Estructuras de Datos

Fecha entrega: 15/05/2024

Para la entrega del TP resuelto arme un único archivo comprimido (.tar.gz) y envielo a través del siguiente formulario: https://forms.gle/E5c8v4m3ip23Vstk9 el cual se encontrará habilitado hasta la fecha de entrega establecida.
Bibliografía sugerida: MIR [1] Capítulos 8, Croft [2] Capítulo 5, MAN [3] Capítulos 4.

1) Codifique un script que indexe una colección[1] que requiera el volcado parcial a disco (asumiendo que existe un límite de memoria). Su script debe recibir un parámetro *n* que indica cada cuántos documentos se debe hacer el volcado a disco. Al finalizar, debe unir (merge) los índices parciales. Para las pruebas use la colección snapshot de Wikipedia[2] y varios valores de *n* (por ejemplo, n = 10% del tamaño de la colección). Registre los tiempos de indexación y de merge por separado. Grafique la distribución de tamaños de las posting lists. Calcule el overhead de su índice respecto de la colección. Calcule el overhead para cada documento. ¿Qué conclusiones se pueden extraer?

2) Codifique un script que empleando la estrategia TAAT sobre el índice creado en el ejercicio 1 y operaciones sobre conjuntos permíta buscar por dos o tres términos utilizando los operadores AND, OR y NOT.

3) Utilizando el código e índice anteriores ejecute corridas con el siguiente subset de queries[3] (filtre solo los de 2 y 3 términos que estén en el vocabulario de su colección) y mida el tiempo de ejecución en cada caso. Para ello, utilice los siguientes patrones booleanos:

   a) Queries |q| = 2
      - t1 AND t2
      - t1 OR t2
      - t1 NOT t2

   b) Queries |q| = 3
      - t1 AND t2 AND t3
      - (t1 OR t2) NOT t3
      - (t1 AND t2) OR t3

[1] Almacene docID o docID+frecuencia de acuerdo a un parámetro de entrada.
[2] http://dg3rtljvitrle.cloudfront.net/wiki-small.tar.gz(debug), http://dg3rtljvitrle.cloudfront.net/wiki-large.tar.gz (run)
[3] https://www.labredes.unlu.edu.ar/sites/www.labredes.unlu.edu.ar/files/site/data/ri/EFF-10K-queries.txt

¿Puede relacionar los tiempos de ejecución con los tamaños de las listas? (pruebe con el índice en disco o cargándolo completamente en memoria antes). ¿Qué conclusiones se pueden extraer?

4) Codifique un script que empleando la estrategia DAAT sobre el índice creado en el ejercicio 1 resuelva consultas usando el modelo vectorial y retorne los top-k documentos de score mayor.

5) Agregue *skip lists* a su índice del ejercicio 1 y ejecute un conjunto de consultas AND sobre el índice original y luego usando los punteros. Compare los tiempos de ejecución con los del ejercicio 3. Luego, agregue un script que permita recuperar las skips list para un término dado. En este caso la salida deberá ser la lista ordenada por docID.

6) Sobre la colección Dump10k escriba un programa que realice una evaluación TAAT y otro usando DAAT. Compare los tiempos de ejecución para un conjunto de queries dados[4]. Separe su análisis por longitud de queries y de posting lists.

7) Comprima el índice del ejercicio 1 utilizando Variable-Length Codes para los docIDs y Elias-gamma para las frecuencias (almacene docIDs y frecuencias en archivos separados). Calcule tiempos de compresión/descompresión del índice completo y tamaño resultante en cada caso. Realice dos experimentos, con y sin DGaps. Compare los tamaños de los índices resultantes.

Bibliografía

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval: The Concepts and Technology Behind Search. Addison-Wesley Publishing Company, USA, 2nd edition, 2008.

[2] Bruce Croft, Donald Metzler, and Trevor Strohman. Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company, USA, 1st edition, 2009.

[3] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Sch ¨utze. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA, 2008.

[4] http://www.labredes.unlu.edu.ar/sites/www.labredes.unlu.edu.ar/files/site/data/ri/queriesDump10K.txt.tar.gz