- Modelos de RI
- Tipos de interrogación
- Ficheros comprimidos

Slide 1

Modelos de RI

- Modelos clásicos
- Variantes

Todos se basan en una interpretación del texto como una secuencia de palabras.

¿Qué es una palabra?

Slide 2

¿Todos los documentos son así?

Tener en cuenta la estructura del documento

secciones, titulos . . .

Tener en cuenta el objetivo del usuario

Respuesta o curiosear (browsing)

Modelos de RI para textos estructurados

same-page(near ('atomic holocaust', Figure(label('earth'))))

- Listas sin superposición
- Proximal nodes

match point posición inicial en el texto de una sequencia que satisface la condición.

región porción de texto

Slide 3

nodo componente estructural del texto, seccion, capítulo ...

Un nodo es una región con propiedades predefinidas que son conocidas tanto por el autor como por el usuario.

Listas sin superposición

- Dividir el texto de cada documento en regiones sin superposición que se almacenan en una lista
- Mantener diferentes listas por cada tipo de nodo
  capítulos, secciones, figuras, ...
- Cada lista es una estructura de datos diferente.
- Las regiones en una misma lista no tienen superposición, pero regiones en diferentes listas si que pueden superponerse.

Slide 4

Listas sin superposición

- Hay que aumentar la información almacenada en la representación del documento. En vez de un vector con pesos, se parece más a un índice.
- Cada document tiene una representación en cada nivel de estructura.
- La información que se guarda por término es una lista de ocurrencias del término a nivel de las regiones que se esten considerando.

Slide 5

- La información se extrapola del texto plano a la región correspondiente.
- Información de las ocurrencias a cada nivel de estructura.

- Permite dar respuesta a interrogaciones del tipo:

  selecciona una región A que contenga esta palabra y que no contiene otras regiones de tipo B

  selecciona una región A que no esté contenida en ninguna otra región

Proximal nodes

- Definición de una estructura jerárquica adicional en los niveles de indexación.
- Cada uno de los niveles de estructura forma parte de una jerarquía compuesta de nodos, por ejemplo.

  capítulos, secciones, párrafos, páginas y líneas
- Cada nodo está asociado a una región de texto, que se obtiene a través de la jerarquía

Slide 6

- Dos regiones se superponen si y solo si están en diferentes niveles.
- La representación es idéntica a la del modelo anterior.
- Permite preguntas como

  [(*section) with ('holocaust')]

  que busca secciones, subsecciones, párrafos ... que contengan 'holocaust'
- Permite, en algunos casos, hacer algo más conveniente que buscar todas las ocurrencias de un término.

Modelos de IR para "browsing"

El usuario no tiene una pregunta concreta en mente y utiliza los documentos recuperados para obtener más información o plantear nuevas interrogaciones.

Contempla la posibilidad de recuperación interactiva

- Plano
- Guiado por estructura
- Modelo hipertexto

Slide 7

En este modelo no siempre hay que contestar a una pregunta!

Plano

- El documento no tiene ningún tipo de organización. Una lista de palabras
- El usuario lee los documentos recuperados y extrae palabras que se incorporan a la interrogación inicial

  relevance feedback
- ¿Dónde está?

  Es fácil perder la referencia, si accedemos a la página 15 de una novela. ¿En qué capítulo estamos?
- Se suele incorporar una indicación jerarquíca del camino seguido (mirar Yahoo! ojo que no es plano)

Slide 8

Guiado por estructura

- Incorporar los elementos estructurales para ayudar en la búsqueda
- Añade un índice de términos

- Es importante que el sistema incorpore mecanismos de visualización para evitar volver al mismo sitio o mantner información de la historia de la búsqueda.

