Universidad Simón Bolívar
Departamento de Computación y Tecnología de la Información
CI2692 - Laboratorio de Algoritmos y Estructuras de Datos II
Trimestre Abril-Julio 2023

TAD Diccionario y tablas de hash

1. Introducción

El objetivo de este laboratorio es presentar una especificación del TAD Diccionario y realizar dos implementaciones concretas basadas en tablas de hash. La primera implementación consiste en una tabla de hash basada basada en encadenamiento. La segunda implementación es una tabla de hash basada en direccionamiento abierto, en específico, la cuco hashing. Para garantizar la consistencia entre la especificación del TAD Diccionario y las implementaciones, para cada una de las implementaciones se va a definir la *relación de acoplamiento* entre el TAD Diccionario y las tablas de hash.

2. TAD Diccionario

La Figura 1 muestra una especificación del TAD Diccionario, basada en la presentada en [3] y en la clase de teoría de Algoritmos y Estructuras II.

3. Tabla de hash basada en encadenamiento

Se quiere que realice una primera implementación del TAD Diccionario por medio de la estructura de datos tabla de hash, usando el método de encadenamiento para la resolución de colisiones. Usted debe implementar una tabla de hash en donde las claves son de tipo entero, y el valor a almacenar que esta asociado a cada clave, es un elemento de tipo String. El encadenamiento debe ser llevado a cabo mediante una lista doblemente enlazada. La Figura 2 muestra un ejemplo de la tabla de hash, la cual es semejante a la que se presenta en el capítulo 11 de [1]. Como en la tabla de hash tenemos claves asociadas a valores, entonces se puede establecer la siguiente *relación de acoplamiento* entre la tabla de hash basada en encadenamiento y el TAD Diccionario:

$$\begin{aligned} conocidos = \{&claves \mid \text{Las } claves \text{ corresponden a las claves en la tabla de hash}\} \\ &\wedge \\ tabla = \{&(clave, valor) \mid \text{Pares de claves y valores asociados en la tabla de hash}\} \end{aligned}$$

Debe implementar una tabla de hash **dinámica**, es decir, el tamaño de la tabla crece si la tabla tiene un factor de carga mayor o igual a un límite establecido. Para este laboratorio, el tamaño inicial de la tabla es de siete casillas. Si el factor de carga llega a ser igual o mayor a 0.7 durante las operaciones de la tabla, entonces tamaño n de la tabla se incrementa aplicando la función $incr(n) = int((n + 16) * 3/2)$, donde int es una función que redondea un número real a un número entero. Esta operación se conoce como *rehashing*.

La lista doblemente enlazada estará constituida por elementos de tipo `HashTableEntry`. En la figura 2 podemos observar que el tipo `HashTableEntry` debe contener al menos cuatro campos: uno para la clave, uno para el valor tipo String y dos para las referencias a otros dos elementos de tipo `HashTableEntry`. El tipo `HashTableEntry` debe ser implementado

Especificación $\mathbb{A}$ del TAD Diccionario

Modelo de Representación

var $conocidas$: set of $T0$

var $tabla$: $T0 \nrightarrow T1$

Invariante de Representación

$conocidas = \text{dom}(tabla)$

Operaciones

proc *crear* (**out** d : Diccionario)

{ **Pre:** true }

{ **Post:** $d.conocidos = \varnothing \quad \wedge \quad d.tabla = \varnothing$ }

proc *agregar* (**in-out** d : Diccionario, **in** $clave$: $T0$, **in** $valor$: $T1$)

{ **Pre:** $clave \notin d.conocidas$ }

{ **Post:** $d.conocidas = d_0.conocidas \cup \{clave\} \ \wedge$
$d.tabla = d_0.tabla \cup \{(clave, valor)\}$ }

proc *eliminar* (**in-out** d : Diccionario, **in** $clave$: $T0$)

{ **Pre:** $clave \in d.conocidas$ }

{ **Post:** $d.conocidas = d_0.conocidas - \{clave\} \ \wedge$
$d.tabla = d_0.tabla - \{(clave, d_0.tabla(clave))\}$ }

proc *buscar* (**in** d : Diccionario, **in** $clave$: $T0$, **out** $valor$: $T1$)

