Isai Daniel Chacón Silva - 201912015

Nicolás Aparicio Claros - 201911357

Estructuras de datos y algoritmos

Laboratorio 6: Mecanismos de Colisión

Paso 1: Responder preguntas preparatorias

a) Teniendo en cuenta cada uno de los requerimientos ¿Cuántos índices implementaría en el Reto? y ¿Por qué?

En principio, el número de índices dependerá de la información necesaria para dar solución a los requerimientos. De acuerdo con lo anterior, se tiene pensado el uso de 5 índices los cuales corresponden a:

Un índice para listar cronológicamente los artistas con el fin de cumplir el requerimiento 1, en donde el *key* vendría dado por el *BeginDate* , y como *value* una TAD lista de artistas para el año en cuestión, con lo que se puede verificar si el artista nació dentro de un rango de años dados por parámetro por el usuario simplemente revisando las llaves.

De manera análoga, se utilizaría un índice para listar cronológicamente las obras de arte según su fecha de adquisición con el fin de cumplir el requerimiento 2, en donde el *key* vendría dado por el *DateAcquired* y como *value* se tendría una TAD lista de obras de arte para el año en cuestión.

Para el requerimiento 3, se necesitaría un índice nuevo con el nombre del artista *(DisplayName)* como llave y sus respectivas obras de arte como una TAD lista. Así pues, sería necesario que cada obra de arte tuviera una referencia al medio o técnica con la que fue realizada.

Por su parte, para el requerimiento 4, dado que la entrada son todas las obras del museo, se podría cargar un índice cuya llave sea la nacionalidad del creador de la obra*(Nationality),* y como *value*, una TAD lista que posea todas las obras asociadas a dicha nacionalidad, no solo de un artista sino de todos los de un mismo país.

Por último, en el quinto requerimiento, se implementaría un nuevo índice con llaves por cada Departamento (*Department*), en donde cada uno posea una TAD lista como *value* con todas las obras de dicho departamento.

En síntesis, se observa que la justificación de usar 5 índices es válida ya que cada uno responde a un requerimiento del reto en específico. Así, es importante asegurarse que cada *value* posea toda la información necesaria ya sea de la obra o del artista para llevar a cabo las tareas de manera exitosa con todos los *prints* requeridos.

b) Según los índices propuestos ¿en qué caso usaría **Linear Probing** o **Separate Chaining** en estos índices? y ¿Por qué?

De acuerdo con los índices que se van a crear para cumplir los requerimientos del reto 2 y las pruebas de eficiencia realizadas posteriormente en este informe, consideramos que el mejor método de colisiones a implementar corresponde a **Linear Probing**, ya que este algoritmo funciona muy bien cuando el número de datos no es tan grande dado que se podría requerir duplicar el tamaño de la tabla y realizar algoritmos de re-hashing. Además, en la práctica,

Linear Probing, suele ser significativamente más rápido que el **encadenamiento** debido a la localidad de referencia, ya que los accesos realizados por **Linear Probing** tienen a estar más cercanos en memoria que aquellos de **Separate Chaining** [1].

Como estamos trabajando con índices para cada una de las categorías de la base de datos, se espera que, en el peor de los casos, el número de datos o la pareja llave valor sea igual a al número de datos que se tiene en el archivo *.csv* tanto para artworks como para artists. De lo contrario, el número de datos siempre será menor si la categoría no es única por obra o artista (e.g. *Department*). Así pues, la única categoría con características relativamente únicas y poco repetidas sería la de *DisplayName*, por lo que solo para este utilizaremos **Separate Chaining**, de modo que no incurramos en costos computacionales excesivos de RAM.

c) Dado el número de elementos de los archivos MoMA, ¿Cuál sería el factor de carga para estos índices según su mecanismo de colisión?

El número total de artistas que se encuentra en el archivo *"Artists-utf8-large.csv"* es de 15,223. Por otro lado, el número de obras de arte en el archivo *"Artworks-utf8-large.csv"* corresponde a 138,150.

De acuerdo con lo anterior y considerando el mecanismo de **Linear Probing**, como se mencionó anteriormente, con el fin de minimizar las colisiones y a priori no considerar limitaciones por espacio en memoria, se utilizará un factor de carga de 0.5 para todos los requerimientos excepto el 3, así como se mencionó anteriormente. Para el caso de los nombres, se utilizará por tanto un factor de carga de 8 aproximadamente con **Separate Chaining**, según lo encontrado en la experimentación de este laboratorio.

Paso 2: Implementar modificaciones en el Reto No. 2

Tabla 1. Tiempos de carga dependiendo del método de colisiones, así como del factor de carga.

Método de colisiones	Factor carga	Tiempo
Chaining	4	828.125 ms
Probing	0.5	812.500 ms
Chaining	2	828.125 ms
Chaining	8	796.875 ms
Probing	0.2	796.875 ms
Probing	0.8	843.75 ms

d) ¿Qué diferencias en el tiempo de ejecución notan al ejecutar la cargar los datos al cambiar la configuración de Linear Probing a Separate Chaining?

Principalmente, como se puede observar en la Tabla. 1, entre los métodos de colisiones, el que presentó un menor tiempo de carga de los datos fue el método de **Linear Probing**, lo cual tiene sentido como se menciona en el inciso b). En cuanto a la diferencia de **Linear Probing** dependiendo de un factor de carga dado, se puede concluir que entre mayor sea el factor de carga, mayor será el tiempo de ejecución. Lo cual es inverso a lo que sucede en **Separate Chaining**, en donde un mayor factor de carga resulta en un menor tiempo de carga. Finalmente, vale aclarar que estos datos fueron medidos para el requerimiento de las nacionalidades y la carga de los artistas con mayor presencia en la base de datos, es decir *American.*

e) ¿Qué configuración de ADT Map escogería para el índice de técnicas o medios?, especifique el mecanismo de colisión, el factor de carga y el numero inicial de elementos.

Para el ADT Map del índice de técnicas utilizaríamos el mecanismo de colisión **Linear Probing** con factor de carga de 0.5 y número inicial de elementos de sería de 138,157 que es el siguiente número primo para 138,150 para asegurarnos que la tabla sea lo suficientemente grande para almacenar todas las obras.

f) ¿Qué configuración de ADT Map escogería para el índice de nacionalidades?, especifique el mecanismo de colisión, el factor de carga y el numero inicial de elementos.

Para el ADT Map del índice de nacionalidades utilizaríamos el mecanismo de colisión **Linear Probing** con factor de carga de 0.5 y número inicial de elementos de 15,227 que es el siguiente número primo para 15,223 para asegurarnos que la tabla sea lo suficientemente grande para almacenar a los artistas.

Referencias

[1] W. lists?, A. Adil and B. Barker, "Why do we use linear probing in hash tables when there is separate chaining linked with lists?", *Stack Overflow*, 2021. [Online]. Available: https://stackoverflow.com/questions/23821764/why-do-we-use-linear-probing-in-hash-tables-when-there-is-separate-chaining-link#:~:text=Separate%20chaining%20%231%20clearly%20uses%20more%20memory%20than,additional%202%20pointers%20floating%20around%20for%20every%20element. [Accessed: 13- Oct- 2021].