

FICHEROS Y BASES DE DATOS (E44)

3º INGENIERÍA EN INFORMÁTICA

Tema 2.

Estructura de un Fichero. Operaciones Básicas.

- 1.- Introducción.
- 2.- Organización: Campos y Registros.
- 3.- Acceso a la Información.
- 4.- Actualización de la Información.
Fragmentación.

(Capítulos 4 y 5 del Folk)

(Capítulo 4 del Elmasri)

INTRODUCCIÓN

Visión Inicial

- Los datos en memoria secundaria aparecen almacenados en ficheros.
- Desde el punto de vista del dispositivo, los ficheros aparecen como una secuencia de bytes.
- Pero los bytes de un fichero representan cierta información.
- Es por esto, que los bytes que constituyen el fichero aparecen agrupados para formar unidades de información.
- Así, en un fichero aparecen un conjunto de entidades caracterizadas por una serie de atributos.
- Las entidades se asocian a Registros, y los atributos a Campos.
- La cuestión se centra en el modo en el que se puede localizar un determinado atributo de una entidad dentro del fichero.
- Una alternativa es conocer de antemano la posición de inicio de cada una de las informaciones.
- Otra posibilidad es marcar de algún modo el principio y el fin de cada información, de modo que no sea necesario prefijar su posición en el fichero.

CAMPOS Y REGISTROS

Definición y Tipos de Campos

- Un Campo se define como la menor unidad de información con significado lógico en un fichero.
- Existen diferentes alternativas para localizar los campos en un fichero, tal y como se muestra.
- Campos de Tamaño Fijo.
- Independientemente del número de bytes de un dato, cada campo ocupa en disco el mismo número de bytes.
- La lectura de la informaciones es muy sencilla, ya que sólo hace falta un pequeño cálculo.
- Para datos con longitud muy variable, puede producir una derroche, bastante importante, de la capacidad del dispositivo.
- Comenzar un Campo con un Indicador de Longitud.
- Este indicador incluye el número de bytes del dato incluido en el campo.
- Suele ser suficiente un indicador de un byte.
- Colocar un Delimitador al Final del Campo.
- Se elige un carácter especial que se escribe a continuación de los bytes del campo.
- La cuestión es la elección del carácter que forma este delimitador, ya que no debe estar en ninguno de los campos del fichero.

CAMPOS Y REGISTROS

Definición y Tipos de Registros

- Un Registro se define como un conjunto de campos que permanecen juntos cuando se observa el fichero con un mayor nivel de abstracción.
- Para poder agrupar los campos de un registro, resulta necesario determinar los límites de un registro, apareciendo diferentes opciones.
- Registros de Tamaño Fijo, a Nivel de Bytes.
- Los campos de longitud fija se suele reunir para formar registros de tamaño fijo.
- Pero puede estar constituido de un número variable de campos de tamaño variable.
- Registros de Tamaño Fijo, a Nivel de Campos.
- Es la forma más sencilla de constituir registros, y tampoco depende del tipo de los campos.
- Comenzar cada Registro con un Indicador de Longitud.
- Similar al equivalente para campos, siendo muy común para registros compuestos por campos de longitud variable.
- Definir un Fichero Índice.
- Se utiliza un Fichero Índice cuyos registros indican el byte de inicio de los registros del fichero original.
- Colocar un Delimitador al final de los Registros.
- La cuestión es la elección del delimitador.

CAMPOS Y REGISTROS

Elección de la Estructura

- Existen diferentes aspectos que influyen en la elección de la estructura de un fichero.
- Una se refiere a las posibilidades del lenguaje de programación utilizado en el desarrollo de la aplicación.
 - Operaciones Predefinidas sobre Ficheros.
 - Modo de Almacenamiento de los Datos.
- También tiene una gran influencia que tipo de operaciones se implemente sobre el fichero.
 - El modo en el que se accede a los datos en el fichero.
 - Aspectos relacionados con las operaciones de actualización como las inserciones, las modificaciones y los borrados.
 - Frecuencia de aparición.
 - Modo de actualización de los datos: En Línea o Fuera de Línea.
- El programador debe de elegir una estructura que facilite el acceso rápido a la información que aparece en el fichero.
- Además, debe posibilitar que la actualización de los datos se realice de modo eficiente, sin influir en exceso en su coste de acceso.
- Lógicamente, una actualización en línea tiene una mayor influencia en el coste de acceso, pero permitir manejar datos más realistas.