Slide 9

Modelo hipertexto

- La estructura tradicional de una novela es secuencial
- Sin embargo en otros contextos puede que se contengan diferentes niveles de información para los que no es necesario leer todo el documento
- Para ello se utiliza el hipertexto

- Compromiso de diseño
- El autor tiene que hacer un estudio cuidadoso de los diferentes niveles de lectura y escribir el hipertexto de acuerdo con el plan
- El lector tiene que quedar satisfecho con la estructura

- El proceso de navegación es un paseo por un grafo dirigido
- Necesita selección del conjunto de documentos, interrogación normal, con el añadido de que debe permitir al usuario acceder a la zona adecuada del texto
- Necesita herramientas que permitan visualizar la historia de la búsqueda

Slide 10

Modelo hipertexto versus Web

Web en un sentido amplio es hipertexto

Si embargo carece de un plan de accion visible (y recuperable) en los hiperenlaces

Veremos más adelante los métodos que se usan en la web para recuperar información

Slide 11

Tipos de interrogación

Repasaremos los modelos más comunes

Una interrogación es la expresión de una necesidad de información

Habrá variaciones y limitaciones dependiendo del modelo de RI seleccionado

Repasaremos los principios más comunes

- Interrogación basada en palabras clave
- Pattern matching
- Interrogaciones estructurales

Slide 12

Interrogación basada en palabras clave

- Palabra simple
- Contexto
- Boolean
- Lenguaje natural

Slide 13

Interrogación basada en palabras clave: Palabra simple

- Es la interrogación elemental se espera recuperar los documentos que contengan la palabra
- Utilizando otros criterios (por ejemplo frecuencia) para ordenar la respuesta
- Hay que definir el concepto de *palabra*
- En algunos modelos queremos acceso a las posiciones del texto donde aparece el término

Slide 14

Interrogación basada en palabras clave:Contexto

Consideramos dos contextos: frase y proximidad

- Frase
  - Interrogación es una secuencia de palabras 'enhance retrieval'
  - Queremos recuperar documentos que contienen la secuencia
  - Hay una cierta indefinición en la definición de "contiene"
  - La misma secuencia pero con diferentes separadores?
    'enhance retrieval'
    'enhance-retrieval'
  - Alguna palabra muy frecuente adicional?
    'enhance the retrieval'
  - La respuesta no es necesariamente la misma que la intersección de las respuestas correspondientes a cada uno de los términos
  - Se puede hacer un ranking similar al de palabras simples

Slide 15

- Proximidad:
  - Es una versión relajada de frase
  - Interrogación: lista de palabras y una distancia
    enhance retrieval WITHIN 4
  - Se recuperan los documentos que contienen las palabras a una distancia (número de palabras) menor o igual que el valor expecificado.
    'enhance the power of retrieval'
    'the retrieval of energy enhance the power'
  - No hay una definición clara de ranking, aunque se puede extender el asociado a términos

Slide 16

Interrogación basada en palabras clave: Boolean

- Ya lo comentamos hace tiempo
- No hay una noción asociada de ranking
- El significado de las operaciones lógicas no es transparente al usuario
- Suele dar mejores resultados si se utiliza una noción de lógica difusa (fuzzy logic)

Slide 17

Interrogación basada en palabras clave: Lenguaje natural

- Extender la noción de difuso a la lógica
- Interrogación es una lista de palabras y preguntas de contexto
- Se recuperan todos los documentos que contienen alguna parte de la interrogación
- Se valoran más los documentos que contienen más partes de la interrogación
- Negación?

  Se permite al usuario incorporar palabras que no deben aparecer en el documento

  Esto se penaliza en la función de valoración
- De forma natural se deben incorporar

  búsqueda de documentos similares

  refinamiento de la interrogación

Slide 18

Pattern matching

Patrón = expresión de propiedades de un texto

Expresión de la sintáxis buscada

- Palabras
- Prefijos
- Sufijos
- Subcadenas
- Rangos
- Permitir errores
- Exresion regular
- Patrones extendidos

  Subconjunto de las expresiones regulares user friendly

Slide 19

Interrogaciones estructurales

- Estructura fija

  Interrogación término y estructura

- Hipertexto

  Recuperación sigue un sistema navegacional, en particular en la web

  La interrogación suele ser en lenguaje natural

  Analizaremos más adelante como hacer ranking en la web incorporando la información que se deriva de la estructura de hiperenlaces.

