

Proyecto 2

Compresor distribuido
Eddy Ramrez

30 de octubre de 2018

1. COMPRESOR MEDIANTE HUFFMAN

Este documento contiene la especificación del proyecto sobre un compresor distribuido. Un árbol de Huffman es método de compresión desarrollado en 1952.

1.1. Descripción del algoritmo

El algoritmo consiste en la creación de un árbol binario que tiene cada uno de los símbolos (normalmente bytes) por hoja, y construido de tal forma que siguiéndolo desde la raíz a cada una de sus hojas se obtiene el código (binario) Huffman asociado.

Los pasos para construir el árbol son:

1. Se crean varios árboles, uno por cada uno de los símbolos del alfabeto (los elementos presentes en el archivo que se desea comprimir), considerando cada uno de los árboles en un nodo sin hijos, y etiquetado cada uno con su símbolo asociado y su frecuencia de aparición.
2. Se toman los dos árboles de menor frecuencia, y se unen creando un nuevo árbol. La etiqueta de la raíz será la suma de las frecuencias de las raíces de los dos árboles que se unen, y cada uno de estos árboles será un hijo del nuevo árbol. También se etiquetan las dos ramas del nuevo árbol: con un 0 la de la izquierda, y con un 1 la de la derecha (o viceversa, ese orden es irrelevante).
3. Se repite el paso 2 hasta que sólo quede un árbol único.

Una vez construido este árbol se puede conocer el código asociado a un símbolo, así como obtener el símbolo asociado a un determinado código.

Para obtener el código asociado a un símbolo se debe proceder del siguiente modo:

1. Comenzar con un código vacío
2. Iniciar el recorrido del árbol en la hoja asociada al símbolo
3. Comenzar un recorrido del árbol hacia arriba
4. Cada vez que se suba un nivel, añadir al código la etiqueta de la rama que se ha recorrido
5. Tras llegar a la raíz, invertir el código
6. El resultado es el código Huffman deseado

Otra posibilidad es hacerlo por backtracking. Para obtener un símbolo a partir de un código se debe hacer así:

1. Comenzar el recorrido del árbol en la raíz de éste
2. Extraer el primer símbolo del código a descodificar
3. Descender por la rama etiquetada con ese símbolo
4. Volver al paso 2 hasta que se llegue a una hoja, que será el símbolo asociado al código
5. En la práctica, casi siempre se utiliza el árbol para obtener

2. Descompresión

Se espera que un programa haga la descompresión del archivo comprimido en el archivo total final, a modo de comprobación del trabajo realizado, para ello puede contar con la tabla de frecuencias y archivo comprimido. Se debe de reconstruir el árbol a partir de la tabla de frecuencias.

3. Aspectos técnicos

Su programa consiste en la distribución del proceso de compresión, el cual consiste en dado un archivo (previamente separado en partes), donde le asigno a cada hilo de ejecución distribuida para que genere la tabla de frecuencias independiente, se centrarán en un solo servidor, que sumará las

frecuencias totales y que finalmente le enviará a cada hilo el total y se pueda comprimir su parte del archivo.

Una vez que se ha hecho su compresión, se vuelve a enviar al servidor para poder tener un único archivo comprimido.

El programa que descomprime debe ser independiente y debe leer el archivo comprimido (que tiene en el encabezado la tabla de frecuencias y los bits, luego). Este descompresor recibe como parámetro el nombre del archivo comprimido y el nombre del archivo descomprimido que debe guardarse.

4. Aspectos Administrativos

La fecha de entrega es el 11 de noviembre y puede ser desarrollado en parejas.