Κώδικες Huffman

Εξετάζουμε το πρόβλημα του σχεδιασμού ενός δυαδικού κώδικα για χαρακτήρες, όπου κάθε χαρακτήρας αναπαρίσταται από μία μοναδική δυαδική συμβολοσειρά (binary string).

- Κώδικες σταθερού μήκους (Fixed-length codeword).
- Κώδικες μεταβλητού μήκους(Variable-length codeword).

	a	b	c	d	e	f
Συχνότητα (σε χιλιάδες)	45	13	12	16	9	5
Κώδικας σταθερού μήκους	000	001	010	011	100	101
Κώδικας μεταβλητού μήκους	0	101	100	111	1101	1100

- Χρειαζόμαστε 300,000 bits για την κωδικοποίηση ενός αρχείου που αποτελείται από 100,000 χαρακτήρες (μόνο από a-f) εάν χρησιμοποιήσουμε τον κώδικα σταθερού μήκους.
- Εάν χρησιμοποιηθεί ο κώδικας μεταβλητού μήκους χρειαζόμαστε συνολικό αριθμό bits ίσο με:

$$1000 \cdot (45 \cdot 1 \ + \ 13 \cdot 3 \ + \ 12 \cdot 3 \ + \ 16 \cdot 3 \ + \ 9 \cdot 4 \ + \ 5 \cdot 4) = 224,000.$$

Περίπου 25% οικονομία χώρου.

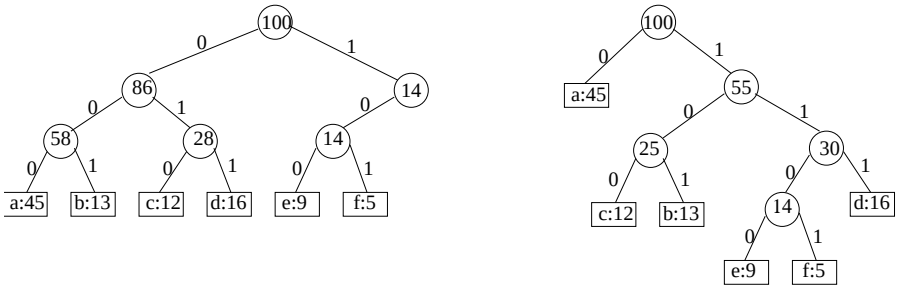
Κώδικες χωρίς πρόθεμα (Prefix-free codes): Είναι κώδικες με την ιδιότητα ότι καμία λέξη (codeword) δεν είναι το πρόθεμα (prefix) κάποιας άλλης.

- Οι κώδικες χωρίς προθέματα απλοποιούν την κωδικοποίηση και την αποκωδικοποίηση.

Κωδικοποίηση: 'abc' : 000 001 010 (κώδικας σταθερού μήκους)
: 0 101 100 (κώδικας μεταβλητού μήκους)

Αποκωδικοποίηση: Χρησιμοποιούμε δυαδικά δένδρα.

Σταθερού μήκους (μη-βέλτιστος) Μεταβλητού μήκους



- Ένας βέλτιστος κώδικας για ένα αρχείο πάντα αναπαρίσταται από ένα πλήρες δυαδικό δένδρο του οποίου κάθε εσωτερικός κόμβος έχει ακριβώς δύο παιδιά.
- Το δένδρο που αναπαριστά ένα βέλτιστο prefix-free κώδικα για ένα σύνολο από C χαρακτήρες έχει ακριβώς $|C|$ φύλλα και $|C| - 1$ εσωτερικούς κόμβους.

Θεωρήστε ένα δένδρο T που αντιστοιχεί σε ένα prefix-free κώδικα για το αλφάβητο C. Επίσης θεωρήστε ένα αρχείο αποτελούμενο από χαρακτήρες του C και έστω:

- $f(c)$ υποδηλώνει τη συχνότητα του χαρακτήρα c μέσα στο αρχείο, και
- $d_T(c)$ υποδηλώνει το βάθος του φύλλου που αναπαριστά το c στο δένδρο T.