- Estructura jerárquica

  Representación en árbol

  Interrogación etiquetas en los nodos IN, WITH etiquetas en las hojas término o tipo de estructura

  buscar una figura dentro de una seccion en la que hay un título con la palabra 'structure'

Slide 20

Ficheros comprimidos

- Compresión de texto
- Proceso de recuperación

Slide 21

Compresión de texto

Objetivo: crear una representación de tamaño menor a partir de la cual se pueda reconstruir el texto original

Métodos

- Asignar un código a cada palabra códigos de Huffman, códigos aritméticos. Reemplazar palabras por código.
- Tecnicas de diccionario: sustituir una secuencia de símbolos repetidos por un puntero a la ocurrencia previa. Ziv-Lempel

Slide 22

Huffman coding

Un símbolo puede corresponder a una palabra o a un conjunto de símbolos, en nuestro caso debe corresponder con el bloque de texto.

Idea: Asignar una codificación en bits de longitud variable a cada símbolo.

La longitud del código es inversa a la frecuencia de aparición del símbolo.

Asegurando que ningún código es prefijo de otro se garantiza la decompresión correcta del texto.

Slide 23 La asignación de códigos se representa mediante un árbol binario con símbolos en las hojas. Se necesita

- determinar los símbolos
- analizar su frecuencia de aparición en el texto
- crear un nodo por cada símbolo con la frecuencia asociada: bosque inicial

for each rose, a rose is a rose, the rose

symbol	frec	symbol	frec
for	1	each	1
rose	4	,	2
a	1	is	1
the	1		

Slide 24

for 1/11 | each 1/11 | rose 4/11 | , 2/11 | a 1/11 | is 1/11 | the 1/11

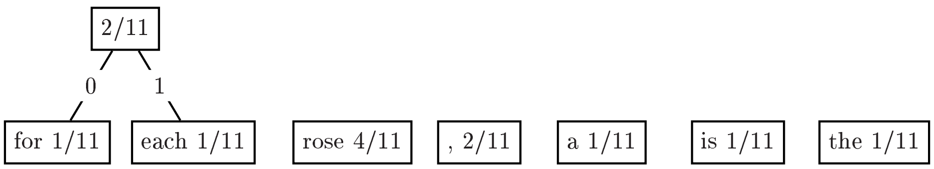
Construir un árbol aplicando las reglas siguientes

- Seleccionar los dos árboles con menor frecuencia en la raíz.
- Unirlos creando una raíz nueva con frecuencia la suma de las frecuencias en las raíces de los dos árboles.
- Las dos ramas del nuevo árbol reciben las etiquetas 0/1.

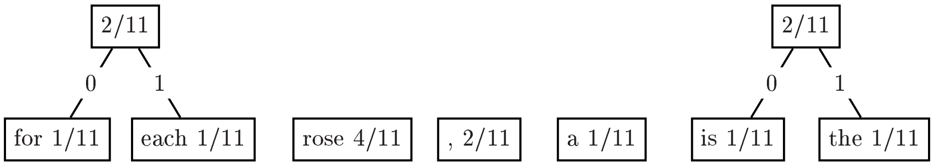
El código asociado con un símbolo es la secuencia de etiquetas encontradas en el camino desde la raíz a la hoja que contiene el símbolo.

Slide 25 Es neceario que no haya palabras que sean prefijo de otras palabras. Esto se consigue con un símbolo de *fin de palabra*