ACCESO A LA INFORMACIÓN

Modos de Acceso

- La información almacenada en un fichero, debe de poder recuperarse con un coste de acceso limitado.
- La recuperación de la información puede ser de dos tipos:
 - Global, que recupera toda la información.
 - Selectiva, en donde se desea recuperar una parte de la información del fichero.
- La resolución de ambos tipos de recuperación se puede realizar mediante dos tipos de acceso concretos:
 - Acceso Secuencial, en el que se accede a todos los registros del fichero.
 - Acceso Directo, que permite acceder a un registro dentro de un fichero.
- Mediante la utilización de la primera opción, el coste de los dos tipos de recuperación es equivalente, (n) , siendo n el número de registros del fichero.
- Este coste puede ser mejorado mediante la lectura de un bloque de registros.
- Por su parte, la segunda opción debe conocer la posición de un registro en el fichero, el Número Relativo de Registro.
- La recuperación selectiva con esta opción requiere el manejo de registros de tamaño fijo, obteniendo un coste (1) .

ACCESO A LA INFORMACIÓN

Definición de Agrupamientos

- El agrupamiento pretende almacenar lo más próximamente posible los registros que, de modo habitual, se acceden conjuntamente.
- De este modo se intenta reducir el coste de las operaciones asociadas,
 - Si se sitúan en el mismo bloque, no será necesario realizar ninguna operación de E/S adicional.
 - Si aparecen en bloques consecutivos, el coste adicional no será demasiado alto.
- El agrupamiento se puede realizar dentro de un fichero o entre ficheros,
 - Intrafichero. Cuando la secuencia física coincide por la secuencia lógica definida por el campo más utilizado en los criterios de búsqueda.
 - Interfichero. Se intercalan los registros de varios ficheros que tengan columnas comunes, y que son habitualmente utilizadas de modo conjunto.
- El agrupamiento interfichero también puede aparecer sobre un único fichero.
- Permite reducir el coste de almacenamiento, ya que las columnas comunes sólo aparecen una vez.
- En cambio, la consulta de los valores de uno sólo de los ficheros y la actualización de las columnas comunes es más costosa.

ACCESO A LA INFORMACIÓN

Técnicas de Compresión

- Estas técnicas permiten sacar partido de la falta de aleatoriedad de los datos.
- La interpretación de la información leída tiene un coste temporal adicional, pero dicho coste es mucho menor que el ahorro producido en número de operaciones de E/S.
- Existen diferentes variantes, cada una de ellas, válido en un entorno determinado,

Compresión Diferencial.

- Se sustituye cada valor por una representación de la diferencia entre éste y el valor anterior.
- La descompresión debe acceder a todos los valores, por lo que sólo resulta aconsejable si se realiza un acceso de tipo global.
- Se puede realizar una Compresión Frontal, en la que se eliminan los caracteres que sean idénticos a los del dato anterior.
- También se puede realizar una Compresión Posterior, en la que sólo se almacenen los caracteres que permite distinguir el dato.

Datos	Comp. Frontal	Front+Post+Perd
ROBERTON	0 - ROBERTONbbbb	0 - 7 - ROBERTO
ROBERTSON	6 - SONbbb	6 - 2- SO
ROBERTSTONE	7 - TONEb	7 - 1 - T
ROBINSON	3 - INSONbbbb	3 - 1- I

ACTUALIZACIÓN DE LA INFORMACIÓN

Reducción del Espacio Ocupado

- La actualización de la información modifica el contenido del fichero,
 - Se eliminan registros.
 - Aparecen nuevos registros.
 - La modificación de la información puede producir un borrado y una inserción.
- Cuando se realizan diferentes operaciones de eliminación e inserción de información, puede aparecer espacio no utilizado en el fichero.
- Su reutilización es muy importante, para reducir el espacio ocupado en memoria secundaria.

Compactado de la Información

- Esta opción es la más sencilla, siendo útil tanto para registros de tamaño fijo como variable.
- Una cuestión es la frecuencia de la operación, que está directamente relacionada con la frecuencia de la actualización del fichero.
- Aún siendo útil para cualquier tipo de registro, puede no ser una solución eficiente para ficheros con un alto grado de actualización.
- Es por ello, que se deben de buscar técnicas alternativas que aseguren una reutilización de la información lo más eficiente posible.

ACTUALIZACIÓN DE LA INFORMACIÓN

Lista de Registros Disponibles

- El borrado de registros debe de asegurar que,
 - El registro quede perfectamente marcado.
 - Se puedan reutilizar estos registros.
- La primera condición es sencilla de cumplir, ya que sólo hace falta elegir un carácter que se escriba al principio del registro.
- El acceso a los registros borrados para su reutilización se puede realizar de diferentes formas:
 - Acceso Secuencial al fichero.
 - Enlazar los registros borrados mediante una Lista.
- La primera opción suele tener un coste bastante alto, por lo que no se suele utilizar.
- La segunda opción parece la más adecuada, pudiendo mejorar su rendimiento mediante su manejo como una Pila,
 - Al principio del fichero aparece un registro de encabezado, en el que se muestra el primer registro de la pila.
 - El resto de registros contienen la localización del siguiente registro en la lista.
 - La inserción y el borrado de registros en la lista se pueden realizar al principio de la pila.

