# INF221 – Algoritmos y Complejidad

# Clase #10
# Código Huffman II

Aldo Berrios Valenzuela

Miércoles 31 de agosto de 2016

**Resumen**

Realizamos otro ejemplo de Código Huffman y posteriormente, demostramos que es un Algoritmo Voraz.

## 1. Código Huffman (continuación)

Tenemos un texto $T$ formado por símbolos de $\Sigma = \{x_1, \ldots, x_n\}$ tales que $x_i$ aparece $f_i$ veces. Queremos minimizar

$$B(R) = \sum_{1 \leq i \leq n} f_i \, d(x_i),$$

donde $d(x_i)$ es el largo del código binario para $x_i$ obtenido en el árbol binario de $R$.

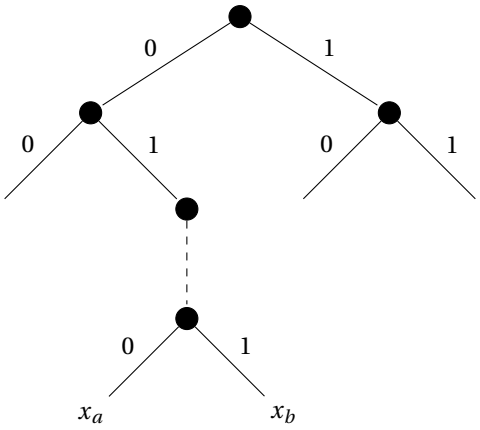


Figura 1: Las hojas que se encuentran a mayor profundidad representan aquellos símbolos que se repiten con menor frecuencia. De la misma forma, aquellos que se encuentren a menor profundidad son símbolos muy frecuentes.

Observaciones:

- Si $R$ es óptimo, todo nodo interno tiene dos hijos.
- Si $d(x_i)$ es la profundidad de $x_i$ hay dos hojas $x_a$, $x_b$ a la profundidad máxima que son hermanos.

### 1.1. Algoritmo

Sucesivamente:

1. Tome los dos símbolos con menos frecuencia de su tabla y reemplacelos por un nuevo símbolo que representa a ambos. Supongamos que estos símbolos son $x_a$ y $x_b$, entonces el nuevo símbolo es $x_{ab}$. La frecuencia de este símbolo conjunto será la suma de la frecuencia de $x_a$ y $x_b$.

2. Cree un árbol que tenga como raíz al símbolo conjunto $x_{ab}$ con $x_a$ y $x_b$ como hojas.
3. Volver a 1 hasta que nuestra tabla esté formada por sólo 1 símbolo conjunto, que representará a todos los símbolos de $\Sigma$.

**Ejemplo 1.1.** Consideremos el Cuadro 1. Nuestro algoritmo nos dice que debemos tomar 2 símbolos con menor

| Símbolo | Frecuencia |
|---|---|
| $a$ | 2 |
| $b$ | 6 |
| $c$ | 3 |
| $d$ | 3 |
| $e$ | 21 |
| $f$ | 5 |
| $g$ | 15 |

Cuadro 1: Dada una determinada secuencia de palabras, se detectó que aparecen los símbolos $a, b, c, d, e, f, g$ con las frecuencias que se muestran arriba.

frecuencia en el Cuadro 1, que en este caso serían $a$ y $c$ o $a$ y $d$ y lo reemplazamos con uno nuevo $ac$ o $ad$ (para este ejercicio ejercicio escogeremos $ac$). La frecuencia de este símbolo será la suma de las frecuencias de $a$ y $c$. Es decir:

$$f_{ac} = f_a + f_b = 2 + 3 = 5$$

Luego, eliminamos los símbolos $a$ y $c$ del Cuadro 1 y agregamos al símbolo conjunto $ac$ dando como origen al Cuadro 2. En seguida, representamos este símbolo conjunto a través de un árbol binario cuya raíz sería $ac$ y sus

| Símbolo | Frecuencia |
|---|---|
| $b$ | 6 |
| $d$ | 3 |
| $e$ | 21 |
| $f$ | 5 |
| $g$ | 15 |
| $ac$ | 5 |

Cuadro 2: Eliminamos los símbolos $a$ y $c$ del Cuadro 1 y lo reemplazamos por $ac$, con frecuencia $f(ac) = 5$.

hojas $a$ y $c$, tal como se muestra en la Figura 2.

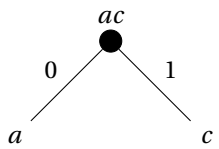


Figura 2: Este árbol tiene como hoja a aquellos símbolos que menos se repiten, generando un peso mínimo hasta el momento.