{ **Pre:** $clave \in d.conocidas$ }

{ **Post:** $valor = d.tabla(clave)$ }

proc *existe* (**in** d : Diccionario, **in** $clave$: $T0$, **out** r : Boolean)

{ **Pre:** true }

{ **Post:** $r \equiv (clave \in d.conocidas)$ }

proc *toString* (**in** d : Diccionario, **out** s : String)

{ **Pre:** True }

{ **Post:** s = String que es una representación de los pares
$(clave, valor)$ contenidos en $d.tabla$ }

proc *numElementos* (**in** d : Diccionario, **out** n : Int)

{ **Pre:** True }

{ **Post:** $n = \#conocidas$ }

Fin del TAD Diccionario

Figura 1: Especificación del TAD Diccionario

como una clase de Kotlin, cuyo constructor recibe una clave de tipo entero y un valor de tipo String.

La función de hash que debe ser usada es el método de la división que se explicó en el curso de teoría de Algoritmos y Estructuras II. De esta implementación son al menos tres los archivos que debe entregar:

`HashTableEntry.kt`: Contiene una clase con la implementación del tipo de datos `HashTableEntry`.

`CircularList.kt`: En este archivo debe estar implementado la lista doblemente enlazada con la clase `CircularList`. La lista debe contener elementos de tipo `HashTableEntry`.

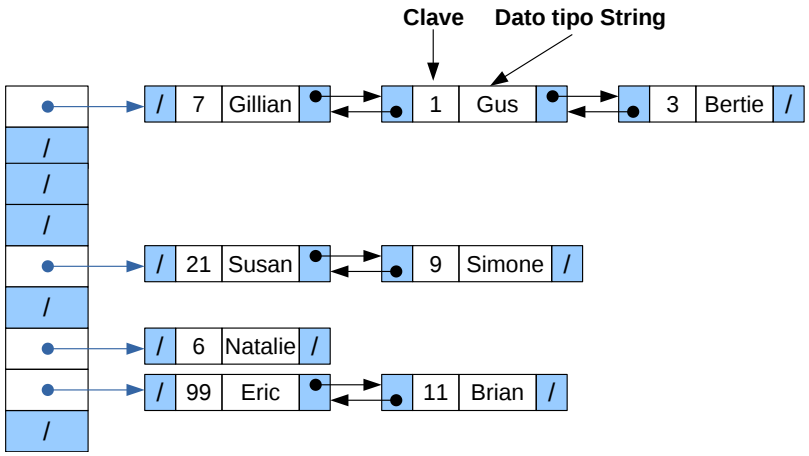

Figura 2: Ejemplo de la tabla de hash donde el encadenamiento es realizado con una lista doblemente enlazada. Las claves son elementos de tipo entero, y el valor asociado a cada clave es de tipo String.

`HashTableChaining.kt`: En este archivo se implementa el tipo de datos tabla de hash basada en encadenamiento, en la clase `HashTableChaining`. La clase debe hacer uso de los tipo de datos `HashTableEntry` y `CircularList`.

4. Cuco hashing

La segunda implementación concreta del TAD Diccionario esta basada en la estructura de datos tabla de hash que usa el método de cuco hashing [2] para la resolución de colisiones. Debe implementar una tabla que usa cuco hashing en donde las claves son de tipo entero, y el valor que esta asociado a cada clave es un elemento de tipo String. La Figura 3 muestra un ejemplo de la tabla de hash.