ACTUALIZACIÓN DE LA INFORMACIÓN

Lista en Registros de Longitud Fija

- El carácter que marca el registro como borrado, aparece al principio del registro.
- En este tipo de Registros, la lista se forma a partir del número relativo de registro, grabado a continuación del carácter de borrado.
- El manejo como pila, acelera significativamente las prestaciones del método.

Lista en Registros de Longitud Variable

- Como los registros son de longitud variable, la lista debe de contener el tamaño de cada uno de los registros que la componen.
- Por esta razón, es necesario utilizar el formato de contador de bytes.
- El carácter de borrado aparece como primer carácter del primer campo del registro, es decir, a continuación del tamaño del registro.
- El enlace entre los registros de la lista se realiza mediante la utilización de la posición física de los registros.
- Dicha información aparece a continuación del carácter de borrado.
- La reutilización de los registros puede llevar a una infrautilización de la memoria secundaria, ya que no todos los registros tienen el mismo tamaño.

ACTUALIZACIÓN DE LA INFORMACIÓN

Fragmentación Interna y Externa

- La Fragmentación se define como el espacio asignado a un fichero en memoria secundaria que no puede ser utilizado para almacenar información.
- La fragmentación puede ser de dos tipos,
 - Interna, si el espacio aparece asignado a un registro.
 - Externa, si el espacio aparece entre el asignado a dos registros.
- El primer tipo aparece cuando se utilizan registros de longitud fija para almacenar datos de tamaño variable.
- Si los datos no tienen un tamaño muy similar, se aconseja la utilización de registros de tamaño variable para resolver el problema, ya que permite eliminar la fragmentación interna.
- Pero, esta opción sólo resuelve el problema cuando el fichero no se modifica, ya que el manejo de la lista de disponibles puede hacer reaparecer el problema.
- Para resolver el problema, la inserción de un nuevo registro sólo asigna al nuevo registro el espacio necesario, mientras que el resto aparece en la lista de disponibles.
- Esta forma de actuar puede llegar a producir fragmentación externa, que se puede reducir mediante la aplicación de diferentes técnicas de manejo de la lista de disponibles.

ACTUALIZACIÓN DE LA INFORMACIÓN

Reducción de la Fragmentación Externa

- La solución más sencilla es la Compactación de los ficheros, aunque es una solución muy costosa.
- Otra posibilidad es la Unión de Huecos, pero su desarrollo requiere una gestión bastante compleja de la lista de disponibles.
- La solución más interesante es la utilización de Estrategias de Colocación que seleccionen el Hueco más adecuado en la lista.

Estrategias de Colocación

- La estrategia más sencilla es la del Primer Ajuste, que toma el primer espacio en el que quepa el nuevo registro, aunque no resuelve la problemática de la fragmentación externa.
- Una solución es ordenar los espacios en orden de tamaño ascendente, de modo que en la búsqueda se obtenga el espacio de tamaño más similar.
- Esta estrategia del Mejor Ajuste, puede tener un coste de inserción y borrado de la lista muy alto, y además los espacios libres puede ser demasiado pequeños para su reutilización.
- La ordenación descendente reduce el coste del borrado, y además el tamaño del espacio libre resultante puede ser reutilizado.
- Esta opción da lugar a la estrategia del Peor Ajuste.

ACTUALIZACIÓN DE LA INFORMACIÓN

Conclusiones

- La actualización de la información debe de resolver el problema de la reasignación de espacio.
- Para registros de tamaño fijo, la utilización de una lista de disponibles resuelve el problema de modo conveniente.
- En el caso de registros de tamaño variable, se debe asegurar que la reutilización de espacio es adecuada, mediante la elección de una estrategia de colocación.
- No existe reglas establecidas en este sentido, aunque si es posible definir un conjunto de recomendaciones.
- Éstas se fundamentan en la frecuencia de las operaciones de actualización y en el tipo de fragmentación que aparece en el fichero.
- Si la frecuencia es baja, la estrategia del primer ajuste es suficiente.
- Cuando el problema se relaciona con la fragmentación interna, no debe ser utilizada la estrategia del peor ajuste.
- Si aparece un problema de fragmentación externa, no debe utilizarse la estrategia del mejor ajuste.
- La utilización de la compactación del fichero y de la unión de huecos, puede aparecer como complemento de la estrategia seleccionada.