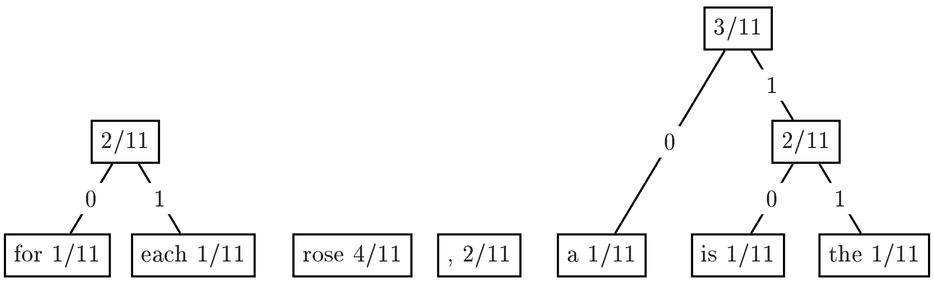
Slide 26



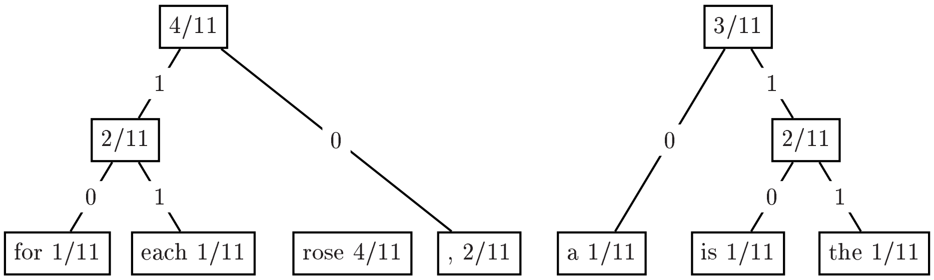
Slide 27



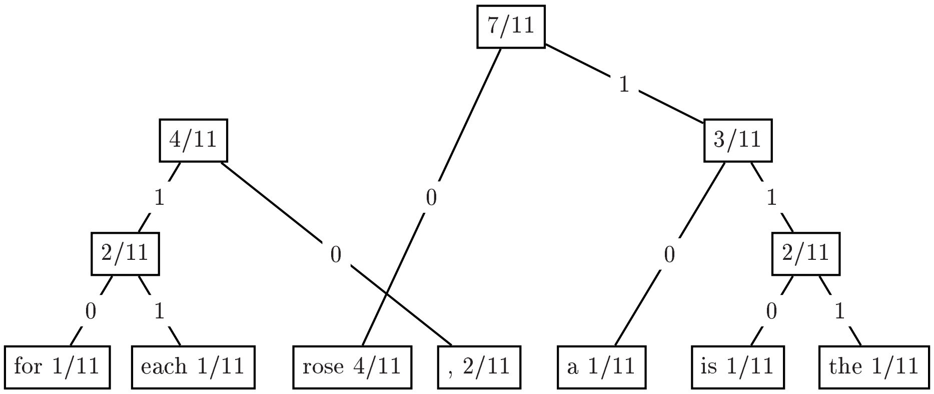
Slide 28

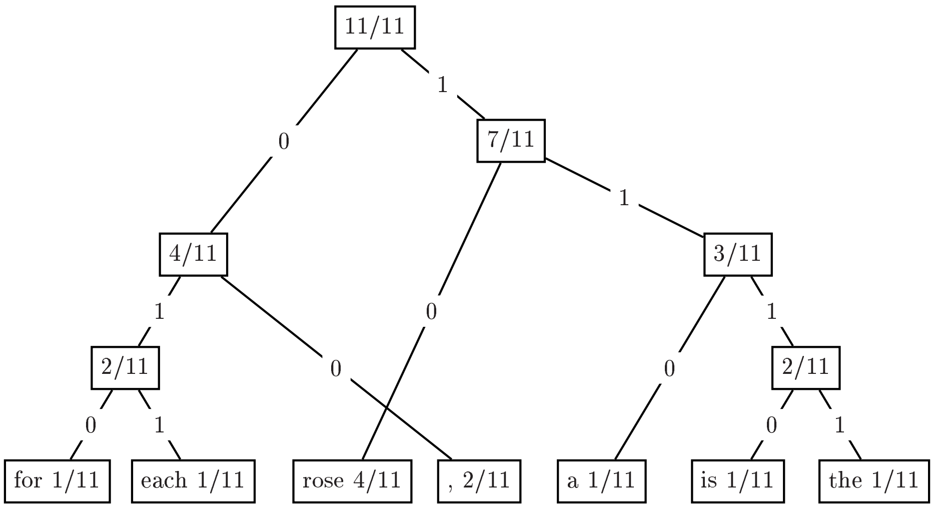


Slide 29



Slide 30





Slide 31

symbol	frec	code	symbol	frec	code
for	1	010	each	1	011
rose	4	10	,	2	00
a	1	110	is	1	1110
the	1	1111			

Slide 32

Recuperación

símbolos = palabras

Generar un código Huffman para cada palabra

Comprimir el texto.

Obtener todos los códigos Huffman de palabras que se desean buscar en el texto.

Buscar los códigos.

Slide 33