AED Tema 8. Tablas Hash

Martín González Dios

18 de noviembre de 2024

El problema de búsqueda es uno de los problemas más importantes y repetidos. Se quiere buscar un elemento entre un conjunto. En función de si el conjunto está ordenado o no, se puede realizar una **búsqueda lineal** (recorre todos los elementos del primero al último buscando el elemento buscado, por lo que el orden de complejidad es de O(n), una **búsqueda binaria** (solo posible para arrays ordenados, pero reduciendo el orden de complejidad a O (log_2(n)), o una **búsqueda por clave** (también llamada direccionamiento directo) (con la que se accede a cualquier elemento por medio de su clave, teniendo una complejidad O(1)).

Siguiendo este último método, se han implementado las **tablas hash**, que proporcionan un tiempo de búsqueda constante, independiente del número de elementos y aunque éstos no estén ordenados (al acceder a cada uno a través de su clave). Además la inserción y el borrado son también constantes. (Las **claves**, elemento identificativo de los datos que queremos ordenar, **pueden ser número o cadenas de caracteres**)

Esto se consigue obteniendo la posición de cada elemento por medio de una fórmula matemática llamada **función hash**. Ejecutar dicha fórmula para obtener la posición se hace de manera constante, independientemente del tamaño de la tabla, y una vez se sabe la posición, se puede acceder al elemento también de forma constante. Estas son claras **ventajas** sobre el resto de estructuras de datos para almacenar conjuntos de datos.

Desventajas:

- Los datos **se almacenan en un array** (estructura estática) por lo que debe fijarse un tamaño máximo desde el principio, que no se podrá sobrepasar. Si el tamaño es demasiado grande, se desperdiciará mucho espacio.
- **No hay forma de recorrer los elementos directamente**, ya que no tienen porque encontrarse en posiciones consecutivas, pudiendo haber varias posiciones vacías de array entre cada par de elementos.
- Tampoco se pueden almacenar los elementos ordenados, ya que su posición viene dada por el resultado devuelto por la fórmula matemática.
- Se puede dar el caso en el que la función hash asigne a un elemento una posición ya ocupada en el array. A esto se le conoce como **colisión**, y es necesario plantear una solución para redirigir el elemento a otra posición, y tener forma de localizarlo después.

1. Funciones Hash

Existen diversas funciones hash con sus ventajas e inconvenientes.

1.1. Función hash por módulo

Es muy simple de entender y de implementar: **h(K) = K mod N** , siendo **h** la función hash, **K** la clave del dato, **mod** el operando módulo (resto de la división), y **N** el tamaño de la tabla (tamaño del array en el que se insertan los datos).

Sin embargo, a pesar de su sencillez, tiene el gran inconveniente de que ciertos valores de N pueden producir muchas colisiones para muchas claves (aquellos que tengan una factorizaci´on con muchos factores). Una forma de solucionarlo es seleccionar para el valor de N un número primo. Solo sirve para **claves** formadas por **números enteros**.

1.2. Función hash cuadrado

Se eleva la clave al cuadrado y del resultado se seleccionan los dígitos centrales: **h(K) = dígitos centrales** (K^2). Para determinar cuantos dígitos centrales se cogen, se puede utilizar una función matemática en función del tamaño de la tabla, por ejemplo: dígitos centrales = log(N).

En este caso el número de colisiones no va a depender de los factores del valor del tamaño de la tabla tan directamente, sino que toma un carácter un poco más aleatorio, aunque por lo general se reducen para tamaños de tabla grandes. Solo sirve para **claves** formadas por **número enteros**.

1.3. Función hash por plegamiento

Se divide la clave en partes con el mismo número de dígitos (salvo la última, que puede tener menos si la división no es exacta) y se realiza una operación de suma o multiplicación sobre las partes, y sobre la solución, se queda con las cifras menos significativas.

El número de cifras para realizar las divisiones en partes y el número de cifras menos significativas que se escogen de la solución dependen también del tamaño de la tabla. Solo sirve para **claves** formadas por **números enteros**.

1.4. Función hash por el método de la división

Para claves formadas por **cadenas de caracteres**, una opción es sumar los valores ascii de cada uno de los caracteres que forman las cadena y al valor entero obtenido calcularle la función hash por módulo.

Como en el caso de la función hash por módulo, uno de los **inconvenientes** puede ser el valor del tamaño de la tabla escogido. Además, también presenta problemas con valores de tamaño muy grandes si las cadenas tienen todas pocos caracteres, ya que no se llenará gran parte de la tabla, desperdiciando espacio mientras se pueden estar provocando colisiones en un número reducido de posiciones.

