

Práctica de recuperación de la información

Ejercicio 0: Dé una definición general de recuperación de información (RI). ¿Qué involucra el retorno de la información cuando consideramos información en la Web?

Ejercicio 1: Supongamos que desea encontrar documentos que contengan al menos k de un conjunto dado de n palabras clave. Suponga también que dispone de un índice de palabras clave que le proporciona una lista (ordenada) de identificadores de documentos que contienen una palabra clave determinada. Dé un algoritmo eficiente para encontrar el conjunto de documentos deseado.

Ejercicio 2: Describa las componentes principales de un sistema de retorno de la información.

Ejercicio 3: ¿Qué es el modelo de espacio vectorial de retorno de la información? ¿Cómo se construye un vector para representar un documento?

Ejercicio 4: ¿Qué es un tesauro? ¿Cómo es beneficioso para el retorno de la información?

Ejercicio 5: Supongamos que desea realizar una consulta de palabras clave en un conjunto de tuplas en una base de datos, donde cada tupla tiene sólo unos pocos atributos, cada uno de los cuales contiene sólo unas pocas palabras. ¿Tiene sentido el concepto de frecuencia de términos en este contexto? ¿Y el de frecuencia inversa de documentos? Explique su respuesta. Sugiera también cómo puede definir la similitud de dos tuplas utilizando los conceptos TF-IDF.

Ejercicio 6: Reúna cinco documentos que contengan unas tres oraciones cada uno, y cada uno contiene algún contenido relacionado. Construya un índice invertido de todas las palabras clave de estos documentos.

Ejercicio 7: Describa el proceso de construir el resultado de un pedido de búsqueda usando un índice invertido. Asuma las consultas son basadas en proximidad. ¿Cómo debe ser el índice invertido? Se usan sentencias *ADJ*, *PRE* y *NEAR*. Considerar el significado de estas sentencias de la filmina 6 del archivo sobre retorno de la información. Dar algoritmo de cómo se procesa cada una de las sentencias para dar una respuesta a una búsqueda.

Ejercicio 8: compare retorno de la información para un conjunto de archivos en su PC con retorno de la información en la web. Definir al menos 5 criterios de comparación en que son diferentes y construya una tabla para ello.

Ejercicio 9: ¿Cuál es la idea por detrás del algoritmo PageRank? ¿Tiene alguna limitación? ¿Qué se puede hacer para superarla?

Ejercicio 10: enuncie 3 problemas que debe resolver un rastreador web.

Ejercicio 11: Comparar un sistema de retorno de la información basado en relevancia con un sistema de BD relacional considerando los siguientes criterios:

- Datos
- Índices (similitudes y diferencias)
- Informaciones acerca de los datos que no son ni los datos ni los índices.
- Ordenamiento de los resultados de las consultas
- Inputs para procesar una consulta