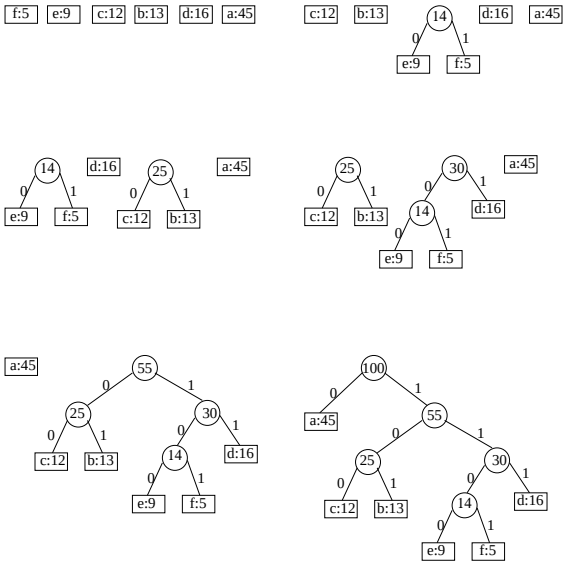
Ο αριθμός των bits που απαιτείται για την κωδικοποίηση του αρχείου είναι:

$$B(T) = \sum_{c \in C} f(c) \cdot d_T(c)$$

Το $B(T)$ ορίζεται ως το κόστος του δένδρου T.

Πρόβλημα Με δεδομένο ένα αλφάβητο C και ένα σύνολο από συχνότητες για τους χαρακτήρες του C, να βρεθεί ένα σχήμα κωδικοποίησης (δένδρο) ελάχιστου κόστους.

$Huffman(C)$
/* Q is a priority queue */
$n = |C|$
$Q = C$
for $i = 1$ **to** $n - 1$ **do**
  $z = allocate\_node()$
  $x = extract\_min(Q)$
  $y = extract\_min(Q)$
  $left[z] = x$
  $right[z] = y$
  $f[z] = f[x] + f[y]$
  $insert(Q, z)$
return $extract\_min(Q)$



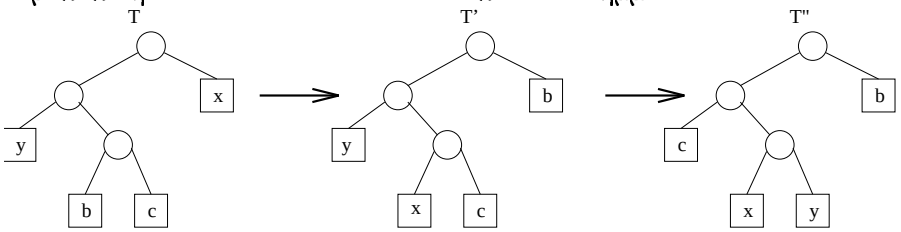
Πολυπλοκότητα: $O(n \log n)$ (Εάν η Q έχει υλοποιηθεί σαν σωρός (binary heap)).

Η ορθότητα του αλγόριθμου του Huffman

Θεώρημα 1 Ο αλγόριθμος $Huffman$ παράγει ένα βέλτιστο prefix-free κώδικα.

Λήμμα 1 Έστω x και y οι δύο χαρακτήρες του C με τις μικρότερες συχνότητες. Τότε, υπάρχει ένας βέλτιστος prefix-free κώδικας για το C στον οποίο οι λέξεις που αντιστοιχούν στα x και y έχουν το ίδιο μήκος και διαφέρουν μόνο στο τελευταίο bit.

Απόδειξη Έστω ένα δένδρο T που αναπαριστά έναν αυθαίρετο βέλτιστο prefix-free κώδικα. Τροποποιήστε το έτσι ώστε να ικανοποιεί το λήμμα.