Nuevamente, escogemos dos símbolos con menor frecuencia en el Cuadro 2. Entre ellos, podemos escoger

- $d$ y $f$
- $d$ y $ac$

En esta ocasión, escogeremos los símbolos $d$ y $f$ (queda como tarea averiguar qué es lo que ocurre si escogemos $d$ y $ac$). Entonces, eliminamos los símbolos $d$ y $f$ del Cuadro 2 y lo reemplazamos por $df$, que tiene una frecuencia de $f_{fd} = f_d + f_f = 8$. El resultado de esto queda representado en el Cuadro 3. En seguida, representamos este símbolo

| Símbolo | Frecuencia |
|---|---|
| $b$ | 6 |
| $e$ | 21 |
| $g$ | 15 |
| $ac$ | 5 |
| $df$ | 8 |

Cuadro 3: Eliminamos los símbolos $d$ y $f$ del Cuadro 2 y en su lugar, agregamos el símbolo $df$ con frecuencia $f_{df} = 8$.

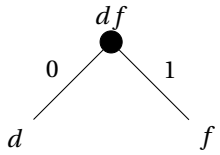


Figura 3: Árbol generado por la segunda iteración. Tiene un peso de $f_{df} = 8$

conjunto a través de un árbol binario con raíz $df$ y hojas $d$ y $f$. Esto está representado en la Figura 3.

Pasamos a la siguiente iteración y buscamos en el Cuadro 3 los dos símbolos de menor frecuencia. Estos son $ac$ y $b$. Por lo tanto, removemos estos dos y agregamos un nuevo símbolo llamado $bac$ (véase el Cuadro 4). Luego,

| Símbolo | Frecuencia |
|---|---|
| $e$ | 21 |
| $g$ | 15 |
| $bac$ | 11 |
| $df$ | 8 |

Cuadro 4: Eliminamos los símbolos $b$ y $ac$ del cuadro 3 y lo reemplazamos por $bac$. Este símbolo tiene una frecuencia de $f_{bac} = f_b + f_{ac} = 11$.

como es costumbre, creamos un árbol con raíz el nodo $bac$ y hojas $b$ y $ac$ (Figura 4).

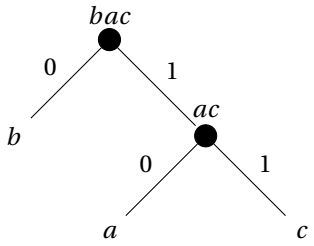


Figura 4: Árbol generado en la tercera iteración. Tiene raíz a $bac$ y hojas $b$ y $ac$. Pero la hoja $ac$ puede representarse en el árbol de la Figura 2.

Continuamos en la cuarta iteración. En ella, buscamos los dos símbolos que tengan la menor frecuencia, los eliminamos y en su lugar agregamos otro símbolo que represente a ambos y cuya frecuencia será la suma de los dos eliminados. Mirando el Cuadro 4, vemos que estos son $bac$ y $df$ con frecuencias $f_{bac} = 11$ y $f_{df} = 8$ respectivamente. Eliminamos estos dos y agregamos $dfbac$ con una frecuencia $f_{dfbac} = f_{df} + f_{bac} = 19$. El resultado se encuentra en el Cuadro 5. En seguida, creamos un nuevo árbol con raíz $dfbac$ y hojas $df$ y $bac$ (Figura 5).

Vamos por la quinta iteración. Buscamos en el Cuadro <>los dos símbolos con menor frecuencia, y estos son $g$ y $dfbac$ cuya suma de frecuencias es de 34. Entonces, quitamos estos símbolos y agregamos $gdfbac$ en su lugar (resultados en el Cuadro 6). Luego, creamos un árbol con raíz $gdfbac$ con hojas $g$ y $dfabc$, tal cual como se muestra en la Figura 6.

| Símbolo | Frecuencia |
|---|---|
| $e$ | 21 |
| $g$ | 15 |
| $dfbac$ | 19 |

Cuadro 5: Eliminamos los símbolos $bac$ y $df$ del Cuadro 4 y agregamos $dfbac$ con una frecuencia de $f_{dfbac} = 19$.

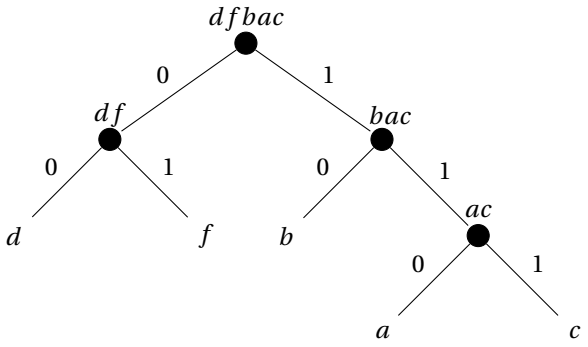


Figura 5: Árbol generado en la cuarta iteración. Tiene raíz $dfbac$ y hojas $df$ y $bac$, que en el fondo son los árboles representados en las Figuras 3 y 4 respectivamente.

| Símbolo | Frecuencia |
|---|---|
| $e$ | 21 |
| $gdfbac$ | 34 |

Cuadro 6: Del cuadro 5 eliminamos los símbolos $g$ y $dfbac$ y en su lugar agregamos $gdfbac$ con una frecuencia de 34.

