Juan Menduiña

Trabajo Práctico N° 5:
Hashing.

Ejercicio 1.

Definir el concepto de hashing (o dispersión). ¿Cómo se relaciona este concepto con archivos?

Hashing (o dispersión) es una técnica utilizada para transformar una clave (como una palabra, un número o un conjunto de datos) en una dirección o índice dentro de una estructura de datos, generalmente una tabla o un archivo. Esta transformación se realiza mediante una función de dispersión (función *hash*), que toma la clave como entrada y produce un número entero, la dirección *hash* o el código *hash*, que se utiliza como índice para acceder, rápidamente, a los datos.

Hashing (o dispersión) es un método de asignación de claves a posiciones en una tabla o archivo mediante una función de dispersión (función *hash*), con el objetivo de facilitar el almacenamiento y recuperación eficiente de información.

- Técnica para generar una direción base única para una llave dada. La dispersión se usa cuando se requiere acceso rápido a una llave.
- Técnica que convierte la llave del registro en un número aleatorio, el que sirve, después, para determinar dónde se almacena el registro.
- Técnica de almacenamiento y recuperación que usa una función de *hash* para mapear registros en dirección de almacenamiento.

Ejercicio 2.

Explicar el concepto de función de dispersión. Enumerar, al menos, tres funciones de dispersión y explicar, brevemente, cómo funciona cada una.

Una función de dispersión (función *hash*) es un procedimiento que toma una clave (como una palabra, un número o un conjunto de datos) como entrada y la convierte en un número entero, la dirección *hash* o el código *hash*, que se utiliza como índice en una tabla o en un archivo para acceder, rápidamente, a los datos.

El objetivo principal es asignar claves a posiciones de manera uniforme para facilitar el acceso rápido a los datos, minimizando las colisiones (casos en los que diferentes claves generan el mismo índice).

Tres funciones de dispersión son:

- Método de la división: Toma la clave *k* (por ejemplo, un número entero) y la divide por un número *m* (preferentemente primo), y se usa el resto como la dirección *hash*.
  hash(k)= k mod m.
- Método de la multiplicación: Toma la clave *k* (por ejemplo, un númeor entero), la multiplica por un número fraccionario (A= 0,618), toma la parte decimal, y la multiplica por m para obtener un índice.
  hash(k)= floor(m * ((k * A) mod 1)).
- Método de la suma de caracteres (para cadenas): Suma los valores ASCII (o numéricos) de todos los caracteres de la clave y, luego, aplica una operación como mod m.

Ejercicio 3.

Explicar los conceptos de sinónimo, colisión y desborde (overflow). ¿Qué condición es necesaria en el archivo directo para que pueda ocurrir una colisión y no un desborde?

Sinónimo: Es una clave diferente que, al ser procesada por la función de dispersión, produce la misma dirección *hash* que otra clave. Es decir, dos claves distintas generan el mismo índice.

Colisión: Situación en la que un registro es asignado a una dirección que está utilizada por otro registro. Se produce cuando dos o más claves se asignan a la misma posición en la tabla o en el archivo.

Desborde (*overflow*): Situación en la que un registro es asignado a una dirección que está utilizada por otro registro y no queda espacio para este nuevo. Se produce cuando no hay espacio disponible en la posición calculada por la función de dispersión ni en las posiciones alternativas previstas para manejar colisiones.

La condición que es necesaria en el archivo directo para que pueda ocurrir una colisión y no un desborde es que exista, al menos, una posición disponible donde se pueda reubicar el sinónimo.

Ejercicio 4.

¿Qué alternativas existen para reducir el número de colisiones (y, por ende, de desbordes) en un archivo organizado mediante la técnica de hashing?

Las alternativas que existen para reducir el número de colisiones (y, por ende, de desbordes) en un archivo organizado mediante la técnica de *hashing* son:

- Algoritmos de dispersión sin colisiones o que éstas nunca produzcan *overflow* (perfectos e imposibles de conseguir).
- Buscar métodos que distribuyan los registros de la forma más aleatoria posible.
- Distribuir pocos registros en muchas direcciones.
- Colocar más de un registro por dirección.

Ejercicio 5.

Explicar, brevemente, qué es la densidad de empaquetamiento. ¿Cuáles son las consecuencias de tener una menor densidad de empaquetamiento en un archivo directo?

La densidad de empaquetamiento es la proporción de espacio del archivo asignado que, en realidad, almacena registros. Las consecuencias de tener una menor densidad de empaquetamiento en un archivo directo es menos *overflow* y más desperdicio de espacio.

Ejercicio 6.

Explicar, brevemente, cómo funcionan las siguientes técnicas de resolución de desbordes que se pueden utilizar en hashing estático: saturación progresiva, saturación encadenada, saturación progresiva encadenada con área de desborde separada y dispersión doble.

Saturación progresiva: Cuando se completa el nodo, se busca el próximo hasta encontrar uno libre.

Saturación encadenada: Es similar a la saturación progresiva, pero los registros de saturación se encadenan y no ocupan, necesariamente, posiciones contiguas.

