Trabajo Práctico de Organización de datos (75.06)

Indexador de archivos

Entrega de diseño

**1. Enunciado.**

**2. Introducción**

Para el desarrollo de este trabajo práctico se cuenta con un conjunto de archivos de texto, en los cuales debe ser posible realizarse búsquedas p untuales por frases en alguno o varios de ellos. Por lo tanto el trabajo práctico estará divido en dos partes, un proceso de construcción y almacenamiento de un índice, y otro, un proceso de consulta, basándonos en el índice anteriormente construido. Dado que se requiere un tamaño del índice lo más pequeño posible y una alta performance en las búsquedas se intentará utilizar modelos que se adecúen a estas pautas.

**3. Proceso de construcción del índice**

Se utilizará el modelo de índices invertidos para guardar el índice en memoria, éste consta de:

1. Almacenamiento de términos (palabras)

2. Almacenamiento de punteros a los archivos donde aparece dicho término.

3. Almacenamiento de el/los offset de cada término en cada archivo (esto es necesario porque la búsqueda a realizarse será por frases)

El índice se compondrá de dos archivos:

El almacenamiento de punteros se realizará a base de la distancia de una aparición con la siguiente, es decir que si el término aparece en los documentos 5,10,15,16 se almacenará 5,5,5,1. Para codificar los punteros se hará uso de los Códigos Delta.

Dado que tratamos con términos de longitud fija, utilizaremos otro archivo donde se almacenará el término, su frecuencia y su offset respectivo en el archivo de punteros.

Construccion del índice

Primero se recorre el archivo secuencialmente, obteniéndose así su lista de términos y frecuencias de cada uno. Se volcarán los datos en un archivo, aumentando la frecuencia de un término si está repetido, y manteniendose este archivo ordenado.

Se tomará un tope fijo de memoria y se guardará un subíndice

Se utilizará sort externo (replacement selection) y finalmente un optimal merge para obtener el índice.

Estructura de archivos utilizados.

Constará de dos archivos

|  |  |
| --- | --- |
| Termino | Offset a punteros |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Long documentos | Doc 1,.. | Docn | Freq1,.. | Freqn | Offset en archivo 1,.. | Offset en archivo n |

**4. Búsqueda por frases**

Para buscar una frase, se parseará la búsqueda con el mismo criterio utilizado para parsear el índice, se buscará en el índice mediante búsqueda binaria los términos involucrados y luego se hará una intersección entre ellos hasta encontrar los matches adecuados de la siguiente manera:

Se realiza la interseccion ubicando los documentos donde aparece toda la lista de palabras.

Por cada documento recuperado se obtiene la frecuencia del término en el mismo y la lista de posiciones. A estas posiciones de la lista se les resta la posición del término en la frase, esto se requiere para asegurarnos que las palabras aparezcan en el mismo orden que se buscaron. El proceso de intersecciones debe comenzar siempre por la palabra de menor frecuencia, si al final de la intersección la lista contiene una o más posiciones significa que la frase existe en dicho documento en la cantidad de veces que aparece en la lista. En otro caso la frase no está en el documento en el órden buscado.