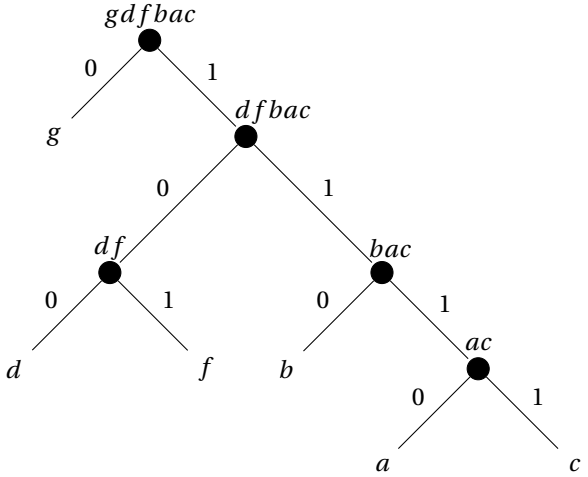


Figura 6: Árbol generado en la quinta iteración. Tiene raíz $gdfbac$ y hojas $g$ y $dfbac$. Esta última hoja corresponde al árbol de la Figura 5.

En la última iteración, vemos que el Cuadro 6 sólo tiene dos símbolos, por lo tanto, es claro que el árbol generado por el Algoritmo de Huffman para la determinada secuencia de palabras del Cuadro 1 es aquel que está representado en la Figura 7.

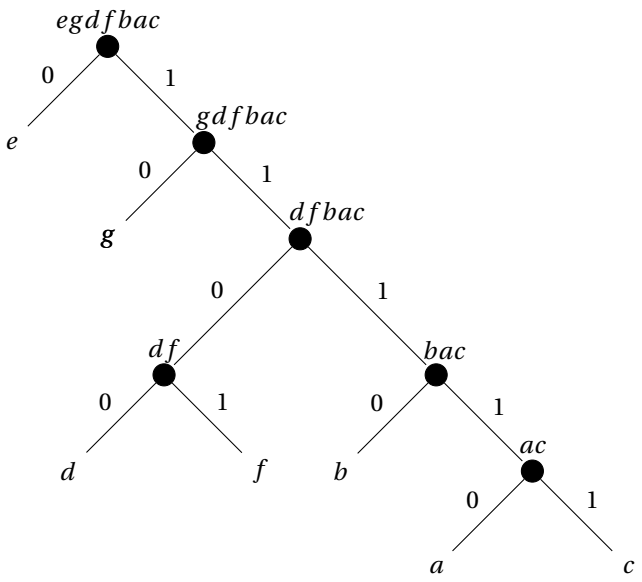


Figura 7: Árbol generado por el Algoritmo de Huffman de la secuencia de palabras del Cuadro 1.

Finalmente, hacemos una tabla para representar la codificación de cada símbolo siguiendo el árbol de la Figura 7. El resultado se encuentra en el cuadro 7.

| Símbolo | Código arrojado por Huffman |
|---|---|
| $a$ | 11110 |
| $b$ | 1110 |
| $c$ | 11111 |
| $d$ | 1100 |
| $e$ | 0 |
| $f$ | 1101 |
| $g$ | 10 |

Cuadro 7: Cada uno de los símbolos de la presente tabla está codificado de tal manera, que no existen ambigüedades al momento de decodificar una secuencia de símbolos. Por otro lado, todos aquellos símbolos que menos se repiten son los que requieren más bits para su representación.

■

Llegamos a la parte entretenida: tenemos que demostrar que el algoritmo de Huffman es óptimo.

*Demostración.* Para demostrar que es óptimo:

- *Elección Voraz:* Lo demostramos la clase pasada.
- *Estructura inductiva:* Elegir un (sub)árbol no interfiere con los demás.
- *Optimal Subestructure:* Recordemos que:
  - $L$: instancia original. En otras palabras, el texto $T$ que nos entregan o la tabla de símbolos con frecuencias respectivas.
  - $x_a, x_b$: símbolos de frecuencia mínima, $f_a + f_b$.
  - $R'$: árbol óptimo para $L'$
  - $L'$: instancia $L \setminus \{x_a, x_b\} \cup \{x_{ab}\}$, frecuencias, ...

- $R$: óptimo para $L$ no es peor que $R' \setminus \{x_{ab}\} \cup x_a \cup x_b$

□

**EVQQTEAR (Ejercicio voluntario para aquellos que tengan la esperanza de aprobar el ramo):** Demostrar que los algoritmos de Kruskal, Prim y Heapsort, son óptimos.