Saturación progresiva encadenada con área de desborde separada: No utiliza nodos de direcciones para los *overflows*, éstos van a nodos especiales.

Dispersión doble: Las técnicas de saturación tienden a agrupar en zonas contiguas y generan búsquedas largas cuando la densidad tiende a uno. La solución de esta técnica de resolución de colisiones es almacenar los registros de *overflow* en zonas no relacionadas, aplicándoles una segunda función *hash* a la llave para producir un número entero, el cual se suma a la dirección original tantas veces como sea necesario hasta encontrar una dirección con espacio.

Ejercicio 7.

Para las siguientes claves, realizar el proceso de dispersión mediante el método de hashing extensible, sabiendo que cada nodo tiene capacidad para dos registros. El número natural indica el orden de llegada de las operaciones. Se debe mostrar el estado del archivo para cada operación. Justificar, brevemente, ante colisión y desborde los pasos que se realizan.

1	*+ Darín*	*00111111*	*2*	*+ Alterio*	*11110100*
3	*+ Sbaraglia*	*10100101*	*4*	*+ De la Serna*	*01010111*
5	*+ Altavista*	*01101011*	*6*	*+ Grandinetti*	*10101010*
7	*- Altavista*	*01101011*	*8*	*- Sbaraglia*	*10100101*

Ejercicio 8.

Realizar el proceso de dispersión mediante el método de hashing extensible, sabiendo que cada registro tiene capacidad para dos claves. El número natural indica el orden de llegada de las operaciones. Se debe mostrar el estado del archivo para cada operación. Justificar, brevemente, ante colisión y desborde los pasos que se realizan.

1	*+ Buenos Aires*	*...1001*	*2*	*+ San Juan*	*...0100*
3	*+ Entre Ríos*	*...1110*	*4*	*+ Corrientes*	*...0010*
5	*+ San Luis*	*...0101*	*6*	*+ Tucumán*	*...0111*
7	*+ Río Negro*	*...0011*	*8*	*+ Jujuy*	*...1111*
9	*+ Salta*	*...1010*	*10*	*- Río Negro*	*...0011*

Ejercicio 9.

Para las siguientes claves, realizar el proceso de dispersión mediante el método de hashing extensible, sabiendo que cada nodo tiene capacidad para dos registros. El número natural indica el orden de llegada de las operaciones. Se debe mostrar el estado del archivo para cada operación. Justificar, brevemente, ante colisión y desborde los pasos que se realizan.

1	*+ Tristana*	*11110010*	*2*	*+ Jarvan IV*	*00111010*
3	*+ Teemo*	*01010100*	*4*	*+ Annie*	*10100101*
5	*+ Ryze*	*10101110*	*6*	*+ Morgana*	*01101011*
7	*+ Garen*	*11001011*	*8*	*- Teemo*	*01010100*

Ejercicio 10.

Para las siguientes claves, realizar el proceso de dispersión mediante el método de hashing extensible, sabiendo que cada nodo tiene capacidad para dos registros. El número natural indica el orden de llegada de las operaciones. Se debe mostrar el estado del archivo para cada operación. Justificar, brevemente, ante colisión y desborde los pasos que se realizan.

1	*+ Guillermo.B*	*01100011*	*2*	*+ Gómez*	*00000001*
3	*+ Gustavo.B*	*01010110*	*4*	*+ Sosa*	*11110100*
5	*+ Enría*	*00110101*	*6*	*+ Guli*	*00101000*
7	*- Gustavo.B*	*01010110*	*8*	*- Sosa*	*11110100*

Ejercicio 11.

Para las siguientes claves, realizar el proceso de dispersión mediante el método de hashing extensible, sabiendo que cada nodo tiene capacidad para dos registros. El número natural indica el orden de llegada de las operaciones. Se debe mostrar el estado del archivo para cada operación. Justificar, brevemente, ante colisión y desborde los pasos que se realizan.

1	*+ Mansilla*	*01100010*	*2*	*+ Cetré*	*10001000*
3	*+ Ascacibar*	*01010111*	*4*	*+ Carrillo*	*11110101*
5	*+ Manyoma*	*00110100*	*6*	*+ Méndez*	*00101001*
7	*+ Alario*	*11000101*	*8*	*- Mansilla*	*01100010*

Ejercicio 12.

Realizar el proceso de dispersión mediante el método de hashing extensible, sabiendo que cada nodo tiene capacidad para dos claves. El número natural indica el orden de llegada de las operaciones. Se deberán explicar los pasos que se realizan en cada operación y dibujar los estados sucesivos correspondientes (inclusive el estado inicial).

1	*+ Aconcagua*	*10100111*	*2*	*+ Kilimanjaro*	*10101010*
3	*+ Mont Blanc*	*00111110*	*4*	*+ Cervino*	*01101111*
5	*+ Etna*	*00110101*	*6*	*+ Chañi*	*11110000*
7	*+ Cho Oyu*	*01011101*	*8*	*+ Vinicunca*	*01011011*
9	*- Chañi*	*11110000*	*10*	*- Cervino*	*01101111*