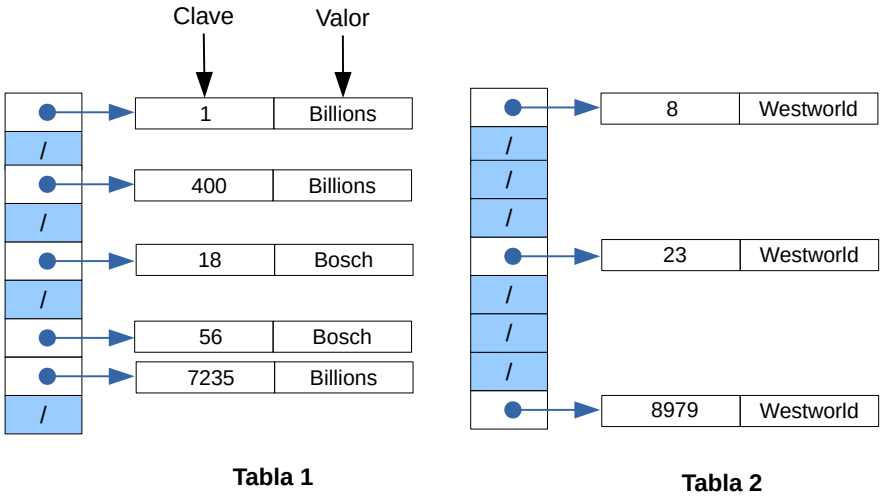

Figura 3: Ejemplo de la tabla de hash usando cuco hashing. Las claves son elementos de tipo entero, y el valor asociado a cada clave es de tipo String.

Como en el cuco hashing tenemos claves asociadas a valores, entonces se puede esta-

blecer la siguiente *relación de acoplamiento* con el TAD Diccionario:

$$\begin{aligned} \textit{conocidos} &= \{\textit{claves} \mid \text{Las } \textit{claves} \text{ corresponden a las claves en la cuco hashing}\} \\ &\wedge \\ \textit{tabla} &= \{(\textit{clave}, \textit{valor}) \mid \text{Pares de claves y valores asociados en la cuco hashing}\} \end{aligned}$$

La cuco hashing es dinámica. En este caso, se debe hacer *rehashing* cuando el factor de carga de la tabla es igual o mayor a 0.7. De debe usar la misma función indicada para la tabla de hash basada en encadenamiento, para el cálculo de tamaño de las nuevas tablas cuando se hace *rehashing*. Cuando se crea una cuco hashing, su tamaño inicial es de siete.

Cada par clave-valor en el cuco hashing, debe estar contenido en una estructura de datos llamada `CuckooHashTableEntry`. El tipo `CuckooHashTableEntry` debe poseer dos campos: el primero se llama clave, de tipo entero, y el segundo se llama valor de tipo String. El tipo `CuckooHashTableEntry` debe ser implementado como una clase de Kotlin, en el archivo `CuckooHashTableEntry.kt`. El constructor de la clase `CuckooHashTableEntry` recibe como entrada un clave, de tipo entero y un valor asociado con la clave, que es de tipo String.

La tabla de hash usando cuco hashing debe ser implementada como una clase de Kotlin, llamada `CuckooHashTable`. El desarrollo de tabla debe tener como base el pseudocódigo presentado en la clase del curso de teoría de Algoritmos y Estructuras de Datos II. La tabla almacenará elementos de tipo `CuckooHashTableEntry`. La `CuckooHashTable` debe poseer dos arreglos que contienen a los elementos de tipo `CuckooHashTableEntry`.

El método de cuco hashing hace uso de dos funciones de hash h_1 y h_2. La función h_1 corresponde al método de la división que se explicó en el curso de teoría de Algoritmos y Estructuras II. Para la función h_2 se debe usar el método de la multiplicación, también dado en clase de teoría, usando como constante A el valor sugerido por Knuth [1] de $A = 0{,}6180339887$.

Su implementación de la cuco hashing debe contener, al menos, los siguientes archivos:

`CuckooHashTableEntry.kt:` Implementación de la clase `CuckooHashTableEntry`.

`CuckooHashTable.kt:` Implementación de la la cuco hashing en la clase `CuckooHashTable`.

4.1. Programa de pruebas de las tablas de hash

Se desea hacer una comparación experimental del rendimiento de la tabla de hash basada en cuco hashing y la tabla hash basada en encadenamiento. El objetivo es comparar el rendimiento de ambas tablas, bajo un conjunto de operaciones, y usando diferente cantidad de datos.