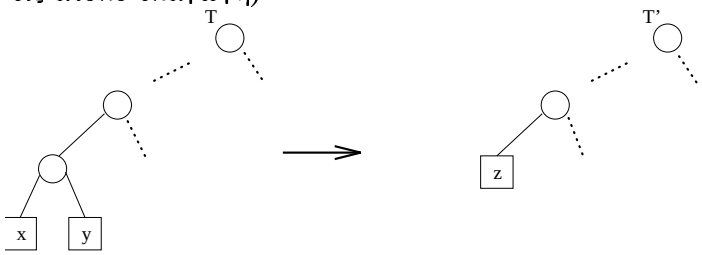


- Έστω b, c είναι δύο χαρακτήρες που είναι 'αδέλφια' μέγιστου βάθους στο T.
- Υποθέστε ότι $f(b) \leq f(c)$ και $f(x) \leq f(y)$.
- Ανταλλάξτε το b με το x. $(\Longrightarrow B(T) \geq B(T'))$
- Ανταλλάξτε το c με το y. $(\Longrightarrow B(T') \geq B(T''))$

□

Λήμμα 2 Έστω ότι το T είναι ένα πλήρες δυαδικό δένδρο που αναπαριστά ένα βέλτιστο prefix-free κώδικα για το αλφάβητο C. Θεωρείστε οποιουσδήποτε 2 χαρακτήρες x και y που εμφανίζονται ως φύλλα 'αδέλφια' στο T, και έστω z ο πατέρας τους. Τότε, θεωρώντας το z σαν ένα χαρακτήρα με συχνότητα $f(z) = f(x) + f(y)$, το δένδρο $T' = T - \{x, y\}$ αναπαριστά ένα βέλτιστο prefix-free κώδικα για το αλφάβητο $C' = C - \{x, y\} \cup \{z\}$.

Απόδειξη (με εις άτοπο επαγωγή)



- $B(T) = B(T') + f(x) + f(y)$.
- Υποθέστε ότι το T' δεν είναι βέλτιστο. $\Longrightarrow$ $\exists$ T" for C' : $B(T") < B(T')$.
- Μπορούμε να δημιουργήσουμε από το T" ένα prefix-free κώδικα για το C με κόστος $B(T") + f(x) + f(y) < B(T)$. Αυτό είναι μία καθαρή αντίφαση λόγω του ότι υποθέσαμε πως το $B(T)$ είναι βέλτιστο.

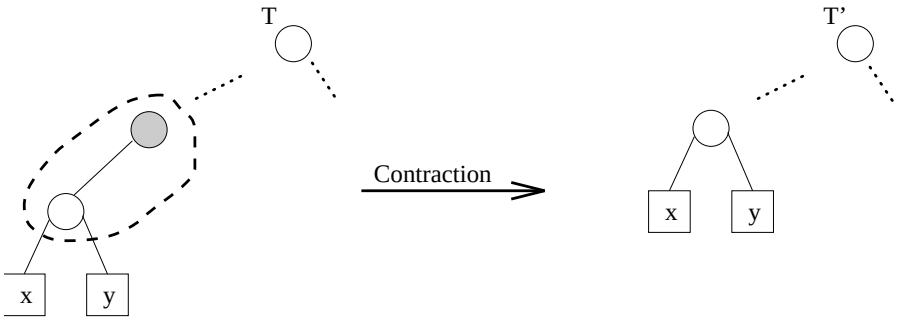
□

Απόδειξη του θεωρήματος 1 Έπεται από τα λήμματα 1 και 2. □

Άσκηση Να αποδειχθεί ότι ένα δυαδικό δένδρο T το οποίο δεν είναι πλήρες, δεν αναπαριστά ένα βέλτιστο prefix-free κώδικα.

Απόδειξη (με εις άτοπο επαγωγή)

- Υποθέστε ένα μη-πλήρες δυαδικό δένδρο το οποίο αναπαριστά ένα βέλτιστο prefix-free κώδικα.
- Τότε, υπάρχει ένας εσωτερικός κόμβος με ένα μόνο παιδί.
- Η συσπείρωση (contraction) του κόμβου αυτού με το παιδί του, παράγει ένα δυαδικό δένδρο T' που αναπαριστά τον κώδικα για το ίδιο αρχείο και επιπλέον έχει μικρότερο κόστος.



□