Ejercicio Formativo #5

CC4102 - Diseño y Análisis de Algoritmos - Otoño 2021

Profesor: Jérémy Barbay - Auxiliar: Javier Oliva S. - Ayudante a cargo: Juan Manuel Saez
Alumno: Matías Vergara

A.

Analiza y compara los ordenes de tamaño de la cantidad de accesos a memoria externa generada por la secuencia entre:

Un árbol B

En el caso de un árbol B, sabemos que la inserción cuesta $O(\log_B(N))$ I/O[1]. En nuestro caso, $N = MB$. Luego requerimos de MB operaciones $O(\log_B(MB))$, lo cual resulta en $O(MB \log_B(MB)) = O(MB \log_B(B) + MB \log_B(M)) = O(MB + MB \log_B(M)) = O(MB(1 + \log_B(M))$

Hashing extendible

Sabemos que ambos modelos de hashing funcionan en $O(1)$ en promedio, y que el Hashing Extendible funciona bien en el caso $N = O(MB)$, que es justamente la situación de la secuencia dada. Aún más, sabemos que por la forma en que dicho modelo maneja la información (índices del árbol de búsqueda en memoria principal y páginas en memoria secundaria), la inserción es prácticamente inmediata, pues basta calcular el hash del elemento (en memoria principal) y ver (según los bits del resultado) a qué página pertenece (la cual tenemos indexada en memoria principal). Una inserción cuesta entre 2 y 3 I/O: una lectura (traer la página) y escritura (reescribirla con el nuevo elemento) en el buen caso (no rebalse), y una lectura y dos escrituras (escribir las dos páginas resultantes de la división de la página rebalsada) de lo contrario. Esto de por sí ya es bueno, pues podemos esperar que la mayoría de las inserciones recaigan en solo 2 I/O, asumiendo que escogimos un buen tamaño para las páginas. Sin embargo, esto puede ser aún mejor si consideramos que conocemos con exactitud la cantidad de llaves a insertar, pues podemos decidir de antemano guardar un dato por página de disco - lo cual es factible pues nos mantenemos en $N = O(MB)$ y con ello la inserción de cada llave tomaría siempre 2 I/O, lo cual se reduciría a $MB \cdot O(1) = O(MB)$ para $N = MB$, con exactamente 2 MB I/O.

Hashing lineal

En el caso del hashing lineal, la ventaja de este modelo frente a sus pares dependerá de una variante sumamente importante y para la cual no se indica al respecto: si podemos *asumir que el elemento no está o si debemos verificarlo*. Si podemos realizar dicha asunción, bastará un acceso para insertar el elemento, con un orden $O(1)$. Sin embargo, si debemos verificar la presencia del elemento, nos veremos obligados a traer cada una de las listas indexadas con el mismo hash, lo cual aumentará la inserción - en promedio - a $1 + L$ accesos, con $\mathcal{L}$ el largo promedio de las listas, suponiendo aleatoriedad en los últimos t bits de la función de hashing. Con lo anterior, tenemos que, en el buen caso, podemos esperar un comportamiento muy similar al de hashing extendible, con MB operaciones $O(1) = O(MB)$, y una cantidad exacta de 2 MB I/O. En el mal caso, por otro lado, el comportamiento dependerá del valor del treshold α escogido y del valor inicial de t, sin embargo, será siempre peor. Si bien esta aseveración

[1] Página 37 del apunte del curso

puede sonar un tanto errática dado que hashing lineal debiera mostrar ventajas sobre hashing extendible, dichas ventajas vienen de la búsqueda y del nulo uso de memoria principal (lo cual suprime la restricción de $N = O(MB)$), mas no aporta en cuanto al costo en I/O de insertar.

Conclusión

En general, podemos esperar que para una secuencia como la entregada, el orden de estructuras de datos (de más conveniente a menos conveniente) sea Hashing extendible $\geqslant$ Hashing Lineal $>$Árbol B. La igualdad entre los modelos de hashing vendrá determinada por la posibilidad de conocer la cantidad exacta de llaves a insertar (para el hashing extendible) y la no-necesidad de verificar la presencia previa de una llave en una página (para el hashing lineal). Además, es importante mencionar que la última desigualdad no será siempre cierta: si el valor de α (treshold de ocupación) se escoge de manera poco sabia, se puede llegar a un modelo en que cada inserción implica una expansión, y en dicho caso la desigualdad se invertiría. Sin embargo, este caso sería bastante absurdo.