Martínez Orozco Victor Manuel 20/03/23

Investigacion 05

Acceso alcatorio a un archivo mediante dispersión "Hash"

Esta es una técnica de acceso a archivos que permite acceder a una parte específica de un archivo de manera más eficiente que el acceso secuencial. Es decir, en lugar de leer el archivo desde el principio al final, con el fin de encontrar la información que se necesita, el acceso aleatorio mediante dispersión utiliza una función matemática llamada "Hash" para calcular la ubicación en el archivo donde se encuentra la información deseada.

Es decir, la función hash toma una clave (un valor único que identifica la información) como entrada y devuelve un valor de dispersión que se utiliza para buscar la información en el archivo.
Esta función esta diseñada de forma que el valor de dispersión termina siendo único para dada clave y que, idealmente distribuya los valores de dispersión de manera uniforme en todo el archivo.

Manejo de colisiones

Una colisión ocurre dos claves diferentes tienen el mismo valor de dispersión, lo que hace que apunten a la misma ubicación en el archivo.

Ocurrida esta situación, es necesario encontrar una forma de manejar la colisión y almacenar la información de manera que se pueda acceder de forma eficiente.

Técnicas para manejar las colisiones

- Encadenamiento: Esta técnica consiste en crear una lista enlazada para cada valor de dispersión y agregar los elementos con la misma clave a la lista correspondiente. De manera que cada valor de dispersión apunta a una lista de elementos en lugar de una única ubicación en el archivo.
- Doble dispersión: Esta técnica utiliza una doble función hash para encontrar una ubicación alternativa en el archivo cuando se produce una colisión. Es decir, si se produce una colisión en la ubicación actual original, se utiliza la segunda función hash para encontrar una nueva ubicación en el archivo.

Exploracion lineal

- Esta técnica implica buscar una nueva ubicacion en el archivo al desplazarse hacia adelante en el archivo desde la ubicación original hasta que se encuentra una ubicación libre. Esta tecnica es simple pero puede ser menos eficiente que otras tecnicas, especialmente si hay colisiones. en gran cantidad.

Exploracion cuadrática: Esta tecnica es similar a la lineal pero utilica una función cuadrática para Calcular la nueva ubicación en el archivo. Esto
- ayuda a reducir la probabilidad de colisiones secundarias y los requisitos de rendimiento.

La dispersión: Esta es una técnica que permite los datos en fragmentos y almacenarlos en diferentes ubicaciones, lo que permite una recuperación más rápida de los datos en caso de fallas o errores. La saturación se da cuando la cantidad de datos aumenta en el sistema. Por el hecho de que los
- fragmentos de los datos se almacenan en un número limitado de ubicaciones, provocando aumento en tiempo de acceso a los datos, dando bajo rendimiento.

Martinez Orozco Victor Manuel 25/03/23

Compartimientos...

En la dispersión, los datos son asignados a compartimientos mediante la función propia de la dispersión que sirve como ya se mencionó anteriormente para ubicar en la tabla hash la entrada de datos. Cada compartimiento contiene una lista de bloques que tienen la misma clave del ~~dispersa~~ bloque y esta sirve para ubicar el compartimiento correspondiente.

Lo anterior en pocas palabras, los compartimientos son los espacios en donde se asignan los datos mediante la función de dispersión.

Tipos de llaves (Primarias y secundarias)

- Primarias: Estas son aquellas que se utilizan para calcular la ubicación en un archivo en el sistema de archivos. Mediante la función de dispersión la llave es transformada a una dirección de memoria que indica la ubicación del archivo en disco.

La función de dispersion es siempre determinista, es decir, siempre devuelve la misma dirección de memoria para la misma llave.

Martinez Victor Manuel 25/03/23

- Llaves secundarias: Estas se utilican para acceder a archivos que se ubican en la misma direccion física. Por ejemplo, si varios se encuentran en el mismo bloque o sector del disco, se puede utilizar una llave secundaria para seleccionar el archivo correcto dentro de ese bloque. Es decir, las llaves secundarias se utilizan para realizar una busqueda dentro de una ubicación específica.
- Índice... Este se interpreta entonces como una tabla que contiene una lista de temas (llaves) y el lugar donde pueden ser encontrados (campos de referencia)

De aspectos relevantes se pueden mencionar:

- Un índice permite tener un orden en un archivo sin tener la necesidad de reorganizarlo.
- Permite que las de escritura y lectura sean menos costosas.
- Los índices simples son representados utilizando estructuras simples de arreglos que contienen las llaves y los campos de referencia

Índices secundarios con arreglos

Añadir un registro al archivo significa añadir una entrada al índice secndario. De forma contraria, para remover un registro del archivo se requiere climinar tanto indices primarios como secundarios que referencian la entrada en el índice primario.

Una forma de implementar estos métodos es utilizar un arreglo adicional que contenga los indices de los elementos del arreglo principal en orden ascendente o descendente según el valor del campo que se está indexando.

Índices secundarios con listas ligadas e invertidas

- Con listas invertidas, viendo la poca eficiencia de escribir en el archivo todas las llaves primarias y secundarias con las que estén relacionadas surge la posibilidad de manejar un arreglo de tamaño Fijo con las llaves primarias que correspondan a una llave secundaria. A este manejo se le conoce como listas invertidas ya que una una llave secundaria hace referencia a una llave primaria.
- Con listas ligadas en cambio, se da la opción de separar el archivo de índices secundarios en dos archivos:
  - Uno contiene solo las llaves secundarias y una referencia a la primer llave primaria en otro archivo
  - El otro contiene una estructura similar a una lista ligada.

Índice selectivo: Este es un indice que se crea en una o varias columnas de una tabla de tal manera que solo indexa una porción de los datos de la tabla, por ejemplo, solo aquellos datos que cumplen con una determinada condición, evitando así tener que acceder a toda la tabla y escanearla.

Bibliografía

- Deitel, P. J., Deitel, H. M., & Nieto, T. R. (2002). Cómo programar en C/C++ y Java. Pearson Educación.
- Microsoft Corporation. (2021). Overview of Marshaling in C++. Microsoft Visual C++ Documentation. Recuperado el 29 de abril de 2023, de https://learn.microsoft.com/es-es/cpp/dotnet/overview-of-marshaling-in-cpp?view=msvc-170
- Zhang, L., Wang, K., & Zhang, C. (2014). Research on file random access based on hash algorithm. Journal of Software, 25(8), 1661-1674. https://doi.org/10.13328/j.cnki.jos.004937
- Ouadoudi, Z., & Rebbani, M. (2019). A new hash table based approach for indexing XML documents. International Journal of Computer Science and Information Security, 17(9), 212-219. https://doi.org/10.47611/ijsi.2019.199