Debe realizar un programa cliente llamado `Main.kt`, el cual compara las dos tablas de hash mediante la siguiente prueba:

1. Se crea un arreglo, el cual va a almacenar n números enteros generados aleatoriamente que se encuentran en el intervalo $\left[0, \frac{n}{3}\right]$. Este arreglo contiene las claves que van a ser agregadas en las tablas. Es decir, las dos tablas de hash deben ser probadas sobre la misma secuencia de claves.

2. El valor asociado a cada clave se obtiene convirtiendo en String cada una de las claves, de esa manera se debe crear un arreglo de pares $(\textit{clave}, \textit{valor})$ de elementos a insertar en la tabla de hash.

3. Para cada uno de los elementos del arreglo de pares $(clave, valor)$, se **busca** si el elemento existe en la tabla de hash, entonces si existe el elemento se **elimina**, de lo contrario se **agrega**.

4. Se debe medir el tiempo usado por la tabla de hash para procesar el arreglo de pares $(clave, valor)$, es decir, los pasos 2 y 3 indicados anteriormente. No se debe tomar en cuenta el tiempo usado para la creación de este arreglo.

El procedimiento descrito anteriormente debe ser aplicado a las dos tablas de hash a comparar. El cliente `Main.kt` se debe ejecutar desde un script llamado `runTestHashTables.sh`, el cual se ejecuta con la siguiente línea de comando:

```
>./runTestHashTables.sh <n>
```

Donde n es el número de elementos que contiene el arreglo de pares $(clave, valor)$, que va a ser procesado por las tablas de hash. Como resultado el programa debe mostrar por la salida estándar, el tiempo de ejecución usado por cada tabla.

La entrega del programa de pruebas, junto con el resto de los códigos fuentes, debe incluir un archivo Makefile que compila todos los archivos Kotlin.

4.2. Evaluación experimental y reporte

Se debe realizar un estudio sobre el rendimiento de las tablas de hash. Se debe ejecutar el programa `runTestHashTables.sh` 5 veces con un valor de 5.000.000 y se debe obtener el tiempo promedio y la desviación estándar. De esta forma podemos comparar el tiempo de ejecución de ambas tablas de hash. Debe realizar un breve reporte con los resultados obtenidos y su análisis. El reporte, **que debe estar en formato PDF**, debe contener los siguientes elementos:

1. La identificación de los estudiantes autores del trabajo.

2. Los datos de la plataforma donde se ejecutaron los algoritmos: sistema de operación, modelo de CPU, cantidad de memoria RAM del computador, versión del compilador Kotlin y versión de la JVM utilizada.

3. Una tabla con el tiempo promedio de cada tabla de hash y su desviación estándar.

4. Un gráfico del tiempo promedio indicado en la tabla anterior, en donde se pueda observar la magnitud de la diferencia en los tiempos de las tablas de hash.

5. Un análisis de los resultados obtenidos.

5. Condiciones de entrega

Los códigos del laboratorio, el informe y la declaración de autenticidad firmada, deben estar contenidas en un archivo comprimido, con formato *tar.xz*, llamado *LabSem8_X_Y.tar.xz*, donde X y Y, son los números de carné de los estudiantes. La entrega del archivo *LabSem8_X_Y.tar.xz*, debe hacerse por la plataforma Classroom, antes de las 11:59 pm del día domingo 2 de julio de 2023.

Referencias

[1] Cormen, T., Leiserson, C., Rivest, R., and Stein, C. *Introduction to algorithms*, 3rd ed. MIT press, 2009.

[2] Pagh, R., and Rodler, F. F. Cuckoo hashing. *Journal of Algorithms 51*, 2 (2004), 122–144.

[3] Ravelo, J. Especificación e implementación de tipos abstractos de datos. `http://ldc.usb.ve/~jravelo/docencia/algoritmos/material/tads.pdf`, 2012.

Guillermo Palma / gvpalma@usb.ve / Junio de 2023