1.5. Función hash por suma ponderada

Otro método para calcular la posición cuando las claves están formadas por **cadenas de caracteres** es obtener los valores ascii de cada carácter, multiplicarlo por un número en función de su posición en la cadena y luego sumar los valores. Para evitar obtener valores no válidos, hay que realizar el módulo del valor del tamaño de la tabla para cada valor de cada carácter.

Como el código ascii recoge 256 caracteres, usar una numeración en base 256 puede ser una buena opción. En ese caso, el último carácter de la cadena tendrá simplemente su valor ascii, el siguiente tendrá su valor ascii más 256, el posterior tendrá su valor ascii más 256*2, y así sucesivamente, sumando después todos esos valores.

2. Resolución de colisiones

Cuando una clave es asignada a una posición ya ocupada se produce una **colisión**, que es necesario resolver para poder asignar una posición válida a dicha clave y poder localizarla después.

Una forma de **evitar las colisiones** primeramente es seleccionar una **buena función hash**. Sin embargo, siempre hay colisiones que son inevitables. Para resolver estas existen dos alternativas: **recolocación** y **encadenamiento**.

2.1. Recolocación

Si la posición asignada al elemento está ya ocupada, se busca otra posición.

Hay 3 tipos de recolocaciones:

- **Recolocación simple**: **si la posición asignada está ocupada se prueba en la siguiente**, y si también está ocupada, en la siguiente, y así hasta llegar a una posición vacía.

  A la hora de **buscar elementos**, se sigue el **mismo proceso**: si no está en la posición correspondiente, se recorren las siguientes hasta encontrarlo, o en el caso de que no esté, hasta recorrer un espacio vacío, o en caso contrario, recorrer toda la tabla.

  Esto supone tener que hacer una distinción entre las posiciones que están vacías porque se ha borrado un elemento y las que lo están porque todavía no se insertó ningún elemento en ellas.

  Desventajas: el coste de la inserción, búsqueda y borrado en el caso de tener que recorrer muchas posiciones, además, se pueden formar grandes bloques de posiciones ocupadas seguidas, lo que repercute directamente en la anterior desventaja.

- **Recolocación lineal**: similar a la recolocación simple, pero en vez de avanzar por las posiciones de una en una, **se avanza de a en a**, siendo a un **número entre 2 y N-1** (con N como valor del tamaño de la tabla)

  Un gran **inconveniente** es que, en función del valor de a, a pesar de haber posiciones vacías, al recorrer la tabla puede no encontrarse ninguna. La forma de solucionarlo es seleccionar un **valor a primo con N**.

  Se conoce como **factor de carga** a la relación entre el número de datos almacenados y el tamaño de la tabla. Cuando este valor se mantiene **por debajo de 1/2**, se tiene una media de colisiones constante.

- **Recolocación cuadrática**: cada vez que se produce una colisión, se busca en la posición resultante de sumar el número de colisión elevado al cuadrado a la posición original. De esta forma se evitan formar bloques de posiciones llenas consecutivas.

  Sin embargo, también es posible no encontrar un posición libre a pesar de que si que las haya, aunque esto solo ocurre cuando la tabla está casi llena. Por ello, es necesario mantener el **factor de carga por debajo de 1/2**, lo que se consigue con la **redispersión**.

2.2. Encadenamiento

Cada posición del vector de datos contiene una lista enlazada, de forma que aunque ya haya elementos en la posición, se puede insertar uno nuevo añadiéndolo al final de la lista sin producir ninguna colisión. Por tanto, la **inserción** se realiza siempre en un **tiempo constante**.

Por otra parte, **la búsqueda y la eliminación** requieren calcular la función hash y posteriormente recorrer la lista hasta encontrar el elemento, lo que supone un **coste lineal** en el peor de los casos.

A mayores de la reducción general en el tiempo de inserción, otra gran **ventaja** es que no existe un número máximo de elementos, ya que las listas pueden crecer tanto como se necesite.

No obstante, si todos los elementos se distribuyen en unas pocas posiciones del array, se acabarán teniendo listas muy largas en ciertas posiciones, mientras otras se mantienen vacías, reduciendo la eficiencia, al asimilarse a trabajar simplemente con arrays.

Para evitar este problema se debe mantener el **factor de carga por debajo de 3/4**, usando la **redispersión**.

3. Redispersión

Cuando en una tabla hash, tanto con recolocación como con encadenamiento, se supera el **factor de carga máximo establecido**, es necesario reducirlo para no empeorar la eficiencia de la tabla hash.

Para ello se realiza la operación de **redispersión**, que **aumenta el tamaño de la tabla**, generalmente al doble, y **recoloca los elementos en sus nuevas posiciones**. Esto supone que sea una **operación costosa**, pero al realizarse pocas veces y mantener la eficiencia de la tabla hash, es **aceptable**.