos en Contraste a Recuperar Información:

- SELECT title

  FROM titles

  WHERE price > 10000 AND type = ‘business’

  versus

- Solicito información sobre especies de mamíferos en bosques tropicales.

Conocimiento Lingüístico

- Lenguaje Natural.
- Por ejemplo: un estudiante de Biología, que necesita información sobre “animales extintos” esperaría recuperar documentos con frases como:

  “animal extinto”

  “especie de animal extinta”

  “la extinción de los animales”
- En otro caso, si se busca sobre:

  “guerra fría”

  se quisiera recuperar un documento que hablara sobre:

  “la crisis de los mísiles cubanos”

Conocimiento Lingüístico (cont.)

- El sistema no tiene idea de que ambas cosas están relacionadas y la intersección de vocabulario puede ser nula.
- La solución a este problema pasa por el análisis lingüístico:
  - Lematizar
  - Etiquetar
  - Detectar frases comunes
- Otro elemento importante es el uso de tesauros y sinónimos.

Modelos

- Es la primera etapa para abordar el tema de la RI.
- Un buen modelo debe de “adivinar” lo que el usuario realmente quiso preguntar.
- Existen modelos más especializados para recuperar información:
  - Modelo Booleano
  - Modelo Vectorial
  - Modelo Probabilístico

Modelo Booleano

- Utiliza los operadores lógicos.
- La relevancia en este modelo es binaria.
- Un ejemplo que ilustra los operadores:

  Maradona AND Mundial AND ((México'86 OR Italia'90) BUTNOT U.S.A.'94)
- Es el más primitivo y malo para recuperar información.
- Es bastante popular.

Modelo Vectorial

- Selecciona un conjunto de palabras útiles para discriminar (términos o keywords).
- Toda palabra del texto es un término, excepto posiblemente las palabras vacías o stopwords.
- Se puede enriquecer esto con procesos de lematización, etiquetando, e identificación de frases.
- Se tiene un conjunto de términos (t1, .., tk) y un conjunto de documentos (d1, .., dN). Un documento di se modela como un vector

$$\rightarrow$$

$$di \rightarrow di = (w(t1,di), \ldots, w(tk,di))$$

- Es el más popular hoy en día.

Modelo Probabilístico

- Reconoce que existe exactamente un subconjunto de documentos que son relevantes para una consulta dada.
- Para cada documento, se intenta evaluar la probabilidad de que el usuario lo considere relevante.
- La relevancia de un documento d se calcula como:

$$\frac{P(\text{d relevante para q})}{P(\text{d no relevante para q})}$$

- Recupera los documentos que con mayor probabilidad son relevantes.
- Es poco popular.

Indización

- Es la segunda etapa para abordar el tema de la RI.
- En un índice estándar cada palabra distinta se almacena con su referencia numérica, que especifica el lugar del documento donde ésta se encuentra:

  experimentar 34, 90

  experimentación 50

  animal 14

  animales 15

Indización (cont.)

- El tamaño de dicho índice puede ser notablemente reducido por medio de técnicas basadas en conocimiento lingüístico.
- La reducción de este tipo de índices comprende desde una octava hasta una décima parte de su tamaño.
- Un índice estándar como el anterior, tomaría la siguiente forma:

  experimentar 34, 50, 90

  animal 14, 15

Índices Analíticos

- Se encuentra al final de una obra.
- Este índice reúne frases y palabras claves en orden alfabético junto a las páginas donde pueden ser encontradas.
- Ha sido adoptado en la documentación de software.
- La implementación de este tipo de índices es costosa.
- Puede ser elaborado eficaz y automáticamente, si se recurre a tecnología basada en conocimiento lingüístico.

Índices Invertidos

- En su versión más básica, consta de dos partes :
  - Vocabulario: Conjunto de términos distintos del texto.
  - Posteo: Para cada término, la lista de documentos donde aparece.
- La construcción de un índice invertido comienza:
  - Recorre la colección de textos secuencialmente.
  - Cada término leído es buscado en el vocabulario.
  - Si no existe se agrega con una lista de posteo vacía.
  - El documento que se está leyendo se agrega al final de la lista de posteo del término.
  - Una vez completada, el índice es grabado al disco.

Truncamiento en la Gestión de Bases de Datos

- Los algoritmos de truncamiento gestionan automáticamente las distintas formas de una palabra.
- Extrae los sufijos, creando una raíz de la misma.
- Ejemplo:
  - Generalizaciones
  - Generalizando
  - Generalizar
  - Generalizas
- Se extraría automáticamente la siguiente cadena: “GENER”.

Truncamiento en la Gestión de Bases de Datos (cont.)

- Las técnicas de truncamiento no brindan una aproximación óptima para la comprensión textual.
- M.F Porter es el creador de uno de los algoritmos de truncamiento más usado para el idioma inglés.
- Las reglas de construcción de palabras se complica (como en lenguas de origen latino).

Conclusiones

- El afán por la búsqueda de mejores y más óptimas soluciones.
- Se trata una visible diferencia, incluso dentro del mundo de las bases de datos, entre recuperar información de documentos y la recuperación de datos.
- La adopción del conocimiento lingüístico, ha sido adoptado para el desarrollo de algoritmos.
- No existe un modelo y un método de indización ideal para la implementación de estos sistemas.